JACOBI-ÄHNLICHE BLOCKVERFAHREN ZUR NUMERISCHEN BEHANDLUNG DES J-SYMMETRISCHEN ALGEBRAISCHEN EIGENWERTPROBLEMS

DISSERTATION

zur

Erlangung des Grades eines Dr. rer. nat.

des Fachbereichs Mathematik und Informatik

der Fernuniversität - Gesamthochschule - Hagen

vorgelegt von

PETER HOPPE

aus Schwerte

Hagen 1984

INHALTSVERZEICHNIS

Einleitung

1. Grundlagen 1
1.1 Allgemeine Bezeichnungen und Definitionen 1
1.2 Jacobi-ähnliche Verfahren 6
1.3 J-symmetrische Matrizen 12

2. Normreduzierende Ähnlichkeitstransformationen 24
2.1 Normreduzierung auf der Blockdiagonalen 25
2.1.1 Eigenschaften der normreduzierenden Funktionen $f(x)$ und $g(t)$ 27
2.1.2 Newton-Iteration zur Minimierung von $f(x)$ und $g(t)$ 34
2.1.3 Exakte Minimierung von $f(x)$ und $g(t)$ in Spezialfällen 40
2.2 Normreduzierung an Außerdiagonalblöcken 42
2.2.1 Eigenschaften der normreduzierenden Funktionen $F(x_1,x_2)$ und $G(t_1,t_2)$ 45
2.2.2 Newton-Iteration zur Minimierung von $F(x_1,x_2)$ und $G(t_1,t_2)$ 57

3. Elementare Transformationsmatrizen für Jacobi-ähnliche Blockverfahren 66
3.1 Die elementare Eberlein-Matrix 67
3.2 Die elementare Eberlein-Veselić-Matrix 74
3.3 Die elementare Jacobi-Paardekooper-Matrix 99

4. Konvergenzbeweise 123
4.1 Zur asymptotisch quadratischen Konvergenz der Jacobi-ähnlichen Blockverfahren 124
4.2 Zur globalen Konvergenz der Jacobi-ähnlichen Blockverfahren 142

5. Numerische Anwendungen 143
5.1 Numerische Resultate der Jacobi-ähnlichen Blockverfahren.. 144
5.2 Numerischer Vergleich mit Standard-Algorithmen zur Eigenwertberechnung 160
5.3 Kombination der Jacobi-ähnlichen Blockverfahren mit dem Verfahren der Simultanen Iteration 164
5.4 Numerische Anwendungen in physikalischen Problemstellungen (Parameteranalyse) 167

Literaturverzeichnis

Anhang

EINLEITUNG

In der angewandten Mathematik hat das algebraische Eigenwertproblem

$$Ax = \lambda x$$

für die Klasse der sogenannten J-symmetrischen Matrizen in den letzten Jahren ständig an Bedeutung gewonnen. Formal ist eine J-symmetrische Matrix A durch die Bedingung

$$A^T = JAJ$$

definiert, wobei J als reelle orthogonale Matrix gegeben ist. Diese Definition stellt eine Verallgemeinerung des üblichen Symmetriebegriffs dar, und es ist klar, daß auch J-symmetrische Matrizen (abhängig von der speziellen Struktur der Matrix J) gewisse Symmetrieeigenschaften aufweisen, die sie wegen des reduzierten Speicherplatzbedarfs für numerische Zwecke interessant werden lassen.

Der Ursprung der J-Symmetrie liegt nun in bestimmten praktischen Anwendungen des mathematischen Matrizenkalküls. Insbesondere können verallgemeinerte Eigenwertprobleme, die aus physikalisch-technischen Problemstellungen resultieren, auf J-symmetrische Eigenwertprobleme zurückgeführt werden. So reduziert sich beispielsweise das quadratische Eigenwertproblem

$$(\lambda^2 M + \lambda D + K)\ x = 0$$

mit reellen symmetrischen positiv definiten Matrizen M,D und K, welches sich in der Mechanik bei der Berechnung der Frequenzen gedämpfter Schwingungen ergibt, durch Linearisierung auf ein J-symmetrisches gewöhnliches Eigenwertproblem doppelter Dimension mit einer Matrix J der Gestalt diag(1,-1,...,1,-1).

Bekanntermaßen haben sich zur Lösung des (allgemeinen) algebraischen Eigenwertproblems in der numerischen Praxis zwei Klassen von Iterationsverfahren durchgesetzt. Dies ist zum einen die Familie der

Jacobi-ähnlichen Verfahren, die auf der klassischen Jacobi-Methode ([28]) von 1846 zur Diagonalisierung reeller symmetrischer Matrizen basieren und Variationen dieser Methode zur Blockdiagonalisierung schiefsymmetrischer, beliebiger normaler und beliebiger nicht-normaler Matrizen darstellen, und zum anderen die Familie der QR-Verfahren, die durch Weiterentwicklung und Verallgemeinerung des QR-Algorithmus von Francis ([13],1961) zur iterativen Transformation einer beliebigen Matrix auf Quasidreiecksform entstanden sind. Natürlich können all diese Verfahren direkt auf das J-symmetrische Eigenwertproblem angewandt werden, jedoch zerstören die meisten von ihnen die J-Symmetrie des Problems. Daher erscheint es sinnvoll, die Verfahren so zu modifizieren, daß sie die J-Symmetrie während der Iteration erhalten. Dies wird möglich, wenn die Ähnlichkeitstransformationen mit Hilfe sogenannter J-orthogonaler Matrizen durchgeführt werden. Die numerische Attraktivität solch modifizierter Algorithmen ist dann offensichtlich, da sie im Vergleich zu den Originalverfahren nur fast die Hälfte des Speicherplatzes und der Rechenzeit benötigen.

Brebner, Grad ([5]) und Bunse-Gerstner ([6]) schlugen J-symmetrische Varianten des QR-Verfahrens vor, und Veselić ([49],[51],[27], [53]) konstruierte mehrere Jacobi-ähnliche Verfahren für J-symmetrische Matrizen. Die vorliegende Arbeit soll einen Beitrag für die "Jacobi-Richtung" leisten. Wir beschäftigen uns mit der Konstruktion und Anwendung Jacobi-ähnlicher Blockverfahren zur Berechnung der Eigenwerte und Eigenvektoren reeller J-symmetrischer Matrizen.

Der Begriff Blockverfahren bedarf hier einer kurzen Erläuterung. Es ist bekannt, daß die klassischen sogenannten normreduzierenden Verfahren der Jacobi-Familie von Eberlein ([9]) und Hari ([18]), welche eine reelle nicht-normale Matrix durch gleichzeitige orthogonale Transformationen zur Diagonalisierung und nicht-orthogonale Transformationen zur Normreduzierung iterativ in reeller Arithmetik auf eine Blockdiagonalform bringen, nicht notwendig quadratisch konvergieren, wenn nicht-reelle Eigenwerte auftreten. Dies gilt auch für die J-symmetrischen Varianten dieser beiden Methoden, die von Veselić ([49]) formuliert wurden. Nun sind aber J-symmetrische Matrizen in der Regel nicht-normal, so daß man auf normreduzierende Methoden angewiesen ist, und für gewöhnlich besitzen J-symmetrische Matri-

zen nicht-reelle Eigenwerte, insbesondere wenn die Matrizen aus den oben zitierten Problemen der Mechanik stammen.

Um nun reelle Jacobi-ähnliche Verfahren für J-symmetrische Matrizen zu konstruieren, die auch in diesen Fällen quadratisch konvergieren, muß eine 2x2-Blockpartition der J-symmetrischen Matrix und eine analoge Blockteilung der Transformationsmatrizen erfolgen. Veselić ([53]) konzipierte zwei solche Verfahren für Matrizen, die J-symmetrisch bezüglich

$$J = \begin{pmatrix} I & O \\ O & -I \end{pmatrix}$$

sind. Auf dem Grundkonzept dieser Methoden aufbauend formulieren und untersuchen wir in dieser Arbeit mehrere reelle Jacobi-ähnliche Blockverfahren für Matrizen, die J-symmetrisch bezüglich

$$J = \mathrm{diag}(1,-1,\ldots,1,-1)$$

sind. Wir haben diese J-Symmetrie gewählt, da sie gegenüber der obigen gewisse Vorteile hinsichtlich algorithmischer Transparenz besitzt.

In § 1.2 geben wir zunächst zum besseren Verständnis unserer Verfahren ein kurzes Resümee der gebräuchlichsten Jacobi-ähnlichen Algorithmen.

Alsdann zeigen wir in § 1.3 einige wesentliche Eigenschaften J-symmetrischer Matrizen auf. Unter anderem beweisen wir einen wichtigen Satz, der ein J-symmetrisches Analogon zum Satz von Mirsky ([34]) darstellt und die theoretische Basis unserer Verfahren bildet:

Für eine beliebige reelle J-symmetrische Matrix A mit den Eigenwerten $\lambda_1,\ldots,\lambda_n$ *gilt*

$$\inf_{R \text{ reell, } J\text{-orthogonal}} \| R^{-1}AR \|^2 = \sum_{i=1}^{n} |\lambda_i|^2 \quad .$$

In Kapitel 2 definieren wir zwei elementare Blocktransformationsmatrizen $S(x)$ und $T(x_1,x_2)$, die uns zur Normreduzierung an Diagonalpivots bzw. Außerdiagonalpivots dienen. Wie in [8] und [53] erhalten wir durch Anwendung der beiden Transformationen gewisse normreduzierende Funktionen, die es für eine geeignete Ähnlichkeitstransformation zu minimieren gilt. Wir untersuchen die Eigenschaften dieser Funktionen insbesondere hinsichtlich eines Zusammenhangs zwischen nicht-endlichen Minima und mehrfachen Eigenwerten der J-symmetrischen Matrix. Da eine exakte Minimierung der normreduzierenden Funktionen nur in Ausnahmefällen möglich ist, diskutieren wir die approximative Minimumbestimmung mittels Newton-ähnlicher Iterationsverfahren, die analog zu den Methoden von Eberlein ([8],[9]), Sacks-Davis ([44], [45]) und Veselić ([53]) durchgeführt wird.

In Kapitel 3 definieren wir eine elementare orthogonale Blocktransformationsmatrix $U(y_1,y_2)$, mit deren Hilfe sich wahlweise der 2x2-Pivotblock des symmetrischen bzw. des schiefsymmetrischen Teils der Matrix eliminieren läßt. Wir können dann verschiedene Jacobiähnliche Blockverfahren konstruieren, indem wir die elementaren Transformationsmatrizen S,T und U mit alternativen Parametern nach dem "Baukastenprinzip" kombinieren. Die weiteren Untersuchungen der Matrizen S,T und U in Kapitel 3 zielen auf zwei spezielle Jacobiähnliche Blockverfahren ab, die in § 4.1 formuliert werden.

In § 4.1 beweisen wir dann nach einer kurzen algorithmischen Beschreibung dieser beiden Verfahren deren asymptotisch quadratische Konvergenz gegen Murnaghan-Form bei zeilenzyklischer Pivotstrategie (unter der Voraussetzung, daß die J-symmetrische Matrix keine mehrfachen Eigenwerte besitzt). Wir folgen dabei den Ideen von Veselić ([53]), der für seine Methoden eine grobe Beweisskizze der quadratischen Konvergenz gab. Ähnliche Beweise führte Wenzel ([56]) für zwei reelle Jacobi-ähnliche Blockverfahren, die für beliebige nichtnormale Matrizen konzipiert sind und mit ihren orthogonalen Transformationen den schiefsymmetrischen Teil der Matrix blockdiagonalisieren. Unsere Beweise erhalten hier eine neue Qualität, da die Verfahren wahlweise den symmetrischen Teil der Matrix diagonalisieren oder den schiefsymmetrischen blockdiagonalisieren. In § 4.2 folgen dann noch einige Anmerkungen zum Problem der globalen Konvergenz un-

serer Methoden.

In Kapitel 5 haben wir die numerischen Resultate unserer Verfahren zusammengestellt. Wir haben alle entwickelten Algorithmen als ALGOL60-Programme formuliert und auf der Rechenanlage IBM 3031 des Rechenzentrums der Fernuniversität Hagen implementiert und ausgetestet. Zunächst diskutieren wir die Ergebnisse unserer Verfahren in einer internen Gegenüberstellung und ziehen dann einen direkten Vergleich mit dem komplexen Eberlein-Verfahren ([8],[10]) und dem Standard-QR-Algorithmus ([13],[31],[40]). Alsdann resümieren wir die Ergebnisse einer wichtigen Anwendung unserer Verfahren als Unterprozedur in einem Prozess der sogenannten Simultanen Iteration, in welchem sie auf blockdiagonaldominanten Matrizen arbeiten (man beachte, daß unsere Algorithmen wie alle Jacobi-ähnlichen Verfahren auf solchen Matrizen besonders schnell sind). Zum Schluß analysieren wir eine für die numerische Praxis typische Problemstellung, in der unsere Verfahren sukzessive auf eine Klasse von parameterabhängigen Matrizen angewandt werden, die auch wiederum eine gewisse Blockdiagonaldominanz aufweisen.

Mein besonderer Dank gilt Herrn Prof. Dr. K. Veselić für die Anregung zu dieser Arbeit, für seine wertvollen Hinweise und Ratschläge sowie seine nimmermüde Bereitschaft zu zahlreichen kritischen Diskussionen.

1. GRUNDLAGEN

1.1 ALLGEMEINE BEZEICHNUNGEN UND DEFINITIONEN

In diesem Abschnitt wollen wir einige grundlegende Bezeichnungen und Definitionen einführen, wobei eine gewisse Vertrautheit im Umgang mit Matrizen vorausgesetzt wird.

Definition 1.1.1:
Es sei $n = 2m$, $m \in \mathbb{N}$, $m \geq 2$. Die reelle nxn-Matrix

$$A = (a_{ij})_{i,j=1(1)n} \tag{1.1.1}$$

sei blockgeteilt in 2x2-Untermatrizen

$$A_{ij} = \begin{pmatrix} a_{2i-1,2j-1} & a_{2i-1,2j} \\ a_{2i,2j-1} & a_{2i,2j} \end{pmatrix} , \tag{1.1.2}$$

d.h. es gelte

$$A = (A_{ij})_{i,j=1(1)m} . \tag{1.1.3}$$

Dann bezeichnen wir A als *Blockmatrix der Dimension n.*

Die Forderung, daß die Dimension n von A geradzahlig ist, stellt keine wesentliche Einschränkung dar, da jede Matrix ungerader Dimension durch Anfügen einer Nullzeile und -spalte zu einer Matrix gerader Dimension mit einem zusätzlichen Eigenwert 0 erweitert werden kann. Außerdem sind im Hinblick auf die Theorie und Anwendung J-symmetrischer Matrizen nur gerade Dimensionen von praktischem Interesse (vgl. § 1.3). Somit sei im folgenden die Dimension n der Blockmatrix A stets implizit als geradzahlig vorausgesetzt.

Der Fall m = 1, der in der Definition ausgeschlossen ist, bedarf keiner weiteren Beachtung, da die Eigenwerte einer 2x2-Matrix durch Auflösen einer einzigen quadratischen Gleichung berechnet werden

können.

Definition 1.1.2:
A sei eine Blockmatrix der Dimension n = 2m. Dann heißt

$$D = \mathrm{diag}(A_{11}, A_{22}, \ldots, A_{mm}) = \bigoplus_{i=1}^{m} A_{ii} \qquad (1.1.4)$$

die *Blockdiagonale* von A und die 4x4-Untermatrix

$$\hat{A}_{pq} = \begin{pmatrix} A_{pp} & A_{pq} \\ A_{qp} & A_{qq} \end{pmatrix} \quad , \; 1 \le p < q \le m \qquad (1.1.5)$$

die *(p,q)-Restriktion* von A. Der 2x2-Block A_{pq} wird als *Pivotblock* bezeichnet.

Definition 1.1.3:
Für die Blockmatrix A (1.1.3) seien A^+ und A^- durch

$$A^+ = (A^+_{ij})_{i,j=1(1)m} = \frac{1}{2}\,(A+A^T) \;, \qquad (1.1.6)$$

$$A^- = (A^-_{ij})_{i,j=1(1)m} = \frac{1}{2}\,(A-A^T) \qquad (1.1.7)$$

definiert. Dann heißt A^+ der *symmetrische Teil* von A und A^- der *schiefsymmetrische Teil* von A.

Definition 1.1.4:
A sei eine reelle quadratische Matrix. Der *Kommutator* von A wird definiert durch

$$C(A) = A^TA - AA^T \;. \qquad (1.1.8)$$

Wir setzen für die Blockmatrix A (1.1.3)

$$C = (C_{ij})_{i,j=1(1)m} = (c_{ij})_{i,j=1(1)n} = C(A) \quad . \qquad (1.1.9)$$

Den Kommutator der (p,q)-Restriktion von A schreiben wir als

$$C(\hat{A}_{pq}) = \begin{pmatrix} \hat{c}_{pp} & \hat{c}_{pq} \\ \hat{c}_{qp} & \hat{c}_{qq} \end{pmatrix} \quad , 1 \leq p < q \leq m \tag{1.1.10}$$

mit

$$\hat{c}_{ij} = \begin{pmatrix} \hat{c}_{2i-1,2j-1} & \hat{c}_{2i-1,2j} \\ \hat{c}_{2i,2j-1} & \hat{c}_{2i,2j} \end{pmatrix} \quad , i,j \in \{p,q\} \tag{1.1.11}$$

und den Kommutator des Diagonalblocks A_{pp} als

$$C(A_{pp}) = \begin{pmatrix} \hat{\hat{c}}_{2p-1,2p-1} & \hat{\hat{c}}_{2p-1,2p} \\ \hat{\hat{c}}_{2p,2p-1} & \hat{\hat{c}}_{2p,2p} \end{pmatrix} \quad , 1 \leq p \leq m. \tag{1.1.12}$$

Definition 1.1.5:
Es sei

$$R = (R_{ij})_{i,j=1(1)m} = (r_{ij})_{i,j=1(1)n} \tag{1.1.13}$$

eine reelle Blockmatrix der Dimension $n = 2m$, und es gelte $1 \leq p < q \leq m$. Wenn R sich nur in der Untermatrix $\hat{R}_{pq}$ von der Einheitsmatrix unterscheidet, d.h. wenn

$$R_{ij} = I\,\delta_{ij},\ i,j = 1(1)m,\ \{i,j\} \cap \{p,q\} = \emptyset \tag{1.1.14}$$

gilt, so bezeichnen wir R als *elementare Matrix*. Eine Ähnlichkeitstransformation mit einer elementaren Matrix R heißt R-*Transformation*.

Definition 1.1.6:
Für $1 \leq i \leq n$ ist

$$V^{(i)} = \mathrm{diag}(1,\ldots,1,\overset{i}{\underset{\downarrow}{-1}},1,\ldots,1), \tag{1.1.15}$$

und für $1 \leq i,j \leq n$ ist

$$P^{(i,j)} = \begin{array}{c} \\ i\rightarrow \\ \\ j\rightarrow \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} & & i & & j & & \\ & & \downarrow & & \downarrow & & \end{array} \\ \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & 0 & & 1 & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & 1 & & 0 & \\ & & & & & & 1 \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{pmatrix} \end{array} \quad . \qquad (1.1.16)$$

Bemerkung 1.1.7:
Wir bezeichnen mit

$\| \cdot \|$ die Euklidische Norm,

$\| \cdot \|_2$ die Spektral-Norm,

$\| \cdot \|_1$ die Spaltensummen-Norm,

$\| \cdot \|_\infty$ die Zeilensummen-Norm,

$\| \cdot \|_M$ die Maximum-Norm

für beliebige reelle Rechtecksmatrizen (vgl. [57]).

Definition 1.1.8:
A sei eine reelle Blockmatrix der Dimension $n = 2m$. Die *Abweichung von der Blockdiagonalgestalt* $S(A)$ ist definiert durch

$$S^2(A) = \sum_{i \neq j}^{m} \| A_{ij} \|^2 \quad . \qquad (1.1.17)$$

Definition 1.1.9:
A sei eine reelle Blockmatrix der Dimension $n = 2m$. Es seien λ_i, $i = 1(1)n$ die Eigenwerte von A und $\mu_k^{(i)}$, $i = 1(1)m$, $k = 1,2$ die Eigenwerte der Diagonalblöcke A_{ii}, und es gelte $1 \leq p < q \leq m$. Dann ist

$$\delta = \delta(A) = \min \{ |\lambda_i - \lambda_j| \mid 1 \leq i < j \leq n \}, \qquad (1.1.18)$$

$$\delta_D(A) = \min \{ |\mu_k^{(i)} - \mu_l^{(j)}| \mid k,l \in \{1,2\}, \ 1 \leq i < j \leq m \}, \qquad (1.1.19)$$

$$\delta_D^{(p,q)}(A) = \min \{ |\mu_k^{(p)} - \mu_l^{(q)}| \mid k,l \in \{1,2\} \}, \tag{1.1.20}$$

$$\begin{aligned} d^{(p,q)}(A) = \min \{ & |a_{2p-1,2p-1} - a_{2q-1,2q-1}|, |a_{2p-1,2p-1} - a_{2q,2q}|, \\ & |a_{2p,2p} - a_{2q-1,2q-1}|, |a_{2p,2p} - a_{2q,2q}| \} \\ & + \big| |a_{2p-1,2p}| - |a_{2q-1,2q}| \big| . \end{aligned} \tag{1.1.21}$$

Man beachte, daß $\delta = \delta(A)$ invariant unter Ähnlichkeitstransformationen ist, während dies nicht notwendig für $\delta_D(A)$, $\delta_D^{(p,q)}(A)$ und $d^{(p,q)}(A)$ gilt.

Definition 1.1.10:
Eine reelle Matrix M ist in *Murnaghan-Form*, wenn sie sich als direkte Summe von Matrizen der Form (λ) für jeden reellen Eigenwert λ und 2x2-Blöcken der Form

$$\begin{pmatrix} \operatorname{Re}\lambda & \operatorname{Im}\lambda \\ -\operatorname{Im}\lambda & \operatorname{Re}\lambda \end{pmatrix} \tag{1.1.22}$$

für jedes Paar konjugiert komplexer Eigenwerte $(\lambda,\overline{\lambda})$ darstellen läßt. M heißt *Murnaghan-Form einer Matrix* A, wenn es eine reguläre Matrix R gibt, so daß $M = R^{-1}AR$ Murnaghan-Form besitzt (s.[36]).

Das Ende eines Beweises werden wir im folgenden mit dem Symbol ■ am rechten Rand kennzeichnen.

1.2 JACOBI-ÄHNLICHE VERFAHREN

Zum besseren Verständnis unserer Algorithmen geben wir in diesem Paragraphen einen kurzen Überblick über die gebräuchlichsten Jacobi-ähnlichen Verfahren, erläutern ihre Wirkensweise und skizzieren den aktuellen Stand ihrer Konvergenzbeweise.

Unter dem Sammelbegriff *Jacobi-ähnliche Verfahren* verstehen wir Verallgemeinerungen des klassischen Jacobi-Verfahrens ([28]) von 1846. Dieses Verfahren zur Berechnung der Eigenwerte einer reellen symmetrischen Matrix erzeugt eine Folge von Matrizen $A_o=A,A_1,A_2,\ldots$ mit Hilfe orthogonaler Ähnlichkeitstransformationen. Jede Transformationsmatrix R ist dabei eine Ebenenrotation, d.h. eine Matrix, die nur in den Komponenten

$$\begin{pmatrix} r_{pp} & r_{pq} \\ r_{qp} & r_{qq} \end{pmatrix} = \begin{pmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{pmatrix} \quad , x \in \mathbb{R} \qquad (1.2.1)$$

von der Einheitsmatrix abweicht, wobei der Winkel x in jedem Schritt so gewählt wird, daß die Summe der Quadrate der Außerdiagonalelemente von A_k durch Eliminieren des Pivotelementes a_{pq} minimiert wird. Die Matrizenfolge (A_k) konvergiert dann gegen eine Diagonalmatrix mit den Eigenwerten von A auf der Diagonalen.

Wegen des hohen Rechenaufwandes erwies sich das Jacobi-Verfahren für lange Zeit als unpraktikabel. Erst durch den verstärkten Einsatz elektronischer Rechenautomaten in den fünfziger Jahren dieses Jahrhunderts wurde es für die Numerik wieder interessant. Nach seiner Wiederentdeckung durch Gregory ([17]) im Jahre 1953 wurde das Verfahren aus Effektivitätsgründen nach der sogenannten *zyklischen* Pivotstrategie angewandt. Bei dieser Strategie werden die Pivotelemente zur Rotation in einer bestimmten zyklischen Folge und nicht, wie im klassischen Verfahren, der Betragsgröße nach (*optimale* Pivotstrategie) gewählt. Diese Anwendungen erwiesen sich als sehr effizient, jedoch mußten die Konvergenzbeweise "nachgeliefert" werden. Die globale Konvergenz der Methode unter zeilenzyklischer Pivotstrategie wurde von Forsythe und Henrici ([12],1960) nachgewiesen, und Henrici

([24],1958), Schönhage ([46],1961,[47],1964) und Wilkinson ([58], 1962) bewiesen die asymptotisch quadratische Konvergenz bei allgemeiner zyklischer Pivotstrategie.

Der Jacobischen Idee folgend entwickelte Paardekooper ([39],1971) ein Verfahren zur Berechnung der Eigenwerte einer reellen schiefsymmetrischen Matrix. In diesem Verfahren wird in jedem Schritt ein 2x2-Pivotblock durch vier nacheinander ausgeführte Rotationen vom Typ (1.2.1) eliminiert, und die Matrizenfolge konvergiert dann gegen die Murnaghan-Form der Ausgangsmatrix. Paardekooper bewies die globale Konvergenz der Methode unter optimaler Pivotstrategie, und Hari ([22],1982) erweiterte den Beweis auf gewisse zeilen- und spaltenoptimale Pivotstrategien. Hari zeigte dabei auch die Probleme eines Beweises der globalen Konvergenz unter zyklischer Pivotstrategie auf. Die asymptotisch quadratische Konvergenz des Verfahrens unter allgemeiner zyklischer Pivotstrategie wurde wiederum von Hari ([20], [21],1982) nachgewiesen.

Zur Konstruktion Jacobi-ähnlicher Verfahren für nicht-normale Matrizen werden Ähnlichkeitstransformationen benötigt, welche die Euklidische Norm der Matrix reduzieren, denn nach dem Satz von Schur (s.[59]) gilt für die Eigenwerte $\lambda_1,\ldots,\lambda_n$ einer Matrix A

$$\| A \|^2 \geq \sum_{i=1}^{n} |\lambda_i|^2 \quad , \tag{1.2.2}$$

und Gleichheit tritt in (1.2.2) genau dann auf, wenn A normal ist. Da nun die Euklidische Norm unter unitären Ähnlichkeitstransformationen invariant bleibt, ist es nicht möglich, mittels solcher Transformationen eine nicht-normale Matrix zu diagonalisieren. Eberlein ([8]) führte daher nicht-unitäre Scherungsmatrizen R ein, d.h. Matrizen, die sich nur in den Komponenten

$$\begin{pmatrix} r_{pp} & r_{pq} \\ r_{qp} & r_{qq} \end{pmatrix} = \begin{pmatrix} \cosh x & -i\, e^{iy} \sinh x \\ i\, e^{-iy} \sinh x & \cosh x \end{pmatrix} \quad , x,y \in \mathbb{R} \tag{1.2.3}$$

von der Einheitsmatrix unterscheiden, und Voevodin ([54]) definierte nicht-unitäre Matrizen R, die lediglich in einer Komponente von I abweichen:

$$\begin{pmatrix} r_{pp} & r_{pq} \\ r_{qp} & r_{qq} \end{pmatrix} = \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \quad , x \in \mathbb{R} . \tag{1.2.4}$$

Die Grundidee der sogenannten *normreduzierenden Verfahren* ist nun die, in jedem Schritt die Euklidische Norm der Matrix zu reduzieren, d.h. mittels nicht-unitärer Ähnlichkeitstransformationen eine Folge $A_o=A,A_1,A_2,\ldots$ zu konstruieren, so daß

$$\| A_{k+1} \| \leq \| A_k \| \ , \ \lim_{k\to\infty} \| A_k \|^2 = \sum_{i=1}^{n} |\lambda_i|^2 \tag{1.2.5}$$

gilt. Das theoretische Fundament zu diesem Vorgehen liefert der Satz von Mirsky ([34]):

$$\inf_{R \text{ regulär}} \| R^{-1}AR \|^2 = \sum_{i=1}^{n} |\lambda_i|^2 \ . \tag{1.2.6}$$

Dabei wird das Infimum in (1.2.6) genau dann angenommen, wenn A diagonalisierbar ist, d.h. genau dann, wenn eine nicht-singuläre Matrix R existiert, so daß $R^{-1}AR$ normal ist. Eine normale Matrix ist nun wiederum unitär (orthogonal) ähnlich zu einer Matrix in Diagonalform (Murnaghan-Form) (s.[59],[38]). Aus diesem Grund kombinieren die normreduzierenden Verfahren einen normreduzierenden und einen diagonalisierenden Prozess in dem Sinne, daß in jedem Iterationsschritt

$$A_{k+1} = T_k^{-1} \, S_k^{-1} \, A_k \, S_k \, T_k \quad , k=0,1,\ldots$$

eine nicht-unitäre Matrix S_k bestimmt wird, so daß (A_k) gegen Normalität konvergiert (vgl.[38]), und eine unitäre bzw. orthogonale Matrix T_k gewählt wird, die eine normale Matrix diagonalisiert bzw. in Murnaghan-Form überführt.

Die wichtigsten Verfahren dieser Klasse sind das sogenannte *reelle* Eberlein-Verfahren ([9],1968) und das sogenannte *komplexe* Eberlein-Verfahren ([10],1970). Das erstere arbeitet mit reellen Scherungen (siehe (1.2.3) mit $y = \frac{\pi}{2}$) und orthogonalen Rotationen (1.2.1) zur Annullierung des Pivotelements des symmetrischen Teils der Matrix. Hari ([19],1982) bewies die globale Konvergenz der Methode bei zeilen- und spaltenzyklischer Pivotstrategie, wobei die Konvergenz so

zu verstehen ist, daß die Folge (A_k) gegen Normalität konvergiert und der symmetrische Teil von (A_k) gegen eine Diagonalmatrix strebt. Im Falle getrennter Eigenwerte konvergiert dann (A_k) gegen die Murnaghan-Form von A. Jedoch ist diese Methode nicht notwendig quadratisch konvergent, sobald nicht-reelle Eigenwerte auftreten.

Hari ([18],1976) schlug eine Modifikation des reellen Eberlein-Verfahrens vor, in der statt des symmetrischen Teils der schiefsymmetrische Teil der Matrix (block-) diagonalisiert wird. Dies geschieht nach der Strategie von Paardekooper ([39]) mittels vier orthogonaler Rotationen (1.2.1), und Hari zeigte, daß dieses Verfahren unter optimaler Pivotstrategie global konvergiert, d.h. die Folge (A_k) strebt gegen Normalität und der schiefsymmetrische Teil von (A_k) konvergiert gegen eine Matrix in Murnaghan-Form. Im Falle getrennter Eigenwerte konvergiert (A_k) gegen die Murnaghan-Form von A. Der Beweis der globalen Konvergenz unter zyklischer Strategie steht noch aus, und analog zu der reellen Eberlein-Methode ist auch dieses Verfahren nicht notwendig quadratisch konvergent.

Das komplexe Eberlein-Verfahren ist für beliebige nicht-normale komplexe Matrizen konzipiert. Es benutzt komplexe Scherungen (1.2.3) zur Normreduzierung und komplexe Rotationen (d.h. komplexe Verallgemeinerungen von (1.2.1)) zur wahlweisen Annullierung des Pivotelements des hermiteschen oder schiefhermiteschen Teils der Matrix. Die Matrizenfolge (A_k) konvergiert hier gegen eine komplexe Diagonalmatrix mit den Eigenwerten auf der Diagonalen, und die Konvergenz erweist sich als asymptotisch quadratisch, jedoch konnten bislang selbst unter optimaler Pivotstrategie keine entsprechenden Beweise geführt werden. In der numerischen Praxis wird dieses Verfahren aufgrund seiner schnellen Konvergenz häufig angewandt, es hat aber durch die komplexe Arithmetik gegenüber den reellen Algorithmen den Nachteil des doppelten Speicherplatzbedarfs und des zumindest zweifachen Aufwandes für die arithmetischen Operationen.

Um nun ein reelles quadratisch konvergentes Jacobi-ähnliches Verfahren für nicht-normale Matrizen zu konstruieren, bedarf es einer Blockteilung der Matrix in 2x2-Blöcke und einer entsprechenden 4x4-Verallgemeinerung der Transformationen (1.2.1),(1.2.3) und (1.2.4). Veselić ([50],[51]) führte elementare Transformationsmatrizen dieses Typs ein und formulierte ein solches Blockverfahren für beliebige

reelle Matrizen ([50],[52],1979). Dieses Verfahren kombiniert orthogonale Rotationen vom Paardekooperschen Typ zur Blockdiagonalisierung des schiefsymmetrischen Teils der Matrix (vgl.[39]) mit normreduzierenden Transformationen, die mit Hilfe der vierparametrigen Blockverallgemeinerung von (1.2.4)

$$\begin{pmatrix} I & X \\ O & I \end{pmatrix} \quad , \; X = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \tag{1.2.7}$$

ausgeführt werden. Wenzel ([56],1983) bewies die globale Konvergenz der Methode (im Sinne der Konvergenz des Verfahrens [18]) unter optimaler Pivotstrategie und die asymptotisch quadratische Konvergenz unter zeilenzyklischer Pivotstrategie, jedoch ist das Problem der globalen Konvergenz unter zyklischer Strategie noch offen.

Viele der zitierten Algorithmen lassen sich nun für J-symmetrische Eigenwertprobleme so modifizieren, daß die Ähnlichkeitstransformationen J-orthogonal ausgeführt werden und somit die J-Symmetrie erhalten bleibt. Da J-symmetrische Matrizen in der Regel nicht-normal sind (vgl. § 1.3), müssen diese modifizierten Verfahren normreduzierend sein. Die theoretische Grundlage hierzu ist durch den Satz 1.3.4 im nächsten Paragraphen gegeben, der eine zu (1.2.6) analoge Aussage für J-symmetrische Matrizen darstellt.

Veselić ([49],1976) konstruierte zwei Algorithmen dieses Typs für Matrizen, die J-symmetrisch bezüglich $J = \begin{pmatrix} I & O \\ O & -I \end{pmatrix}$ sind. Diese sind im wesentlichen Modifikationen des reellen Eberlein-Verfahrens ([9]) und des Verfahrens von Hari ([18]). Veselić bewies die globale Konvergenz beider Methoden unter optimaler Pivotstrategie, und Hari ([19],[23],1982) erweiterte die globale Konvergenz des erstgenannten Verfahrens auf den Fall zeilen- bzw. spaltenzyklischer Pivotstrategie. Jedoch sind diese Verfahren (wie die ursprünglichen Methoden) nicht notwendig quadratisch konvergent, wenn nicht-reelle Eigenwerte auftreten.

Auch für J-symmetrische Matrizen muß man nun eine 2x2-Blockpartition vornehmen, um reelle Jacobi-ähnliche Verfahren zu formulieren, die eine asymptotisch quadratische Konvergenz ermöglichen. Dabei er-

weisen sich die Transformationen (1.2.7) für diesen Zweck als ungeeignet, da sie die J-Symmetrie zerstören. Veselić ([53],1983) konstruierte mit Hilfe von zweiparametrigen Verallgemeinerungen der reellen hyperbolischen Transformationen (1.2.3) zwei normreduzierende Blockverfahren für Matrizen, die J-symmetrisch bezüglich $J = \begin{pmatrix} I & O \\ O & -I \end{pmatrix}$ sind. In diesen Verfahren wird im orthogonalen Teil mittels zwei Rotationen (1.2.1) wahlweise der 2x2-Pivotblock des symmetrischen oder des schiefsymmetrischen Teils der Matrix eliminiert. In den folgenden Kapiteln modifizieren wir nun diese Verfahren, indem wir sie für eine J-Symmetrie bezüglich

$$J = \mathrm{diag}(1,-1,\ldots,1,-1)$$

formulieren und einige kleinere algorithmische Veränderungen vornehmen, und führen dann für diese Modifikationen die Konvergenzuntersuchungen durch.

1.3 J-SYMMETRISCHE MATRIZEN

In diesem Paragraphen wollen wir eine kurze Einführung in die Theorie der J-symmetrischen Matrizen geben und einige wichtige Anwendungen in physikalisch-technischen Problemstellungen aufzeigen.

Definition 1.3.1:
A sei eine reelle quadratische Matrix, und J sei eine reelle orthogonale Matrix gleicher Dimension. A heißt *J-symmetrisch*, wenn gilt

$$A^T = JAJ \ . \qquad (1.3.1)$$

Wie man sich leicht überlegt, ist die Bedingung (1.3.1) gleichbedeutend damit, daß die Matrix JA symmetrisch ist. Die gebräuchlichsten Formen von J sind

$$J_1 = \begin{pmatrix} I & O \\ O & -I \end{pmatrix} \qquad \text{(vgl.[49],[53])},$$

$$J_2 = \mathrm{diag}(1,-1,\ldots,1,-1) \qquad \text{(vgl.[51],[27])},$$

$$J_3 = \begin{pmatrix} 0 & & & 1 \\ & \ddots & & \\ & & \ddots & \\ 1 & & & 0 \end{pmatrix} \qquad \text{(vgl.[14])}.$$

Die Betrachtung J-symmetrischer Matrizen ist nun keineswegs nur von theoretischem Interesse, da viele verallgemeinerte Eigenwertprobleme mit symmetrischen Matrizen auf gewöhnliche Eigenwertprobleme mit J-symmetrischen Matrizen zurückgeführt werden können. Wichtige Anwendungen ergeben sich beispielsweise aus der Untersuchung linearer mechanischer Systeme. So führt die Berechnung der Frequenzen gedämpfter Schwingungen auf ein quadratisches Eigenwertproblem der Form

$$(\lambda^2 M + \lambda D + K)\ x = O \qquad (1.3.2)$$

mit symmetrischen positiv definiten Matrizen M,D,K. In der Regel besitzen diese Matrizen eine Bandstruktur, wobei M und D in vielen Fäl-

len sogar Diagonalgestalt haben. *)

Das quadratische Eigenwertproblem wird dann linearisiert. Mit dem Ansatz (vgl.[3],[53])

$$\mu = \frac{1}{\lambda},\ z = \begin{pmatrix} \lambda x \\ x \end{pmatrix},\ S = \begin{pmatrix} O & M \\ M & D \end{pmatrix},\ T = \begin{pmatrix} M & O \\ O & -K \end{pmatrix} \qquad (1.3.3)$$

erhält man zunächst ein symmetrisches lineares Eigenwertproblem

$$Sz = \mu Tz \qquad (1.3.4)$$

doppelter Dimension. Da M und K als positiv definit vorausgesetzt sind, existiert eine Zerlegung

$$T = CJ_1C^T ,$$

wobei diese beispielsweise mit dem gewöhnlichen Cholesky-Verfahren (s.[59]) auf stabile Weise durchgeführt werden kann. Das Problem (1.3.4) ist dann äquivalent zu

$$Ay = \mu y \qquad (1.3.5)$$

mit

$$A = J_1C^{-1}S(C^T)^{-1},\ y = C^Tz\ . \qquad (1.3.6)$$

Die Matrix A ist hier J-symmetrisch bezüglich J_1, d.h. von der Form

$$A = \begin{pmatrix} F & G \\ -G^T & H \end{pmatrix}\ , \qquad (1.3.7)$$

wobei die Blockpartition wie in (1.3.3) ist und F und H symmetrisch sind. Die J-Symmetrie bezüglich J_1 entspricht damit, grob gesagt, einer Symmetrie bis auf das Vorzeichen.

*) Der interessierte Leser sei hier bezüglich der Struktur der Matrizen auf [35], [55] verwiesen, die algebraischen Eigenschaften des Problems (1.3.2) werden ausführlich in [29] diskutiert.

Man beachte, daß die Eigenwerte μ_i des Problems (1.3.5) reziprok zu den Eigenwerten λ_i des ursprünglichen Problems (1.3.2) sind. Eine andere Art der Linearisierung transformiert das quadratische Eigenwertproblem auf ein J-symmetrisches Problem der Form

$$Ay = \lambda y$$

(vgl.[53],[26]). Sei dazu $M = M_1 M_1^T$ die Cholesky-Zerlegung von M (diese existiert, da M positiv definit ist). Mit

$$D' = M_1^{-1} D (M_1^T)^{-1}, \quad K' = M_1^{-1} K (M_1^T)^{-1}, \quad z = M_1^T x \tag{1.3.8}$$

ist (1.3.2) äquivalent zu

$$(\lambda^2 I + \lambda D' + K')\, z = 0 .$$

Die Matrix K' ist hier wieder positiv definit, also existiert auch die Cholesky-Zerlegung $K' = LL^T$. Der Ansatz

$$y_1 = L^T z, \quad y_2 = \lambda z$$

führt dann auf ein gewöhnliches Eigenwertproblem der Form

$$\begin{pmatrix} 0 & L^T \\ -L & -D' \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \lambda \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} , \tag{1.3.9}$$

wobei die Matrix wieder J-symmetrisch bezüglich J_1 ist.

Es sei hier am Rande vermerkt, daß gewöhnlich die Bandstruktur der Matrizen M,D,K durch die Linearisierung zerstört wird. So sind die Matrizen C^{-1} in (1.3.6) und M_1^{-1} in (1.3.8) in der Regel voll besetzt. Wenn jedoch M eine Diagonalmatrix ist, so behalten bei der zweiten Linearisierung D' und K' weiterhin ihre Bandstruktur, und diese überträgt sich auf die Matrix (1.3.9).

Wie man sich leicht überlegt, läßt sich nun jede reelle quadratische Matrix A der Dimension $n = 2m$, die J-symmetrisch bezüglich J_1

ist, mit Hilfe der gleichzeitig auf Zeilen und Spalten angewandten Permutation

$$\begin{pmatrix} 1,2,\ldots, & m, & m+1,m+2,\ldots,2m \\ 1,3,\ldots,2m-1, & 2, & 4, \ldots,2m \end{pmatrix}$$

in eine Matrix, die J-symmetrisch bezüglich J_2 ist, transformieren. Da diese Permutation eine Ähnlichkeitstransformation darstellt, bleiben die Eigenwerte dabei invariant. Die Symmetrieeigenschaften einer solchen Matrix lassen sich dann durch die "Schachbrett"-Struktur

$$A = \begin{pmatrix} + & - & + & - & + & - & \cdots \\ - & + & - & + & - & + & \\ + & - & + & - & + & - & \\ - & + & - & + & - & + & \\ + & - & + & - & + & - & \\ - & + & - & + & - & + & \\ \vdots & & & & & & \end{pmatrix}$$

illustrieren, wobei die "+"-Positionen den symmetrischen Teil und die "-"-Positionen den schiefsymmetrischen Teil von A darstellen, d.h. es gilt

$$a_{ij} = (-1)^{i+j} a_{ji}, \quad i,j = 1(1)n.$$

Die J-symmetrische Darstellung bezüglich J_2 weist, wie wir später sehen werden, einige Vorteile hinsichtlich der rechentechnischen Transparenz auf. Daher wollen wir im folgenden unter J-Symmetrie stets J-Symmetrie bezüglich J_2 verstehen, es sei denn, es wird ausdrücklich eine andere J-Symmetrie definiert.

Trotz der formalen Symmetrieeigenschaft (1.3.1) besitzen J-symmetrische Matrizen einige "unschöne" Attribute. So sind die Eigenwerte nicht notwendig reell, und die Matrizen selbst sind nicht notwendig normal und können defektiv sein. Jedoch bedarf es auch hier wie im Fall symmetrischer Matrizen keiner gesonderten Berechnung der Linkseigenvektoren. Wie man leicht verifiziert, berechnet sich zu einem Rechtseigenvektor x der zugehörige Linkseigenvektor als $y = Jx$. *)

*) Der interessierte Leser sei hinsichtlich einer ausführlichen Einführung in die Theorie der J-symmetrischen Matrizen auf [30] und [14] verwiesen.

Um nun die J-Symmetrie einer Matrix unter einer Ähnlichkeitstransformation zu erhalten, benötigt man geeignete Transformationsmatrizen. Dies führt zu der

Definition 1.3.2:
Eine reelle quadratische Matrix R heißt *J-orthogonal*, wenn gilt

$$R^T J R = J \ . \tag{1.3.10}$$

Wie man sofort sieht, ist eine J-orthogonale Matrix R nicht-singulär, und die Inverse berechnet sich als

$$R^{-1} = J R^T J \ . \tag{1.3.11}$$

Damit ist klar, daß eine J-orthogonale Ähnlichkeitstransformation

$$A' = R^{-1} A R \tag{1.3.12}$$

die J-Symmetrie von A erhält. Hier liegt nun die numerische Attraktivität der J-Symmetrie begründet: Ein Verfahren zur Berechnung der Eigenwerte von A, welches auf J-orthogonalen Ähnlichkeitstransformationen basiert, reduziert den Speicherplatzbedarf und die Rechenzeit auf fast die Hälfte. Beispielsweise können die Rechnungen ausschließlich auf dem rechtsoberen Dreieck von A ausgeführt werden. Zudem bedarf es wegen (1.3.11) keiner expliziten Berechnung der linken Transformationsmatrix R^{-1} in (1.3.12), und die Kondition von R kann für jede Matrixnorm durch

$$K(R) = \|R\| \, \|R^{-1}\| \leq \|J\|^2 \, \|R\| \, \|R^T\| \tag{1.3.13}$$

abgeschätzt werden.

Das Ziel des genannten Verfahrens wird es nun sein, die J-symmetrische Matrix A iterativ auf eine Gestalt zu bringen, aus der die Eigenwerte ohne viel Mühe berechnet werden können (man bedenke, daß beispielsweise die Murnaghan-Form von A wieder J-symmetrisch ist). Wir formulieren daher im folgenden zwei wichtige Sätze, die die theoretische Grundlage unserer numerischen Verfahren aus § 4.1 bilden.

Satz 1.3.3:
A sei eine reelle J-symmetrische Matrix der Dimension n. Falls die Eigenwerte von A nicht-defektiv sind, so existiert eine reelle J-orthogonale Ähnlichkeitstransformation von A auf Murnaghan-Form.

Der Beweis dieses Satzes wurde in [26,Theorem 2.2.11] unter der Voraussetzung getrennter Eigenwerte geführt. Er überträgt sich sinngemäß auf den Fall nicht-defektiver Eigenwerte.

Satz 1.3.4:
A sei eine reelle J-symmetrische Matrix der Dimension n, und λ_i, $i = 1(1)n$ seien die Eigenwerte von A (die Vielfachheiten mitgezählt). Dann gilt

$$\inf_{R \text{ reell,J-orthogonal}} \| R^{-1}AR \|^2 = \sum_{i=1}^{n} |\lambda_i|^2 \quad . \tag{1.3.14}$$

Die Aussage dieses Satzes stellt ein J-symmetrisches Analogon zu dem allgemeineren Resultat von Mirsky ([34],vgl.(1.2.6)) dar. Der Beweis stützt sich wesentlich auf einen Satz von Gohberg, Lancaster und Rodman ([14]).

Beweis:
Aus [34] ergibt sich zunächst

$$\inf_{R \text{ reell,J-orthogonal}} \|R^{-1}AR\|^2 \geq \inf_{R \text{ regulär}} \|R^{-1}AR\|^2 = \sum_{i=1}^{n} |\lambda_i|^2. \tag{1.3.15}$$

Alsdann existiert nach [14,Theorem 5.3] eine reelle invertierbare Matrix S mit

$$S^{-1}AS = N, \qquad S^{T}JS = \tilde{J}, \tag{1.3.16}$$

wobei

$$N = N_1 \oplus \ldots \oplus N_r \oplus N_{r+1} \oplus \ldots \oplus N_{r+s}$$

die reelle Jordansche Normalform von A ist, d.h. $N_1,\ldots,N_r$ sind vom Typ

$$\begin{pmatrix} \lambda & 1 & & \mathbf{0} \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ \mathbf{0} & & & \lambda \end{pmatrix}, \quad \lambda \in \mathbb{R} \tag{1.3.17}$$

und $N_{r+1},\ldots,N_{r+s}$ vom Typ

$$\begin{pmatrix} \begin{bmatrix} \sigma & \tau \\ -\tau & \sigma \end{bmatrix} & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & & \mathbf{0} \\ & \ddots & \ddots & \\ & & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & \\ \mathbf{0} & & & \begin{bmatrix} \sigma & \tau \\ -\tau & \sigma \end{bmatrix} \end{pmatrix}, \quad \sigma,\tau \in \mathbb{R}. \tag{1.3.18}$$

Weiter ist die Matrix $\tilde{J}$ definiert durch

$$\tilde{J} = \rho_1\tilde{J}_1 \oplus \ldots \oplus \rho_r\tilde{J}_r \oplus \tilde{J}_{r+1} \oplus \ldots \oplus \tilde{J}_{r+s}, \quad \rho_k \in \{+1,-1\},$$

wobei die $\tilde{J}_k$, $k = 1(1)r+s$ vom Typ J_3 sind und dieselbe Dimension wie die N_k besitzen (es sind nach [14] auch die Vorzeichen der ρ_k eindeutig bestimmt, dies ist jedoch für unseren Beweis unerheblich).

Wie man sich leicht überlegt, sind die Jordanblöcke N_k J-symmetrisch bezüglich $\tilde{J}_k$. Damit ist dann die gesamte Matrix N J-symmetrisch bezüglich $\tilde{J}$.

Mit Hilfe dieser Aussagen beweisen wir nun die folgende Zwischenbehauptung:

$$\inf_{T \text{ reell},\tilde{J}\text{-orthogonal}} \| T^{-1}NT \|^2 = \sum_{i=1}^{n} |\lambda_i|^2 . \tag{1.3.19}$$

Zunächst gilt wieder wie in (1.3.15)

$$\inf_{T \text{ reell},\tilde{J}\text{-orthogonal}} \| T^{-1}NT \|^2 \geq \sum_{i=1}^{n} |\lambda_i|^2 . \tag{1.3.20}$$

Wir definieren dann die Diagonalmatrizen

$$\varrho^{(2t)} = \operatorname{diag}(\omega^t,\ldots,\omega^2,\omega,\omega^{-1},\omega^{-2},\ldots,\omega^{-t}),$$

$$Q^{(2t+1)} = \mathrm{diag}(\omega^t,\ldots,\omega^2,\omega,1,\omega^{-1},\omega^{-2},\ldots,\omega^{-t}),$$
$$Q^{(4t)} = \mathrm{diag}(\omega^t,\omega^t,\ldots,\omega,\omega,\omega^{-1},\omega^{-1},\ldots,\omega^{-t},\omega^{-t}),$$
$$Q^{(4t+2)} = \mathrm{diag}(\omega^t,\omega^t,\ldots,\omega,\omega,1,1,\omega^{-1},\omega^{-1},\ldots,\omega^{-t},\omega^{-t})$$

mit $t \in \mathbb{N}$, $\omega \in \mathbb{R}$, $\omega > 0$. Der Index j bestimmt dabei die Dimension von $Q^{(j)}$, und wir ordnen $Q^{(2t)}, Q^{(2t+1)}$ den Jordanblöcken vom Typ (1.3.17) und $Q^{(4t)}, Q^{(4t+2)}$ den Jordanblöcken vom Typ (1.3.18) zu. Somit existiert zu jedem Jordanblock N_k eine Matrix

$$Q_k \in \{Q^{(2t)}, Q^{(2t+1)}, Q^{(4t)}, Q^{(4t+2)}\},\ k = 1(1)r+s,$$

und man weist leicht nach, daß diese J-orthogonal bezüglich $\tilde{J}_k$ ist. Hieraus folgt dann sofort, daß

$$Q = Q_1 \oplus \ldots \oplus Q_r \oplus Q_{r+1} \oplus \ldots \oplus Q_{r+s}$$

J-orthogonal bezüglich $\tilde{J}$ ist.

Alsdann betrachten wir die Matrix $Q^{-1}NQ$. Diese Matrix ist eine direkte Summe von Blöcken der Art

$$\begin{pmatrix} \lambda & \omega^{-1} & & \mathbf{0} \\ & \ddots & \ddots & \\ & & \ddots & \omega^{-1} \\ \mathbf{0} & & & \lambda \end{pmatrix} \quad \text{bzw.} \quad \begin{pmatrix} \lambda & \omega^{-1} & & & & \mathbf{0} \\ & \ddots & \ddots & & & \\ & & \omega^{-1} & & & \\ & & & \omega^{-2} & & \\ & & & & \omega^{-1} & \\ & & & & \ddots & \omega^{-1} \\ \mathbf{0} & & & & & \lambda \end{pmatrix}$$

und

$$\begin{pmatrix} \begin{bmatrix} \sigma & \tau \\ -\tau & \sigma \end{bmatrix} \begin{bmatrix} \omega^{-1} & 0 \\ 0 & \omega^{-1} \end{bmatrix} & & \mathbf{0} \\ & \ddots & \\ & & \begin{bmatrix} \omega^{-1} & 0 \\ 0 & \omega^{-1} \end{bmatrix} \\ \mathbf{0} & & \begin{bmatrix} \sigma & \tau \\ -\tau & \sigma \end{bmatrix} \end{pmatrix}, \begin{pmatrix} \begin{bmatrix} \sigma & \tau \\ -\tau & \sigma \end{bmatrix} \begin{bmatrix} \omega^{-1} & 0 \\ 0 & \omega^{-1} \end{bmatrix} & & & & & \mathbf{0} \\ & \ddots & & & & & \\ & & \begin{bmatrix} \omega^{-1} & 0 \\ 0 & \omega^{-1} \end{bmatrix} & & & & \\ & & & \begin{bmatrix} \sigma & \tau \\ -\tau & \sigma \end{bmatrix} \begin{bmatrix} \omega^{-2} & 0 \\ 0 & \omega^{-2} \end{bmatrix} & & & \\ & & & & \begin{bmatrix} \omega^{-1} & 0 \\ 0 & \omega^{-1} \end{bmatrix} & & \\ & & & & & \ddots & \\ & & & & & & \begin{bmatrix} \omega^{-1} & 0 \\ 0 & \omega^{-1} \end{bmatrix} \\ \mathbf{0} & & & & & & \begin{bmatrix} \sigma & \tau \\ -\tau & \sigma \end{bmatrix} \end{pmatrix}.$$

Damit gilt

$$\| Q^{-1}NQ \|^2 = \sum_{i=1}^{n} |\lambda_i|^2 + l_1\omega^{-2} + l_2\omega^{-4},\ l_1,l_2 \in \mathbb{N}.$$

Nun ist $Q = Q(\omega)$ eine (matrixwertige) Funktion von ω, und es folgt

$$\begin{aligned} \inf_{T \text{ reell},\tilde{J}\text{-orthogonal}} \| T^{-1}NT \|^2 &\leq \inf_{Q(\omega)} \| Q(\omega)^{-1}NQ(\omega) \|^2 \\ &\leq \lim_{\omega\to\infty} \| Q(\omega)^{-1}NQ(\omega) \|^2 \\ &= \sum_{i=1}^{n} |\lambda_i|^2 . \end{aligned}$$

Zusammen mit (1.3.20) ergibt sich hieraus die Behauptung (1.3.19).

Wir zeigen als nächstes, daß die Matrix $\tilde{J}$ orthogonal ähnlich zu J ist. Wie man sofort sieht, läßt sich jede Matrix vom Typ J_3 durch eine Ähnlichkeitstransformation mit

$$U_1 = \frac{\sqrt{2}}{2}\begin{pmatrix} 1 & & & & & -1 \\ & \ddots & & & \iddots & \\ & & 1 & -1 & & \\ & & 1 & 1 & & \\ & \iddots & & & \ddots & \\ 1 & & & & & 1 \end{pmatrix} \quad \text{bzw.} \quad U_2 = \frac{\sqrt{2}}{2}\begin{pmatrix} 1 & & & & & -1 \\ & \ddots & & & \iddots & \\ & & 1 & & & \\ & & & \sqrt{2} & & \\ & \iddots & & & \ddots & \\ 1 & & & & & 1 \end{pmatrix}$$

(U_1 bei gerader, U_2 bei ungerader Dimension, die übrigen Komponenten in U_1,U_2 sind jeweils 0) auf die Gestalt diag(1,...,1,-1,...,-1) bringen. Da die Matrizen U_1 und U_2 eine direkte Summe von Jacobischen Ebenenrotationen um 45° (und der Matrix (1)) darstellen, sind sie orthogonal (vgl.[59]). Somit existiert eine orthogonale Matrix U als direkte Summe von Matrizen vom Typ U_1,U_2, so daß $U^T\tilde{J}U$ gleich einer Diagonalmatrix mit Elementen +1,-1 ist. Die Signatur von $\tilde{J}$ ist wegen (1.3.16) gleich $(\frac{n}{2},\frac{n}{2})$, daher besitzt $U^T\tilde{J}U$ nach dem Trägheitssatz von Sylvester (s.[62]) je $\frac{n}{2}$ Elemente +1 und -1. Mit Hilfe einer geeigneten Permutationsmatrix P erhalten wir dann

$$P^TU^T\tilde{J}UP = J, \tag{1.3.21}$$

und da P orthogonal ist (vgl.[59]), folgt hieraus die zweite Zwischenbehauptung.

Aus (1.3.21) ergeben sich nun zwei wichtige Folgerungen. Aufgrund von (1.3.16) erhält man zunächst mit $V = UP$

$$V^T S^T J S V = J,$$

d.h. die Matrix SV ist J-orthogonal. Zum anderen existiert zu jeder $\tilde{J}$-orthogonalen Matrix T eine J-orthogonale Matrix $W = V^{-1}TV$ (vgl. [14,Proposition 2.3]). Damit gilt für eine beliebige $\tilde{J}$-orthogonale Matrix T

$$\begin{aligned} \| T^{-1}NT \|^2 &= \| T^{-1}S^{-1}AST \|^2 \\ &= \| VW^{-1}V^{-1}S^{-1}ASVWV^{-1} \|^2 \\ &= \| W^{-1}V^{-1}S^{-1}ASVW \|^2 , \end{aligned}$$

wobei die letzte Gleichung aus der Orthogonalität von V folgt. Da nun SV und W J-orthogonal sind, so folgt

$$\inf_{R \text{ reell,J-orthogonal}} \| R^{-1}AR \|^2 \leq \| T^{-1}NT \|^2$$

und hieraus mit Hilfe von (1.3.19)

$$\begin{aligned} \inf_{R \text{ reell,J-orthogonal}} \| R^{-1}AR \|^2 &\leq \inf_{T \text{ reell,}\tilde{J}\text{-orthogonal}} \| T^{-1}NT \|^2 \\ &= \sum_{i=1}^{n} |\lambda_i|^2 . \end{aligned}$$

Zusammen mit (1.3.15) ergibt sich dann die Behauptung des Satzes. ■

Es soll hier nicht unerwähnt bleiben, daß der Satz 1.3.4 auch mit Hilfe der Konvergenzaussagen für die Verfahren von Veselić ([49]) bewiesen werden kann. Die nach diesen Verfahren konstruierten Matrixfolgen $A_k = R_k^{-1}AR_k$, R_k reell und J-orthogonal bezüglich J_1, konvergieren gegen Normalität (vgl.[49,Theorem 1 und Remark 4]), und es folgt

$$\| R_k^{-1}AR_k \|^2 \to \sum_{i=1}^{n} |\lambda_i|^2 \qquad \text{für } k \to \infty ,$$

was wiederum (1.3.14) beweist.

Ferner sei hier noch angemerkt, daß mit Hilfe des Theorems von Gohberg, Lancaster, Rodman ein alternativer Beweis des Satzes 1.3.3 geführt werden kann. Zudem läßt sich dieser Satz damit auf den defektiven Fall erweitern. Es existiert dann eine reelle J-orthogonale Ähnlichkeitstransformation auf eine allgemeine Blockdiagonalgestalt, wobei die einzelnen Blöcke die Dimension der Jordanblöcke besitzen. Jedoch ist diese Erweiterung für unsere Anwendungen nicht interessant.

Zum Schluß dieses Paragraphen wollen wir noch einen weiteren angenehmen Aspekt der J-Symmetrie aufzeigen. Wie man leicht nachweist, hat der Kommutator einer J-symmetrischen Blockmatrix A der Dimension $n = 2m$ die Darstellung (vgl.(1.1.9))

$$C = C(A) = \begin{pmatrix} 0 & c_{12} & 0 & c_{14} & \cdots & 0 & c_{1n} \\ c_{12} & 0 & c_{23} & 0 & \cdots & c_{2,n-1} & 0 \\ \vdots & & & & & & \vdots \\ c_{1n} & 0 & c_{3n} & 0 & \cdots & c_{n-1,n} & 0 \end{pmatrix} \tag{1.3.22}$$

mit

$$c_{ij} = -2 \sum_{k=1}^{n} a_{ik} a_{jk}. \tag{1.3.23}$$

Damit besitzt C nur $\frac{n^2}{2}$ von null verschiedene Elemente. Analog weist die Matrix $C(\hat{A}_{pq})$ (vgl.(1.1.10)) die Struktur

$$C(\hat{A}_{pq}) = \begin{pmatrix} 0 & \hat{c}_{2p-1,2p} & 0 & \hat{c}_{2p-1,2q} \\ \hat{c}_{2p-1,2p} & 0 & \hat{c}_{2p,2q-1} & 0 \\ 0 & \hat{c}_{2p,2q-1} & 0 & \hat{c}_{2q-1,2q} \\ \hat{c}_{2p-1,2q} & 0 & \hat{c}_{2q-1,2q} & 0 \end{pmatrix} \tag{1.3.24}$$

auf, wobei sich die zwei Elemente von $\hat{C}_{pq}$ als

$$\begin{aligned}
\hat{c}_{2p-1,2q} &= 2a_{2p-1,2p-1}\,a_{2p-1,2q} - 2a_{2p-1,2p}\,a_{2p,2q} \\
&\quad + 2a_{2p-1,2q-1}\,a_{2q-1,2q} - 2a_{2p-1,2q}\,a_{2q,2q} \quad , \\
\hat{c}_{2p,2q-1} &= 2a_{2p-1,2p}\,a_{2p-1,2q-1} + 2a_{2p,2p}\,a_{2p,2q-1} \\
&\quad - 2a_{2p,2q-1}\,a_{2q-1,2q-1} - 2a_{2p,2q}\,a_{2q-1,2q}
\end{aligned} \tag{1.3.25}$$

berechnen. Weiterhin gilt (vgl.(1.1.12))

$$C(A_{pp}) = \begin{pmatrix} 0 & \hat{\hat{c}}_{2p-1,2p} \\ \hat{\hat{c}}_{2p-1,2p} & 0 \end{pmatrix} \tag{1.3.26}$$

mit

$$\hat{\hat{c}}_{2p-1,2p} = 2a_{2p-1,2p-1}\,a_{2p-1,2p} - 2a_{2p-1,2p}\,a_{2p,2p} \, . \tag{1.3.27}$$

Hieraus erhalten wir dann noch das folgende

Lemma 1.3.5:
A sei eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$. Falls $S(A) = 0$ und $\| C(D) \| = 0$ gilt, so besitzt A Murnaghan-Form.

Beweis:
Aus $\| C(D) \| = 0$ folgt nach (1.3.26) und (1.3.27) sofort

$$a_{2i-1,2i}\,(a_{2i-1,2i-1} - a_{2i,2i}) = 0, \quad i = 1(1)m.$$

Wegen $S(A) = 0$ hat A damit Murnaghan-Form. ■

2. NORMREDUZIERENDE ÄHNLICHKEITSTRANSFORMATIONEN

Wie wir in § 1.3 gesehen haben, sind J-symmetrische Matrizen nicht notwendig normal, und aus der Diskussion in § 1.2 wissen wir, daß Jacobi-ähnliche Verfahren für J-symmetrische Matrizen aus normreduzierenden und diagonalisierenden Transformationen bestehen.

Für Blockmatrizen ist es nun notwendig, sowohl eine Normreduzierung der Gesamtmatrix A mit Pivotblöcken im außerdiagonalen Teil von A durchzuführen als auch die Blockdiagonale D zu normalisieren. Wir untersuchen daher in § 2.1 zunächst Ähnlichkeitstransformationen zur Normreduzierung auf der Blockdiagonalen und diskutieren dann in § 2.2 die Eigenschaften von Ähnlichkeitstransformationen zur Normreduzierung an Außerdiagonalblöcken.

2.1 NORMREDUZIERUNG AUF DER BLOCKDIAGONALEN

A sei eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$, und es gelte $1 \le p \le m$, p fest. Die elementare Matrix $S = S(x)$ ist definiert durch

$$S_{pp} = \begin{pmatrix} \cosh x & \sinh x \\ \sinh x & \cosh x \end{pmatrix}, \quad x \in \mathbb{R}, \tag{2.1.1}$$

$$S_{ij} = I\delta_{ij} \quad \text{für } i,j = 1(1)m,\ (i,j) \neq (p,p).$$

S ist damit im Gegensatz zu den anderen elementaren Matrizen, die wir später untersuchen, nur in einem 2x2-Diagonalblock von der Einheitsmatrix verschieden. Diese Matrix wurde erstmalig von Eberlein ([8]) zur Normreduzierung beliebiger Matrizen benutzt. Uns dient sie hier zur Normreduzierung auf den Diagonalblöcken. Wir nennen S die *elementare Eberlein-Matrix*.

Mit Hilfe der bekannten Eigenschaften hyperbolischer Funktionen verifiziert man leicht die folgenden Aussagen:

S ist J-orthogonal und damit nicht-singulär. Für $x \neq 0$ ist S nicht-orthogonal. Es gilt

$$\begin{aligned} &S(0) = I, \quad S(x+y) = S(x)\,S(y), \\ &S(x)^{-1} = S(-x), \quad S(x)^{T} = S(x). \end{aligned} \tag{2.1.2}$$

Es sei für festes $x \in \mathbb{R}$

$$A' = S(x)^{-1} A S(x) . \tag{2.1.3}$$

Dann ist A' aufgrund der J-Orthogonalität von S wieder J-symmetrisch. Die Ähnlichkeitstransformation verändert nur die p-te Blockzeile und -spalte von A, die restlichen Matrixblöcke bleiben invariant unter (2.1.3). Die p-te Blockzeile berechnet sich dabei als

$$
\begin{aligned}
A_{pi}' &= S_{pp}(-x)A_{pi}, \quad i = 1(1)m, \; i \neq p, \\
A_{pp}' &= S_{pp}(-x)A_{pp}S_{pp}(x),
\end{aligned}
\tag{2.1.4}
$$

und die p-te Blockspalte ergibt sich aus der J-Symmetrie von A'. In elementweiser Notation erhalten wir aus (2.1.4)

$$
\left.
\begin{aligned}
a_{2p-1,i}' &= a_{2p-1,i}c - a_{2p,i}s \\
a_{2p,i}' &= a_{2p,i}c - a_{2p-1,i}s
\end{aligned}
\right\} \quad i = 1(1)n, \; i \neq 2p-1, 2p,
$$

$$
\begin{aligned}
a_{2p-1,2p-1}' &= a_{2p-1,2p-1}c^2 + 2a_{2p-1,2p}cs - a_{2p,2p}s^2, \\
a_{2p-1,2p}' &= a_{2p-1,2p}c^2 + (a_{2p-1,2p-1} - a_{2p,2p})cs + a_{2p-1,2p}s^2, \\
a_{2p,2p}' &= a_{2p,2p}c^2 - 2a_{2p-1,2p}cs - a_{2p-1,2p-1}s^2,
\end{aligned}
\tag{2.1.5}
$$

wobei wir $c = \cosh x$, $s = \sinh x$ gesetzt haben.

In den folgenden Untersuchungen gilt es nun, den Parameter x für die S-Transformation geeignet zu wählen.

2.1.1 EIGENSCHAFTEN DER NORMREDUZIERENDEN FUNKTIONEN f(x) UND g(t)

Wir betrachten noch einmal die Ähnlichkeitstransformation $S^{-1}AS$ (vgl.(2.1.3)). Es sei nun p fest, $1 \le p \le m$ und $x \in \mathbb{R}$ beliebig. Wir definieren dann

$$\tilde{A} = \tilde{A}(x) = S(x)^{-1}AS(x), \quad x \in \mathbb{R}. \tag{2.1.6}$$

$\tilde{A}$ ist somit bei vorgegebener Matrix A eine Funktion von x. Wir definieren weiter

$$f(x) = \frac{1}{4}(\| \tilde{A} \|^2 - \| A \|^2), \quad x \in \mathbb{R}. \tag{2.1.7}$$

Mit Hilfe von (2.1.5) und den Additionstheoremen für hyperbolische Funktionen erhalten wir nach einer "Straightforward"-Rechnung die folgende Darstellung von f:

$$f(x) = \alpha(\cosh 4x - 1) + \beta \sinh 4x + a(\cosh 2x - 1) + b \sinh 2x \tag{2.1.8}$$

mit

$$\alpha = \frac{1}{2} a_{2p-1,2p}^2 + \frac{1}{8}(a_{2p-1,2p-1} - a_{2p,2p})^2, \quad \beta = \frac{1}{2} a_{2p-1,2p}\,(a_{2p-1,2p-1} - a_{2p,2p}),$$
$$a = \frac{1}{2} \sum_{i \neq 2p-1,2p}^{n} (a_{2p-1,i}^2 + a_{2p,i}^2), \quad b = -\sum_{i \neq 2p-1,2p}^{n} a_{2p-1,i}\, a_{2p,i} \tag{2.1.9}$$

(siehe [8],[45],[53]). Die Ableitungen von f berechnen sich als

$$f'(x) = 4\tilde{\beta} + 2\tilde{b}, \quad f''(x) = 16\tilde{\alpha} + 4\tilde{a} \tag{2.1.10}$$

mit

$$\begin{aligned} \tilde{\alpha} &= \tilde{\alpha}(x) = \alpha \cosh 4x + \beta \sinh 4x, \\ \tilde{\beta} &= \tilde{\beta}(x) = \alpha \sinh 4x + \beta \cosh 4x, \\ \tilde{a} &= \tilde{a}(x) = a \cosh 2x + b \sinh 2x, \\ \tilde{b} &= \tilde{b}(x) = a \sinh 2x + b \cosh 2x, \end{aligned} \tag{2.1.11}$$

und es gilt das folgende Lemma, welches analog zu [53,Lemma 2.1] mit Hilfe der Eigenschaften (2.1.2) bewiesen wird:

Lemma 2.1.1:
Es seien die Terme (2.1.9) als Funktionen von A aufgefaßt, d.h.es gelte

$$\alpha = \alpha(A),\ \beta = \beta(A),\ a = a(A),\ b = b(A).$$

Dann folgt

$$\tilde{\alpha} = \alpha(\tilde{A}),\ \tilde{\beta} = \beta(\tilde{A}),\ \tilde{a} = a(\tilde{A}),\ \tilde{b} = b(\tilde{A}), \qquad (2.1.12)$$

wobei $\tilde{A}$ durch (2.1.6) gegeben ist.

Für die Kommutatormatrix $\tilde{C} = C(\tilde{A})$ folgt hieraus mit (1.3.23), (2.1.9) und (2.1.10) die Identität

$$\tilde{c}_{2p-1,2p} = \tilde{c}_{2p-1,2p}(x) = f'(x)\ . \qquad (2.1.13)$$

Unser Ziel ist es nun, durch die Ähnlichkeitstransformation (2.1.3) eine Normreduzierung der Matrix A zu erreichen, d.h. es soll

$$\| A' \| \leq \| A \|$$

gelten. Da die Normabnahme möglichst stark sein soll, sind wir daran interessiert, einen Minimalpunkt der Funktion exakt oder zumindest angenähert zu bestimmen, um anschließend mit diesem als Transformationsparameter die S-Transformation auszuführen. Dabei beachte man, daß wegen (2.1.13) bei exakter Bestimmung eines Minimalpunktes von f das Kommutatorelement $c_{2p-1,2p}$ nach der Transformation verschwindet (vgl. auch [42]).

Bevor wir nun die Minimumbestimmung näher untersuchen, wollen wir einige interessante Eigenschaften der Funktion anführen, von deren Richtigkeit man sich leicht überzeugt (vgl.[53],[37]):

f ist auf $\mathbb{R}$ definiert, und für die Koeffizienten (2.1.9) gelten die

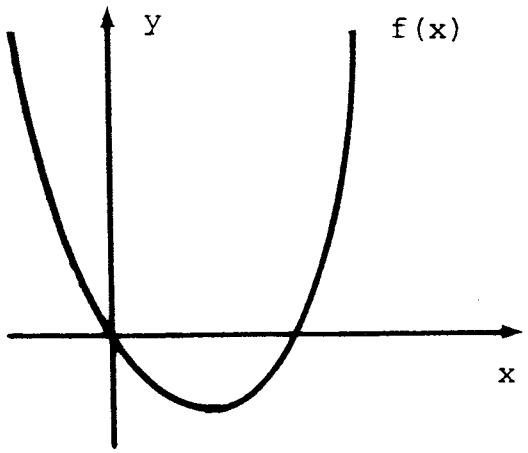
Ungleichungen

$$|\beta| \leq \alpha \ , \quad |b| \leq a \ . \tag{2.1.14}$$

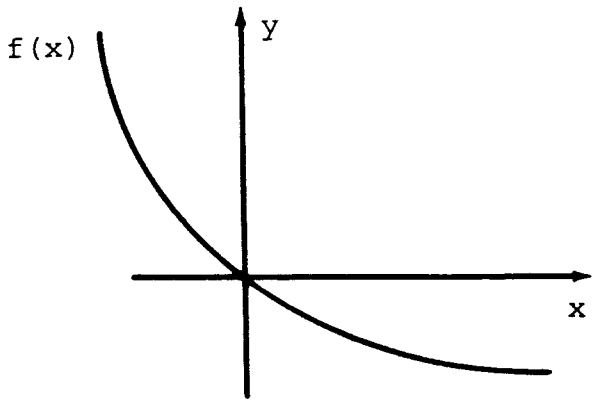
Weiterhin ist f konvex und nach unten durch $-\alpha - a$ beschränkt. Ist mindestens einer der Terme α, a echt größer als 0, so ist f streng konvex, und falls mindestens eine der Ungleichungen (2.1.14) streng ist, so ist f gleichmäßig konvex. Im letzteren Fall hat f ein eindeutiges globales Minimum, und es gilt

$$\lim_{x \to \pm \infty} f(x) = \infty \quad .$$

Der typische Funktionsverlauf sieht dann wegen $f(0) = 0$ folgendermaßen aus:



Ist f nicht gleichmäßig konvex, aber streng konvex, so hat f als Summe von Exponentialfunktionen die folgende typische Gestalt:



In diesem Fall besitzt f kein Minimum. Wir sprechen hier von einem "*Minimum in* $\pm\infty$". Der einzige Fall, in dem f konvex, aber nicht streng konvex ist, ergibt sich als $f \equiv 0$.

Es liegt nun die Vermutung nahe, daß bei einem "Minimum in $\pm\infty$" die Matrix A defektive Eigenwerte besitzt. Wir untersuchen daher im folgenden die Fälle, in denen f streng konvex, aber nicht gleichmäßig konvex ist, im Hinblick auf Vielfachheiten und Defektivitäten der Eigenwerte von A.

Es gelte dazu o.B.d.A. $p = 1$. Wir können uns auf die Fälle

$$\begin{aligned} &\text{(i)} && \beta = \alpha > 0, && b = a = 0, \\ &\text{(ii)} && \beta = \alpha = 0, && b = a > 0, \\ &\text{(iii)} && \beta = \alpha > 0, && b = a > 0 \end{aligned} \qquad (2.1.15)$$

beschränken, denn die übrigen Konstellationen lassen sich durch eine J-orthogonale Ähnlichkeitstransformation mit $V^{(1)}$ auf diese Fälle zurückführen. Es ergibt sich jeweils, daß f ein "Minimum in $-\infty$" besitzt, und aus (2.1.9) erhalten wir nach einigen einfachen Rechnungen, daß in allen drei Fällen die Eigenwerte $\lambda_{1,2}$ von A mit den Eigenwerten des 2x2-Blocks A_{11} identisch und diese jeweils reell und doppelt sind, es gilt

$$\lambda_{1,2} = \frac{a_{11}+a_{22}}{2} \quad .$$

Im Fall (i) weist man leicht nach, daß sich das (1,2)-Element der Resolvente $(A - \lambda I)^{-1}$ von A als

$$- \frac{a_{12}}{(\lambda_{1,2} - \lambda)^2}$$

berechnet. Wegen $a_{12} \neq 0$ besitzt die Resolvente damit in $\lambda_{1,2}$ einen Pol der Ordnung 2, d.h. $\lambda_{1,2}$ ist hier ein defektiver Eigenwert (vgl. [7,Theorem VII.18]).

In den Fällen (ii) und (iii) können jedoch nicht-defektive Eigen-

werte auftreten, wie die Beispiele

$$A(\omega) = \begin{pmatrix} 1 & 0 & 0 & \omega \\ 0 & 1 & 0 & -\omega \\ 0 & 0 & 1 & 1 \\ -\omega & -\omega & -1 & 1 \end{pmatrix} , \ \omega \in \mathbb{R} , \ \lambda_{1,2} = 1, \ \lambda_{3,4} = 1 \pm i,$$

$$A(\omega) = \begin{pmatrix} 1+\omega & \omega & 0 & \omega \\ -\omega & 1-\omega & 0 & -\omega \\ 0 & 0 & 2 & 0 \\ -\omega & -\omega & 0 & 1-\omega \end{pmatrix} , \ \omega \in \mathbb{R} , \ \lambda_{1,2} = 1, \ \lambda_3 = 2, \ \lambda_4 = 1 - \omega \qquad (2.1.16)$$

zeigen. Dabei beachte man, daß beide Matrizen für $\omega \to 0$ gegen Normalität konvergieren. Somit ist leider die obige Vermutung widerlegt, und die Matrizen (2.1.16) zeigen eine weitere unschöne Eigenschaft von f auf, denn selbst wenn die Matrix A fast-normal ist, kann für f ein "Minimum in $\pm\infty$" vorliegen.

Wir erhalten als Konsequenz aus den obigen Untersuchungen das folgende

Lemma 2.1.2:
A habe getrennte Eigenwerte. Dann folgt

$$\lim_{x \to \pm\infty} f(x) = \infty \quad . \qquad (2.1.17)$$

Man könnte nun wiederum vermuten, daß bei guter Trennung der Eigenwerte das Minimum von f nicht allzu groß sein kann. Jedoch wird auch diese Vermutung durch das folgende Matrixbeispiel widerlegt:

$$A(\omega) = \begin{pmatrix} 2 & \omega & 0 & 0 \\ -\omega & 2\omega & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{pmatrix} , \omega, \lambda_3, \lambda_4 \in \mathbb{R} , \ \omega \geq \frac{1}{2} . \qquad (2.1.18)$$

Die Eigenwerte sind $\lambda_{1,2} = 1 + \omega \pm \sqrt{2\omega - 1}\, i$, λ_3, λ_4, wobei wir λ_3 und λ_4 als gut getrennt voraussetzen. Es gilt hier $\alpha = \omega^2 - \omega + \frac{1}{2}$,

$\beta = -\omega^2 + \omega$, $a = b = 0$ und damit

$$f(x) = \alpha(e^{-4x} - 1) + \frac{1}{2}\sinh 4x \ .$$

Für $\omega = \frac{1}{2}$ ist das Eigenwertpaar $\lambda_{1,2}$ defektiv (vgl.(2.1.15)(i)), während die Eigenwerte $\lambda_1,\ldots,\lambda_4$ für $\omega >> \frac{1}{2}$ gut getrennt sind. Dennoch wird das Minimum von f für $\omega \to \infty$ (und damit $\alpha \to \infty$) beliebig groß.

Es stellt sich daher die generelle Frage, ob bei einem gut-konditionierten Eigenwertproblem mit nicht-defektiven Eigenwerten große Parameter für die S-Transformation zugelassen werden sollen, da diese eine schlechte Kondition der Matrix S bewirken, was wiederum gefährlich für die numerische Stabilität der Eigenvektoren von A sein kann. Wir werden auf diese Frage im folgenden Paragraphen eingehen.

Wir wollen nun noch eine andere Darstellung der Funktion f(x) diskutieren. Ausgehend von (2.1.18) erhalten wir mit Hilfe der Variablentransformation

$$t = \tanh x$$

nach einigen elementaren Umformungen unter Ausnutzung der bekannten Eigenschaften hyperbolischer Funktionen

$$f(x) = g(t) = 8\alpha \frac{t^2}{(1-t^2)^2} + 4\beta \frac{t(1+t^2)}{(1-t^2)^2} + 2a \frac{t^2}{1-t^2} + 2b \frac{t}{1-t^2} \ . \quad (2.1.19)$$

Dies ist eine gebrochen rationale Funktion von t, die wegen $|\tanh x| < 1$ in $(-1,1)$ definiert ist. Bildlich gesprochen ist die Funktion g durch seitliches "Zusammendrücken" der Funktion f entstanden. Mit Hilfe von

$$g'(t) = f'(x) \frac{1}{1-t^2} \ , \quad g''(t) = f''(x) \frac{1}{(1-t^2)^2} + f'(x) \frac{2t}{(1-t^2)^2} \quad (2.1.20)$$

überzeugt man sich leicht davon, daß für g die gleichen Konvexitäts-

aussagen und Minimumeigenschaften wie für f gelten. Auch die Aussagen über die "Minima in $\pm\infty$" gelten analog, nur mit dem Unterschied, daß für g "Minima in ± 1" vorliegen.

Wir nennen die Funktionen f(x) und g(t) aufgrund ihres Charakters *normreduzierende Funktionen*. Der Versuch, im Falle gleichmäßiger Konvexität das Minimum von f bzw. g exakt zu bestimmen, führt auf eine homogene Gleichung 4. bzw. 6. Grades (vgl.[8],[45]). Da diese nur in Spezialfällen oder nur höchst instabil gelöst werden kann, untersuchen wir im folgenden die Minimumbestimmung mit Hilfe numerischer Methoden.

2.1.2 NEWTON-ITERATION ZUR MINIMIERUNG VON f(x) UND g(t)

In Anbetracht der Konvexitätseigenschaften der Funktionen f und g bietet es sich an, ihr Minimum jeweils mit Hilfe des Newton-Verfahrens (vgl.[37]) näherungsweise zu bestimmen. Insbesondere ist dieses Verfahren lokal quadratisch konvergent, was wiederum für die asymptotisch quadratische Konvergenz der gesamten Matrixiteration wichtig ist. Betrachten wir zunächst die Funktion f.

Es sei $c_{2p-1,2p} \neq 0$ und o.B.d.A. f gleichmäßig konvex (ansonsten ist das Minimum bekannt). Wir definieren dann

$$x^{(0)} = 0, \quad x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})} \quad , k = 0,1,\ldots, \qquad (2.1.21)$$

wobei aufgrund der Voraussetzungen stets $f''(x^{(k)}) \neq 0$ gilt.

Diese Art der Minimumapproximation zur Normreduzierung geht auf Sacks-Davis ([44],[45]) zurück. Er unterschied dabei zwei Varianten dieses Prozesses, indem entweder nur 1 Schritt in (2.1.21) ausgeführt oder aber die Iteration fortgesetzt wurde, bis die Differenz der Iterierten hinreichend klein war. Sacks-Davis zeigte, daß für $x = x^{(1)}$ die Abschätzung

$$f(x^{(1)}) - f(x^{(0)}) = \frac{1}{4}(\|A'\|^2 - \|A\|^2) \leq - \frac{c_{2p-1,2p}^2}{12\,\|A\|^2}(1+\eta) \qquad (2.1.22)$$

mit

$$\eta \in \mathbb{R}, \quad |\eta| < 0.0758$$

gilt (vgl.[44]). Somit ist bei einer S-Transformation mit $x = x^{(1)}$ als Transformationsparameter die Normabnahme garantiert. Weiterhin deutete Sacks-Davis ohne Beweis an, daß die Newton-Iteration (2.1.21) global konvergiert. Wir geben hierzu einen kurzen Beweis (analog zu [53,Theorem 2.3]).

Lemma 2.1.3:
Die Funktion f(x) (vgl.(2.1.8)) sei gleichmäßig konvex, und es sei

$x^* \in \mathbb{R}$ das Minimum von f. Die Iterierten $x^{(k)}$, $k \in \mathbb{N}_o$ des Newton-Verfahrens seien durch (2.1.21) bestimmt. Dann gilt

$$\lim_{k \to \infty} x^{(k)} = x^* \quad . \tag{2.1.23}$$

Beweis:
Zunächst einmal folgt aus (2.1.22) aufgrund von Lemma 2.1.1 zusammen mit (2.1.10) und (2.1.13)

$$\begin{aligned} f(x^{(k+1)}) - f(x^{(k)}) &\leq - \frac{\tilde{c}_{2p-1,2p}^{\,2}(x^{(k)})}{12 \, \|\tilde{A}(x^{(k)})\|^2} (1+\eta) \\ &\leq - \frac{\tilde{c}_{2p-1,2p}^{\,2}(x^{(k)})}{12 \, \|A\|^2} (1+\eta) \quad , \; k \in \mathbb{N}_o \, , \end{aligned} \tag{2.1.24}$$

wobei sich die letzte Ungleichung aus $\|\tilde{A}(x^{(k)})\| \leq \|A\|$ ergibt. Damit gilt (es sei denn, der Prozess ist nach einer endlichen Anzahl von Iterationsschritten beendet)

$$f(x^{(k+1)}) > f(x^{(k)}) \geq -\alpha - a \, , \; k \in \mathbb{N}_o$$

und daher

$$f(x^{(k)}) \to \xi \, , \quad \xi \in \mathbb{R} \qquad \text{für } k \to \infty \quad .$$

Aus (2.1.24) und (2.1.13) folgt dann

$$f'(x^{(k)}) \to 0 \quad \text{für } k \to \infty$$

und damit aufgrund der gleichmäßigen Konvexität von f die Behauptung. ■

Eine S-Transformation mit dem optimalen Parameter $x = x^*$ wollen wir dann so verstehen, daß die Transformation mit einem Parameter $x^{(k)}$ ausgeführt wird, für welchen $\|\tilde{c}_{2p-1,2p}(x^{(k)})\|$ eine vorgegebene Schwelle unterschreitet. Wegen

$$f(x^{(k)}) \leq f(x^{(1)}) \, , \; k \in \mathbb{N}$$

ist dabei die Normabnahme zumindest genauso stark wie bei einer S-Transformation mit $x = x^{(1)}$.

Wir bezeichnen allgemein die Variante, in der nur 1 Iterationsschritt (2.1.21) ausgeführt wird, als *Standard-Normreduzierung*. Diesen Typ von Normreduzierung findet man u.a. auch in [8] und [54]. Die zweite Variante, in der mehrere Schritte in (2.1.21) ausgeführt werden, bis das Minimum von f hinreichend genau approximiert ist, nennen wir generell *Optimale Normreduzierung*. Diese Art von Normreduzierung wurde u.a. auch in [56] benutzt.

Die Vorteile der optimalen Normreduzierung liegen auf der Hand. Mit dieser Variante kann eine wesentlich bessere Normabnahme als mit der Standard-Normreduzierung erreicht werden, wobei der Aufwand zur Bestimmung der Iterierten im Vergleich zur Berechnung der Größen (2.1.9) gering ist. Dieser letzte Aspekt verstärkt sich noch mit wachsender Dimension der Matrix A. Durch eine bessere Normabnahme wiederum (insbesondere zu Beginn des Prozesses) kann man eine schnellere Konvergenz der gesamten Matrixiterationsfolge erwarten (siehe dazu § 5.1).

Andrerseits liefert unter Umständen der eine Schritt der Standard-Normreduzierung eine hinreichend gute Normabnahme, so daß die Iteration (2.1.21) nur unnötigen Rechenaufwand bewirkt. Dies ist insbesondere dann möglich, wenn die Matrix A schon annähernd normal ist. In der praktischen Anwendung gilt es daher, diese Aspekte gegeneinander abzuwägen und die jeweils besser geeignete Methode zu benutzen.

Wir betrachten alsdann die Funktion g (vgl.(2.1.19)), wobei wir wieder o.B.d.A. gleichmäßige Konvexität voraussetzen. Bei einer Newton-Iteration für g analog zu (2.1.21) ergibt sich hier das Problem, daß das Verfahren nicht notwendig global konvergiert. Veselić ([53]) schlug daher für diese Funktion einen modifizierten Iterationsprozess vor, der aus einer wiederholten Anwendung des ersten Newton-Schrittes zur Minimierung von g besteht. Für den Startwert $t^{(0)} = 0$ liefert dieser wegen (2.1.20),(2.1.10) und (2.1.13)

$$t^{(1)} = -\frac{g'(0)}{g''(0)} = -\frac{c_{2p-1,2p}}{16\alpha + 4a}\,, \qquad (2.1.25)$$

und die Iteration wird somit unter Beachtung von (2.1.2) durch

$$y^{(0)} = 0, \quad y^{(k+1)} = y^{(k)} + \operatorname{artanh} t^{(k+1)},$$

$$t^{(k+1)} = -\frac{\tilde{c}_{2p-1,2p}(y^{(k)})}{16\tilde{\alpha}(y^{(k)}) + 4\tilde{a}(y^{(k)})}, \quad k = 0,1,... \tag{2.1.26}$$

definiert. Dabei sei darauf hingewiesen, daß sich aufgrund von (2.1.13) die Iterierten direkt aus (2.1.10) und (2.1.11) berechnen, ohne daß die einzelnen S-Transformationen für jedes $y^{(k)}$ explizit ausgeführt werden.

Wir unterscheiden auch hier wieder eine Standard- und eine optimale Variante. Dabei ist die Formel (2.1.25) der Standardvariante mit der von Eberlein ([8],[9]) identisch. Eberlein bewies, ohne einen Hinweis auf die Herkunft ihrer Formel zur Berechnung der Transformationsparameter zu geben, die Abschätzung

$$g(t^{(1)}) - g(t^{(0)}) = \frac{1}{4}(\|A'\|^2 - \|A\|^2) \leq -\frac{c_{2p-1,2p}^2}{12\,\|A\|^2} . \tag{2.1.27}$$

Also ist auch bei einer S-Transformation mit dem Parameter $x = y^{(1)}$ die Normabnahme garantiert.

Für die optimale Variante können wir wieder mit Hilfe von Lemma 2.1.1 die Abschätzung (2.1.27) im Sinne von (2.1.24) verallgemeinern, und es läßt sich analog zum Beweis von Lemma 2.1.3 die folgende Aussage verifizieren:

Lemma 2.1.4:
Die Funktion g(t) (vgl.(2.1.19)) sei gleichmäßig konvex, und es sei t^* das Minimum von g. Die Iterierten $t^{(k)}$, $k \in \mathbb{N}_0$ seien durch (2.1.26) bestimmt. Dann gilt

$$\lim_{k \to \infty} t^{(k)} = t^* . \tag{2.1.28}$$

Alle weiteren Aussagen bezüglich der optimalen Normreduzierung (2.1.21) übertragen sich dann sinngemäß auf diese optimale Variante.

Damit stehen uns vier Möglichkeiten der Parameterbestimmung für die S-Transformation zur Verfügung, die alle eine globale Normreduzierung garantieren. Die optimale Normreduzierung bezüglich g ist dabei durch die Berechnung von $\operatorname{artanh} t^{(k)}$ in jedem Iterationsschritt aufwendiger als die optimale Normreduzierung bezüglich f. Andrerseits erscheint die Standard-Normreduzierung für g zweckmäßiger als die entsprechende für f, denn wie man leicht sieht, sind $x^{(1)}$ in (2.1.21) und $t^{(1)}$ in (2.1.25) identisch, jedoch lassen sich die Komponenten der Matrix S schneller und stabiler aus $t^{(1)}$ als aus $x^{(1)}$ berechnen.

Im vorangegangenen Paragraphen wurde die Frage aufgeworfen, ob große Parameter x für die S-Transformation zugelassen werden sollen. Zunächst einmal ist klar, daß bei den Standard-Normreduzierungen dieses Problem nicht auftritt, da wegen (1.3.23),(2.1.9) und (2.1.14)

$$|x^{(1)}| \leq \frac{1}{2}, \qquad |y^{(1)}| = |\operatorname{artanh} t^{(1)}| \leq 0.55$$

gilt. Jedoch können bei Anwendung der optimalen Normreduzierungen die Transformationsparameter x stark anwachsen, wie die Matrizen in § 2.1.1 gezeigt haben. Sacks-Davis ([45]) untersuchte das Phänomen der großen Parameter experimentell und befand, daß diese die Stabilität der Verfahren (hinsichtlich der Veränderungen der Matrix A) nicht beeinträchtigen. Diese Beobachtung wird durch eine Bemerkung von Wilkinson ([60]) theoretisch untermauert, da die Norm von A abnimmt. Hingegen kann sehr wohl die Stabilität der Eigenvektoren von A darunter leiden, daß aus großen Parametern eine schlechte Kondition der Matrix S resultiert. Außerdem können große Parameter ein Indiz für das Auftreten defektiver bzw. fast-defektiver Eigenwerte sein, wie das Beispiel (2.1.15)(i) dokumentiert. In einem solchen Fall würde eine optimale Normreduzierung den defektiven 2x2-Block zu normalisieren versuchen.

Aus diesen Gründen werden wir uns in den Jacobi-ähnlichen Blockverfahren in § 4.1 auf die Standard-Normreduzierung bezüglich g beschränken (generell für $g \neq 0$, auch wenn g nicht gleichmäßig konvex ist). Man beachte dabei, daß für diese Verfahren das Erreichen einer

2x2-Blockdiagonalgestalt ohne vollständige Normalisierung der Matrix ein sinnvolles Abbruchkriterium darstellt.

Zum Schluß dieses Abschnitts sei vermerkt, daß natürlich auch eine Kombination von Standard- und optimaler Normreduzierung denkbar ist, z.B. in dem Sinne, daß anfänglich bei voller Matrix eine optimale Normreduzierung durchgeführt wird, um dann bei fast-blockdiagonaler Matrix auf die Standard-Normreduzierung umzuschwenken. Experimentelle Untersuchungen dieser Art wurden von Sacks-Davis (s.[45]) angestellt.

2.1.3 EXAKTE MINIMIERUNG VON f(x) UND g(t) IN SPEZIALFÄLLEN

Wie wir in § 2.1.1 schon andeuteten, läßt sich das Minimum von f bzw. g in Spezialfällen exakt bestimmen. Interessant ist insbesondere der Fall $a = b = 0$. Hier erhalten wir für $\alpha \neq 0$ aus (2.1.10) und (2.1.11)

$$\alpha \sinh 4x + \beta \cosh 4x = 0 \ ,$$

d.h. das Minimum von f bzw. g berechnet sich exakt aus

$$\tanh 4x = -\frac{\beta}{\alpha} \ . \tag{2.1.29}$$

Hieraus ergibt sich das folgende

Lemma 2.1.5:
A sei eine reelle J-symmetrische Blockdiagonalmatrix der Dimension $n = 2m$. Falls A nicht-defektiv ist, so existiert eine reelle J-orthogonale Ähnlichkeitstransformation, die A auf Murnaghan-Form transformiert.

Beweis:
Wir erhalten die gesuchte Transformationsmatrix als Produkt von m Elementarmatrizen S (für $p = 1(1)m$, vgl.(2.1.1)), d.h. als direkte Summe von m 2x2-Matrizen der Form

$$\begin{pmatrix} \cosh x & \sinh x \\ \sinh x & \cosh x \end{pmatrix} \ .$$

Die Blockdiagonalgestalt von A impliziert $a = b = 0$ für jede Matrix S, und aus der Nicht-Defektivität von A folgt aufgrund der Untersuchungen für (2.1.15)(i) $|\beta| < \alpha$. Damit existiert für jeden 2x2-Block A_{pp} jeweils ein $x \in \mathbb{R}$, welches die Gleichung (2.1.29) erfüllt, und nach einer S-Transformation mit diesem x gilt wegen (2.1.10) und (2.1.13)

$$\hat{c}_{2p-1,2p}' = c_{2p-1,2p}' = 0 \ , \ 1 \leq p \leq m \ .$$

Also sind alle Blöcke A'_{pp} normal, d.h. die transformierte Matrix hat Murnaghan-Form (vgl. Lemma 1.3.5). ■

Die Existenz dieser Ähnlichkeitstransformation ist im wesentlichen aus zwei Gründen interessant. Zum einen können wir den iterativen Prozess, der die Matrizenfolge auf Murnaghan-Form bringen soll, bei hinreichend genauer Blockdiagonalgestalt stoppen und ihn durch die obige Transformation zum Abschluß bringen. Damit sind dann auch die Eigenvektoren von A sofort bekannt, denn wie man sich leicht überlegt, besteht bei einer Endmatrix in Murnaghan-Form die akkumulierte Transformationsmatrix spaltenweise aus den Real- und Imaginärteilen der Eigenvektoren von A. Zum anderen kann diese Transformation bei fast-blockdiagonalen Startmatrizen mit nicht-normalen Diagonalblöcken angewandt werden. In § 4.1 werden wir sehen, daß die asymptotisch quadratische Konvergenz der dort formulierten Jacobi-ähnlichen Blockverfahren sowohl durch $S(A)$ als auch durch $\| C(D) \|$ bestimmt wird. Wir können dann durch die obige Transformation $\| C(D) \|$ zum Verschwinden bringen, und es ist zu erwarten, daß hierdurch die quadratische Konvergenz eher einsetzt, vorausgesetzt, $S(A)$ wurde nicht zu stark gestört (man beachte, daß die Transformation für die Gesamtmatrix A nicht notwendig normreduzierend wirkt).

In beiden Anwendungen besteht jedoch die Gefahr, daß große Parameter x in (2.1.29) auftreten. Diese können durch einen fast-defektiven 2x2-Block A_{pp} begründet sein, aber auch bei nicht-defektiven Blöcken können die Parameter beliebig groß werden, wie das Beispiel (2.1.18) zeigt. Für die numerische Praxis muß daher der Parameter x sinnvoll beschränkt werden. Wir kommen auf dieses Problem in Kapitel 5 zurück.

2.2 NORMREDUZIERUNG AN AUSSERDIAGONALBLÖCKEN

A sei wieder eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$, und es gelte $1 \le p < q \le m$, p,q fest. Die elementare Matrix $T = T(x_1,x_2)$ ist definiert durch

$$T_{pp} = \begin{pmatrix} \cosh x_1 & 0 \\ 0 & \cosh x_2 \end{pmatrix}, \quad T_{pq} = \begin{pmatrix} 0 & \sinh x_1 \\ \sinh x_2 & 0 \end{pmatrix},$$

$$T_{qp} = \begin{pmatrix} 0 & \sinh x_2 \\ \sinh x_1 & 0 \end{pmatrix}, \quad T_{qq} = \begin{pmatrix} \cosh x_2 & 0 \\ 0 & \cosh x_1 \end{pmatrix}, \qquad x_1,x_2 \in \mathbb{R}, \tag{2.2.1}$$

$$T_{ij} = I\delta_{ij} \quad \text{für } i,j = 1(1)m,\ (i,j) \neq (p,p),(p,q),(q,p),(q,q).$$

Diese zweiparametrige elementare Matrix, die in gewissem Sinne eine Verallgemeinerung der Eberlein-Matrix S darstellt, wurde von Veselić ([53]) eingeführt und zur Normreduzierung von Matrizen, die J-symmetrisch bezüglich $J = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}$ sind, angewandt. Auch wir benutzen T zur Normreduzierung der Gesamtmatrix A, jedoch nach modifizierten Strategien der Parameterwahl für x_1,x_2. Wir nennen T die *elementare Eberlein-Veselić-Matrix*.

Wie in § 2.1 verifiziert man wieder leicht mit Hilfe der bekannten Eigenschaften hyperbolischer Funktionen die folgenden Aussagen:

T ist J-orthogonal und damit nicht-singulär. Für $(x_1,x_2) \neq (0,0)$ ist T nicht-orthogonal. Es gilt

$$T(0,0) = I, \quad T(x_1+y_1,x_2+y_2) = T(x_1,x_2)\,T(y_1,y_2),$$
$$T(x_1,x_2)^{-1} = T(-x_1,-x_2), \quad T(x_1,x_2)^T = T(x_1,x_2). \tag{2.2.2}$$

Für feste $x_1,x_2 \in \mathbb{R}$ sei

$$A' = T(x_1,x_2)^{-1} A T(x_1,x_2)\ . \tag{2.2.3}$$

Dann ist A' wieder J-symmetrisch. Es wird unter (2.2.3) nur die p-te und q-te Blockzeile bzw. -spalte transformiert, die übrigen Matrixblöcke bleiben invariant. Die transformierten Blöcke berechnen sich dabei wegen (2.2.2) als

$$\left.\begin{aligned} A_{pi}' &= T_{pp}A_{pi} - T_{pq}A_{qi} \\ A_{qi}' &= -T_{qp}A_{pi} + T_{qq}A_{qi} \end{aligned}\right\} \quad i = 1(1)m,\ i \neq p,q,$$

$$\begin{aligned} A_{pp}' &= T_{pp}A_{pp}T_{pp} - T_{pq}A_{qp}T_{pp} + T_{pp}A_{pq}T_{qp} - T_{pq}A_{qq}T_{qp}, \\ A_{pq}' &= T_{pp}A_{pp}T_{pq} - T_{pq}A_{qp}T_{pq} + T_{pp}A_{pq}T_{qq} - T_{pq}A_{qq}T_{qq}, \\ A_{qq}' &= -T_{qp}A_{pp}T_{pq} + T_{qq}A_{qp}T_{pq} - T_{qp}A_{pq}T_{qq} + T_{qq}A_{qq}T_{qq}, \end{aligned} \tag{2.2.4}$$

und die p-te und q-te Blockspalte und A_{qp}' ergeben sich aus der J-Symmetrie von A'.

Wenn wir (2.2.4) elementweise notieren, so erhalten wir

$$\left.\begin{aligned} a_{2p-1,i}' &= a_{2p-1,i}\,c_1 - a_{2q,i}\,s_1 \\ a_{2p,i}' &= a_{2p,i}\,c_2 - a_{2q-1,i}\,s_2 \\ a_{2q-1,i}' &= a_{2q-1,i}\,c_2 - a_{2p,i}\,s_2 \\ a_{2q,i}' &= a_{2q,i}\,c_1 - a_{2p-1,i}\,s_1 \end{aligned}\right\} \quad i = 1(1)n,\ i \neq 2p-1,2p,2q-1,2q$$

$$\begin{aligned} a_{2p-1,2p-1}' &= a_{2p-1,2p-1}\,c_1^2 + 2a_{2p-1,2q}\,c_1s_1 - a_{2q,2q}\,s_1^2, \\ a_{2p-1,2p}' &= a_{2p-1,2p}\,c_1c_2 + a_{2p-1,2q-1}\,c_1s_2 - a_{2p,2q}\,c_2s_1 + a_{2q-1,2q}\,s_1s_2, \\ a_{2p,2p}' &= a_{2p,2p}\,c_2^2 + 2a_{2p,2q-1}\,c_2s_2 - a_{2q-1,2q-1}\,s_2^2, \\ a_{2p-1,2q-1}' &= a_{2p-1,2q-1}\,c_1c_2 + a_{2p-1,2p}\,c_1s_2 + a_{2q-1,2q}\,c_2s_1 - a_{2p,2q}\,s_1s_2, \\ a_{2p-1,2q}' &= a_{2p-1,2p-1}\,c_1s_1 + a_{2p-1,2q}\,(c_1^2+s_1^2) - a_{2q,2q}\,c_1s_1, \\ a_{2p,2q-1}' &= a_{2p,2p}\,c_2s_2 + a_{2p,2q-1}\,(c_2^2+s_2^2) - a_{2q-1,2q-1}\,c_2s_2, \\ a_{2p,2q}' &= a_{2p,2q}\,c_1c_2 - a_{2q-1,2q}\,c_1s_2 - a_{2p-1,2p}\,c_2s_1 - a_{2p-1,2q-1}\,s_1s_2, \end{aligned} \tag{2.2.5}$$

$$a'_{2q-1,2q-1} = a_{2q-1,2q-1}c_2^2 - 2a_{2p,2q-1}c_2s_2 - a_{2p,2p}s_2^2,$$

$$a'_{2q-1,2q} = a_{2q-1,2q}c_1c_2 - a_{2p,2q}c_1s_2 + a_{2p-1,2q-1}c_2s_1 + a_{2p-1,2p}s_1s_2,$$

$$a'_{2q,2q} = a_{2q,2q}c_1^2 - 2a_{2p-1,2q}c_1s_1 - a_{2p-1,2p-1}s_1^2,$$

wobei wir $c_i = \cosh x_i$, $s_i = \sinh x_i$, $i = 1,2$ gesetzt haben.

In den beiden folgenden Paragraphen werden nun mehrere Strategien zur Bestimmung der Transformationsparameter x_1, x_2 diskutiert.

2.2.1 EIGENSCHAFTEN DER NORMREDUZIERENDEN FUNKTIONEN $F(x_1,x_2)$ UND $G(t_1,t_2)$

Analog zu § 2.1.1 schreiben wir zunächst die Ähnlichkeitstransformation (2.2.3) in Abhängigkeit von (x_1,x_2). Es sei dazu $1 \le p < q \le m$, p,q fest, dann definieren wir

$$\tilde{A} = T(x_1,x_2)^{-1} A T(x_1,x_2) \quad , \; x_1,x_2 \in \mathbb{R} . \tag{2.2.6}$$

Damit ist $\tilde{A} = \tilde{A}(x_1,x_2)$ wieder bei vorgegebener Matrix A eine Funktion von $x_1,x_2 \in \mathbb{R}$. Weiterhin definieren wir

$$F(x_1,x_2) = \frac{1}{4}(\| \tilde{A} \|^2 - \| A \|^2) \; , \; x_1,x_2 \in \mathbb{R} . \tag{2.2.7}$$

Es gilt (vgl.[53])

$$\| \tilde{A}^+ \|^2 - \| \tilde{A}^- \|^2 = \| A^+ \|^2 - \| A^- \|^2$$

und aufgrund der J-Symmetrie von A und $\tilde{A}$

$$\| A \|^2 = \| A^+ \|^2 + \| A^- \|^2 \; , \quad \| \tilde{A} \|^2 = \| \tilde{A}^+ \|^2 + \| \tilde{A}^- \|^2 .$$

Damit ergibt sich

$$F(x_1,x_2) = \frac{1}{2}(\| \tilde{A}^+ \|^2 - \| A^+ \|^2) = \frac{1}{2}(\| \tilde{A}^- \|^2 - \| A^- \|^2) . \tag{2.2.8}$$

Wie in § 2.1.1 erhalten wir dann mit Hilfe von (2.2.5) und den Additionstheoremen der hyperbolischen Funktionen nach einer recht mühseligen Rechnung die folgende Darstellung von F:

$$\begin{aligned} F(x_1,x_2) = {} & \alpha_1(\cosh 4x_1 - 1) + \beta_1 \sinh 4x_1 + \alpha_{11}(\cosh 2x_1 - 1) + \beta_{11} \sinh 2x_1 \\ & + \alpha_2(\cosh 4x_2 - 1) + \beta_2 \sinh 4x_2 + \alpha_{22}(\cosh 2x_2 - 1) + \beta_{22} \sinh 2x_2 \\ & + \delta_+(\cosh(2x_1+2x_2) - 1) + \varepsilon_+ \sinh(2x_1+2x_2) \\ & + \delta_-(\cosh(2x_1-2x_2) - 1) + \varepsilon_- \sinh(2x_1-2x_2) \end{aligned} \tag{2.2.9}$$

mit

$$\alpha_1 = \frac{1}{2}a_{2p-1,2q}^2 + \frac{1}{8}(a_{2p-1,2p-1} - a_{2q,2q})^2 \,,\quad \beta_1 = \frac{1}{2}a_{2p-1,2q}(a_{2p-1,2p-1} - a_{2q,2q}) \,,$$

$$\alpha_2 = \frac{1}{2}a_{2p,2q-1}^2 + \frac{1}{8}(a_{2p,2p} - a_{2q-1,2q-1})^2 \,,\quad \beta_2 = \frac{1}{2}a_{2p,2q-1}(a_{2p,2p} - a_{2q-1,2q-1}) \,,$$

$$\delta_+ = \frac{1}{4}(a_{2p-1,2q-1} - a_{2p,2q})^2 + \frac{1}{4}(a_{2q-1,2q} + a_{2p-1,2p})^2 \,,$$

$$\varepsilon_+ = \frac{1}{2}(a_{2p-1,2q-1} - a_{2p,2q})\,(a_{2q-1,2q} + a_{2p-1,2p}) \,,$$

$$\delta_- = \frac{1}{4}(a_{2p-1,2q-1} + a_{2p,2q})^2 + \frac{1}{4}(a_{2q-1,2q} - a_{2p-1,2p})^2 \,,$$

$$\varepsilon_- = \frac{1}{2}(a_{2p-1,2q-1} + a_{2p,2q})\,(a_{2q-1,2q} - a_{2p-1,2p}) \,, \qquad (2.2.10)$$

$$\alpha_{11} = \frac{1}{2}\sum_{i\neq 2p-1,2p,2q-1,2q}^{n}(a_{2p-1,i}^2 + a_{2q,i}^2),\quad \beta_{11} = -\sum_{i\neq 2p-1,2p,2q-1,2q}^{n} a_{2p-1,i}a_{2q,i} \,,$$

$$\alpha_{22} = \frac{1}{2}\sum_{i\neq 2p-1,2p,2q-1,2q}^{n}(a_{2p,i}^2 + a_{2q-1,i}^2),\quad \beta_{22} = -\sum_{i\neq 2p-1,2p,2q-1,2q}^{n} a_{2p,i}a_{2q-1,i}$$

(vgl.[53]). Die partiellen Ableitungen von F berechnen sich als

$$\frac{\partial F}{\partial x_i} = 4\tilde{\beta}_i + 2\tilde{\beta}_{ii} + 2\tilde{\varepsilon}_+ - 2(-1)^i\tilde{\varepsilon}_- \quad , i = 1,2 \,,$$

$$\frac{\partial^2 F}{\partial x_i^2} = 16\tilde{\alpha}_i + 4\tilde{\alpha}_{ii} + 4\tilde{\delta}_+ + 4\tilde{\delta}_- \quad , i = 1,2 \,, \qquad (2.2.11)$$

$$\frac{\partial^2 F}{\partial x_1 \partial x_2} = \frac{\partial^2 F}{\partial x_2 \partial x_1} = 4\tilde{\delta}_+ - 4\tilde{\delta}_-$$

mit

$$\begin{aligned}
\tilde{\alpha}_i &= \tilde{\alpha}_i(x_i) = \alpha_i \cosh 4x_i + \beta_i \sinh 4x_i \,,\\
\tilde{\beta}_i &= \tilde{\beta}_i(x_i) = \alpha_i \sinh 4x_i + \beta_i \cosh 4x_i \,,\\
\tilde{\delta}_\pm &= \tilde{\delta}_\pm(x_1,x_2) = \delta_\pm \cosh(2x_1 \mp 2x_2) + \varepsilon_\pm \sinh(2x_1 \mp 2x_2) \,,\\
\tilde{\varepsilon}_\pm &= \tilde{\varepsilon}_\pm(x_1,x_2) = \delta_\pm \sinh(2x_1 \mp 2x_2) + \varepsilon_\pm \cosh(2x_1 \mp 2x_2) \,,\\
\tilde{\alpha}_{ii} &= \tilde{\alpha}_{ii}(x_i) = \alpha_{ii} \cosh 2x_i + \beta_{ii} \sinh 2x_i \,,\\
\tilde{\beta}_{ii} &= \tilde{\beta}_{ii}(x_i) = \alpha_{ii} \sinh 2x_i + \beta_{ii} \cosh 2x_i \,.
\end{aligned} \qquad (2.2.12)$$

Veselić ([53]) bewies unter Anwendung der Eigenschaften (2.2.2) das folgende wichtige

Lemma 2.2.1:
Es seien die Terme (2.2.10) als Funktionen von A aufgefaßt, d.h. es gelte

$$\alpha_1 = \alpha_1(A),\ \beta_1 = \beta_1(A),\ \ldots,\ \beta_{22} = \beta_{22}(A)\ .$$

Dann folgt

$$\tilde{\alpha}_1 = \alpha_1(\tilde{A}),\ \tilde{\beta}_1 = \beta_1(\tilde{A}),\ \ldots,\ \tilde{\beta}_{22} = \beta_{22}(\tilde{A})\ , \tag{2.2.13}$$

wobei $\tilde{A}$ durch (2.2.6) gegeben ist.

Hieraus erhalten wir für die Kommutatormatrix $\tilde{C} = C(\tilde{A})$ nach (1.3.23),(2.2.10) und (2.2.11)

$$\begin{aligned} \tilde{c}_{2p-1,2q} &= \tilde{c}_{2p-1,2q}(x_1,x_2) = \frac{\partial F}{\partial x_1}(x_1,x_2)\ , \\ \tilde{c}_{2p,2q-1} &= \tilde{c}_{2p,2q-1}(x_1,x_2) = \frac{\partial F}{\partial x_2}(x_1,x_2)\ . \end{aligned} \tag{2.2.14}$$

Wie in § 2.1.1 sind wir auch hier daran interessiert, einen Minimalpunkt der Funktion F zu approximieren, um durch die T-Transformation (2.2.3) eine möglichst gute Normreduzierung zu erreichen. Bei exakter Bestimmung dieses Minimalpunktes verschwinden die Kommutatorelemente $c_{2p-1,2q}$,$c_{2p,2q-1}$ nach der Transformation (vgl.(2.2.14)), d.h. es gilt dann

$$c_{pq}' = 0$$

(siehe auch [42]).

Ähnlich wie die normreduzierenden Funktionen f und g hat auch diese Funktion einige interessante Eigenschaften, die es zunächst zu untersuchen gilt, bevor wir uns der analytischen bzw. numerischen Minimumbestimmung widmen. Man verifiziert leicht die folgenden Aussagen (vgl. [53],[37]):

F ist auf $\mathbb{R}^2$ definiert, konvex und nach unten durch

$$-\alpha_1 - \alpha_2 - \alpha_{11} - \alpha_{22} - \delta_+ - \delta_-$$

beschränkt. Für die Koeffizienten (2.2.10) gelten die Ungleichungen

$$|\beta_i| \le \alpha_i \ , \ |\beta_{ii}| \le \alpha_{ii} \ , \ i=1,2, \tag{2.2.15}$$

$$|\varepsilon_+| \le \delta_+ \ , \ |\varepsilon_-| \le \delta_- \ , \ 4|\beta_i| + |\beta_{ii}| \le 4\alpha_i + \alpha_{ii} \ , \ i=1,2. \tag{2.2.16}$$

F ist streng konvex, wenn mindestens zwei der Terme

$$4\alpha_1+\alpha_{11} \ , \ 4\alpha_2+\alpha_{22} \ , \ \delta_+ \ , \ \delta_-$$

größer als 0 sind, und gleichmäßig konvex, wenn zwei der Ungleichungen (2.2.16) streng sind. Die Hessematrix von F

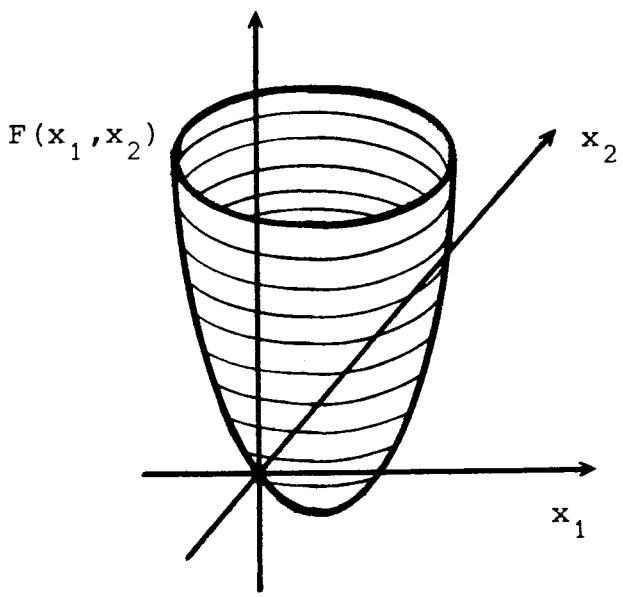
$$H(x_1,x_2) = \left(\frac{\partial^2 F}{\partial x_i \partial x_j} \right)_{i,j=1,2} \tag{2.2.17}$$

ist dann positiv definit bzw. gleichmäßig positiv definit, ansonsten ist H positiv semidefinit.

Im Falle gleichmäßiger Konvexität hat F ein eindeutiges globales Minimum, und es gilt

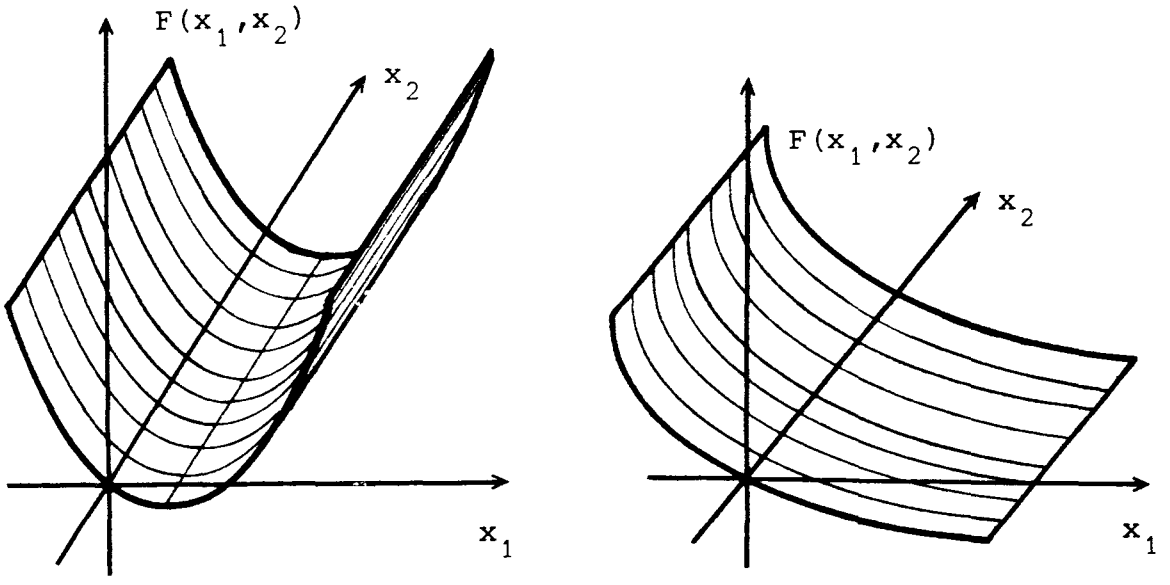
$$F(x_1,x_2) \to \infty \quad \text{für} \quad \| (x_1,x_2)^T \| \to \infty \quad .$$

F hat dann die typische "Kelch"-Gestalt:



Ist F nicht gleichmäßig konvex, so können folgende Fälle auftreten:

(i) F ist nicht streng konvex. F ist dann Funktion nur einer Veränderlichen x_1 bzw. x_2, $x_1 + x_2$ oder $x_1 - x_2$ und als solche wieder streng konvex. Wir unterscheiden die beiden folgenden typischen Gestalten von F:



(2.2.18)

Im ersten Fall sprechen wir von einer *Rinne*, im zweiten von einem *Hang*. Die Funktion F hat in beiden Fällen kein eindeutiges globales Minimum. Liegt eine Rinne vor, so ist jeder Punkt auf der horizontalen Geraden, die am "Boden" der Rinne verläuft, ein Minimum von F. Im Falle eines Hanges ist F Summe von Exponentialfunktionstermen und besitzt daher kein Minimum. Wie in § 2.1.1 sprechen wir hier wieder von einem "*Minimum im Unendlichen*". Die Hessematrix $H(x_1,x_2)$ ist in diesen Fällen singulär, wobei die folgenden fünf verschiedenen Konstellationen für H auftreten können:

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 16\tilde{\alpha}_1+4\tilde{\alpha}_{11} & 0 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ 0 & 16\tilde{\alpha}_2+4\tilde{\alpha}_{22} \end{pmatrix}, \quad \begin{pmatrix} 4\tilde{\delta}_+ & 4\tilde{\delta}_+ \\ 4\tilde{\delta}_+ & 4\tilde{\delta}_+ \end{pmatrix}, \quad \begin{pmatrix} 4\tilde{\delta}_- & -4\tilde{\delta}_- \\ -4\tilde{\delta}_- & 4\tilde{\delta}_- \end{pmatrix}.$$

Die erste Konstellation ergibt $F \equiv 0$, und die übrigen vier entsprechen Ebenendrehungen der in (2.2.18) skizzierten Figuren um jeweils 45°.

Für die Kommutatorelemente $c_{2p-1,2q}$, $c_{2p,2q-1}$ gilt für diese fünf Alternativen der Reihe nach

$$(c_{2p-1,2q}, c_{2p,2q-1}) = \begin{cases} (0,0) \\ (4\beta_1+2\beta_{11}, 0) \\ (0, 4\beta_2+2\beta_{22}) \\ (2\varepsilon_+, 2\varepsilon_+) \\ (2\varepsilon_-, -2\varepsilon_-) \end{cases} . \qquad (2.2.19)$$

(ii) F ist streng konvex. Der Graph von F ist dann eine (additive) Kombination der in (i) beschriebenen Rinnen und Hänge. Dabei sind die Kombinationen von zwei, drei oder vier Hängen verschiedener Richtungen und von einer Rinne mit ein, zwei oder drei Hängen möglich. F besitzt in allen Fällen kein Minimum, und wir sprechen wieder von einem "Minimum im Unendlichen".

Diese Eigenschaften von F legen nun ähnlich wie in § 2.1.1 die Vermutung nahe, daß die Matrix A bei einem "Minimum im Unendlichen" defektive Eigenwerte besitzt, und daß bei einem Minimum am "Boden" einer horizontalen Rinne mehrfache nicht-defektive Eigenwerte von A auftreten. Aus diesem Grund untersuchen wir im folgenden die Fälle, in denen $F \neq 0$ nicht gleichmäßig konvex ist, im Hinblick auf Vielfachheiten und Defektivitäten der Eigenwerte von A.

O.B.d.A. sei dazu $p = 1$, $q = 2$. Wir stellen zunächst die Fälle, in denen der Graph von F eine Kombination von ein bis vier Hängen darstellt, tabellarisch zusammen. Es gilt dabei generell $|\beta_i| = \alpha_i$, $|\beta_{ii}| = \alpha_{ii}$, $i=1,2$ und $|\varepsilon_\pm| = \delta_\pm$, und die möglichen Vorzeichenkonstellationen von $\beta_1, \beta_{11}, \ldots, \varepsilon_-$ sind in den ersten sechs Spalten der nachfolgenden Tabelle aufgeführt. Diese Auflistung ist vollständig in dem Sinne, daß alle nicht erwähnten Konstellationen durch J-orthogonale Ähnlichkeitstransformationen von A mit den Matrizen

$$V^{(1)}, V^{(2)}, V^{(3)}, V^{(4)}, P^{(1,3)}, P^{(2,4)} \qquad (2.2.20)$$

auf die verzeichneten Fälle zurückgeführt werden können.

In allen aufgeführten Fällen besitzt F ein "Minimum im Unendlichen", und mit Hilfe einiger elementarer Rechnungen läßt sich jeweils nachweisen, daß die Eigenwerte $\lambda_1,\ldots,\lambda_4$ von A mit den Eigenwerten des 4x4-Blocks $\hat{A}_{12}$ identisch und dabei paarweise doppelt reell sind, es gilt

$$\lambda_{1,2} = \frac{a_{11}+a_{44}}{2} \quad , \quad \lambda_{3,4} = \frac{a_{22}+a_{33}}{2} \quad .$$

Die Tabelle gibt nun Aufschluß darüber, ob die Eigenwerte notwendig defektiv sind oder auch nicht-defektiv sein können. Ein *D* in der siebten Spalte bedeutet, daß zumindest für eines der beiden Eigenwertpaare die Defektivität bewiesen werden kann, wobei der Beweis analog zu dem Vorgehen in § 2.1.1 unter Betrachtung der Resolvente $(A-\lambda I)^{-1}$ von A geführt wird. Ein *ND* wiederum besagt, daß Beispiele existieren, in denen beide Eigenwertpaare und die übrigen Eigenwerte der Matrix nicht-defektiv sind. Es würde den Rahmen dieser Arbeit sprengen, alle Beispiele anzuführen. Wir haben daher im Anhang einige charakteristische 6x6-Beispiele zusammengestellt.

	Signum							
	β_1	β_{11}	β_2	β_{22}	ε_+	ε_-	D/ND	Anhang
1 Hang	+1	0	0	0	0	0	D	
	0	+1	0	0	0	0	ND	(A1)
	+1	+1	0	0	0	0	ND	
	0	0	0	0	+1	0	ND	(A2)
2 Hänge	+1	0	+1	0	0	0	D	
	+1	0	0	+1	0	0	D	
	0	+1	0	+1	0	0	ND	
	+1	0	+1	+1	0	0	D	
	+1	+1	0	+1	0	0	ND	
	+1	+1	+1	+1	0	0	ND	(A3)
	+1	0	0	0	+1	0	D	
	+1	0	0	0	-1	0	D	
	0	+1	0	0	+1	0	ND	
	0	+1	0	0	-1	0	ND	
	+1	+1	0	0	+1	0	ND	
	+1	+1	0	0	-1	0	ND	
	0	0	0	0	+1	+1	D	
3 Hänge	+1	0	+1	0	+1	0	D	
	+1	0	-1	0	-1	0	D	
	+1	0	0	+1	+1	0	D	
	+1	0	0	-1	+1	0	ND	
	+1	0	0	-1	-1	0	D	
	0	+1	0	+1	+1	0	ND	
	0	+1	0	-1	+1	0	ND	

	Signum							
	β_1	β_{11}	β_2	β_{22}	ε_+	ε_-	D/ND	Anhang
3 Hänge	+1	0	+1	+1	+1	0	D	
	+1	0	-1	-1	+1	0	ND	
	+1	0	-1	-1	-1	0	D	
	+1	+1	0	+1	+1	0	ND	
	+1	+1	0	-1	+1	0	ND	
	+1	+1	0	-1	-1	0	ND	
	+1	+1	+1	+1	+1	0	ND	(A4)
	+1	+1	-1	-1	+1	0	ND	
	+1	0	0	0	+1	+1	ND	
	+1	0	0	0	+1	-1	D	
	0	+1	0	0	+1	+1	ND	
	0	+1	0	0	+1	-1	ND	
	+1	+1	0	0	+1	+1	ND	(A5)
	+1	+1	0	0	+1	-1	ND	
4 Hänge	+1	0	+1	0	+1	-1	D	
	+1	0	0	+1	+1	+1	ND	
	+1	0	0	+1	+1	-1	D	
	0	+1	0	+1	+1	+1	ND	
	+1	0	+1	+1	+1	+1	ND	
	+1	0	+1	+1	+1	-1	D	
	+1	+1	0	+1	+1	+1	ND	
	+1	+1	0	+1	+1	-1	ND	
	+1	+1	+1	+1	+1	+1	ND	(A6)

Somit ist bereits der erste Teil der obigen Vermutung widerlegt. Auch der zweite Teil dieser Vermutung bewahrheitet sich nicht, wie die nachfolgende Untersuchung der Fälle, in denen der Graph von F eine Rinne darstellt, zeigt.

Wir unterscheiden hier die Fälle

$$\begin{aligned} &\text{(i)} && \beta_1 = \alpha_1 > 0 \ , \ -\beta_{11} = \alpha_{11} > 0 \ , \\ &\text{(ii)} && 4|\beta_1| + |\beta_{11}| < 4\alpha_1 + \alpha_{11} \ , \\ &\text{(iii)} && |\varepsilon_+| < \delta_+ \ , \end{aligned} \tag{2.2.21}$$

wobei jeweils die restlichen Koeffizienten (2.2.10) identisch 0 sind. Alle übrigen Konstellationen lassen sich wieder durch J-orthogonale Ähnlichkeitstransformationen mit den Matrizen (2.2.20) auf die Fälle (i) - (iii) zurückführen. Nach einigen einfachen Rechnungen ergibt sich hier für (i) und (ii), daß

$$\lambda_{1,2} = a_{22}$$

doppelter reeller Eigenwert von A ist, und für (iii) erhalten wir wieder, daß die Eigenwerte $\lambda_1,\ldots,\lambda_4$ von A mit den Eigenwerten des 4x4-Blocks $\hat{A}_{12}$ identisch sind. Sie sind paarweise doppelt und berechnen sich als

$$\lambda_{1,2,3,4} = \frac{a_{11}+a_{22}}{2} \pm \sqrt{\frac{1}{4}(a_{11}-a_{22})^2 - a_{12}^2 + a_{13}^2} \ ,$$

d.h. sie können reell und komplex auftreten. Es läßt sich für alle drei Fälle leicht nachweisen, daß die Eigenwerte $\lambda_{1,2}$ bzw. $\lambda_{3,4}$ nicht-defektiv sind, wenn ihre Vielfachheit genau 2 ist. Wenn sie jedoch größer als 2 ist, so können diese Eigenwerte defektiv sein, wie die Beispiele (A7) - (A9) im Anhang zeigen.

Wir schließen die Untersuchung ab, indem wir noch die Fälle betrachten, in denen der Graph von F eine Kombination von einer Rinne mit ein bis drei Hängen darstellt. Die nachfolgende Auflistung ist vollständig, da sich wieder alle nicht verzeichneten Konstellationen durch J-orthogonale Ähnlichkeitstransformationen mit den Matrizen

(2.2.20) aus den aufgeführten ergeben. Die Tabelle ist so zu verstehen, daß entweder generell $|\beta_i| = \alpha_i$, $|\beta_{ii}| = \alpha_{ii}$, $|\varepsilon_\pm| = \delta_\pm$ gilt, wobei dann das Vorzeichen von $\beta_i, \beta_{ii}, \varepsilon_\pm$ angegeben wird, oder aber es gilt $4|\beta_1| + |\beta_{11}| < 4\alpha_1 + \alpha_{11}$ bzw. $|\varepsilon_+| < \delta_+$, was wir durch "<" in der entsprechenden Spalte kennzeichnen.

F besitzt in allen Fällen ein "Minimum im Unendlichen", und es läßt sich nachweisen, daß in den Fällen, die die Konstellationen (2.2.21)(i),(ii) beinhalten, die Matrix A den doppelten reellen Eigenwert

$$\lambda_{1,2} = \frac{a_{22}+a_{33}}{2} \tag{2.2.22}$$

besitzt. In den Fällen mit $|\varepsilon_+| < \delta_+$ kann man zeigen, daß die Eigenwerte $\lambda_1,\ldots,\lambda_4$ von A mit den Eigenwerten von $\hat{A}_{12}$ identisch sind und sich als paarweise doppelt (reell oder komplex) ergeben:

$$\lambda_{1,2,3,4} = \frac{a_{11}+a_{44}}{4} + \frac{a_{22}+a_{33}}{4} \pm \sqrt{\frac{1}{16}(a_{11}+a_{44}-a_{22}-a_{33})^2 - a_{12}a_{34} - a_{13}a_{24}}\,. \tag{2.2.23}$$

Wenn in der Tabelle ein *D* aufgeführt ist, so kann für das Eigenwertpaar (2.2.22) bzw. für mindestens eins der beiden Eigenwertpaare (2.2.23) mit Hilfe der Resolvente von A die Defektivität bewiesen werden, und ein *ND* gibt an, daß Beispiele existieren, in denen die Matrix A ausschließlich nicht-defektive Eigenwerte besitzt. Einige typische Beispiele der Dimension 6 sind wieder im Anhang aufgeführt.

	β_1	β_{11}	β_2	β_{22}	ε_+	ε_-	D/ND	Anhang
	+1	-1	+1	0	0	0	D	
	<		+1	0	0	0	D	
	+1	-1	0	+1	0	0	ND	
	<		0	+1	0	0	ND	
	+1	-1	+1	+1	0	0	ND	(A10)
1 Rinne	<		+1	+1	0	0	ND	
und	+1	-1	0	0	+1	0	ND	
1 Hang	-1	+1	0	0	+1	0	ND	
	<		0	0	+1	0	ND	(A11)
	+1	0	0	0	<	0	D	
	0	+1	0	0	<	0	ND	
	+1	+1	0	0	<	0	ND	(A12)
	0	0	0	0	<	+1	D	

	β_1	β_{11}	β_2	β_{22}	ε_+	ε_-	D/ND	Anhang
	+1	-1	+1	0	+1	0	ND	
	-1	+1	+1	0	+1	0	ND	
	<		+1	0	+1	0	ND	
	+1	-1	0	+1	+1	0	ND	
	-1	+1	0	+1	+1	0	ND	
1 Rinne	<		0	+1	+1	0	ND	
und	+1	-1	+1	+1	+1	0	ND	
2 Hänge	-1	+1	+1	+1	+1	0	ND	
	<		+1	+1	+1	0	ND	
	+1	-1	0	0	+1	-1	ND	
	<		0	0	+1	-1	ND	
	+1	0	-1	0	<	0	ND	
	+1	0	0	-1	<	0	ND	

	β_1	β_{11}	β_2	β_{22}	ε_+	ε_-	D/ND	Anhang
	0	+1	0	-1	<	0	ND	
	+1	0	-1	-1	<	0	ND	
1 Rinne	+1	+1	0	-1	<	0	ND	
und	+1	+1	-1	-1	<	0	ND	
2 Hänge	+1	0	0	0	<	+1	D	
	0	+1	0	0	<	+1	ND	
	+1	+1	0	0	<	+1	ND	(A13)

	β_1	β_{11}	β_2	β_{22}	ε_+	ε_-	D/ND	Anhang
	+1	-1	+1	0	+1	-1	ND	
		<	+1	0	+1	-1	ND	
	+1	-1	0	+1	+1	-1	ND	
		<	0	+1	+1	-1	ND	
	+1	-1	+1	+1	+1	-1	ND	(A14)
1 Rinne		<	+1	+1	+1	-1	ND	
und	+1	0	-1	0	<	+1	ND	
3 Hänge	+1	0	0	-1	<	+1	ND	
	0	+1	0	-1	<	+1	ND	
	+1	0	-1	-1	<	+1	ND	
	+1	+1	0	-1	<	+1	ND	
	+1	+1	-1	-1	<	+1	ND	(A15)

Wir haben damit für alle möglichen Fälle, in denen F nicht gleichmäßig konvex ist, mehrfache Eigenwerte von A nachgewiesen, und wir erhalten als Konsequenz das folgende

Lemma 2.2.2:
A habe getrennte Eigenwerte. Dann ist F gleichmäßig konvex, und es gilt

$$\lim_{\|(x_1,x_2)^T\| \to \infty} F(x_1,x_2) = \infty \quad . \tag{2.2.24}$$

Jedoch hat die Untersuchung unsere obige Vermutung über die Eigenschaften von F im Zusammenhang mit mehrfachen Eigenwerten von A widerlegt, und es stellt sich wie in § 2.1.1 die generelle Frage, ob wir bei einem gut-konditionierten Eigenwertproblem mit nicht-defektiven Eigenwerten große Parameter (x_1,x_2) für die T-Transformation, die sich durch die Approximation eines "Minimums im Unendlichen" ergeben können, zulassen sollen. Dabei beachte man, daß F selbst bei fast-normalen Matrizen A ein "Minimum im Unendlichen" besitzen kann, wie das Beispiel (A1) im Anhang zeigt (diese Matrix konvergiert für $\omega \to 0$ gegen Normalität). Wir werden diese Problematik der großen Parameter im folgenden Paragraphen diskutieren.

Zum Schluß dieses Abschnitts wollen wir nun noch eine zweite Darstellung der Funktion F untersuchen. Mit Hilfe der Variablentransformation

$$t_i = \tanh x_i \quad , \; i = 1,2$$

erhalten wir aus (2.2.9) nach einigen elementaren Umformungen

$$\begin{aligned} F(x_1,x_2) = G(t_1,t_2) = {} & 8\alpha_1 \frac{t_1^2}{(1-t_1^2)^2} + 4\beta_1 \frac{t_1(1+t_1^2)}{(1-t_1^2)^2} + 2\alpha_{11} \frac{t_1^2}{1-t_1^2} + 2\beta_{11} \frac{t_1}{1-t_1^2} \\ & + 8\alpha_2 \frac{t_2^2}{(1-t_2^2)^2} + 4\beta_2 \frac{t_2(1+t_2^2)}{(1-t_2^2)^2} + 2\alpha_{22} \frac{t_2^2}{1-t_2^2} + 2\beta_{22} \frac{t_2}{1-t_2^2} \\ & + 2\delta_+ \frac{(t_1+t_2)^2}{(1-t_1^2)(1-t_2^2)} + 2\varepsilon_+ \frac{(t_1+t_2)(1+t_1t_2)}{(1-t_1^2)(1-t_2^2)} \\ & + 2\delta_- \frac{(t_1-t_2)^2}{(1-t_1^2)(1-t_2^2)} + 2\varepsilon_- \frac{(t_1-t_2)(1-t_1t_2)}{(1-t_1^2)(1-t_2^2)} \quad . \end{aligned} \tag{2.2.25}$$

Wegen $|\tanh x_i| < 1$ ist G auf dem Gebiet $D = (-1,1)\times(-1,1)$ definiert. Analog zu den Betrachtungen der Funktionen f und g in § 2.1.1 ist G bildlich gesprochen wieder durch gleichmäßiges "Zusammendrükken" der Funktion F entstanden. Die partiellen Ableitungen von G berechnen sich als

$$\begin{aligned} & \frac{\partial G}{\partial t_i} = \frac{\partial F}{\partial x_i} \frac{1}{1-t_i^2} \quad , \quad \frac{\partial^2 G}{\partial t_i^2} = \frac{\partial^2 F}{\partial x_i^2} \frac{1}{(1-t_i^2)^2} + \frac{\partial F}{\partial x_i} \frac{2t_i}{(1-t_i^2)^2} \quad , \; i=1,2 \; , \\ & \frac{\partial^2 G}{\partial t_1 \partial t_2} = \frac{\partial^2 G}{\partial t_2 \partial t_1} = \frac{\partial^2 F}{\partial x_1 \partial x_2} \frac{1}{(1-t_1^2)(1-t_2^2)} \quad , \end{aligned} \tag{2.2.26}$$

und man überlegt sich leicht, daß sich die Minimumeigenschaften von F auf G übertragen, auch die Aussagen über die "Minima im Unendlichen". Diese entsprechen hier "Minima auf dem Rand von D". Jedoch ist die Funktion G nicht mehr notwendig konvex, wie das folgende Beispiel zeigt:

Für $\varepsilon_+ = \delta_+ = 1$, alle übrigen Koeffizienten identisch O, gilt

$$F(x_1,x_2) = e^{2x_1+2x_2} - 1 \ , \quad G(t_1,t_2) = 2\,\frac{t_1+t_2}{(1-t_1)(1-t_2)} \ ,$$

F ist konvex, G ist nicht konvex.

Analog zu § 2.1.1 nennen wir die Funktionen $F(x_1,x_2)$ und $G(t_1,t_2)$ aufgrund ihrer Anwendung *normreduzierende Funktionen*. Auch hier ist die exakte Minimumbestimmung für F bzw. G nur in Spezialfällen möglich, da diese generell auf ein homogenes Gleichungssystem (mit 2 Gleichungen und 2 Unbekannten) mindestens 4. Grades führt. Wir untersuchen daher im folgenden Paragraphen die Minimumbestimmung mittels numerischer Methoden.

2.2.2 NEWTON-ITERATION ZUR MINIMIERUNG VON $F(x_1,x_2)$ UND $G(t_1,t_2)$

Wir betrachten zunächst die Funktion F. Aufgrund der Konvexitätseigenschaften von F bietet sich hier wie in § 2.1.2 ein Newton-ähnliches Iterationsverfahren zur Minimumbestimmung an. Man beachte dabei wieder den Aspekt, daß die lokal quadratische Konvergenz eines solchen Verfahrens die Voraussetzungen für die asymptotisch quadratische Konvergenz der gesamten Matrixiteration schafft.

Es sei $\| C_{pq} \| > 0$. Wir definieren dann die Iterierten

$$x^{(k)} = (x_1^{(k)}, x_2^{(k)})^T \quad , \quad k \in \mathbb{N}_0$$

alternativ durch (i) oder (ii):

(i) F ist gleichmäßig konvex:

$$x^{(0)} = (0,0) \quad , \quad x^{(k+1)} = x^{(k)} - \gamma_k \, H(x^{(k)})^{-1} \begin{pmatrix} \tilde{c}_{2p-1,2q}(x^{(k)}) \\ \tilde{c}_{2p,2q-1}(x^{(k)}) \end{pmatrix}, k=0,1,\ldots, \tag{2.2.27}$$

wobei γ_k jeweils als erste Zahl der Folge $(2^{-j})_{j\in\mathbb{N}_0}$ bestimmt wird, so daß

$$F(x^{(k+1)}) - F(x^{(k)}) \leq -\frac{1}{3}\gamma_k \begin{pmatrix} \tilde{c}_{2p-1,2q}(x^{(k)}) \\ \tilde{c}_{2p,2q-1}(x^{(k)}) \end{pmatrix}^T H(x^{(k)})^{-1} \begin{pmatrix} \tilde{c}_{2p-1,2q}(x^{(k)}) \\ \tilde{c}_{2p,2q-1}(x^{(k)}) \end{pmatrix} \tag{2.2.28}$$

gilt.

(ii) F ist nicht gleichmäßig konvex:

$x^{(0)} = (0,0)$. Falls $|\tilde{c}_{2p-1,2q}(x^{(k)})| \geq |\tilde{c}_{2p,2q-1}(x^{(k)})|$, dann

$$x_1^{(k+1)} = x_1^{(k)} + \operatorname{artanh}\left(- \frac{\tilde{c}_{2p-1,2q}(x^{(k)})}{16\tilde{\alpha}_1(x_1^{(k)}) + 4\tilde{\alpha}_{11}(x_1^{(k)}) + 4\tilde{\delta}_+(x_1^{(k)}, x_2^{(k)}) + 4\tilde{\delta}_-(x_1^{(k)}, x_2^{(k)})} \right),$$
$$x_2^{(k+1)} = x_2^{(k)} \quad , \tag{2.2.29}$$

sonst

$$x_1^{(k+1)} = x_1^{(k)} \quad , \tag{2.2.30}$$

$$x_2^{(k+1)} = x_2^{(k)} + \operatorname{artanh}\left(- \frac{\tilde{c}_{2p,2q-1}(x^{(k)})}{16\tilde{\alpha}_2(x_2^{(k)}) + 4\tilde{\alpha}_{22}(x_2^{(k)}) + 4\tilde{\delta}_+(x_1^{(k)}, x_2^{(k)}) + 4\tilde{\delta}_-(x_1^{(k)}, x_2^{(k)})} \right)$$

für $k = 0,1,\ldots$.

Dabei ist aufgrund der Voraussetzungen die Matrix $H(x^{(k)})$ in (2.2.27) stets invertierbar, und die Nenner in (2.2.29) und (2.2.30) sind ungleich 0. Nach (2.2.14) berechnen sich die Iterierten direkt aus (2.2.11) und (2.2.12), ohne daß die einzelnen T-Transformationen für jedes Parameterpaar $x^{(k)}$ explizit ausgeführt werden.

Mit der alternativen Bestimmung der Iterierten in (i) und (ii) tragen wir der Tatsache Rechnung, daß F bei mehrfachen Eigenwerten von A nicht notwendig gleichmäßig konvex sein muß und in diesem Fall die Hessematrix H singulär sein kann (vgl. § 2.2.1). In der numerischen Praxis entscheidet man durch Abprüfen der Terme

$$\delta_+ - |\varepsilon_+|, \ \delta_- - |\varepsilon_-|, \ 4\alpha_1 + \alpha_{11} - 4|\beta_1| - |\beta_{11}|, \ 4\alpha_2 + \alpha_{22} - 4|\beta_2| - |\beta_{22}|,$$

nach welcher Strategie die Iteration durchgeführt wird.

Der Iterationsprozess (2.2.27) wird gemeinhin als *gedämpftes Newton-Verfahren* bezeichnet (vgl.[37, § 8.2]), wobei die spezielle Form der Bestimmung der *Dämpfungsfaktoren* γ_k von Goldstein ([15]), Armijo ([1]) und Elkin([11]) stammt. Die Existenz der γ_k wird durch die gleichmäßige Konvexität von F garantiert (vgl. [37,Lemma 8.3.2 und S.491]), und man beachte, daß ihre Berechnung durch einen endlichen Prozess erfolgt. Die Iteration gemäß (2.2.29) bzw. (2.2.30) geht auf Eberlein ([8],[9]) zurück. Wie man leicht nachweist, entspricht sie der wiederholten Durchführung des ersten Schrittes des gewöhnlichen Newton-Verfahrens zur Minimierung der (einparametrigen) Funktion $G(t_1,0)$ bzw. $G(0,t_2)$ (vgl. auch § 2.1.2).

Beide Prozesse (i) und (ii) garantieren in jedem Iterationsschritt einen Abstieg in Richtung des Infimums von F. Für (i) folgt dies aus (2.2.28), da $H(x^{(k)})$ und damit $H(x^{(k)})^{-1}$ positiv definit ist, und für

(ii) ergibt sich aus einer Abschätzung von Eberlein ([8]) zusammen mit Lemma 2.2.1

$$F(x^{(k+1)}) - F(x^{(k)}) \leq -\frac{1}{3} \frac{\tilde{c}_{2p-1,2q}^{\,2}(x^{(k)})}{16\tilde{\alpha}_1(x_1^{(k)}) + 4\tilde{\alpha}_{11}(x_1^{(k)}) + 4\tilde{\delta}_+(x_1^{(k)}, x_2^{(k)}) + 4\tilde{\delta}_-(x_1^{(k)}, x_2^{(k)})} \quad (2.2.31)$$

bzw.

$$F(x^{(k+1)}) - F(x^{(k)}) \leq -\frac{1}{3} \frac{\tilde{c}_{2p,2q-1}^{\,2}(x^{(k)})}{16\tilde{\alpha}_2(x_2^{(k)}) + 4\tilde{\alpha}_{22}(x_2^{(k)}) + 4\tilde{\delta}_+(x_1^{(k)}, x_2^{(k)}) + 4\tilde{\delta}_-(x_1^{(k)}, x_2^{(k)})} \quad . (2.2.32)$$

Würde man die Iteration (2.2.27) generell mit $\gamma_k = 1$ durchführen, so wäre sie identisch mit der gewöhnlichen Newton-Iteration. Dann jedoch bestünde die Gefahr des sogenannten "*Overshootings*", d.h. die einzelnen Iterationsschritte erfolgen zwar in Abstiegsrichtung, aber man schießt über das Minimum hinaus (bzw. daran vorbei), so daß nicht notwendig $F(x^{(k+1)}) \leq F(x^{(k)})$ gilt.

Lemma 2.2.3:
Für die Funktion $F(x_1, x_2)$ (vgl.(2.2.9)) seien die Iterierten $x^{(k)}$, $k \in \mathbb{N}_0$ des modifizierten Newton-Verfahrens durch (2.2.27) bzw. (2.2.29),(2.2.30) bestimmt. Dann gilt

$$F(x^{(k)}) \to \inf F \quad \text{für } k \to \infty \quad . \qquad (2.2.33)$$

Beweis:
Der Beweis wird getrennt für die Fälle (i) und (ii) geführt. In (i) ist F gleichmäßig konvex und besitzt daher ein eindeutiges globales Minimum x^*. Elkin ([11]) bewies unter diesen Voraussetzungen

$$\lim_{k \to \infty} x^{(k)} = x^* \quad .$$

Mit der Stetigkeit von F folgt hieraus die Behauptung.

In (ii) erhalten wir wie im Beweis von Lemma 2.1.3 aufgrund von (2.2.31) und (2.2.32)

$$F(x^{(k)}) \to \xi, \quad \xi \in \mathbb{R}^2 \quad \text{für} \quad k \to \infty$$

und damit wegen (2.2.14)

$$\text{grad } F(x^{(k)}) \to 0 , \quad k \to \infty .$$

Aus der Konvexität von F ergibt sich hieraus dann wieder die Behauptung. ■

Wie in § 2.1.2 unterscheiden wir auch hier eine *Standard-Normreduzierung*, in der wir nur einen Iterationsschritt (2.2.27) bzw. (2.2.29) oder (2.2.30) durchführen, und eine *Optimale Normreduzierung*, in der die Iteration solange ausgeführt wird, bis das Infimum von F hinreichend genau approximiert wurde, d.h. bis $\| \tilde{C}_{pq}(x^{(k)}) \|$ für ein $k \in \mathbb{N}$ eine vorgegebene Schwelle unterschritten hat. Die garantierte Normabnahme von A bei einer T-Transformation mit den Parametern $(x_1,x_2) = (x_1^{(1)},x_2^{(1)})$ der Standardvariante kann wegen

$$F(x^{(1)}) - F(x^{(0)}) = \frac{1}{4}(\| A' \|^2 - \| A \|^2)$$

durch (2.2.28) bzw. (2.2.31),(2.2.32) abgeschätzt werden, und bei einer T-Transformation mit den Parametern $(x_1^{(k)},x_2^{(k)})$ der optimalen Variante ist die Normabnahme wegen

$$F(x^{(k)}) \leq F(x^{(1)})$$

zumindest genauso stark. Die allgemeinen Vor- und Nachteile dieser beiden Varianten ergeben sich sinngemäß aus der Diskussion der analogen Methoden in § 2.1.2 .

Abgesehen von einigen pathologischen Fällen wird die Funktion F in der Praxis gleichmäßig konvex sein, so daß die Iteration gemäß (2.2.27) erfolgt. Dabei ist die Berechnung der Dämpfungsfaktoren γ_k weniger aufwendig, als es zunächst scheint. Elkin ([11]) bewies, daß abhängig vom Startwert $x^{(0)}$ ein $k_0 \in \mathbb{N}$ existiert, so daß $\gamma_k = 1$ für $k \geq k_0$ gilt, und in § 3.2 zeigen wir, daß bei Anwendung dieser Iteration auf Matrizen, die in einem gewissen Sinne fast Murnaghan-Form besitzen, automatisch $\gamma_k = 1$ folgt.

Es deutet vieles darauf hin, daß durch die spezielle Gestalt von F die Ungleichung (2.2.28) generell mit $\gamma_k = 1$ erfüllt ist und damit immer schon für diesen Wert eine Normabnahme erfolgt. Einen Beweis dieser Vermutung haben wir leider nicht erbringen können, jedoch haben wir auch kein Beispiel konstruieren können, welches die Vermutung widerlegt, und auch experimentelle numerische Tests (wir kommen hierauf in § 5.1 zurück) haben keinen Widerspruch erbracht.

Somit erscheint diese modifizierte Newton-Iteration in der Praxis weniger aufwendig als der analoge Prozess von Veselić ([53]), der in jedem Iterationsschritt eine doppelte Parameterberechnung gemäß (2.2.27) mit $\gamma_k = 1$ und (2.2.29) bzw. (2.2.30) erfordert. In § 5.2 haben wir diese beiden Methoden hinsichtlich ihrer numerischen Effizienz verglichen.

Aus den Untersuchungen des vorangegangenen Paragraphen ergibt sich auch für die T-Transformation das Problem, daß große Parameter (x_1, x_2) auftreten können. Bei der Standard-Normreduzierung ist dies zwar nicht der Fall, da für die Transformationsparameter $(x_1^{(1)}, x_2^{(2)})$ die Abschätzungen

$$|x_i^{(1)}| \leq 1 \quad \text{bzw.} \quad |x_i^{(1)}| \leq \operatorname{artanh} \frac{1}{2} \leq 0.55 \ , \ i=1,2$$

(vgl.[53],[8]) gelten. Jedoch können bei Anwendung der optimalen Normreduzierung die Parameter stark anwachsen, insbesondere dann, wenn die Eigenwerte von A mehrfach bzw. fast-mehrfach, d.h. schlecht getrennt sind. Die schon in § 2.1.2 zitierten Ergebnisse von Sacks-Davis ([45]) und Wilkinson ([60]) besagen, daß hierdurch die numerische Stabilität der Berechnung von A' nicht beeinträchtigt wird, da die T-Transformation eine Normreduzierung bewirkt. Andrerseits sind wir aber auch daran interessiert, die Eigenvektoren von A stabil zu berechnen. Da nun aus großen Transformationsparametern x_i eine große Kondition der Matrix T resultiert, welche wiederum gefährlich für die Stabilität der Eigenvektoren sein kann, werden wir in der Praxis (vgl. Algorithmus 4.1.1)

$$|x_i^{(k)}| \leq 1 \ , \ i=1,2$$

fordern, d.h. die optimale Normreduzierung (egal ob mit (2.2.27) oder (2.2.29),(2.2.30)) dann stoppen, wenn $|x_1^{(k+1)}|$ oder $|x_2^{(k+1)}|$ größer als 1 wird. Man beachte, daß damit zumindest 1 Iterationsschritt ausgeführt werden kann.

Wir betrachten als nächstes die numerische Minimierung der Funktion $G(t_1,t_2)$. Da G nicht notwendig konvex ist, läßt sich die Iteration (2.2.27) nicht ohne weiteres auf diese Funktion übertragen. Wenn wir uns jedoch auf den ersten Schritt dieser Iteration mit $\gamma_k = 1$, d.h. auf den ersten Schritt des gewöhnlichen Newton-Verfahrens beschränken, so erhalten wir wegen (2.2.26) für $t^{(0)} = (0,0)$

$$H(0,0)\begin{pmatrix} t_1^{(1)} \\ t_2^{(1)} \end{pmatrix} = -\begin{pmatrix} c_{2p-1,2q} \\ c_{2p,2q-1} \end{pmatrix}, \qquad (2.2.34)$$

wobei H durch (2.2.17) definiert ist, und damit, vorausgesetzt $H(0,0)$ ist nicht-singulär,

$$t_1^{(1)} = -\frac{(16\alpha_2+4\alpha_{22}+4\delta_+ +4\delta_-)c_{2p-1,2q} - (4\delta_+ -4\delta_-)c_{2p,2q-1}}{\det H(0,0)},$$

$$t_2^{(1)} = -\frac{(16\alpha_1+4\alpha_{11}+4\delta_+ +4\delta_-)c_{2p,2q-1} - (4\delta_+ -4\delta_-)c_{2p-1,2q}}{\det H(0,0)}. \qquad (2.2.35)$$

Also berechnen sich $(t_1^{(1)},t_2^{(1)})$ nach den gleichen Formeln wie $(x_1^{(1)},x_2^{(1)})$ bei der Standard-Normreduzierung für F (mit $\gamma_k = 1$). Der Vorteil der Parameterbestimmung (2.2.35) ist jedoch der, daß die Komponenten der Matrix T schneller und stabiler aus $t_i^{(1)}$ als aus $x_i^{(1)}$ bestimmt werden können.

Diese Art der Parameterbestimmung stellt eine direkte Verallgemeinerung der Strategie von Eberlein ([8],[9]) dar. Damit erhebt sich sofort die Frage, ob auch dieser Ansatz bei einer T-Transformation mit den Parametern $x_i = \operatorname{artanh} t_i^{(1)}$ eine garantierte Normabnahme von A bewirkt, und ob sich die diesbezügliche Abschätzung von [8] analog verallgemeinern läßt.

Diese erwünschte Abschätzung wirft jedoch Probleme auf. Zunächst einmal ergibt sich im Gegensatz zu der Eberlein-Methode keine natürliche Beschränkung der Parameter, es gilt lediglich (vgl.[53])

$$|t_i^{(1)}| \leq 1 \ , \quad i=1,2.$$

Also existieren Fälle, in denen in (2.2.35) $t_i^{(1)} = 1$ gilt und damit $(t_1^{(1)}, t_2^{(1)})$ auf dem Rand des (offenen) Definitionsbereichs D von G liegt, d.h. die Komponenten der Matrix T sind nicht definiert.

Wir haben nachgewiesen, daß hier exakt 8 Konstellationen der Koeffizienten (2.2.10) möglich sind. Diese können sämtlich durch J-orthogonale Ähnlichkeitstransformationen mit den Matrizen (2.2.20) auf den Fall $t_1^{(1)} = 1$, $t_2^{(1)} = \frac{1}{2}$ mit

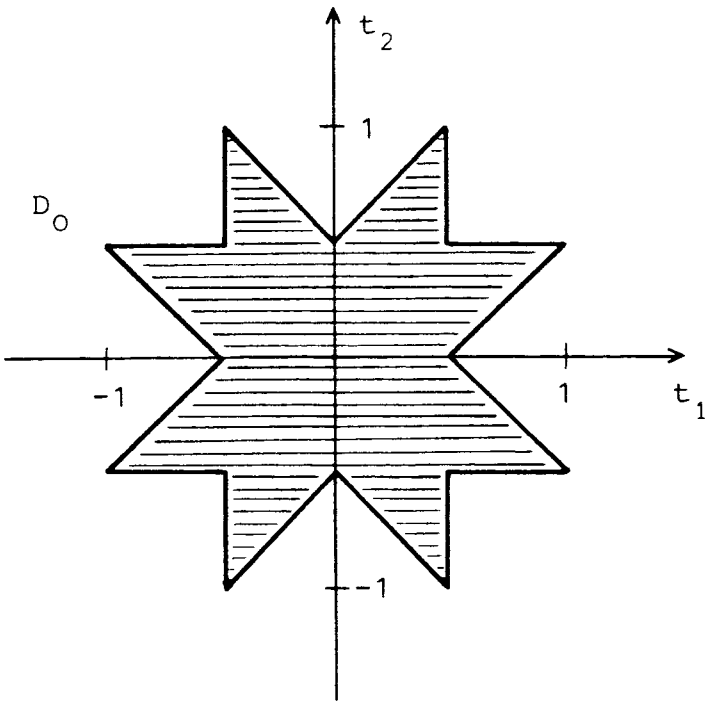
$$-\beta_{22} = \alpha_{22} > 0 \ , \quad -\varepsilon_- = \delta_- > 0 \ ,$$

alle übrigen Koeffizienten identisch 0, zurückgeführt werden. Dabei zeigt das Beispiel der Matrix

$$A(\omega) = \left(\begin{array}{cccc|cc} 1 & \omega & \omega & 0 & 0 & 0 \\ -\omega & 2 & 0 & \omega & \omega & \omega \\ \omega & 0 & 2 & -\omega & \omega & \omega \\ 0 & \omega & \omega & 1 & 0 & 0 \\ \hline 0 & -\omega & \omega & 0 & 0 & 0 \\ 0 & \omega & -\omega & 0 & 0 & 0 \end{array}\right) \ , \ \omega \in \mathbb{R}$$

mit den nicht-defektiven Eigenwerten $\lambda_{1,2} = 1$, $\lambda_{3,4} = 2$, $\lambda_{5,6} = 0$ (vgl. § 2.2.1), daß dieser Fall selbst bei fast-normalen Matrizen auftreten kann (für $\omega \to 0$ konvergiert $A(\omega)$ gegen Normalität).

Wir haben weiterhin nachgewiesen, daß $(t_1^{(1)}, t_2^{(1)})$ notwendig in der abgeschlossenen Menge $D_0 \subset \overline{D}$ liegt, wobei D_0 die folgende "Rosetten"-Gestalt besitzt:



Und wir haben andrerseits gezeigt, daß für jeden beliebigen Punkt $(t_1,t_2) \in D_O$ eine Koeffizientenkonstellation in (2.2.35) existiert, so daß $(t_1^{(1)},t_2^{(1)}) = (t_1,t_2)$ gilt, mit anderen Worten, D_O ist "voll".

Leider haben wir aber keinen generellen Beweis für die garantierte Normreduzierung erbringen können, auch nicht unter einer erzwungenen Parameterbeschränkung $|t_i| \leq \frac{1}{2}$, die die "Hörner" von D_O abschneidet. Es gilt jedoch das folgende

Lemma 2.2.4:
A sei eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$, und für ein festes Pivotpaar (p,q) mit $1 \leq p < q \leq m$ gelte

$$a_{2p-1,2q-1} = a_{2p,2q} = 0 \; . \tag{2.2.36}$$

Es sei G durch (2.2.25) und H durch (2.2.17) definiert, und die Parameter $t_1^{(1)},t_2^{(1)}$ seien durch (2.2.35) bestimmt, dann folgt

$$G(t_1^{(1)},t_2^{(1)}) = \frac{1}{4}\left(\|A'\|^2-\|A\|^2\right) \leq -\frac{1}{27}\begin{pmatrix} c_{2p-1,2q} \\ c_{2p,2q-1} \end{pmatrix}^T H(0,0)^{-1} \begin{pmatrix} c_{2p-1,2q} \\ c_{2p,2q-1} \end{pmatrix} . \tag{2.2.37}$$

Der Beweis soll hier nur angedeutet werden. Aus (2.2.36) erhalten wir $\varepsilon_+ = \varepsilon_- = 0$. Mit Hilfe von (2.2.15) ergibt sich dann nach einigen elementaren Abschätzungen

$$|t_i^{(1)}| \leq \frac{1}{2} \quad , \; i=1,2 \; ,$$

d.h. die Parameter sind wie in [8] natürlich beschränkt. Die Behauptung folgt dann unter Ausnutzung von (2.2.34) aus einer Verallgemeinerung der Beweisschritte von Eberlein ([8]).

Natürlich ist die Voraussetzung (2.2.36) relativ stark, jedoch lassen sich Jacobi-ähnliche Blockverfahren formulieren, in denen diese Bedingung in jedem Pivotblock vor der T-Transformation gilt. So definieren wir in § 3.3 die sogenannte Jacobi-Transformation. Falls wir diese jedesmal direkt vor der T-Transformation ausführen, so ist (2.2.36) stets erfüllt (vgl.(3.3.7),(3.3.8)). In § 5.1 werden wir die Effektivität dieser Verfahren diskutieren.

Es deutet vieles darauf hin, daß eine T-Transformation mit den Parametern $x_i = \operatorname{artanh} t_i^{(1)}$ generell eine Normreduzierung von A bewirkt. Wir haben zwar, wie schon erwähnt, keinen entsprechenden Beweis führen können, es ist uns aber auch nicht gelungen, ein Beispiel zu konstruieren, in dem $\|A\|$ nach der T-Transformation zunimmt. Auch experimentelle Untersuchungen (siehe § 5.1) haben diese Vermutung nicht widerlegen können.

Daher ist es für die numerische Praxis interessant, auch mit Hilfe dieser T-Transformation Jacobi-ähnliche Blockverfahren zu konstruieren. Natürlich bedürfen dann die Fälle, in denen $H(0,0)$ singulär ist bzw. $|t_i^{(1)}|$ zu groß wird, einer Sonderbehandlung, beispielsweise mit einer Eberlein-Strategie wie in (2.2.29) bzw. (2.2.30). In § 5.1 haben wir ein solches Verfahren formuliert und diskutieren dort die numerischen Resultate.

3. ELEMENTARE TRANSFORMATIONSMATRIZEN FÜR JACOBI-ÄHNLICHE BLOCKVERFAHREN

Elementare Transformationsmatrizen sind gemäß Definition 1.1.5 reelle n-dimensionale Blockmatrizen R, die sich höchstens in der (p,q)-Restriktion $\hat{R}_{pq}$ von der Einheitsmatrix unterscheiden. Somit wird eine elementare Matrix durch die Vorgabe der vier 2x2-Blöcke R_{pp}, R_{pq}, R_{qp} und R_{qq} eindeutig definiert.

In § 2.1 und § 2.2 haben wir bereits die elementaren Matrizen S und T, die in den Jacobi-ähnlichen Blockverfahren zur Normreduzierung auf Diagonalblöcken bzw. auf Außerdiagonalpivots dienen, definiert und einige wesentliche Eigenschaften diskutiert. In diesem Kapitel wird für den diagonalisierenden Schritt der Verfahren eine weitere elementare Matrix U eingeführt und untersucht.

Mit Hilfe dieser elementaren Transformationsmatrizen lassen sich nun verschiedene Jacobi-ähnliche Blockverfahren konstruieren, die reelle J-symmetrische Blockmatrizen iterativ auf Murnaghan-Form transformieren. In § 4.1 formulieren wir zwei spezielle Verfahren dieses Typs. Für den Beweis der asymptotisch quadratischen Konvergenz dieser Verfahren stellen wir in diesem Kapitel die Hilfsmittel bereit, indem wir gewisse Abschätzungen für die elementaren Matrizen S,T und U beweisen.

3.1 DIE ELEMENTARE EBERLEIN-MATRIX

Wir verwenden die elementare Eberlein-Matrix S (vgl.(2.1.1)) zur Normreduzierung auf der Blockdiagonalen. In § 4.1 formulieren wir zwei spezielle Jacobi-ähnliche Blockverfahren, welche diesen Schritt für einen festen Diagonalblock A_{pp} mittels einer Standard-Normreduzierung für die Funktion $g(t)$ ausführen.

Wir zeigen daher in Lemma 3.1.2, daß der Kommutator von A_{pp} nach obiger S-Transformation quadratisch klein wird, falls die Matrix A fast Murnaghan-Form besitzt, d.h. A fast-blockdiagonal und die Blockdiagonale D fast-normal ist (vgl. Lemma 1.3.5).

Diese letztere Voraussetzung wollen wir zunächst exakt formulieren. A sei dazu eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$. Es soll dann gelten:

$$s(A) \leq \frac{\sqrt{2}}{500\sqrt{m-1}}\ \delta\ ,$$
$$\|C(A_{ii})\| \leq \frac{1}{500}\ \delta^2\ ,\quad i = 1(1)m\ . \tag{3.1.1}$$

Diese Ungleichungen werden als Generalvoraussetzungen für die nachfolgenden Untersuchungen dieses Paragraphen gebraucht. Wir beginnen mit dem Beweis eines technischen Lemmas.

Lemma 3.1.1:
A sei eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$. Es gelte $\delta > 0$ und die Voraussetzungen (3.1.1). Dann gilt für die Eigenwerte $\mu_k^{(i)}$, $k=1,2$ der Diagonalblöcke A_{ii} von A

$$|\mu_1^{(i)} - \mu_2^{(i)}| \geq \frac{11}{12}\ \delta\ ,\quad i = 1(1)m\ . \tag{3.1.2}$$

Beweis:
Aus einer bekannten Abschätzung für das sogenannte *Henrici-Maß* $\Delta(A)$ für beliebige quadratische Matrizen A (vgl.[25]) folgt mit $\dim A_{ii} = 2$

$$\Delta(A_{ii}) \leq \sqrt[4]{\frac{1}{2}} \sqrt{\| C(A_{ii}) \|} \leq \frac{\sqrt[4]{\frac{1}{2}}}{\sqrt{500}} \delta \quad , \quad i = 1(1)m \ .$$

Weiterhin erhalten wir aus der Cauchy-Schwarzschen Ungleichung und der J-Symmetrie von A

$$\sum_{j \neq i}^{m} \| A_{ij} \| \leq \frac{\sqrt{m-1}}{\sqrt{2}} S(A) \leq \frac{1}{500} \delta \ , \quad i = 1(1)m \ . \qquad (3.1.3)$$

Nach einem Satz von Meyer und Veselić,([33]), der eine Verallgemeinerung des klassischen Satzes von Gerschgorin (s.[59]) darstellt, liegen die Eigenwerte von A in der Vereinigung der Kreisscheiben

$$K_k^{(i)} = \{ z \in \mathbb{C} \mid |\mu_k^{(i)} - z| \leq \Delta(A_{ii}) + \sum_{j \neq i}^{m} \| A_{ij} \| \} \ , k \in \{1,2\} \ , i=1(1)m \ .$$

Falls dabei 1 der Kreise ein zusammenhängendes Gebiet bilden, welches disjunkt zu den übrigen Kreisen $K_k^{(i)}$ ist, so enthält dieses Gebiet genau 1 Eigenwerte von A.

Die Radien dieser Kreise sind nun kleiner als $\frac{1}{24} \delta$, und damit sind die Kreise aufgrund der Definition von δ paarweise disjunkt, und jede Kreisscheibe $K_k^{(i)}$ enthält genau einen Eigenwert von A.

Die Mittelpunkte der $K_k^{(i)}$ sind die Eigenwerte der Diagonalblöcke. Gälte nun für ein i, $1 \leq i \leq m$

$$|\mu_1^{(i)} - \mu_2^{(i)}| < \frac{11}{12} \delta \ ,$$

so wäre der Abstand der zwei Eigenwerte aus den zugehörigen Kreisscheiben $K_1^{(i)}$ und $K_2^{(i)}$ kleiner als δ. Dies ist ein Widerspruch zur Definition von δ, und es folgt die Behauptung. ■

Wir kommen nun zu der wichtigsten Aussage dieses Paragraphen:

Lemma 3.1.2:
A sei eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$. Es gelte $\delta > 0$ und die Voraussetzungen (3.1.1). Es sei $1 \leq p \leq m$, p fest, und die elementare Matrix S sei bestimmt durch

$$\tanh x = - \frac{c_{2p-1,2p}}{16\alpha + 4a} \tag{3.1.4}$$

(vgl.(2.1.25)). Dann gilt nach der S-Transformation für $A' = S^{-1}AS$

$$\| C(A_{pp}') \| \leq 2.533 \frac{\|A\|^2}{\delta^4} \| C(A_{pp}) \|^2 + 3.373 \frac{\|A\|^2}{\delta^2} s^2(A) \ . \tag{3.1.5}$$

Beweis:
Wir setzen o.B.d.A. $p = 1$ und schreiben wieder abkürzend $c = \cosh x$, $s = \sinh x$, $t = \tanh x$. Es gilt wegen (1.3.23) und (2.1.9)

$$t = - \frac{c_{12}}{16\alpha+4a} = - \frac{2\beta+b}{8\alpha+2a} \ . \tag{3.1.6}$$

Dabei ist der Nenner, wie man aus den folgenden Abschätzungen sieht, ungleich 0.

Für die Eigenwerte von A_{11} gilt

$$|\mu_1^{(1)} - \mu_2^{(1)}|^2 = |4a_{12}^2 - (a_{11}-a_{22})^2| \ ,$$

und hieraus folgt mit der Dreiecksungleichung und (3.1.2)

$$4a_{12}^2 + (a_{11}-a_{22})^2 \geq \frac{121}{144} \delta^2 \ . \tag{3.1.7}$$

Es ergibt sich dann mit $a \geq 0$

$$|t| \leq \frac{2|\beta|+|b|}{8\alpha} = \frac{2|\beta|+|b|}{4a_{12}^2+(a_{11}-a_{22})^2} \leq \frac{144}{121} \frac{2|\beta|+|b|}{\delta^2} \ .$$

Mit $|\beta| = \frac{1}{4}|\hat{c}_{12}| = \frac{\sqrt{2}}{8} \| C(A_{11}) \|$, $|b| \leq a$ und

$$a \leq \frac{1}{4} s^2(A) \tag{3.1.8}$$

folgt weiter

$$|t| \leq \frac{36}{121} \frac{1}{\delta^2} (\sqrt{2} \| C(A_{11}) \| + s^2(A)) \tag{3.1.9}$$

und damit wegen (3.1.1)

$$|t| \leq 0.000844 \quad . \tag{3.1.10}$$

Aus $s = \frac{t}{\sqrt{1-t^2}}$ erhalten wir sofort

$$|s| \leq 1.0000004\ |t| \leq 0.000845 \quad , \tag{3.1.11}$$

und aus $c = \frac{1}{\sqrt{1-t^2}}$ und

$$\frac{1}{\sqrt{1-\xi}} - 1 \leq \frac{\xi}{2(1-\xi)} \quad \text{für} \quad 0 \leq \xi < 1 \tag{3.1.12}$$

folgt

$$c = 1 + \gamma \ , \quad \gamma \leq 0.5000004\ t^2 \quad . \tag{3.1.13}$$

Es gilt nun a_{12}' und $a_{11}'-a_{22}'$ abzuschätzen. Wir haben (vgl.(2.1.5))

$$\begin{aligned} a_{12}' &= a_{12}(1+2s^2) + (a_{11}-a_{22})\,c\,s \\ &= a_{12} + (a_{11}-a_{22})\frac{t}{\sqrt{1-t^2}} + 2a_{12}s^2 + (a_{11}-a_{22})\gamma\, s \quad . \end{aligned}$$

Einsetzen von (3.1.6) liefert

$$\begin{aligned} a_{12}' &= a_{12} - \frac{a_{12}(a_{11}-a_{22})^2 + b(a_{11}-a_{22})}{4a_{12}^2 + (a_{11}-a_{22})^2 + 2a} \frac{1}{\sqrt{1-t^2}} + 2a_{12}s^2 + (a_{11}-a_{22})\gamma\, s \\ &= a_{12} - \frac{a_{12}}{\sqrt{1-t^2}} + \frac{4a_{12}^3 + 2a_{12}a - b(a_{11}-a_{22})}{4a_{12}^2 + (a_{11}-a_{22})^2 + 2a} \frac{1}{\sqrt{1-t^2}} + 2a_{12}s^2 + (a_{11}-a_{22})\gamma\, s, \end{aligned}$$

also

$$\begin{aligned} |a_{12}'| \leq |a_{12}| \left(\frac{1}{\sqrt{1-t^2}} - 1\right) + \frac{4|a_{12}|^3 + 2|a_{12}|a + |b|\ |a_{11}-a_{22}|}{4a_{12}^2 + (a_{11}-a_{22})^2 + 2a} \frac{1}{\sqrt{1-t^2}} \\ + 2|a_{12}|s^2 + |a_{11}-a_{22}|\,\gamma\,|s| \quad . \end{aligned}$$

Mit Hilfe der Ungleichung (3.1.12), $a \geq 0$ und (3.1.7) erhalten wir dann

$$|a_{12}'| \leq |a_{12}| \frac{t^2}{2(1-t^2)} + \frac{144}{121} \frac{4|a_{12}|^3 + 2|a_{12}|a + |b|\,|a_{11}-a_{22}|}{\delta^2} \frac{1}{\sqrt{1-t^2}}$$

$$+ 2|a_{12}|s^2 + |a_{11}-a_{22}|\,\gamma\,|s|$$

und hieraus mit $|b| \leq a$, (3.1.8),(3.1.10),(3.1.11) und (3.1.13)

$$|a_{12}'| \leq 2.501|a_{12}|t^2 + 0.000423|a_{11}-a_{22}|t^2 + 0.596 \frac{|a_{12}|}{\delta^2} S^2(A)$$

$$+ 0.298 \frac{|a_{11}-a_{22}|}{\delta^2} S^2(A) + 4.768 \frac{|a_{12}|^3}{\delta^2} .$$

Es gilt $|a_{12}| \leq \frac{\sqrt{2}}{2} \|A\|$ und $|a_{11}-a_{22}| \leq \sqrt{2}\,\|A\|$, wobei die letzte Abschätzung aus der Dreiecksungleichung und der Cauchy-Schwarzschen Ungleichung folgt. Damit haben wir

$$|a_{12}'| \leq 1.770\|A\|t^2 + 0.843 \frac{\|A\|}{\delta^2} S^2(A) + 4.768 \frac{|a_{12}|^3}{\delta^2} . \quad (3.1.14)$$

Analog zeigt man mit vertauschten Rollen von a_{12} und $a_{11}-a_{22}$

$$|a_{11}'-a_{22}'| \leq 3.540\|A\|t^2 + 1.686 \frac{\|A\|}{\delta^2} S^2(A) + 1.192 \frac{|a_{11}-a_{22}|^3}{\delta^2} . \quad (3.1.15)$$

Es gilt (vgl.(1.3.27)) $\hat{c}_{12}' = 2a_{12}'(a_{11}'-a_{22}')$. Zur Abschätzung von $\hat{c}_{12}'$ führen wir nun eine Fallunterscheidung durch.

Es sei zunächst $|a_{11}-a_{22}| \geq 2|a_{12}|$. Dann gilt wegen (3.1.7) und $\sqrt{\xi+\eta} \leq \sqrt{\xi} + \sqrt{\eta}$ für $\xi,\eta \geq 0$

$$|a_{11}-a_{22}| \geq \frac{1}{2}|a_{11}-a_{22}| + |a_{12}| \geq \sqrt{\frac{1}{4}(a_{11}-a_{22})^2 + a_{12}^2} \geq \frac{11}{24}\delta$$

und somit

$$|a_{12}| = \frac{1}{\sqrt{8}} \frac{\| C(A_{11}) \|}{|a_{11}-a_{22}|} \leq 0.772 \frac{\| C(A_{11}) \|}{\delta} \quad .$$

Wegen (2.1.27) gilt weiterhin

$$|a_{11}'-a_{22}'| \leq \sqrt{2}\, \|A'\| \leq \sqrt{2}\, \|A\| \ ,$$

so daß wir insgesamt

$$\begin{aligned} \| C(A_{11}') \| &= \sqrt{2}|\hat{c}_{12}'| = \sqrt{8}|a_{12}'|\ |a_{11}'-a_{22}'| \\ &\leq 16\|A\| \left(1.770\|A\| t^2 + 0.843 \frac{\|A\|}{\delta^2} s^2(A) + 2.194 \frac{\| C(A_{11}) \|^3}{\delta^5} \right) \\ &\leq 2.533 \frac{\|A\|^2}{\delta^4} \| C(A_{11}) \|^2 + 3.373 \frac{\|A\|^2}{\delta^2} s^2(A) \end{aligned}$$

erhalten, wobei in die letzte Abschätzung (3.1.9) mit der Cauchy-Schwarzschen Ungleichung, die Voraussetzungen (3.1.1) und $\delta \leq \sqrt{2}\|A\|$ eingeht.

Analog folgern wir im alternativen Fall $|a_{11}-a_{22}| < 2|a_{12}|$

$$|a_{12}| \geq \frac{11}{48} \delta \ ,$$

also

$$|a_{11}-a_{22}| \leq 1.544 \frac{\| C(A_{11}) \|}{\delta} \quad .$$

Die Abschätzungen $|a_{12}'| \leq \frac{\sqrt{2}}{2} \|A'\| \leq \frac{\sqrt{2}}{2} \|A\|$ und (3.1.15) implizieren dann wieder

$$\| C(A_{11}') \| \leq 2.533 \frac{\|A\|^2}{\delta^4} \| C(A_{11}) \|^2 + 3.373 \frac{\|A\|^2}{\delta^4} s^2(A) \ . \qquad \blacksquare$$

Wir wollen nun noch den Normzuwachs in der Außerdiagonalen nach einer S-Transformation abschätzen:

Lemma 3.1.3:
Unter den Voraussetzungen von Lemma 3.1.2 gilt nach der S-Transformation für $A' = S^{-1}AS$

$$s(A') \leq 1.0017\, s(A) \quad . \tag{3.1.16}$$

Beweis:
Wir schreiben $A = D + E$ und erhalten damit für die transformierte Matrix die Zerlegung

$$A' = S^{-1}DS + S^{-1}ES = D' + E'.$$

Aufgrund der Blockdiagonalgestalt von S (vgl.(2.1.1)) ist D' wieder eine Blockdiagonalmatrix und E' eine Matrix mit verschwindender Blockdiagonale. Also gilt

$$s(A') = \|E'\| \leq \|S^{-1}\|_2 \|E\| \|S\|_2 = \|S^{-1}\|_2 \|S\|_2\, s(A) \quad .$$

Die Eigenwerte der Matrizen S und S^{-1} sind $\lambda = c + |s|$, $\lambda = c - |s|$ und (n-2) mal $\lambda = 1$. Daher ist wegen $c \geq 1$ der betragsgrößte Eigenwert von S und S^{-1} $\lambda = c + |s|$, und es folgt jeweils mit (3.1.10), (3.1.11) und (3.1.13)

$$\|S^{-1}\|_2 = \|S\|_2 = c + |s| \leq 1.000846$$

und hieraus die Behauptung. ■

3.2 DIE ELEMENTARE EBERLEIN-VESELIĆ-MATRIX

Wir haben die elemetare Eberlein-Veselić-Matrix T bereits in § 2.2 definiert. Sie dient zur Normreduzierung der gesamten Matrix A, wobei mehrere Wahlmöglichkeiten für die Transformationsparameter x_1, x_2 existieren. Die speziellen Jacobi-ähnlichen Blockverfahren in § 4.1 führen diesen Schritt am Pivotblock A_{pq} mittels einer Standard-Normreduzierung bzw. einer optimalen Normreduzierung für die Funktion $F(x_1, x_2)$ aus (vgl. § 2.2.2). Wir untersuchen daher im folgenden die Eigenschaften der entsprechenden T-Transformationen.

Analog zu § 3.1 betrachten wir die T-Transformation ausschließlich unter der Voraussetzung, daß die Matrix A fast Murnaghan-Form besitzt, also A fast-blockdiagonal und D fast-normal ist. Wir zeigen, daß dann die gedämpfte Newton-Iteration (2.2.27) zur Normreduzierung mit der gewöhnlichen Newton-Iteration identisch ist, d.h. die Dämpfungsfaktoren $\gamma_k = 1$ gesetzt werden.

Die wichtigsten Aussagen dieses Paragraphen sind in Lemma 3.2.7 formuliert. Wir zeigen hier, daß für das oben zitierte Verfahren mit Standard-Normreduzierung der Pivotblock C_{pq} der Kommutatormatrix von A nach der T-Transformation quadratisch klein wird. Des weiteren wird in diesem Lemma eine Abschätzung für die Transformationsparameter x_1, x_2 bei optimaler Normreduzierung bewiesen.

Zunächst wollen wir die generelle Voraussetzung dieses Paragraphen, daß A fast Murnaghan-Form besitzt, exakt formulieren:

A sei eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$, dann gelte

$$s(A) \leq \|A\| \varepsilon , \quad \varepsilon = \frac{1}{1000\sqrt{m-1}} \frac{\delta^4}{\|A\|^4} ,$$
$$\|C(D)\| \leq \frac{1}{10000} \delta^2 . \qquad (3.2.1)$$

Diese Voraussetzungen sind wesentlich stärker als die Bedingungen (3.1.1) vor einer S-Transformation. Wir benötigen sie im Hinblick auf

die Anwendung des Satzes von Newton-Kantorovich (s.[37]) in Lemma 3.2.7.

Wir beginnen mit dem Beweis zweier technischer Lemmata, die unabhängig von den Voraussetzungen (3.2.1) gelten.

Lemma 3.2.1:
A sei eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$, und es sei $1 \le p < q \le m$. Dann gilt

$$\delta_D^{(p,q)}(A) \le d^{(p,q)}(A) + \frac{1}{\sqrt[4]{2}} \left(\sqrt{\| C(A_{pp}) \|} + \sqrt{\| C(A_{qq}) \|} \right) . \quad (3.2.2)$$

Beweis:
Wir führen den Beweis mittels einer dreifachen Fallunterscheidung. Dazu betrachten wir die Eigenwerte $\mu_k^{(j)}$, $k=1,2$, $j=p,q$ der Diagonalblöcke A_{pp} und A_{qq} und unterscheiden dann die Fälle von 4 nicht-reellen, 4 reellen und je 2 nicht-reellen und 2 reellen Eigenwerten. Zur Vereinfachung setzen wir im folgenden o.B.d.A. $p = 1$, $q = 2$.

Fall 1: (4 nicht-reelle Eigenwerte)
Die Eigenwerte von A_{11} und A_{22} seien jeweils nicht-reell. Es gilt dann

$$|a_{11}-a_{22}| < 2|a_{12}| , \quad |a_{33}-a_{44}| < 2|a_{34}| \quad (3.2.3)$$

und

$$\mu_{1,2}^{(1)} = \frac{a_{11}+a_{22}}{2} \pm \sqrt{a_{12}^2 - \frac{1}{4}(a_{11}-a_{22})^2}\, i , \quad \mu_{1,2}^{(2)} = \frac{a_{33}+a_{44}}{2} \pm \sqrt{a_{34}^2 - \frac{1}{4}(a_{33}-a_{44})^2}\, i .$$

Aus der Definition von $\delta_D^{(p,q)}(A)$ folgt

$$\begin{aligned} \delta_D^{(1,2)}(A) &\le |\mu_1^{(1)} - \mu_1^{(2)}| \\ &\le |\operatorname{Re}\mu_1^{(1)} - \operatorname{Re}\mu_1^{(2)}| + |\operatorname{Im}\mu_1^{(1)} - \operatorname{Im}\mu_1^{(2)}| \\ &= \left| \frac{a_{11}+a_{22}}{2} - \frac{a_{33}+a_{44}}{2} \right| + \left| \sqrt{a_{12}^2 - \frac{1}{4}(a_{11}-a_{22})^2} - \sqrt{a_{34}^2 - \frac{1}{4}(a_{33}-a_{44})^2} \right| . \end{aligned}$$

Man verifiziert leicht die Ungleichung

$$\left| \frac{a_{11}+a_{22}}{2} - \frac{a_{33}+a_{44}}{2} \right| \leq \min \{ |a_{11}-a_{33}|, |a_{11}-a_{44}|, |a_{22}-a_{33}|, |a_{22}-a_{44}| \} + \frac{|a_{11}-a_{22}|}{2} + \frac{|a_{33}-a_{44}|}{2} ,$$

und aus $-\sqrt{\xi-\eta} \leq -\sqrt{\xi}+\sqrt{\eta}$ für $\xi \geq \eta \geq 0$ ergibt sich

$$\left| \sqrt{a_{12}^2 - \tfrac{1}{4}(a_{11}-a_{22})^2} - \sqrt{a_{34}^2 - \tfrac{1}{4}(a_{33}-a_{44})^2} \right| \leq \Big| |a_{12}| - |a_{34}| \Big| + \frac{|a_{11}-a_{22}|}{2} + \frac{|a_{33}-a_{44}|}{2} .$$

Also gilt

$$\delta_D^{(1,2)}(A) \leq d^{(1,2)}(A) + |a_{11}-a_{22}| + |a_{33}-a_{44}|$$

und damit wegen (3.2.3) und (1.3.26),(1.3.27)

$$\begin{aligned} \delta_D^{(1,2)}(A) &\leq d^{(1,2)}(A) + \sqrt{2|a_{12}|\,|a_{11}-a_{22}|} + \sqrt{2\,|a_{34}|\,|a_{33}-a_{44}|} \\ &\leq d^{(1,2)}(A) + \frac{1}{\sqrt[4]{2}} \left(\sqrt{\| C(A_{11}) \|} + \sqrt{\| C(A_{22}) \|} \right) . \end{aligned}$$

Fall 2: (4 reelle Eigenwerte)

Die Eigenwerte von A_{11} und A_{22} seien jeweils reell. Es gilt dann

$$|a_{11}-a_{22}| \geq 2|a_{12}| , \quad |a_{33}-a_{44}| \geq 2|a_{34}| \qquad (3.2.4)$$

und

$$\mu_{1,2}^{(1)} = \frac{a_{11}+a_{22}}{2} \pm \sqrt{\tfrac{1}{4}(a_{11}-a_{22})^2 - a_{12}^2} , \quad \mu_{1,2}^{(2)} = \frac{a_{33}+a_{44}}{2} \pm \sqrt{\tfrac{1}{4}(a_{33}-a_{44})^2 - a_{34}^2} .$$

Wenn wir nun den Satz von Gerschgorin (s.[59]) auf die 2x2-Matrix A_{11} anwenden, so erhalten wir, daß die Eigenwerte $\mu_{1,2}^{(1)}$ von A_{11} in der Vereinigung der Kreisscheiben mit den Mittelpunkten a_{11} und a_{22} und dem jeweiligen Radius $|a_{12}|$ liegen. Nach (3.2.4) können sich diese Kreise höchstens berühren, so daß sich in jedem Kreis genau ein Eigenwert befindet. Analoge Aussagen gelten für die 2x2-Matrix A_{22}. Es

sei nun o.B.d.A.

$$|a_{11}-a_{33}| = \min\{|a_{11}-a_{33}|,|a_{11}-a_{44}|,|a_{22}-a_{33}|,|a_{22}-a_{44}|\} \quad .$$

Wir haben dann in dem Kreis um a_{11} mit dem Radius $|a_{12}|$ genau einen der Eigenwerte $\mu_{1,2}^{(1)}$ und in dem Kreis um a_{33} mit dem Radius $|a_{34}|$ genau einen der Eigenwerte $\mu_{1,2}^{(2)}$. Damit gilt

$$\delta_D^{(1,2)}(A) \le |a_{11}-a_{33}| + |a_{12}| + |a_{34}| \le d^{(1,2)}(A) + |a_{12}| + |a_{34}|$$

und wegen (3.2.4)

$$\begin{aligned}\delta_D^{(1,2)}(A) &\le d^{(1,2)}(A) + \sqrt{\tfrac{1}{2}|a_{12}|\,|a_{11}-a_{22}|} + \sqrt{\tfrac{1}{2}|a_{34}|\,|a_{33}-a_{44}|} \\ &\le d^{(1,2)}(A) + \frac{1}{\sqrt[4]{2}}\left(\sqrt{\| C(A_{11})\|} + \sqrt{\| C(A_{22})\|}\right) \quad .\end{aligned}$$

Fall 3: (2 nicht-reelle und 2 reelle Eigenwerte)

Es seien o.B.d.A. die Eigenwerte von A_{11} reell und die von A_{22} nicht-reell. Es gilt dann

$$|a_{11}-a_{22}| \ge 2|a_{12}| \quad , \quad |a_{33}-a_{44}| < 2|a_{34}| \qquad (3.2.5)$$

und

$$\mu_{1,2}^{(1)} = \frac{a_{11}+a_{22}}{2} \pm \sqrt{\tfrac{1}{4}(a_{11}-a_{22})^2 - a_{12}^2} \quad , \quad \mu_{1,2}^{(2)} = \frac{a_{33}+a_{44}}{2} \pm \sqrt{a_{34}^2 - \tfrac{1}{4}(a_{33}-a_{44})^2}\, i \quad .$$

Es sei wieder o.B.d.A.

$$|a_{11}-a_{33}| = \min\{|a_{11}-a_{33}|,|a_{11}-a_{44}|,|a_{22}-a_{33}|,|a_{22}-a_{44}|\} \quad .$$

Mit der gleichen Argumentation wie im Fall 2 folgern wir nun, daß in dem Kreis um a_{11} mit dem Radius $|a_{12}|$ genau einer der Eigenwerte von A_{11}, sagen wir $\mu_1^{(1)}$, liegt. Also gilt

$$\delta_D^{(1,2)}(A) \le |\mu_1^{(1)} - \mu_1^{(2)}|$$

$$\leq |\mu_1^{(1)} - \operatorname{Re}\mu_1^{(2)}| + |\operatorname{Im}\mu_1^{(2)}|$$

$$\leq |a_{11} - \frac{a_{33}+a_{44}}{2}| + |a_{12}| + \sqrt{a_{34}^2 - \tfrac{1}{4}(a_{33}-a_{44})^2} \quad .$$

Aus $\sqrt{a_{34}^2 - \frac{1}{4}(a_{33}-a_{44})^2} \leq |a_{34}| \leq |a_{12}| + \big||a_{12}| - |a_{34}|\big|$ folgt dann

$$\delta_D^{(1,2)}(A) \leq |a_{11}-a_{33}| + \frac{|a_{33}-a_{44}|}{2} + 2|a_{12}| + \big||a_{12}| - |a_{34}|\big|$$

$$= d^{(1,2)}(A) + \frac{|a_{33}-a_{44}|}{2} + 2|a_{12}|$$

und wieder wegen (3.2.5)

$$\delta_D^{(1,2)}(A) \leq d^{(1,2)}(A) + \sqrt{\tfrac{1}{2}|a_{34}|\,|a_{33}-a_{44}|} + \sqrt{2|a_{12}|\,|a_{11}-a_{22}|}$$

$$\leq d^{(1,2)}(A) + \frac{1}{\sqrt[4]{2}}\left(\sqrt{\|C(A_{11})\|} + \sqrt{\|C(A_{22})\|}\right) . \quad \blacksquare$$

Lemma 3.2.2:

A sei eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$, und es sei $1 \leq p < q \leq m$. Dann gilt

$$d^{(p,q)}(A) \leq \sqrt{16\min(\alpha_1,\alpha_2) + 8\min(\delta_+,\delta_-)} \quad . \qquad (3.2.6)$$

Beweis:

Es sei wieder o.B.d.A. $p = 1$, $q = 2$. Mit Hilfe der Cauchy-Schwarzschen Ungleichung zeigt man sofort

$$d^{(1,2)}(A) \leq \left(2\min\{(a_{11}-a_{33})^2,(a_{11}-a_{44})^2,(a_{22}-a_{33})^2,(a_{22}-a_{44})^2\} + 2(|a_{12}|-|a_{34}|)^2\right)^{\frac{1}{2}}.$$

Weiterhin gilt (vgl.(2.2.10))

$$(a_{11}-a_{44})^2 \leq 8\alpha_1 \ , \quad (a_{22}-a_{33})^2 \leq 8\alpha_2 \ ,$$

$$(|a_{12}| - |a_{34}|)^2 \leq (a_{34}+a_{12})^2 \leq 4\delta_+ \ , \ (|a_{12}| - |a_{34}|)^2 \leq (a_{34}-a_{12})^2 \leq 4\delta_- \ .$$

Zusammen folgt hieraus die Behauptung. $\blacksquare$

Als nächstes beweisen wir zwei weitere technische Lemmata, in die die Generalvoraussetzungen (3.2.1) eingehen.

Lemma 3.2.3:
A sei eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$. Es gelte $\delta > 0$ und die Voraussetzungen (3.2.1). Dann folgt

$$\delta_D(A) \geq \frac{47}{48}\,\delta \quad . \tag{3.2.7}$$

Beweis:
Der Beweis wird mit der gleichen Argumentation wie der von Lemma 3.1.1 geführt. Aus der Abschätzung von Henrici ([25]) folgt zunächst mit $\| C(A_{ii}) \| \leq \| C(D) \|$ und den Voraussetzungen (3.2.1)

$$\Delta(A_{ii}) \leq \frac{\sqrt[4]{\frac{1}{2}}}{100}\,\delta \;, \quad i = 1(1)m \quad .$$

Aus (3.1.3) erhalten wir dann mit (3.2.1) und $\delta \leq \sqrt{2}\|A\|$

$$\sum_{j \neq i}^{m} \|A_{ij}\| \leq \frac{1}{500}\,\delta \;, \quad i = 1(1)m \quad .$$

Die Radien der Kreisscheiben $K_k^{(i)}$ aus dem Satz von Meyer und Veselić ([33]) sind damit kleiner als $\frac{1}{96}\,\delta$, und die $K_k^{(i)}$ sind paarweise disjunkt. Somit enthält jede Kreisscheibe wieder genau einen Eigenwert von A. Gälte nun

$$\delta_D(A) < \frac{47}{48}\,\delta \;,$$

so lägen mindestens zwei Eigenwerte der Diagonalblöcke A_{ii} weniger als $\frac{47}{48}\,\delta$ voneinander entfernt, und damit wäre der Abstand der zwei Eigenwerte von A aus den zugehörigen Kreisscheiben kleiner als δ. Hieraus folgt die Behauptung. ■

Lemma 3.2.4:
Unter den Voraussetzungen von Lemma 3.2.3 gilt

$$d^{(p,q)}(A) \geq \frac{27}{28}\,\delta \quad . \tag{3.2.8}$$

Beweis:
Aus den Lemmata 3.2.3 und 3.2.1 folgt

$$\delta \le \frac{48}{47}\,\delta_D(A) \le \frac{48}{47}\,\delta_D^{(p,q)}(A)$$
$$\le \frac{48}{47}\left(d^{(p,q)}(A) + \frac{1}{\sqrt[4]{2}}\left(\sqrt{\|C(A_{pp})\|} + \sqrt{\|C(A_{qq})\|}\right)\right).$$

Mit $\sqrt{\xi} + \sqrt{\eta} \le \sqrt{2}\sqrt{\xi+\eta}$ für $\xi,\eta \ge 0$ und $\|C(A_{pp})\| + \|C(A_{qq})\| \le \sqrt{2}\|C(D)\|$ gilt dann

$$\sqrt{\|C(A_{pp})\|} + \sqrt{\|C(A_{qq})\|} \le \sqrt[4]{8}\sqrt{\|C(D)\|}\,.$$

Insgesamt erhalten wir damit

$$\delta \le \frac{48}{47}\left(d^{(p,q)}(A) + \frac{\sqrt{2}}{100}\,\delta\right),$$

und hieraus folgt die Behauptung. ■

Bevor wir die wichtigste Aussage dieses Paragraphen in Lemma 3.2.7 mit Hilfe des Satzes von Newton-Kantorovich (s.[37]) beweisen, müssen wir dazu in den folgenden zwei Lemmata einige wesentliche Hilfsmittel bereitstellen.

Lemma 3.2.5:
A sei eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$. Es gelte $\delta > 0$ und die Voraussetzungen (3.2.1). Es sei $1 \le p < q \le m$, p,q fest, und die elementare Matrix T sei durch

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = -H(0,0)^{-1}\begin{pmatrix} c_{2p-1,2q} \\ c_{2p,2q-1} \end{pmatrix} \qquad (3.2.9)$$

bestimmt, wobei H durch (2.2.17) definiert ist. Dann gilt

$$\text{(i)} \quad \|H(0,0)^{-1}\|_2 \le \frac{784}{729}\,\frac{1}{\delta^2}\,, \qquad (3.2.10)$$

$$\text{(ii)} \qquad \left\| \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right\| \leq 3.042 \, \frac{\|A\|}{\delta^2} \, S(A) \leq 0.00609 \quad . \qquad (3.2.11)$$

(Man beachte, daß die Hessematrix $H(0,0)$ wegen $\delta > 0$ und Lemma 2.2.2 invertierbar ist.)

Beweis:

(i) Es gilt allgemein für $y = (y_1, y_2)^T \neq 0$

$$\begin{aligned} y^T H(0,0) y &= (16\alpha_1 + 4\alpha_{11} + 4\delta_+ + 4\delta_-) y_1^2 + (8\delta_+ - 8\delta_-) y_1 y_2 + (16\alpha_2 + 4\alpha_{22} + 4\delta_+ + 4\delta_-) y_2^2 \\ &= (16\alpha_1 + 4\alpha_{11}) y_1^2 + (16\alpha_2 + 4\alpha_{22}) y_2^2 + 4\delta_+ (y_1 + y_2)^2 + 4\delta_- (y_1 - y_2)^2 \\ &\geq 16\alpha_1 y_1^2 + 16\alpha_2 y_2^2 + 4\delta_+ (y_1 + y_2)^2 + 4\delta_- (y_1 - y_2)^2 \qquad (3.2.12) \\ &\geq 16 \min(\alpha_1, \alpha_2) \; (y_1^2 + y_2^2) + 8 \min(\delta_+, \delta_-) \; (y_1^2 + y_2^2) \end{aligned}$$

und damit wegen (3.2.6) und (3.2.8)

$$y^T H(0,0) y \geq (d^{(p,q)}(A))^2 \, y^T y \geq \frac{729}{784} \, \delta^2 \, y^T y \quad .$$

Aus der Cauchy-Schwarzschen Ungleichung folgt nun

$$\| H(0,0) \, y \| \; \|y\| \geq y^T H(0,0) y \quad ,$$

d.h. wir haben

$$\| H(0,0) \, y \| \geq \frac{729}{784} \, \delta^2 \, \|y\| \quad .$$

Hieraus ergibt sich dann sofort

$$\| H(0,0)^{-1} \|_2 = \sup_{y \neq 0} \frac{\|y\|}{\| H(0,0) \, y \|} \leq \frac{784}{729} \; \frac{1}{\delta^2} \quad .$$

(ii) Mit (3.2.9) und (3.2.10) erhalten wir aufgrund der Verträglichkeit der Spektral-Norm und der Euklidischen Norm

$$\left\| \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right\| \leq \frac{784}{729} \frac{1}{\delta^2} \left\| \begin{pmatrix} c_{2p-1,2q} \\ c_{2p,2q-1} \end{pmatrix} \right\| .$$

Es gilt nun

$$\left\| \begin{pmatrix} c_{2p-1,2q} \\ c_{2p,2q-1} \end{pmatrix} \right\| = \| C_{pq} \| \leq \frac{\sqrt{2}}{2} S(C(A)) ,$$

und wegen $S(C(A)) \leq 4\|A\| S(A)$ (vgl.[56]) ergibt sich dann insgesamt

$$\left\| \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right\| \leq 3.042 \frac{\|A\|}{\delta^2} S(A) \leq 0.00609 ,$$

wobei die letzte Ungleichung aus (3.2.1) und $\delta \leq \sqrt{2}\|A\|$ folgt. ■

Lemma 3.2.6:
A sei eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$. Es gelte $\delta > 0$ und die Voraussetzungen (3.2.1). Es sei $1 \leq p < q \leq m$, p,q fest, und $M \subset \mathbb{R}^2$ sei definiert durch

$$M = (-0.0122, 0.0122)^2 .$$

Dann gilt für $x,y \in M$

(i) $$\| (\text{grad } F)(x) - (\text{grad } F)(y) \| \leq 4.00716 \|A\|^2 \|x-y\| , \quad (3.2.13)$$

(ii) $$\| H(x) - H(y) \|_2 \leq 1.893 \|A\|^2 \|x-y\| , \quad (3.2.14)$$

wobei F und H durch (2.2.7) und (2.2.17) gegeben sind.

Beweis:
Zunächst gilt für $\xi = (\xi_1, \xi_2)^T \in M$

$$|\tanh \xi_i| \leq |\xi_i| \leq 0.0122 , \; i=1,2 .$$

Hieraus folgt sukzessive aus den Additionstheoremen der hyperboli-

schen Funktionen

$$|\sinh 2\xi_i| \le 0.0245 \ , \quad \cosh 2\xi_i \le 1.000298 \ , \ i=1,2 \ ,$$
$$|\sinh 4\xi_i| \le 0.0491 \ , \quad \cosh 4\xi_i \le 1.00120 \ , \ i=1,2 \ , \qquad (3.2.15)$$
$$|\sinh(2\xi_1 \pm 2\xi_2)| \le 0.0491 \ , \ \cosh(2\xi_1 \pm 2\xi_2) \le 1.00120 \ .$$

Alsdann erhalten wir aus dem Mittelwertsatz der Differentialrechnung für vektorwertige Funktionen

$$\| (\text{grad } F)(x) - (\text{grad } F)(y) \| \le L \ \| x - y \| , \ x,y \in M$$

mit

$$L = \max_{\xi \in \{ x+\omega(y-x) \mid 0 \le \omega \le 1 \}} \| H(\xi) \|_2 \ ,$$

wobei die Verträglichkeit der Spektral-Norm und der Euklidischen Norm eingeht.

Es gelte nun für ein beliebiges $\xi \in M$ für die Komponenten der Hessematrix $H(\xi)$ o.B.d.A. $4\tilde{\alpha}_1+\tilde{\alpha}_{11} \ge 4\tilde{\alpha}_2+\tilde{\alpha}_{22}$, $\tilde{\delta}_+ \ge \tilde{\delta}_-$, dann folgt

$$\| H(\xi) \|_2 \le 16\tilde{\alpha}_1 + 4\tilde{\alpha}_{11} + 8\tilde{\delta}_+ \ ,$$

denn die Eigenwerte einer symmetrischen 2x2-Matrix

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$$

sind durch

$$\lambda_{1,2} = \frac{a_{11}+a_{22}}{2} \pm \sqrt{\frac{1}{4}(a_{11}-a_{22})^2 + a_{12}^2}$$

gegeben, und somit wird die Spektral-Norm dieser Matrix durch

$$\frac{1}{2}|a_{11}+a_{22}| + \frac{1}{2}|a_{11}-a_{22}| + |a_{12}|$$

majorisiert. Mit (2.2.12) erhalten wir dann

$$\| H(\xi) \|_2 \leq 16\alpha_1 \cosh 4\xi_1 + 16|\beta_1|\,|\sinh 4\xi_1| + 4\alpha_{11}\cosh 2\xi_1 + 4|\beta_{11}|\,|\sinh 2\xi_1|$$
$$+ 8\delta_+ \cosh(2\xi_1+2\xi_2) + 8|\varepsilon_+|\,|\sinh(2\xi_1+2\xi_2)| \quad .$$

Es müssen nun die Koeffizienten der hyperbolischen Terme abgeschätzt werden. Aus $|a_{2p-1,2q}| \leq \frac{\sqrt{2}}{2} S(A)$ und $|a_{2p-1,2p-1} - a_{2q,2q}| \leq \sqrt{2}\,\|A\|$ folgt

$$|\beta_1| \leq \frac{1}{2} S(A)\,\|A\| \quad ,$$

und aus $|\beta_{11}| \leq \alpha_{11}$ erhalten wir

$$|\beta_{11}| \leq \frac{1}{4} S^2(A) \quad .$$

Weiterhin folgt mit $|a_{2p-1,2q-1} - a_{2p,2q}| \leq S(A)$ und $|a_{2q-1,2q} + a_{2p-1,2p}| \leq \|A\|$

$$|\varepsilon_+| \leq \frac{1}{2} S(A)\,\|A\| \quad ,$$

und es gilt

$$4\alpha_1 + \alpha_{11} + 2\delta_+ \leq \|A\|^2 \quad .$$

In diese letzten Abschätzungen ging mehrfach die Cauchy-Schwarzsche Ungleichung ein.

Insgesamt erhalten wir dann mit (3.2.15), den Voraussetzungen (3.2.1) und $\delta \leq \sqrt{2}\,\|A\|$

$$\| H(\xi) \|_2 \leq 1.00120(16\alpha_1 + 4\alpha_{11} + 8\delta_+) + 0.00236\,\|A\|^2 \leq 4.00716\,\|A\|^2$$

für $\xi \in M$. Somit gilt insbesondere

$$L \leq 4.00716\,\|A\|^2 \quad ,$$

und es folgt die Behauptung (i).

Für den Beweis der zweiten Behauptung fassen wir $H(x)$ als Funktion

$H : \mathbb{R}^2 \to \mathbb{R}^4$ auf. Es sei also

$$H = (H_1, H_2, H_3, H_4)^T$$

mit

$$H_1(x) = 16\tilde{\alpha}_1 + 4\tilde{\alpha}_{11} + 4\tilde{\delta}_+ + 4\tilde{\delta}_- \ , \quad H_4(x) = 16\tilde{\alpha}_2 + 4\tilde{\alpha}_{22} + 4\tilde{\delta}_+ + 4\tilde{\delta}_- \ ,$$

$$H_2(x) = H_3(x) = 4\tilde{\delta}_+ - 4\tilde{\delta}_- \ ,$$

und

$$H' = \left(\frac{\partial H_i}{\partial x_j} \right)_{i = 1(1)4,\ j = 1,2}$$

sei die Funktionalmatrix von H. Man verifiziert leicht die folgende Abschätzung für ein beliebiges $\xi \in M$:

$$\| H'(\xi) \|_M = 2 \max_{\substack{1 \le i \le 4 \\ 1 \le j \le 2}} \left| \frac{\partial H_i}{\partial x_j} (\xi) \right|$$

$$\le 128\alpha_1 |\sinh 4\xi_1| + 128|\beta_1| \cosh 4\xi_1 + 16\alpha_{11} |\sinh 2\xi_1| + 16|\beta_{11}| \cosh 2\xi_1$$

$$+ 16\delta_+ |\sinh(2\xi_1 + 2\xi_2)| + 16|\varepsilon_+| \cosh(2\xi_1 + 2\xi_2)$$

$$+ 16\delta_- |\sinh(2\xi_1 - 2\xi_2)| + 16|\varepsilon_-| \cosh(2\xi_1 - 2\xi_2) \quad .$$

Wie im Beweis von (i) zeigt man dann mit (3.2.15) und (3.2.1)

$$\| H'(\xi) \|_M \le 1.893 \, \|A\|^2 \ , \quad \xi \in M \ ,$$

und mit dem Mittelwertsatz der Differentialrechnung folgt dann wieder, da auch die Maximum-Norm und die Euklidische Norm verträglich sind,

$$\|H(x) - H(y)\|_2 \le \|H(x) - H(y)\| \le \max_{\xi \in \{x + \omega(y-x) \,|\, 0 \le \omega \le 1\}} \| H'(\xi) \|_M \, \|x - y\|$$

$$\le 1.893 \, \|A\|^2 \|x - y\| \ , \quad x, y \in M \quad .$$

■

Wir können nun die Hauptaussage dieses Abschnitts beweisen:

Lemma 3.2.7:
A sei eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$. Es gelte $\delta > 0$ und die Voraussetzungen (3.2.1). Es sei $1 \le p < q \le m$, p,q fest. Zur Bestimmung der Parameter $x = (x_1, x_2)^T$ der elementaren Matrix T werde die gedämpfte Newton-Iteration

$$x^{(0)} = 0,\ x^{(k+1)} = x^{(k)} - \gamma_k H(x^{(k)})^{-1} (\text{grad } F)(x^{(k)}),\ k=0,1,\ldots \quad (3.2.16)$$

nach der Strategie (2.2.28) durchgeführt.

Dann gilt $\gamma_k = 1$ für $k = 0,1,\ldots,$ und die Iterierten (3.2.16) konvergieren quadratisch gegen das eindeutige Minimum x^* von F.

Nach einer T-Transformation mit $x^{(1)}$ (Standard-Normreduzierung) folgt für $A' = T^{-1}AT$

$$\| C_{pq}' \| \le 150.982 \frac{\|A\|^6}{\delta^6} s^2(A) , \quad (3.2.17)$$

und für eine T-Transformation mit x^* (optimale Normreduzierung) gilt

$$\|x^*\| \le 6.084 \frac{\|A\|}{\delta^2} s(A) \le 0.0122 \quad (3.2.18)$$

und

$$\| C_{pq}' \| = 0 \quad . \quad (3.2.19)$$

Beweis:
Zunächst einmal sei vermerkt, daß aufgrund der Voraussetzung $\delta > 0$ die Funktion F gleichmäßig konvex ist (vgl. Lemma 2.2.2). Damit ist $H(x^{(k)})$ invertierbar, und F besitzt ein eindeutiges Minimum x^*.

Alsdann sei M wie in Lemma 3.2.6 definiert. Wir setzen vorläufig $x^{(k)} \in M$ für die Iterierten (3.2.16) voraus. Später zeigen wir, daß dies automatisch aus den Voraussetzungen folgt. (Man beachte, daß mit $x^{(k)} \in M$ natürlich auch unsere Zusatzforderung $|x_i^{(k)}| \le 1$, $i=1,2$ erfüllt ist.)

Als erstes gilt es nun zu beweisen, daß für $\gamma_k = 1$

$$F(x^{(k+1)}) - F(x^{(k)}) \leq -\frac{1}{3}(\text{grad } F)(x^{(k)})^T H(x^{(k)})^{-1} (\text{grad } F)(x^{(k)}), \quad k=0,1,\ldots \quad (3.2.20)$$

gilt (vgl.(2.2.28)). Wegen Lemma 2.2.1 genügt es aber o.B.d.A., diese Ungleichung für $k = 0$ zu beweisen.

Aus der Taylorentwicklung von F um 0 erhalten wir allgemein für $x \in M$

$$F(x) = (\text{grad } F)(0)^T x + \frac{1}{2} x^T H(\xi) x \quad , \quad \xi \in M \ .$$

Für $x = x^{(1)}$ folgt hieraus mit (3.2.16)

$$F(x^{(1)}) = -x^{(1)^T} H(0) x^{(1)} + \frac{1}{2} x^{(1)^T} H(\xi) x^{(1)} \ , \quad \xi \in M \ .$$

Es gilt nun für $\xi = (\xi_1, \xi_2)^T \in M$ wegen (3.2.15), (2.2.12) und (2.2.15)

$$\tilde{\alpha}_1(\xi_1) \leq \alpha_1 \cosh 4\xi_1 + \alpha_1 |\sinh 4\xi_1| \leq \frac{20}{19} \alpha_1 \ ,$$

und analog zeigt man

$$\tilde{\alpha}_2(\xi_2) \leq \frac{20}{19} \alpha_2 \ , \quad \tilde{\alpha}_{11}(\xi_1) \leq \frac{20}{19} \alpha_{11} \ , \quad \tilde{\alpha}_{22}(\xi_2) \leq \frac{20}{19} \alpha_{22} \ ,$$

$$\tilde{\delta}_+(\xi_1, \xi_2) \leq \frac{20}{19} \delta_+ \ , \quad \tilde{\delta}_-(\xi_1, \xi_2) \leq \frac{20}{19} \delta_- \ .$$

Somit haben wir für $x^{(1)} = (x_1^{(1)}, x_2^{(1)})^T$ (vgl.(3.2.12))

$$\begin{aligned} x^{(1)^T} H(\xi) x^{(1)} &= (16\tilde{\alpha}_1(\xi_1) + 4\tilde{\alpha}_{11}(\xi_1)) x_1^{(1)^2} + (16\tilde{\alpha}_2(\xi_2) + 4\tilde{\alpha}_{22}(\xi_2)) x_2^{(1)^2} \\ &\quad + 4\tilde{\delta}_+(\xi_1, \xi_2)(x_1^{(1)} + x_2^{(1)})^2 + 4\tilde{\delta}_-(\xi_1, \xi_2)(x_1^{(1)} - x_2^{(1)})^2 \\ &\leq \frac{20}{19} x^{(1)^T} H(0) x^{(1)} \ . \end{aligned}$$

Insgesamt erhalten wir also

$$F(x^{(1)}) \leq -\frac{9}{19} x^{(1)^T} H(0) x^{(1)} = -\frac{9}{19} (\text{grad } F)(0)^T H(0)^{-1} (\text{grad } F)(0) \ ,$$

wobei die letzte Identität wieder aus (3.2.16) folgt. Damit ist (3.2.20) für $k = 0$ bewiesen.

Da die Iteration (3.2.16) nun mit der gewöhnlichen Newton-Iteration identisch ist, kann der Konvergenzsatz von Newton-Kantorovich angewandt werden. Wir setzen

$$K_1 = 1.893 \, \|A\|^2 , \quad K_2 = \| H(0)^{-1} \|_2 , \quad K_3 = \|x^{(1)}\| .$$

Hieraus folgt für

$$K = K_1 K_2 K_3$$

mit (3.2.10),(3.2.11) und den Voraussetzungen (3.2.1)

$$K \leq 6.193 \, \frac{\|A\|^3}{\delta^4} \, s(A) < \frac{1}{2} .$$

Wir setzen weiter

$$r = \frac{1}{K_1 K_2} \left(1 - \sqrt{1 - 2K}\right)$$

und erhalten mit $1 - \sqrt{1 - 2K} \leq 2K$ und (3.2.11)

$$r \leq 2K_3 \leq 0.01218 . \tag{3.2.21}$$

Also ist auch

$$S(0,r) = \{ x \in \mathbb{R}^2 \mid \|x\| \leq r \} \subset M .$$

Da M konvex ist, sind hiermit alle Voraussetzungen des Satzes von Newton-Kantorovich ([37,Theorem 12.6.2]) erfüllt, und es folgt, daß die $x^{(k)}$ quadratisch gegen die Nullstelle von grad F und damit gegen das Minimum von F konvergieren.

Die Abschätzung [37,Theorem 12.6.2,(5)] ergibt für $k = 1$

$$\| x^{(1)} - x^* \| \leq 2K_1 K_2 \|x^{(1)}\|^2 . \tag{3.2.22}$$

Mit (2.2.14),(3.2.13) und $(\text{grad } F)(x^*) = 0$ folgt dann für die Standard-Normreduzierung

$$\| C'_{pq} \| = \left\| \begin{pmatrix} c'_{2p-1,2q} \\ c'_{2p,2q-1} \end{pmatrix} \right\| = \| (\text{grad } F)(x^{(1)}) \| \leq 4.00716 \, \|A\|^2 \| x^{(1)} - x^* \|$$

und hieraus mit (3.2.22),(3.2.10) und (3.2.11)

$$\| C'_{pq} \| \leq 150.982 \, \frac{\|A\|^6}{\delta^6} \, s^2(A) \quad .$$

Wird die $\top$-Transformation mit x^* ausgeführt, so gilt wegen (2.2.14)

$$C'_{pq} = 0 \quad ,$$

und es folgt die Behauptung (3.2.19). Aus [37,Theorem 12.6.2] erhalten wir noch mit (3.2.21)

$$\|x^*\| \leq r \leq 2 \, \|x^{(1)}\|$$

und hieraus sofort mit (3.2.11) die Abschätzung (3.2.18).

Wir hatten zu Anfang des Beweises generell vorausgesetzt, daß die Iterierten $x^{(k)}$ in M liegen. Jetzt können wir induktiv beweisen, daß $x^{(k)} \in M$ für $k = 0,1,\ldots$ gilt.

Trivialerweise ist $x^{(0)} = 0 \in M$. Es seien dann $x^{(0)},\ldots,x^{(k)} \in M$ für ein $k \in \mathbb{N}_0$. $x^{(k+1)}$ wird nun durch (3.2.16) zunächst mit $\gamma_k = 1$ bestimmt. Aus [37,Theorem 12.6.2] folgt, daß $x^{(k+1)}$ in $S(0,r)$ und damit wieder in M liegt. Da jetzt wieder bewiesen werden kann, daß mit $\gamma_k = 1$ die Ungleichung (3.2.20) gilt, ist $x^{(k+1)}$ definitiv bestimmt, und es gilt somit $x^{(k+1)} \in M$. ■

Die Aussagen über eine $\top$-Transformation mit x^* wollen wir im folgenden so verstehen, daß x^* durch die Iteration (3.2.16) hinreichend approximiert wird, d.h. in dem Sinne, daß für ein $k \in \mathbb{N}$ der Term $\| (\text{grad } F)(x^{(k)}) \|$ unter der Rechnergenauigkeit liegt. Nach [37,Theorem 12.6.2] gilt dann für dieses $x^{(k)}$ ebenfalls die Abschätzung

(3.2.18), während die Identität (3.2.19) als Ungleichung

$$\| C_{pq}' \| \leq \text{Rechnergenauigkeit}$$

zu interpretieren ist.

Um die folgenden Rechnungen zu vereinfachen, wollen wir die Parameter x_1, x_2 einer T-Transformation generell mit (3.2.18) abschätzen, obwohl im Falle einer Standard-Normreduzierung eine bessere Abschätzung gilt (vgl.(3.2.11)).

Wir müssen nun die Veränderungen der Matrix A unter einer T-Transformation untersuchen. Dazu beweisen wir zunächst einige Aussagen über die nicht-trivialen Blöcke der Matrix T.

Lemma 3.2.8:
Es gelten die Voraussetzungen von Lemma 3.2.7, und die Parameter x_1, x_2 der elementaren Matrix T seien durch $x^{(1)}$ bzw. x^* bestimmt. Dann folgt

$$\text{(i)} \quad \|T_{pp}\|_2 = \|T_{qq}\|_2 \leq 1.0000745 \ , \tag{3.2.23}$$

$$\text{(ii)} \quad \|T_{pq}\|_2 = \|T_{qp}\|_2 \leq 6.085 \frac{\|A\|}{\delta^2} S(A) \leq 0.0123 \ . \tag{3.2.24}$$

(iii) Für $\overset{\circ}{T}_{pp} = T_{pp} - I$, $\overset{\circ}{T}_{qq} = T_{qq} - I$ gilt

$$\|\overset{\circ}{T}_{pp}\|_2 = \|\overset{\circ}{T}_{qq}\|_2 \ , \tag{3.2.25}$$

$$\|\overset{\circ}{T}_{pp}\|_2^2 + 2\|\overset{\circ}{T}_{pp}\|_2 = \|T_{pq}\|_2^2 \ . \tag{3.2.26}$$

Beweis:
Wir setzen wieder abkürzend $s_i = \sinh x_i$, $c_i = \cosh x_i$, $i=1,2$. Aus (2.2.1) erhalten wir

$$\|T_{pp}\|_2 = \|T_{qq}\|_2 = \max(c_1, c_2) \ ,$$

und mit $\cosh \xi = (1-\tanh^2 \xi)^{-\frac{1}{2}}$, $|\tanh \xi| \leq |\xi|$ für $\xi \in \mathbb{R}$ und (3.2.18) folgt dann sofort Behauptung (i).

Aus (2.2.1) ergibt sich weiterhin

$$\|T_{pq}\|_2 = \|T_{qp}\|_2 = \max(|s_1|,|s_2|) \quad ,$$

und mit $\sinh \xi = \tanh \xi \, (1-\tanh^2 \xi)^{-\frac{1}{2}}$ für $\xi \in \mathbb{R}$ erhalten wir aus (3.2.18)

$$|s_i| \leq 1.0000745 \; |x_i| \leq 6.085 \, \frac{\|A\|}{\delta^2} \, s(A) \leq 0.0123 \quad , \; i=1,2$$

und damit die Behauptung (ii).

Alsdann überlegt man sich leicht

$$\|\overset{\circ}{T}_{pp}\|_2 = \|\overset{\circ}{T}_{qq}\|_2 = \max(c_1-1,c_2-1) \quad .$$

Es sei o.B.d.A. $c_1 = \max(c_1,c_2)$, $|s_1| = \max(|s_1|,|s_2|)$. Hieraus folgt

$$\|\overset{\circ}{T}_{pp}\|_2^2 + 2\|\overset{\circ}{T}_{pp}\|_2 = (c_1-1)^2 + 2(c_1-1) = c_1^2 - 1 = s_1^2 = \|T_{pq}\|_2^2$$

und somit Behauptung (iii). ■

Wir können jetzt die Änderung des Pivotblocks A_{pq} und der Diagonalblöcke A_{pp} und A_{qq} unter einer T-Transformation abschätzen.

Lemma 3.2.9:
Unter den Voraussetzungen von Lemma 3.2.7 gilt nach einer T-Transformation mit $x^{(1)}$ bzw. x^* für $A' = T^{-1}AT$

$$\text{(i)} \quad A_{ii}' = A_{ii} + W_{ii} \; , \; \|W_{ii}\| \leq 69.579 \, \frac{\|A\|^3}{\delta^4} \, s^2(A) \; , \; i=p,q, \qquad (3.2.27)$$

$$\text{(ii)} \quad A_{pq}' = A_{pq} + W_{pq} \; , \; \|W_{pq}\| \leq 8.608 \, \frac{\|A\|^2}{\delta^2} \, s(A) \; . \qquad (3.2.28)$$

Außerdem gilt

$$\text{(iii)} \quad \|A_{pq}'\| \leq 10.023 \, \frac{\|A\|^2}{\delta^2} \, s(A) \leq \frac{1}{70} \, \delta \; . \qquad (3.2.29)$$

Beweis:

(i) Aus (2.2.4) erhält man für A_{pp}' die Darstellung

$$A_{pp}' = A_{pp} + W_{pp}$$

mit

$$W_{pp} = \overset{\circ}{T}_{pp} A_{pp} + A_{pp} \overset{\circ}{T}_{pp} + \overset{\circ}{T}_{pp} A_{pp} \overset{\circ}{T}_{pp} - T_{pq} A_{qp} T_{pp} + T_{pp} A_{pq} T_{qp} - T_{pq} A_{qq} T_{qp} \quad .$$

Es folgt dann aus (3.2.24),(3.2.26) und der J-Symmetrie von A

$$\|W_{pp}\| \leq \|T_{pq}\|_2^2 (\|A_{pp}\| + \|A_{qq}\|) + 2\|T_{pp}\|_2 \|T_{pq}\|_2 \|A_{pq}\| \quad .$$

Mit $\|A_{pp}\| + \|A_{qq}\| \leq \sqrt{2}\|A\|$ und $\|A_{pq}\| \leq \frac{\sqrt{2}}{2} S(A)$ und den Abschätzungen von Lemma 3.2.8 gilt damit

$$\begin{aligned} \|W_{pp}\| &\leq 52.365 \frac{\|A\|^3}{\delta^4} S^2(A) + 8.607 \frac{\|A\|}{\delta^2} S^2(A) \\ &\leq 69.579 \frac{\|A\|^3}{\delta^4} S^2(A) \quad , \end{aligned}$$

wobei in die letzte Ungleichung $\delta \leq \sqrt{2}\|A\|$ eingeht. Analog zeigt man die zweite Aussage von (i).

(ii) Für A_{pq}' erhält man ebenfalls aus (2.2.4) die Darstellung

$$A_{pq}' = A_{pq} + W_{pq}$$

mit

$$W_{pq} = T_{pp} A_{pp} T_{pq} - T_{pq} A_{qp} T_{pq} + \overset{\circ}{T}_{pp} A_{pq} + A_{pq} \overset{\circ}{T}_{qq} + \overset{\circ}{T}_{pp} A_{pq} \overset{\circ}{T}_{qq} - T_{pq} A_{qq} T_{qq} \quad .$$

Es folgt mit (3.2.23),(3.2.25),(3.2.26) und der J-Symmetrie von A

$$\|W_{pq}\| \leq \|T_{pp}\|_2 \|T_{pq}\|_2 (\|A_{pp}\| + \|A_{qq}\|) + 2\|T_{pq}\|_2^2 \|A_{pq}\| \quad .$$

Wie im Beweis von (i) ergibt sich dann unter Anwendung von Lemma 3.2.8 und $\delta \leq \sqrt{2}\|A\|$

$$\|W_{pq}\| \leq 8.608 \frac{\|A\|^2}{\delta^2} s^2(A) \quad .$$

(iii) Wegen $\|A_{pq}\| \leq \frac{\sqrt{2}}{2} s(A)$ und $\delta \leq \sqrt{2}\|A\|$ folgt aus (3.2.28) sofort

$$\|A_{pq}'\| \leq \frac{\sqrt{2}}{2} s(A) + 8.608 \frac{\|A\|^2}{\delta^2} s(A) \leq 10.023 \frac{\|A\|^2}{\delta^2} s(A)$$

und damit aus den Voraussetzungen (3.2.1)

$$\|A_{pq}'\| \leq \frac{1}{70} \delta \quad .$$ ■

Als nächstes schätzen wir die Normzunahme in der Außerdiagonalen nach einer T-Transformation ab.

Lemma 3.2.10:
Unter den Voraussetzungen von Lemma 3.2.7 gilt nach einer T-Transformation mit $x^{(1)}$ bzw. x^* für $A' = T^{-1}AT$

$$s(A') \leq 14.315 \frac{\|A\|^2}{\delta^2} s(A) \quad . \tag{3.2.30}$$

Beweis:
Nach (2.2.4) gilt wegen der J-Symmetrie von A und A'

$$s^2(A') - s^2(A) = 2 \sum_{i \neq p,q}^{m} (\|A_{pi}'\|^2 + \|A_{qi}'\|^2) - 2 \sum_{i \neq p,q}^{m} (\|A_{pi}\|^2 + \|A_{qi}\|^2)$$
$$+ 2\|A_{pq}'\|^2 - 2\|A_{pq}\|^2 \quad .$$

Weiter folgt aus (2.2.4) für $i \neq p,q$

$$\|A_{pi}'\| \leq \|T_{pp}\|_2 \|A_{pi}\| + \|T_{pq}\|_2 \|A_{qi}\| \quad ,$$

$$\|A_{qi}'\| \leq \|T_{qp}\|_2 \|A_{pi}\| + \|T_{qq}\|_2 \|A_{qi}\|$$

und damit nach Anwendung der Cauchy-Schwarzschen Ungleichung und und Lemma 3.2.8

$$\|A_{pi}'\|^2 \le 2.000299\,\|A_{pi}\|^2 + 0.000303\,\|A_{qi}\|^2 ,$$

$$\|A_{qi}'\|^2 \le 0.000303\,\|A_{pi}\|^2 + 2.000299\,\|A_{qi}\|^2 .$$

Hieraus ergibt sich

$$\begin{aligned} 2\sum_{i\neq p,q}^{m} (\|A_{pi}'\|^2+\|A_{qi}'\|^2) - 2\sum_{i\neq p,q}^{m} (\|A_{pi}\|^2+\|A_{qi}\|^2) & \\ \le 1.000602 \sum_{i\neq p,q}^{m} (2\|A_{pi}\|^2+2\|A_{qi}\|^2) & \\ \le 1.000602\, s^2(A) \quad . & \qquad (3.2.31) \end{aligned}$$

Aus Lemma 3.2.9 erhalten wir

$$\|A_{pq}'\|^2 \le \|A_{pq}\|^2 + 2\|A_{pq}\|\,\|W_{pq}\| + \|W_{pq}\|^2$$

und damit wegen $\|A_{pq}\| \le \frac{\sqrt{2}}{2}\, s(A)$ und (3.2.28)

$$\|A_{pq}'\|^2 - \|A_{pq}\|^2 \le 12.174\, \frac{\|A\|^2}{\delta^2}\, s^2(A) + 74.098\, \frac{\|A\|^4}{\delta^4}\, s^2(A) .$$

Insgesamt ergibt sich dann mit (3.2.31) und $\delta \le \sqrt{2}\,\|A\|$

$$s^2(A') \le 204.895\, \frac{\|A\|^4}{\delta^4}\, s^2(A)$$

und daraus die Behauptung. ■

Wir müssen jetzt noch für die nachfolgenden Rechnungen untersuchen, wie sich der Kommutator der Diagonalblöcke und der (p,q)-Restriktion von A unter einer T-Transformation verhält.

Lemma 3.2.11:
Unter den Voraussetzungen von Lemma 3.2.7 gilt nach einer T-Transformation mit $x^{(1)}$ bzw. x^* für $A' = T^{-1}AT$

$$\text{(i)} \quad \|C(A_{ii}')\| \le \frac{1}{1522}\,\delta^2 \ , \ i=p,q \ , \qquad (3.2.32)$$

$$\text{(ii)}\quad \|C(A_{pp}')\| + \|C(A_{qq}')\| \le \|C(A_{pp})\| + \|C(A_{qq})\| + 393.599\,\frac{\|A\|^4}{\delta^4}\,s^2(A)$$

$$\le \frac{1}{1076}\,\delta^2 \quad . \tag{3.2.33}$$

Beweis:

(i) Nach (3.2.27) können wir die transformierten Diagonalblöcke A_{pp}' und A_{qq}' in der Form

$$A_{ii}' = A_{ii} + W_{ii} \quad , \; i = p,q$$

darstellen. Hieraus erhalten wir nach einer einfachen Rechnung

$$C(A_{ii}') = A_{ii}'^{T}A_{ii}' - A_{ii}'A_{ii}'^{T}$$

$$= C(A_{ii}) + W_{ii}^{T}A_{ii}' - A_{ii}'W_{ii}^{T} + A_{ii}^{T}W_{ii} - W_{ii}A_{ii}^{T} \quad , \; i=p,q$$

und damit

$$\|C(A_{ii}')\| \le \|C(A_{ii})\| + 2\|A_{ii}'\|\,\|W_{ii}\| + 2\|A_{ii}\|\,\|W_{ii}\| \quad , i=p,q \; . \tag{3.2.34}$$

Es gilt $\|A_{ii}\| \le \|A\|$ und wegen (2.2.28) $\|A_{ii}'\| \le \|A'\| \le \|A\|$, also haben wir

$$\|C(A_{ii}')\| \le \|C(A_{ii})\| + 4\|A\|\,\|W_{ii}\| \quad , \; i=p,q \; .$$

Mit $\|C(A_{ii})\| \le \|C(D)\|$, (3.2.27), den Voraussetzungen (3.2.1) und $\delta \le \sqrt{2}\,\|A\|$ folgt dann die Behauptung.

(ii) Aus (3.2.34) erhalten wir mit $\|A_{pp}\| + \|A_{qq}\| \le \sqrt{2}\,\|A\|$, $\|A_{pp}'\| + \|A_{qq}'\| \le \sqrt{2}\,\|A'\| \le \sqrt{2}\,\|A\|$ und (3.2.27)

$$\|C(A_{pp}')\| + \|C(A_{qq}')\| \le \|C(A_{pp})\| + \|C(A_{qq})\| + 393.599\,\frac{\|A\|^4}{\delta^4}\,s^2(A)$$

und hieraus mit $\|C(A_{pp})\| + \|C(A_{qq})\| \le \sqrt{2}\,\|C(D)\|$, (3.2.1) und wieder $\delta \le \sqrt{2}\,\|A\|$

$$\| C(A_{pp}') \| + \| C(A_{qq}') \| \leq \frac{1}{1076} \delta^2 \ .$$

Lemma 3.2.12:
Unter den Voraussetzungen von Lemma 3.2.7 gilt nach einer T-Transformation mit $x^{(1)}$ bzw. x^* für $A' = T^{-1}AT$

$$\text{(i)} \quad \| \hat{C}_{pq}' \| \leq \| C_{pq}' \| + 0.513\ S^2(A) \ , \tag{3.2.35}$$

$$\text{(ii)} \quad \| \hat{C}_{pp}' \| + \| \hat{C}_{qq}' \| \leq \| C(A_{pp}) \| + \| C(A_{qq}) \| + 795.442 \frac{\|A\|^4}{\delta^4} S^2(A) \ . \tag{3.2.36}$$

Beweis:
(i) Nach (1.3.23),(1.3.25) und (2.2.10) haben wir

$$\hat{C}_{pq}' = C_{pq}' - \begin{pmatrix} 0 & 2\beta_{11}' \\ 2\beta_{22}' & 0 \end{pmatrix} \ .$$

Aus (2.2.12),(2.2.15),(3.2.18) und (3.2.15) folgt

$$|\beta_{ii}'| \leq \alpha_{ii} |\sinh 2x_i| + \alpha_{ii} \cosh 2x_i \leq 1.0248\ \alpha_{ii} \ , \ i=1,2 \ ,$$

und mit $\sqrt{\xi+\eta} \leq \sqrt{\xi} + \sqrt{\eta}$ für $\xi,\eta \geq 0$ und $\alpha_{11} + \alpha_{22} \leq \frac{1}{4} S^2(A)$ erhalten wir dann

$$\| \hat{C}_{pq}' \| \leq \| C_{pq}' \| + 2\sqrt{\beta_{11}'^2 + \beta_{22}'^2} \leq \| C_{pq}' \| + 0.513\ S^2(A) \ .$$

(ii) Wie man leicht nachweist, gelten die Identitäten

$$\hat{C}_{pp} = C(A_{pp}) + A_{qp}^T A_{qp} - A_{pq} A_{pq}^T \ ,$$

$$\hat{C}_{qq} = C(A_{qq}) + A_{pq}^T A_{pq} - A_{qp} A_{qp}^T \ .$$

Hieraus folgt dann aufgrund der J-Symmetrie von A'

$$\| \hat{C}_{pp}' \| + \| \hat{C}_{qq}' \| \leq \| C(A_{pp}') \| + \| C(A_{qq}') \| + 4 \| A_{pq}' \|^2 \ ,$$

und mit (3.2.33) und (3.2.29) ergibt sich hieraus die Behauptung. ■

Zum Schluß dieses Abschnitts beweisen wir noch zwei technische Aussagen, die die Änderung von $\delta_D(A)$ und $d^{(p,q)}(A)$ unter einer T-Transformation betreffen.

Lemma 3.2.13:
Unter den Voraussetzungen von Lemma 3.2.7 gilt nach einer T-Transformation mit $x^{(1)}$ bzw. x^* für $A' = T^{-1}AT$

$$\delta_D(A') \geq \frac{12}{13}\,\delta \quad . \tag{3.2.37}$$

Beweis:
Der Beweis wird analog zum Beweis von Lemma 3.2.3 geführt. Wir haben hier

$$\Delta(A_{ii}') \leq \sqrt[4]{\tfrac{1}{2}}\,\sqrt{\|C(A_{ii}')\|} \leq \frac{\sqrt[4]{\frac{1}{2}}}{\sqrt{1522}}\,\delta\,,\quad i = 1(1)m.$$

Dabei werden $\|C(A_{pp}')\|$ und $\|C(A_{qq}')\|$ nach (3.2.32) abgeschätzt, während die Diagonalblöcke A_{ii} für $i \neq p,q$ unter der T-Transformation invariant bleiben und daher die Abschätzung für $\|C(A_{ii}')\|$, $i \neq p,q$ aus den Generalvoraussetzungen (3.2.1) folgt.

Weiter erhalten wir aus der Cauchy-Schwarzschen Ungleichung, (3.2.30), (3.2.1) und $\delta \leq \sqrt{2}\,\|A\|$

$$\sum_{j\neq i}^{m} \|A_{ij}'\| \leq \frac{\sqrt{m-1}}{\sqrt{2}}\,S(A') \leq \frac{1}{69}\,\delta\,,\quad i = 1(1)m\,.$$

Es folgt dann

$$\Delta(A_{ii}') + \sum_{j\neq i}^{m} \|A_{ij}'\| < \frac{1}{26}\,\delta\,,\quad i = 1(1)m$$

und daraus mit derselben Argumentation wie in Lemma 3.2.3 die Behauptung. ■

Lemma 3.2.14:
Unter den Voraussetzungen von Lemma 3.2.7 gilt nach einer T-Transformation mit $x^{(1)}$ bzw. x^* für $A' = T^{-1}AT$

$$d^{(p,q)}(A') \geq \frac{7}{8}\delta . \tag{3.2.38}$$

Beweis:
Der Beweis verläuft analog zum Beweis von Lemma 3.2.4. Es gilt hier wegen (3.2.37)

$$\delta \leq \frac{13}{12}\left(d^{(p,q)}(A') + \frac{1}{\sqrt[4]{2}}\left(\sqrt{\| C(A'_{pp})\|} + \sqrt{\| C(A'_{qq})\|}\right)\right)$$

und damit wegen $\sqrt{\xi}+\sqrt{\eta} \leq \sqrt{2}\sqrt{\xi+\eta}$ für $\xi,\eta \geq 0$ und (3.2.33)

$$\delta \leq \frac{13}{12}\left(d^{(p,q)}(A') + \frac{\sqrt[4]{2}}{\sqrt{1076}}\,\delta\right) .$$

Hieraus folgt die Behauptung. ■

3.3 DIE ELEMENTARE JACOBI-PAARDEKOOPER-MATRIX

Es sei A' eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$, und es gelte $1 \le p < q \le m$, p,q fest. Wir definieren die elementare Matrix $U = U(y_1,y_2)$ durch

$$U_{pp} = \begin{pmatrix} \cos y_1 & 0 \\ 0 & \cos y_2 \end{pmatrix}, \quad U_{pq} = \begin{pmatrix} -\sin y_1 & 0 \\ 0 & -\sin y_2 \end{pmatrix},$$

$$U_{qp} = \begin{pmatrix} \sin y_1 & 0 \\ 0 & \sin y_2 \end{pmatrix}, \quad U_{qq} = \begin{pmatrix} \cos y_1 & 0 \\ 0 & \cos y_2 \end{pmatrix}, \qquad y_1,y_2 \in (-\tfrac{\pi}{2},\tfrac{\pi}{2}],$$

$$U_{ij} = I\delta_{ij} \quad \text{für} \quad i,j = 1(1)m,\ (i,j) \neq (p,p),(p,q),(q,p),(q,q)\ . \tag{3.3.1}$$

Diese elementare Matrix kann man als zweiparametrige Verallgemeinerung der klassischen Rotationsmatrix von Jacobi ([28]) auffassen. Andrerseits stellt sie auch einen Spezialfall der vierparametrigen Rotationsmatrix von Paardekooper ([39]) dar. Jacobi benutzte seine Matrix zur Eliminierung eines Außerdiagonalelements einer symmetrischen Matrix, während Paardekooper durch Anwendung seiner Matrix einen außerdiagonalen 2x2-Block einer schiefsymmetrischen Matrix annullierte.

In § 4.1 werden Jacobi-ähnliche Blockverfahren formuliert, welche die elementare Matrix U zu zwei verschiedenen Zwecken verwenden. Einerseits dient sie zur Annullierung des Pivotblocks $A_{pq}^{'+}$ des symmetrischen Teils $A^{'+}$ von A', andrerseits eliminiert sie den Pivotblock $A_{pq}^{'-}$ des schiefsymmetrischen Teils $A^{'-}$ von A'. Dies ist mit zwei Rotationen möglich, da die Pivotblöcke $A_{pq}^{'+}$ und $A_{pq}^{'-}$ aufgrund der J-Symmetrie von A' jeweils nur aus zwei Elementen bestehen.

Wir nennen U die *elementare Jacobi-Paardekooper-Matrix*. Eine Ähnlichkeitstransformation mit U nennen wir *Jacobi-Transformation*, wenn $A_{pq}^{'+}$ eliminiert wird, und *Paardekooper-Transformation*, wenn $A_{pq}^{'-}$ eliminiert wird.

Die Verfahren aus § 4.1 führen an einem festen Pivotblock A_{pq} zunächst eine T-Transformation und anschließend eine U-Transformation

aus. Wir setzen daher im folgenden stets voraus, daß sich für ein festes Pivotpaar (p,q), $1 \le p < q \le m$ die Eingangsmatrix A' für die U-Transformation als Ausgangsmatrix einer T-Transformation ergibt. Die Eingangsmatrix A der T-Transformation erfülle dabei die Voraussetzungen (3.2.1), d.h. sie besitze fast Murnaghan-Form. Wegen (3.2.30) und (3.2.32) hat dann aber auch A' fast Murnaghan-Form.

Die wichtigste Aussage dieses Paragraphen wird in Lemma 3.3.8 formuliert. Wir zeigen hier, daß unter den obigen Voraussetzungen der Pivotblock A_{pq} nach einer U-Transformation quadratisch klein wird.

Zunächst aber konstatieren wir einige grundlegende Eigenschaften der Jacobi-Paardekooper-Matrix, von deren Richtigkeit man sich leicht überzeugt:

U ist orthogonal und J-orthogonal und damit nicht-singulär. Es gilt

$$\begin{aligned} &U(0,0) = I \ , \quad U(y_1,y_2)^{-1} = U(-y_1,-y_2) \ , \\ &U_{pp} = U_{qq} \ , \quad U_{pq} = -U_{qp} \ . \end{aligned} \tag{3.3.2}$$

Für feste $y_1,y_2 \in (-\frac{\pi}{2},\frac{\pi}{2}]$ sei

$$A'' = U(y_1,y_2)^{-1} A' U(y_1,y_2) \ . \tag{3.3.3}$$

Dann ist A'' wieder J-symmetrisch. Es wird unter (3.3.3) nur die p-te und q-te Blockzeile und -spalte von A' transformiert. Wegen (3.3.2) berechnet sich A'' als

$$\begin{aligned} &A''_{ij} = A'_{ij} \ , \quad i \neq p,q \ , \ j \neq p,q \ , \\ &\left.\begin{aligned} A''_{pi} &= U_{pp}A'_{pi} - U_{pq}A'_{qi} \\ A''_{qi} &= -U_{qp}A'_{pi} + U_{qq}A'_{qi} \end{aligned}\right\} \quad i = 1(1)m, \ i \neq p,q \ , \\ &A''_{pp} = U_{pp}A'_{pp}U_{pp} - U_{pq}A'_{qp}U_{pp} + U_{pp}A'_{pq}U_{qp} - U_{pq}A'_{qq}U_{qp} \ , \\ &A''_{pq} = U_{pp}A'_{pp}U_{pq} - U_{pq}A'_{qp}U_{pq} + U_{pp}A'_{pq}U_{qq} - U_{pq}A'_{qq}U_{qq} \ , \\ &A''_{qq} = -U_{qp}A'_{pp}U_{pq} + U_{qq}A'_{qp}U_{pq} - U_{qp}A'_{pq}U_{qq} + U_{qq}A'_{qq}U_{qq} \ , \end{aligned} \tag{3.3.4}$$

und die p-te und q-te Blockspalte und A_{qp}'' ergeben sich aus der J-Symmetrie von A''.

Aus der Orthogonalität von U folgt

$$\left.\begin{aligned} \|A_{pi}''\|^2 + \|A_{qi}''\|^2 &= \|A_{pi}'\|^2 + \|A_{qi}'\|^2 \\ \|A_{ip}''\|^2 + \|A_{iq}''\|^2 &= \|A_{ip}'\|^2 + \|A_{iq}'\|^2 \end{aligned}\right\} \quad i = 1(1)m,\ i \neq p,q \qquad (3.3.5)$$

und daraus aufgrund der J-Symmetrie von A' und A''

$$S^2(A'') - S^2(A') = 2\|A_{pq}''\|^2 - 2\|A_{pq}'\|^2 \ . \qquad (3.3.6)$$

Die Parameter y_1, y_2 (auch *Winkel* genannt) der Jacobi-Paardekooper-Matrix werden nun folgendermaßen bestimmt:

Zur Eliminierung von $A_{pq}'^{+}$ (Jacobi-Transformation) berechnen sich y_1, y_2 aus (vgl.[28])

$$\tan 2y_1 = \frac{2a'_{2p-1,2q-1}}{a'_{2p-1,2p-1} - a'_{2q-1,2q-1}} \ , \ y_1 \in (-\tfrac{\pi}{4}, \tfrac{\pi}{4}] \ , \qquad (3.3.7)$$

$$\tan 2y_2 = \frac{2a'_{2p,2q}}{a'_{2p,2p} - a'_{2q,2q}} \ , \ y_2 \in (-\tfrac{\pi}{2}, \tfrac{\pi}{2}] \ . \qquad (3.3.8)$$

Dabei ist y_1 durch (3.3.7) eindeutig definiert, während für (3.3.8) zwei Lösungen $y_2^{(i)}$, i=1,2 existieren. So ist aufgrund der π-Periodizität des Tangens mit $y_2^{(1)} \in (-\frac{\pi}{4}, 0]$ auch

$$y_2^{(2)} = y_2^{(1)} + \frac{\pi}{2}$$

und mit $y_2^{(1)} \in (0, \frac{\pi}{4}]$ auch

$$y_2^{(2)} = y_2^{(1)} - \frac{\pi}{2}$$

eine Lösung von (3.3.8). Es gilt dann für $y_2^{(1)} \in (-\frac{\pi}{4}, 0]$

$$\cos y_2^{(2)} = -\sin y_2^{(1)} , \quad \sin y_2^{(2)} = \cos y_2^{(1)}$$

und damit

$$U(y_1, y_2^{(2)}) = U(y_1, y_2^{(1)}) V^{(2p,2q)} D^{(2q)} ,$$

und für $y_2^{(1)} \in (0, \frac{\pi}{4}]$

$$\cos y_2^{(2)} = \sin y_2^{(1)} , \quad \sin y_2^{(2)} = -\cos y_2^{(1)} ,$$

$$U(y_1, y_2^{(2)}) = U(y_1, y_2^{(1)}) V^{(2p,2q)} D^{(2p)} .$$

Eine Jacobi-Transformation mit den Winkeln $y_1, y_2^{(2)}$ entspricht daher einer Jacobi-Transformation mit $y_1, y_2^{(1)}$, gefolgt von einer Permutation der 2p-ten und 2q-ten Zeile und Spalte von A'' und einer Skalierung der 2p-ten oder 2q-ten Zeile und Spalte von A'' mit -1.

Diese alternative Parameterwahl ermöglicht uns, das Phänomen der "gestreuten" Diagonalblöcke zu verhindern. So kann die 4x4-Matrix $\hat{A}_{pq}''$ nach einer Jacobi-Transformation mit $y_1, y_2^{(1)}$ das Aussehen

$$\begin{pmatrix} * & \bullet & 0 & * \\ \bullet & * & * & 0 \\ 0 & * & * & \bullet \\ * & 0 & \bullet & * \end{pmatrix}$$

haben, wobei "*" für ein "großes" und "•" für ein "kleines" Element steht. Die oben beschriebene Permutation ergibt dann

$$\begin{pmatrix} * & * & 0 & \bullet \\ * & * & \bullet & 0 \\ 0 & \bullet & * & * \\ \bullet & 0 & * & * \end{pmatrix}$$

für $\hat{A}_{pq}''$ (vgl.[53],[26],[27]). Die folgende Abfrage stellt ein sinnvolles Kriterium für die Entscheidung dar, ob eine Vertauschung mit anschließender Skalierung vorgenommen werden soll:

Es sei $A'' = U(y_1, y_2^{(1)})^{-1} A' U(y_1, y_2^{(1)})$. Falls

$$|a''_{2p-1,2q}| + |a''_{2p,2q-1}| > |a''_{2p-1,2p}| + |a''_{2q-1,2q}| \tag{3.3.9}$$
$$\wedge\ (|a''_{2p-1,2p-1} - a''_{2q,2q}| < 2|a''_{2p-1,2q}| \vee |a''_{2p,2p} - a''_{2q-1,2q-1}| < 2|a''_{2p,2q-1}|),$$

so berechne A'' als $U(y_1, y_2^{(2)})^{-1} A' U(y_1, y_2^{(2)})$ (vgl.[53]).

Hierbei stellt die zweite Bedingung von (3.3.9) sicher, daß die Eigenwerte der gestreuten Blöcke nicht alle reell sind. In einem solchen Fall ist eine Permutation unnötig.

Nach der Jacobi-Transformation gilt dann $A''^{+}_{pq} = 0$ und (s.[28])

$$S^2(A''^{+}) = S^2(A'^{+}) - 2\|A'^{+}_{pq}\|^2 \ . \tag{3.3.10}$$

Zur Annullierung von A'^{-}_{pq} (Paardekooper-Transformation) berechnen sich y_1, y_2 aus (vgl.[39])

$$\tan 2y_1 = -2\frac{a'_{2p-1,2p}a'_{2p,2q-1} - a'_{2q-1,2q}a'_{2p-1,2q}}{a'^2_{2p-1,2p} - a'^2_{2q-1,2q} + a'^2_{2p-1,2q} - a'^2_{2p,2q-1}}\ , \ y_1 \in (-\tfrac{\pi}{4}, \tfrac{\pi}{4}], \tag{3.3.11}$$

$$\tan y_2 = \begin{cases} \dfrac{a'_{2p-1,2p}\cos y_1 + a'_{2q-1,2q}\sin y_1}{a'_{2p-1,2p}\cos y_1 - a'_{2p,2q-1}\sin y_1} \\ \\ \dfrac{a'_{2p,2q-1}\cos y_1 + a'_{2p-1,2p}\sin y_1}{a'_{2q-1,2q}\cos y_1 - a'_{2p-1,2q}\sin y_1} \end{cases}, \ y_2 \in (-\tfrac{\pi}{2}, \tfrac{\pi}{2}]. \tag{3.3.12}$$

Hierbei wählt man die obere Formel von (3.3.12), wenn

$$|a'_{2p-1,2p}\cos y_1 - a'_{2p,2q-1}\sin y_1| > |a'_{2q-1,2q}\cos y_1 - a'_{2p-1,2q}\sin y_1| \tag{3.3.13}$$

gilt, ansonsten die untere.

Durch (3.3.11),(3.3.12) und (3.3.13) sind y_1 und y_2 eindeutig definiert, und nach der Paardekooper-Transformation gilt $A''^{-}_{pq} = 0$ und (s.[39])

$$s^2(A''^-) = s^2(A'^-) - 2\|A_{pq}'^-\|^2 \quad . \tag{3.3.14}$$

Die Entscheidung, ob wir eine Jacobi- oder Paardekooper-Transformation durchführen, wird mit Hilfe des folgenden Kriteriums getroffen:

Falls

$$\begin{aligned}\|A_{pq}'^+\| + \min\{&|a'_{2p-1,2p-1}-a'_{2q-1,2q-1}|, |a'_{2p-1,2p-1}-a'_{2q,2q}|,\\ &|a'_{2p,2p}-a'_{2q-1,2q-1}|, |a'_{2p,2p}-a'_{2q,2q}|\}\\ > \|A_{pq}'^-\| + &\big||a'_{2p-1,2p}| - |a'_{2q-1,2q}|\big|\end{aligned} \tag{3.3.15}$$

gilt, so wird eine Jacobi-Transformation ausgeführt, sonst eine Paardekooper-Transformation.

Diesem Kriterium liegt die folgende Idee zugrunde: Solange noch keine Blockdiagonaldominanz vorliegt, ist es wegen (3.3.10) und (3.3.14) günstiger, den Teil von A'_{pq} zu eliminieren, der größer ist. Daher rühren die jeweils ersten Summanden in (3.3.15). Hat die Matrix A' jedoch schon fast Murnaghan-Form, so liefern die zweiten Summanden Näherungen für die Trennung der Real- bzw. Imaginärteile der Eigenwerte von A'. Diese bestimmen dann, da die Terme $\|A_{pq}'^+\|$ und $\|A_{pq}'^-\|$ unbedeutend klein sind, welche Transformation durchgeführt wird. Ähnliche Kriterien wurden von Eberlein ([10]), Veselić ([53]) und Wenzel ([56]) benutzt.

Für die nachfolgenden Untersuchungen müssen wir nun die Winkel der Jacobi- bzw. Paardekooper-Transformation abschätzen. Dabei setzen wir für den Winkel y_2 der Jacobi-Transformation zunächst $y_2 \in (-\frac{\pi}{4}, \frac{\pi}{4}]$ voraus. Später zeigen wir dann, daß diese Einschränkung unter den gegebenen Voraussetzungen immer gilt.

Lemma 3.3.1:
A sei eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$. Es gelte $\delta > 0$ und die Voraussetzungen (3.2.1). Es sei $1 \le p < q \le m$, p,q fest, und A' sei durch eine T-Transformation für den Pivotblock

A_{pq} mit den Parametern $x^{(1)}$ bzw. x^* (vgl. Lemma 3.2.7) gegeben.

Die elementare Matrix U sei durch (3.3.7) und (3.3.8) bestimmt, wobei $y_2 \in (-\frac{\pi}{4}, \frac{\pi}{4}]$ gelte. Dann genügen die Winkel y_1, y_2 den Ungleichungen

$$|\sin y_i| \leq 2.324 \frac{\|A_{pq}'^{+}\|}{\delta} \quad , \; i = 1,2 \; . \tag{3.3.16}$$

Beweis:
Es gilt wegen $y_1 \in (-\frac{\pi}{4}, \frac{\pi}{4}]$

$$|\sin y_1| \leq |\tan y_1| \leq \frac{1}{2}|\tan 2y_1|$$

und damit nach (3.3.7)

$$|\sin y_1| \leq \frac{|a'_{2p-1,2q-1}|}{|a'_{2p-1,2p-1} - a'_{2q-1,2q-1}|} \; .$$

Aufgrund von (3.3.15) gilt weiter

$$\begin{aligned} |a'_{2p-1,2p-1} - a'_{2q-1,2q-1}| &\geq \frac{1}{2} d^{(p,q)}(A') + \frac{1}{2}\|A_{pq}'^{-}\| - \frac{1}{2}\|A_{pq}'^{+}\| \\ &\geq \frac{1}{2} d^{(p,q)}(A') - \frac{1}{2}\|A_{pq}'\| \qquad (3.3.17) \\ &\geq \frac{241}{560}\delta \; , \end{aligned}$$

wobei die letzte Abschätzung aus (3.2.29) und (3.2.38) folgt. Mit $|a'_{2p-1,2q-1}| \leq \|A_{pq}'^{+}\|$ erhalten wir dann schließlich

$$|\sin y_1| \leq 2.324 \frac{\|A_{pq}'^{+}\|}{\delta} \; .$$

Die Abschätzung für $\sin y_2$ wird analog bewiesen, da wir $y_2 \in (-\frac{\pi}{4}, \frac{\pi}{4}]$ vorausgesetzt haben. ■

Lemma 3.3.2:
Es gelten die Voraussetzungen von Lemma 3.3.1. Die elementare Matrix U sei durch (3.3.11), (3.3.12) und (3.3.13) bestimmt. Dann genügen die

Winkel y_1, y_2 den Ungleichungen

$$|\sin y_1| \leq 2.327 \frac{\|A_{pq}'^{-}\|}{\delta} \quad , \tag{3.3.18}$$

$$|\sin y_2| \leq 4.678 \frac{\|A_{pq}'^{-}\|}{\delta} \quad . \tag{3.3.19}$$

Beweis:

Der Beweis folgt im wesentlichen den Ideen von Hari ([20]). Jedoch läßt Hari die Entscheidung für eine der beiden Formeln (3.3.12) generell offen und wählt dann im Beweis die jeweils geeignete. Durch Einbeziehung des Kriteriums (3.3.13) zur Berechnung von y_2 wird die Beweisführung bei uns aufwendiger.

Wir setzen zur Vereinfachung wieder o.B.d.A. $p = 1$, $q = 2$, und es sei $c_i = \cos y_i$, $s_i = \sin y_i$, $i=1,2$. Zunächst einmal gilt wegen (3.3.15),(3.2.29) und (3.2.38)

$$\begin{aligned}
\sqrt{a_{12}'^2 + a_{34}'^2} &\geq \max\{|a_{12}'|, |a_{34}'|\} \\
&\geq \big||a_{12}'| - |a_{34}'|\big| \\
&\geq \tfrac{1}{2} d^{(1,2)}(A') + \tfrac{1}{2}\|A_{12}'^{+}\| - \tfrac{1}{2}\|A_{12}'^{-}\| \qquad (3.3.20) \\
&\geq \tfrac{1}{2} d^{(1,2)}(A') - \tfrac{1}{2}\|A_{12}'\| \\
&\geq \frac{241}{560}\,\delta \quad .
\end{aligned}$$

Alsdann erhalten wir mit Hilfe der Dreiecksungleichung und der Cauchy-Schwarzschen Ungleichung

$$|\tan 2y_1| \leq 2 \frac{|a_{12}'||a_{23}'| + |a_{34}'||a_{14}'|}{\big||a_{12}'^2 - a_{34}'^2| - |a_{14}'^2 - a_{23}'^2|\big|}$$

$$\leq 2 \frac{\sqrt{a_{12}'^2 + a_{34}'^2}\ \sqrt{a_{14}'^2 + a_{23}'^2}}{\big||a_{12}'^2 - a_{34}'^2| - |a_{14}'^2 - a_{23}'^2|\big|}$$

$$= 2\,\frac{\|A_{12}'^{-}\|}{\left|\frac{|a_{12}'|+|a_{34}'|}{\sqrt{a_{12}'^2+a_{34}'^2}}\,\big||a_{12}'|-|a_{34}'|\big| - \frac{|a_{14}'^2-a_{23}'^2|}{\sqrt{a_{12}'^2+a_{34}'^2}}\right|}\quad,$$

und es folgt mit (3.3.20) und $\sqrt{\xi+\eta} \le \sqrt{\xi}+\sqrt{\eta}$ für $\xi,\eta \ge 0$

$$\frac{|a_{12}'|+|a_{34}'|}{\sqrt{a_{12}'^2+a_{34}'^2}}\,\big||a_{12}'|-|a_{34}'|\big| \ge \frac{241}{560}\,\delta$$

und wieder mit (3.3.20) und $|a_{14}'^2-a_{23}'^2| \le a_{14}'^2+a_{23}'^2 \le \|A_{12}'\|^2 \le \frac{1}{4900}\,\delta^2$ (vgl.(3.2.29))

$$\frac{|a_{14}'^2-a_{23}'^2|}{\sqrt{a_{12}'^2+a_{34}'^2}} \le \frac{4}{8435}\,\delta\quad.$$

Hieraus ergibt sich dann insgesamt wegen $y_1 \in (-\frac{\pi}{4},\frac{\pi}{4}]$

$$|\sin y_1| \le |\tan y_1| \le \frac{1}{2}|\tan 2y_1| \le 2.327\,\frac{\|A_{12}'^{-}\|}{\delta}\quad. \qquad (3.3.21)$$

Es sei nun die Ungleichung (3.3.13) erfüllt. Dann gilt

$$\begin{aligned}|a_{12}'|c_1 + |a_{23}'||s_1| &\ge |a_{12}'c_1 - a_{23}'s_1|\\ &\ge |a_{34}'c_1 - a_{14}'s_1|\\ &\ge |a_{34}'|c_1 - |a_{14}'||s_1|\end{aligned}$$

und damit

$$|a_{34}'| - |a_{12}'| \le (|a_{14}'| + |a_{23}'|)\,|\tan y_1|.$$

Aus $|a_{14}'|+|a_{23}'| \le \sqrt{2}\,\|A_{12}'^{-}\|$ und (3.3.21) folgt dann mit $\|A_{12}'^{-}\| \le \|A_{12}'\|$ und (3.2.29)

$$|a_{34}'| - |a_{12}'| \leq \frac{1}{1470}\,\delta\ .$$

Zusammen mit (3.3.20) impliziert diese Ungleichung

$$|a_{12}'| \geq |a_{34}'| \quad . \tag{3.3.22}$$

Zur weiteren Beweisführung wird nun eine Fallunterscheidung durchgeführt. Es sei zunächst $|a_{34}'| \leq \frac{241}{560}\,\delta$.

Nach (3.3.13) berechnet sich y_2 aus der oberen Formel von (3.3.12). Es gilt somit

$$|\sin y_2| \leq |\tan y_2| \leq \frac{|a_{14}'|c_1 + |a_{34}'||s_1|}{\big||a_{12}'|c_1 - |a_{23}'||s_1|\big|} \quad . \tag{3.3.23}$$

Aus (3.3.21) folgt wieder mit $\|A_{12}'^{-}\| \leq \|A_{12}'\|$ und (3.2.29)

$$|s_1| \leq 0.0333 \tag{3.3.24}$$

und hieraus mit $\sqrt{1-\xi'} \geq 1 - \frac{\xi}{2\sqrt{1-\xi'}}$ für $0 \leq \xi < 1$

$$c_1 \geq 0.999445. \tag{3.3.25}$$

Wegen (3.3.20),(3.3.22) und $|a_{23}'| \leq \|A_{12}'\| \leq \frac{1}{70}\,\delta$ (vgl.(3.2.29)) ergibt sich dann

$$\big||a_{12}'|c_1 - |a_{23}'||s_1|\big| \geq 0.429\,\delta\ ,$$

und aus (3.3.21) und $|a_{14}'| \leq \|A_{12}'\|$ erhalten wir insgesamt

$$|\sin y_2| \leq 4.678\,\frac{\|A_{12}'^{-}\|}{\delta} \quad .$$

Alsdann sei $|a_{34}'| > \frac{241}{560}\,\delta$. Wie in (3.3.23) gilt

$$|\sin y_2| \leq \frac{\frac{|a_{14}'|}{|a_{34}'|} c_1 + |s_1|}{\left| \frac{|a_{12}'|}{|a_{34}'|} c_1 - \frac{|a_{23}'|}{|a_{34}'|} |s_1| \right|} \quad .$$

Mit Hilfe von (3.3.22),(3.3.24),(3.3.25) und erneut mit $|a_{23}'| \leq \frac{1}{70} \delta$ erhalten wir

$$\left| \frac{|a_{12}'|}{|a_{34}'|} c_1 - \frac{|a_{23}'|}{|a_{34}'|} |s_1| \right| \geq 0.99833 \ ,$$

und aus (3.3.21) und $|a_{14}'| \leq \|A_{12}'^{-}\|$ folgt dann auch in diesem Fall die Behauptung (3.3.19).

Wenn nun die Ungleichung (3.3.13) nicht erfüllt ist, so zeigt man analog

$$|a_{34}'| \geq |a_{12}'| \ .$$

Die Abschätzung (3.3.19) wird dann mit der unteren Formel von (3.3.12) und vertauschten Rollen von a_{12}' und a_{34}' bzw. a_{14}' und a_{23}' bewiesen. ■

Wie man sieht, sind die Abschätzungen für die Jacobi-Transformation wesentlich schärfer als die für die Paardekooper-Transformation, insbesondere für den Winkel y_2. Jedoch wollen wir im folgenden, um die weiteren Rechnungen einheitlich durchführen zu können, die Winkel y_i immer mit (3.3.18) und (3.3.19) abschätzen, unabhängig davon, welche der beiden Transformationen ausgeführt wird.

Analog zum Vorgehen in § 3.2 untersuchen wir nun die Veränderungen der Matrix A' unter einer U-Transformation. Zu diesem Zweck beweisen wir zunächst einige Aussagen über die nicht-trivialen Blöcke von U.

Lemma 3.3.3:
Es gelten die Voraussetzungen von Lemma 3.3.1. Die Parameter y_1, y_2 der elementaren Matrix U seien durch (3.3.7) und (3.3.8) bzw. (3.3.11),

(3.3.12) und (3.3.13) bestimmt. Dann folgt

(i) $$\|U_{pp}\|_2 \leq 1 \quad , \tag{3.3.26}$$

(ii) $$\|U_{pq}\|_2 \leq 46.888 \frac{\|A\|^2}{\delta^3} S(A) \leq 0.0669 \quad . \tag{3.3.27}$$

(iii) Für $\overset{\circ}{U}_{pp} = U_{pp} - I$ gilt

$$\|\overset{\circ}{U}_{pp}\|_2^2 + 2\|\overset{\circ}{U}_{pp}\|_2 \leq 1.00450 \, \|U_{pq}\|_2^2 \quad . \tag{3.3.28}$$

Beweis:
Wir setzen wieder abkürzend $s_i = \sin y_i$, $c_i = \cos y_i$, i=1,2. Aus (3.3.1) folgt dann sofort

$$\|U_{pp}\|_2 = \max(c_1, c_2) \leq 1.$$

Weiterhin gilt

$$\|U_{pq}\|_2 = \max(|s_1|, |s_2|)$$

und damit nach Lemma 3.3.1, Lemma 3.3.2 und (3.2.29)

$$\|U_{pq}\|_2 \leq 46.888 \frac{\|A\|^2}{\delta^3} S(A) \leq 0.0669 \quad .$$

Alsdann verifiziert man leicht

$$\|\overset{\circ}{U}_{pp}\|_2 = \max(1-c_1, 1-c_2) \quad .$$

Es sei nun o.B.d.A. $1-c_2 = \max(1-c_1, 1-c_2)$, $|s_2| = \max(|s_1|, |s_2|)$. Hieraus ergibt sich dann wegen $1 - \sqrt{1-\xi} \leq \frac{\xi}{2\sqrt{1-\xi}}$ für $0 \leq \xi < 1$ und (3.3.27)

$$\|\overset{\circ}{U}_{pp}\|_2^2 + 2\|\overset{\circ}{U}_{pp}\|_2 = (1-c_2)^2 + 2(1-c_2) = 4(1-c_2) - s_2^2$$

$$\leq 2\frac{s_2^2}{\sqrt{1-s_2^2}} - s_2^2 \leq 1.00450 \, \|U_{pq}\|_2^2 \quad . \qquad \blacksquare$$

In den nächsten zwei Lemmata schätzen wir die Änderung der Blöcke A_{pp}', A_{qq}' und A_{pq}' und die Normzunahme in der Außerdiagonalen unter einer U-Transformation ab.

Lemma 3.3.4:
Unter den Voraussetzungen von Lemma 3.3.1 gilt nach einer Jacobi- bzw. Paardekooper-Transformation für $A'' = U^{-1}A'U$

$$\text{(i)} \quad A_{ii}'' = A_{ii}' + W_{ii}' \ , \ \|W_{ii}'\| \le 4453 \frac{\|A\|^5}{\delta^6} s^2(A) \ , \ i = p,q \ , \tag{3.3.29}$$

$$\text{(ii)} \quad A_{pq}'' = A_{pq}' + W_{pq}' \ , \ \|W_{pq}'\| \le 66.437 \frac{\|A\|^3}{\delta^3} s(A) \ . \tag{3.3.30}$$

Beweis:
(i) Aus (3.3.4) erhalten wir

$$A_{pp}'' = A_{pp}' + W_{pp}'$$

mit

$$W_{pp}' = \overset{\circ}{U}_{pp}A_{pp}' + A_{pp}'\overset{\circ}{U}_{pp} + \overset{\circ}{U}_{pp}A_{pp}'\overset{\circ}{U}_{pp} - U_{pq}A_{qp}'U_{pp} + U_{pp}A_{pq}'U_{qp} - U_{pq}A_{qq}'U_{qp}.$$

Aufgrund der J-Symmetrie von A' können wir W_{pp}' mit Hilfe von (3.3.2) und (3.3.28) folgendermaßen abschätzen:

$$\|W_{pp}'\| \le 1.00450\|U_{pq}\|_2^2(\|A_{pp}'\|+\|A_{qq}'\|) + 2\|U_{pp}\|_2 \|U_{pq}\|_2 \|A_{pq}'\| \ .$$

Hieraus folgt dann mit $\|A_{pp}'\|+\|A_{qq}'\| \le \sqrt{2}\|A'\| \le \sqrt{2}\|A\|$, mit (3.2.29), den Abschätzungen von Lemma 3.3.3 und $\delta \le \sqrt{2}\|A\|$

$$\|W_{pp}'\| \le 4453 \frac{\|A\|^5}{\delta^6} s^2(A) \ .$$

Die zweite Aussage von (i) wird analog bewiesen.

(ii) Für A_{pq}'' erhalten wir nach (3.3.4) die Darstellung

$$A_{pq}'' = A_{pq}' + W_{pq}'$$

mit

$$W_{pq}' = U_{pp}A_{pp}'U_{pq} - U_{pq}A_{qp}'U_{pq} + \overset{\circ}{U}_{pp}A_{pq}' + A_{pq}'\overset{\circ}{U}_{qq} + \overset{\circ}{U}_{pp}A_{pq}'\overset{\circ}{U}_{qq} - U_{pq}A_{qq}'U_{qq} .$$

Es folgt dann wieder mit (3.3.2),(3.3.28) und der J-Symmetrie von A'

$$\|W_{pq}'\| \leq \|U_{pp}\|_2 \|U_{pq}\|_2 (\|A_{pp}'\| + \|A_{qq}'\|) + 2.00450 \|U_{pq}\|_2^2 \|A_{pq}'\|$$

und hieraus mit den gleichen Abschätzungen wie im Beweis von (i)

$$\|W_{pq}'\| \leq 66.437 \frac{\|A\|^3}{\delta^3} s(A) . \qquad \blacksquare$$

Lemma 3.3.5:
Unter den Voraussetzungen von Lemma 3.3.1 gilt nach einer Jacobi- bzw. Paardekooper-Transformation für $A'' = U^{-1}A'U$

$$s(A'') \leq 114.044 \frac{\|A\|^3}{\delta^3} s(A) . \qquad (3.3.31)$$

Beweis:
Nach (3.3.6) gilt

$$s^2(A'') = s^2(A') + 2\|A_{pq}''\|^2 - 2\|A_{pq}'\|^2 .$$

Weiterhin gilt nach (3.3.30)

$$\|A_{pq}''\| \leq \|A_{pq}'\| + \|W_{pq}'\|$$

und damit

$$\|A_{pq}''\|^2 - \|A_{pq}'\|^2 \leq 2\|A_{pq}'\| \|W_{pq}'\| + \|W_{pq}'\|^2 \leq 6298 \frac{\|A\|^6}{\delta^6} s^2(A) ,$$

wobei die letzte Ungleichung aus (3.2.29),(3.3.30) und $\delta \leq \sqrt{2}\|A\|$ folgt. Mit (3.2.30) und nochmals $\delta \leq \sqrt{2}\|A\|$ erhalten wir dann

$$s^2(A'') \leq 13006 \frac{\|A\|^6}{\delta^6} s^2(A)$$

und somit die Behauptung. ■

In den folgenden zwei Lemmata untersuchen wir, wie sich der Kommutator der Diagonalblöcke und der (p,q)-Restriktion von A' unter einer U-Transformation verändert.

Lemma 3.3.6:
Unter den Voraussetzungen von Lemma 3.3.1 gilt nach einer Jacobi- bzw. Paardekooper-Transformation für $A'' = U^{-1}A'U$

$$\text{(i)} \quad \| C(A_{ii}'') \| \leq \frac{1}{54} \delta^2 \ , \ i = p,q \ , \tag{3.3.32}$$

$$\text{(ii)} \quad \| C(D'') \| \leq \sqrt{2} \| C(D) \| + 36744 \frac{\|A\|^6}{\delta^6} s^2(A) \ . \tag{3.3.33}$$

Beweis:
(i) Zunächst einmal erhalten wir aus den Lemmata 3.2.9 und 3.3.4

$$A_{ii}'' = A_{ii} + W_{ii}'' \ , \quad \|W_{ii}''\| \leq 4593 \frac{\|A\|^5}{\delta^6} s^2(A) \ , \ i = p,q \ ,$$

wobei $\delta \leq \sqrt{2}\|A\|$ in die Abschätzung eingeht. Dann gilt analog zu (3.2.34)

$$\| C(A_{ii}'') \| \leq \| C(A_{ii}) \| + 2\|A_{ii}''\| \, \|W_{ii}''\| + 2\|A_{ii}\| \, \|W_{ii}''\| \ , \ i = p,q \ . \tag{3.3.34}$$

Aus $\|A_{ii}\| \leq \|A\|$ und $\|A_{ii}''\| \leq \|A''\| = \|A'\| \leq \|A\|$ (vgl.(2.2.28)) ergibt sich dann insgesamt

$$\| C(A_{ii}'') \| \leq \| C(A_{ii}) \| + 18372 \frac{\|A\|^6}{\delta^6} s^2(A) \ , \ i = p,q \ , \tag{3.3.35}$$

und hieraus folgt mit $\| C(A_{ii}) \| \leq \| C(D) \|$, den Voraussetzungen (3.2.1) und $\delta \leq \sqrt{2}\|A\|$ die Behauptung (3.3.32).

(ii) Es gilt

$$\| C(D'') \|^2 = \sum_{i=1}^{m} \| C(A_{ii}'') \|^2 = \sum_{i \neq p,q}^{m} \| C(A_{ii}) \|^2 + \| C(A_{pp}'') \|^2 + \|C(A_{qq}'')\|^2 \ ,$$

da die Diagonalblöcke A_{ii} für $i \neq p,q$ unter der T- und U-Transformation invariant bleiben. Daraus ergibt sich mit (3.3.35) und der Cauchy-Schwarzschen Ungleichung

$$\| C(D'') \|^2 \leq 2 \| C(D) \|^2 + 4 \cdot 18372^2 \frac{\|A\|^{12}}{\delta^{12}} s^4(A)$$

und hieraus mit $\sqrt{\xi+\eta} \leq \sqrt{\xi} + \sqrt{\eta}$ für $\xi,\eta \geq 0$ die Behauptung (3.3.33). ■

Lemma 3.3.7:
Unter den Voraussetzungen von Lemma 3.3.1 gilt nach einer Jacobi- bzw. Paardekooper-Transformation für $A'' = U^{-1}A'U$

$$\| \hat{C}_{pq}'' \| \leq 261.272 \frac{\|A\|^6}{\delta^6} s^2(A) + 66.310 \frac{\|A\|^2}{\delta^3} \| C(D) \| s(A). \quad (3.3.36)$$

Beweis:
Aus der Struktur der elementaren Matrix U ergibt sich

$$\hat{A}_{pq}'' = \hat{U}_{pq}^{-1} \hat{A}_{pq}' \hat{U}_{pq} \quad .$$

Aufgrund der Orthogonalität von U ist auch $\hat{U}_{pq}$ wieder orthogonal, und da für jede orthogonale Matrix R die Identität

$$C(R^T A R) = R^T C(A) R$$

gilt, folgt

$$C(\hat{A}_{pq}'') = \hat{U}_{pq}^{-1} C(\hat{A}_{pq}') \hat{U}_{pq} \quad ,$$

d.h. wir erhalten $C(\hat{A}_{pq}'')$ aus $C(\hat{A}_{pq}')$ mit denselben Transformationen, die $\hat{A}_{pq}'$ zu $\hat{A}_{pq}''$ machen. Insbesondere gilt (vgl.(3.3.4))

$$\hat{C}_{pq}'' = U_{pp}\hat{C}_{pp}'U_{pq} - U_{pq}\hat{C}_{qp}'U_{pq} + U_{pp}\hat{C}_{pq}'U_{qq} - U_{pq}\hat{C}_{qq}'U_{qq} \quad .$$

Hieraus folgt wegen (3.3.2), der Symmetrie von $C(\hat{A}_{pq}')$ und Lemma 3.3.3

$$\|\hat{C}_{pq}''\| \le (\|U_{pq}\|_2^2 + \|U_{pp}\|_2^2)\|\hat{C}_{pq}'\| + \|U_{pp}\|_2 \|U_{pq}\|_2 (\|\hat{C}_{pp}'\| + \|\hat{C}_{qq}'\|)$$

$$\le 1.00448 \|\hat{C}_{pq}'\| + 46.888 (\|\hat{C}_{pp}'\| + \|\hat{C}_{qq}'\|) \frac{\|A\|^2}{\delta^3} s(A) .$$

Mit Hilfe der Lemmata 3.2.12 und 3.2.7 erhalten wir dann unter Beachtung von $\| C(A_{pp}) \| + \| C(A_{qq}) \| \le \sqrt{2} \| C(D) \|$

$$\|\hat{C}_{pq}''\| \le 1.00448 \left(150.982 \frac{\|A\|^6}{\delta^6} s^2(A) + 0.513\ s^2(A)\right)$$

$$+ 46.888 \left(\sqrt{2} \| C(D) \| + 795.442 \frac{\|A\|^4}{\delta^4} s^2(A)\right) \frac{\|A\|^2}{\delta^3} s(A)$$

und hieraus mit (3.2.1) und $\delta \le \sqrt{2}\|A\|$ die Behauptung. ■

Es sei hier noch einmal betont, daß wir auch im Fall $C_{pq}' = 0$ (optimale Normreduzierung) mit der Abschätzung (3.2.17) gearbeitet haben, um einheitlich rechnen zu können. Ansonsten ergäben sich bessere Konstanten in (3.3.36).

Wir kommen nun zu der Hauptaussage dieses Paragraphen:

Lemma 3.3.8:
Unter den Voraussetzungen von Lemma 3.3.1 gilt nach einer Jacobi- bzw. Paardekooper-Transformation für $A'' = U^{-1}A'U$

$$\|A_{pq}''\| \le 313.004 \frac{\|A\|^6}{\delta^7} s^2(A) + 79.440 \frac{\|A\|^2}{\delta^4} \| C(D) \| s(A) . \quad (3.3.37)$$

Beweis:
Für den Beweis setzen wir wieder zur Vereinfachung o.B.d.A. $p = 1, q = 2$. Wir untersuchen die Jacobi- und die Paardekooper-Transformation getrennt. Zunächst sei A'' die Matrix nach der Jacobi-Transformation. Es gilt dann $a_{13}'' = a_{24}'' = 0$ und damit $\varepsilon_+'' = \varepsilon_-'' = 0$. Aus (1.3.25) und (2.2.10) folgt somit

$$\hat{c}_{14}'' = 4\beta_1'' = 2a_{14}''(a_{11}'' - a_{44}'') , \quad \hat{c}_{23}'' = 4\beta_2'' = 2a_{23}''(a_{22}'' - a_{33}'') .$$

Hieraus erhalten wir mit (1.3.24)

$$\|A_{12}''\| = \sqrt{a_{14}''^2 + a_{23}''^2} = \frac{1}{2}\left(\frac{1}{(a_{11}''-a_{44}'')^2}\hat{c}_{14}''^2 + \frac{1}{(a_{22}''-a_{33}'')^2}\hat{c}_{23}''^2\right)^{\frac{1}{2}}$$

$$\leq \frac{\|\hat{C}_{12}''\|}{2\min(|a_{11}''-a_{44}''|,|a_{22}''-a_{33}''|)} \quad .$$

Es sei nun o.B.d.A. $|a_{11}''-a_{44}''| = \min(|a_{11}''-a_{44}''|,|a_{22}''-a_{33}''|)$. Aus der Dreiecksungleichung ergibt sich

$$\begin{aligned}|a_{11}'-a_{44}'| - |a_{11}''-a_{44}''| &\leq |a_{11}'-a_{44}'-(a_{11}''-a_{44}'')| \\ &\leq |a_{11}''-a_{11}'| + |a_{44}''-a_{44}'| \\ &\leq \|A_{11}''-A_{11}'\| + \|A_{22}''-A_{22}'\|\end{aligned}$$

und hieraus mit (3.3.29),(3.2.1) und $\delta \leq \sqrt{2}\|A\|$

$$|a_{11}'-a_{44}'| - |a_{11}''-a_{44}''| \leq \frac{1}{79}\delta \quad . \tag{3.3.38}$$

Wie in (3.3.17) folgt dann

$$|a_{11}''-a_{44}''| \geq |a_{11}'-a_{44}'| - \frac{1}{79}\delta \geq \frac{241}{560}\delta - \frac{1}{79}\delta \quad .$$

Insgesamt erhalten wir somit

$$\|A_{12}''\| \leq 1.198\,\frac{\|\hat{C}_{12}''\|}{\delta}$$

und wegen (3.3.36)

$$\|A_{12}''\| \leq 313.004\,\frac{\|A\|^6}{\delta^7}s^2(A) + 79.440\,\frac{\|A\|^2}{\delta^4}\|C(D)\|s(A) \quad . \tag{3.3.39}$$

Es sei nun A'' die Matrix nach der Paardekooper-Transformation. Aus $a_{14}'' = a_{23}'' = 0$ erhalten wir hier $\beta_1'' = \beta_2'' = 0$ und damit aus (1.3.25) und (2.2.10)

$$\hat{c}_{14}'' = 2\varepsilon_+'' + 2\varepsilon_-'' = 2a_{13}'' a_{34}'' - 2a_{24}'' a_{12}'' ,$$

$$\hat{c}_{23}'' = 2\varepsilon_+'' - 2\varepsilon_-'' = 2a_{13}'' a_{12}'' - 2a_{24}'' a_{34}'' .$$

Diese beiden Identitäten stellen ein lineares Gleichungssystem für a_{13}'' und a_{24}'' dar, und es gilt

$$\begin{pmatrix} a_{13}'' \\ a_{24}'' \end{pmatrix} = \frac{1}{2} \begin{pmatrix} a_{34}'' & -a_{12}'' \\ a_{12}'' & -a_{34}'' \end{pmatrix}^{-1} \begin{pmatrix} \hat{c}_{14}'' \\ \hat{c}_{23}'' \end{pmatrix}$$

$$= \frac{1}{2(a_{12}''^2 - a_{34}''^2)} \begin{pmatrix} -a_{34}'' & a_{12}'' \\ -a_{12}'' & a_{34}'' \end{pmatrix} \begin{pmatrix} \hat{c}_{14}'' \\ \hat{c}_{23}'' \end{pmatrix} .$$

Hieraus folgt

$$\|A_{12}''\| = \left\| \begin{pmatrix} a_{13}'' \\ a_{24}'' \end{pmatrix} \right\| \leq \frac{1}{2|a_{12}''^2 - a_{34}''^2|} \left\| \begin{pmatrix} -a_{34}'' & a_{12}'' \\ -a_{12}'' & a_{34}'' \end{pmatrix} \right\|_2 \|\hat{C}_{12}''\| .$$

Man verifiziert leicht

$$\left\| \begin{pmatrix} -a_{34}'' & a_{12}'' \\ -a_{12}'' & a_{34}'' \end{pmatrix} \right\|_2 = |a_{12}''| + |a_{34}''| ,$$

und wir erhalten damit

$$\|A_{12}''\| \leq \frac{1}{2\big||a_{12}''| - |a_{34}''|\big|} \|\hat{C}_{12}''\| .$$

Mehrmaliges Anwenden der Dreiecksungleichung ergibt wieder

$$\begin{aligned}\big|\,|a_{12}'|-|a_{34}'|\,\big| - \big|\,|a_{12}''|-|a_{34}''|\,\big| &\le \big|\,|a_{12}'|-|a_{34}'| - (|a_{12}''|-|a_{34}''|)\,\big| \\ &\le \big|\,|a_{12}''|-|a_{12}'|\,\big| + \big|\,|a_{34}''|-|a_{34}'|\,\big| \\ &\le |a_{12}''-a_{12}'| + |a_{34}''-a_{34}'| \\ &\le \frac{\sqrt{2}}{2}\,\|A_{11}''-A_{11}'\| + \frac{\sqrt{2}}{2}\,\|A_{22}''-A_{22}'\|\,,\end{aligned}$$

und daraus folgt wie in (3.3.38)

$$\big|\,|a_{12}'|-|a_{34}'|\,\big| - \big|\,|a_{12}''|-|a_{34}''|\,\big| \le \frac{1}{112}\,\delta\,.$$

Mit Hilfe von (3.3.20) erhalten wir dann

$$\big|\,|a_{12}''|-|a_{34}''|\,\big| \ge \big|\,|a_{12}'|-|a_{34}'|\,\big| - \frac{1}{112}\,\delta \ge \frac{241}{560}\,\delta - \frac{1}{112}\,\delta = \frac{59}{140}\,\delta\,.$$

Insgesamt folgt hieraus mit (3.3.36)

$$\begin{aligned}\|A_{12}''\| &\le \frac{70}{59}\,\frac{\|\hat{C}_{12}''\|}{\delta} \\ &\le 309.984\,\frac{\|A\|^6}{\delta^7}\,s^2(A) + 78.673\,\frac{\|A\|^2}{\delta^4}\,\|C(D)\|\,s(A)\,.\end{aligned} \tag{3.3.40}$$

Die Abschätzungen (3.3.39) und (3.3.40) ergeben dann zusammen die Behauptung. ■

Bislang hatten wir vorausgesetzt, daß für den zweiten Winkel der Jacobi-Transformation die Einschränkung $y_2 \in (-\frac{\pi}{4},\frac{\pi}{4}]$ gilt. Wir haben jetzt die Hilfsmittel bereitgestellt, um zu zeigen, daß dies unter den Voraussetzungen (3.2.1) immer gilt.

Lemma 3.3.9:
Es gelten die Voraussetzungen von Lemma 3.3.1. Die Parameter y_1, y_2 der elementaren Matrix U seien durch (3.3.7) und (3.3.8) bestimmt.

Dann gilt

$$y_2 \in (-\tfrac{\pi}{4},\tfrac{\pi}{4}] \quad . \tag{3.3.41}$$

Beweis:
Es sei $y_2^{(1)} \in (-\frac{\pi}{4},\frac{\pi}{4}]$. Wir betrachten dann

$$A'' = U(y_1,y_2^{(1)})^{-1} A' U(y_1,y_2^{(1)})$$

und zeigen, daß das Permutationskriterium (3.3.9) nicht erfüllt ist.

Zunächst einmal gilt für die Matrix A''

$$\delta_D(A'') \geq \tfrac{3}{5}\delta \quad . \tag{3.3.42}$$

Denn analog zum Beweis von Lemma 3.2.13 erhalten wir mit (3.3.32) und (3.2.1)

$$\Delta(A_{ii}'') \leq \sqrt[4]{\tfrac{1}{2}} \sqrt{\| C(A_{ii}'') \|} \leq \frac{\sqrt[4]{\frac{1}{2}}}{\sqrt{54}}\delta \ , \quad i = 1(1)m$$

und mit Hilfe der Cauchy-Schwarzschen Ungleichung, (3.3.31) und wieder (3.2.1)

$$\sum_{j\neq i}^{m} \|A_{ij}''\| \leq \frac{\sqrt{m-1}}{\sqrt{2}} S(A'') \leq \tfrac{1}{12}\delta \ , \ i = 1(1)m \ .$$

Zusammen ergibt sich

$$\Delta(A_{ii}'') + \sum_{j\neq i}^{m} \|A_{ij}''\| < \tfrac{1}{5}\delta \ , \quad i = 1(1)m$$

und hiermit die Ungleichung (3.3.42).

Alsdann folgt wie im Beweis von Lemma 3.2.14 mit (3.3.32)

$$\delta \leq \tfrac{5}{3}\left(d^{(p,q)}(A'') + \frac{1}{\sqrt[4]{2}} \left(\sqrt{\| C(A_{pp}'') \|} + \sqrt{\| C(A_{qq}'') \|} \right) \right)$$

$$\leq \tfrac{5}{3}\left(d^{(p,q)}(A'') + \frac{\sqrt[4]{8}}{\sqrt{54}}\delta \right)$$

und hieraus

$$\delta \leq \frac{27}{10}\, d^{(p,q)}(A'') \quad . \tag{3.3.43}$$

Wir nehmen nun an, es sei (3.3.9) erfüllt. O.B.d.A. gelte dabei

$$|a''_{2p-1,2p-1} - a''_{2q,2q}| < 2|a''_{2p-1,2q}| \quad .$$

Dann haben wir einerseits aufgrund der Cauchy-Schwarzschen Ungleichung, (3.3.37),(3.2.1) und $\delta \leq \sqrt{2}\,\|A\|$

$$|a''_{2p-1,2q}| + |a''_{2p,2q-1}| \leq \sqrt{2}\,\|A''_{pq}\| = \sqrt{2}\,\|A''_{pq}\| \leq \frac{1}{2180}\,\delta.$$

Andrerseits erhalten wir aus (3.3.9) und der Dreiecksungleichung

$$\begin{aligned} |a''_{2p-1,2q}| + |a''_{2p,2q-1}| &> \tfrac{1}{3}(|a''_{2p-1,2p}| + |a''_{2q-1,2q}|) + \tfrac{1}{3}|a''_{2p-1,2p-1} - a''_{2q,2q}| \\ &\geq \tfrac{1}{3}\big||a''_{2p-1,2p}| - |a''_{2q-1,2q}|\big| + \tfrac{1}{3}|a''_{2p-1,2p-1} - a''_{2q,2q}| \\ &\geq \tfrac{1}{3}\, d^{(p,q)}(A'') \end{aligned}$$

und damit nach (3.3.43)

$$|a''_{2p-1,2q}| + |a''_{2p,2q-1}| > \frac{10}{81}\delta \quad .$$

Also ist das Kriterium (3.3.9) nicht erfüllt, und es gilt $y_2 = y_2^{(1)}$, d.h. $y_2 \in (-\frac{\pi}{4},\frac{\pi}{4}]$. ■

Wir haben damit gezeigt, daß die Jacobi-Transformation im asymptotischen Bereich keine Vertauschungen von Zeilen und Spalten erfordert.

Zum Abschluß dieses Paragraphen beweisen wir noch ein technisches Lemma über die maximale Änderung der Außerblockdiagonalen von A nach einer T- und U-Transformation.

Lemma 3.3.10:
Unter den Voraussetzungen von Lemma 3.3.1 gilt nach einer Jacobi- bzw. Paardekooper-Transformation für $A'' = U^{-1}A'U$

$$\text{(i)} \quad \|A_{ip}''\|^2 + \|A_{iq}''\|^2 \le 2.001(\|A_{ip}\|^2 + \|A_{iq}\|^2) \,, \; i = 1(1)m \,, \; i \neq p,q \,, \tag{3.3.44}$$

$$\text{(ii)} \quad \|A_{pi}''\| \le 1.000898\,\|A_{pi}\| + 55.497\,\frac{\|A\|^2}{\delta^3}\,s(A)\,\|A_{qi}\| \,, \; i = 1(1)m, \; i \neq p,q. \tag{3.3.45}$$

Beweis:
(i) Zunächst einmal gilt nach (3.3.5)

$$\|A_{ip}''\|^2 + \|A_{iq}''\|^2 = \|A_{ip}'\|^2 + \|A_{iq}'\|^2 \,, \; i = 1(1)m \,, \; i \neq p,q$$

und dann wegen (2.2.4), der Cauchy-Schwarzschen Ungleichung und Lemma 3.2.8

$$\begin{aligned} \|A_{ip}'\|^2 &\le 2\|T_{pp}\|_2^2\,\|A_{ip}\|^2 + 2\|T_{pq}\|_2^2\,\|A_{iq}\|^2 \\ &\le 2.000299\,\|A_{ip}\|^2 + 0.000303\,\|A_{iq}\|^2 \,, \end{aligned}$$

$$\|A_{iq}'\|^2 \le 0.000303\,\|A_{ip}\|^2 + 2.000299\,\|A_{iq}\|^2 \,, \; i = 1(1)m, \; i \neq p,q.$$

Zusammen ergibt sich die Behauptung (3.3.44).

(ii) Aus (2.2.4) folgt wieder für $i \neq p,q$

$$\|A_{pi}'\| \le \|T_{pp}\|_2\,\|A_{pi}\| + \|T_{pq}\|_2\,\|A_{qi}\| \,,$$

$$\|A_{qi}'\| \le \|T_{qp}\|_2\,\|A_{pi}\| + \|T_{qq}\|_2\,\|A_{qi}\| \,.$$

Mit (3.3.4) erhalten wir dann

$$\begin{aligned} \|A_{pi}''\| &\le \|U_{pp}\|_2\,\|A_{pi}'\| + \|U_{pq}\|_2\,\|A_{qi}'\| \\ &\le \|U_{pp}\|_2\,\|T_{pp}\|_2\,\|A_{pi}\| + \|U_{pp}\|_2\,\|T_{pq}\|_2\,\|A_{qi}\| \\ &\quad + \|U_{pq}\|_2\,\|T_{qp}\|_2\,\|A_{pi}\| + \|U_{pq}\|_2\,\|T_{qq}\|_2\,\|A_{qi}\| \end{aligned}$$

für $i = 1(1)m$, $i \neq p,q$ und hieraus mit den Lemmata 3.2.8 und 3.3.3 und $\delta \leq \sqrt{2}\,\|A\|$ die Behauptung (3.3.45). ■

4. KONVERGENZBEWEISE

Wir haben in den vorangegangenen Kapiteln die Eigenschaften elementarer Transformationsmatrizen studiert, mit deren Hilfe nun zahlreiche Jacobi-ähnliche Blockverfahren zur Blockdiagonalisierung reeller J-symmetrischer Matrizen konstruiert werden können. In § 4.1 formulieren wir zwei Verfahren, die bei zeilenzyklischer Pivotstrategie neben gewissen normreduzierenden Schritten wahlweise Transformationen zur Reduktion des außerdiagonalen symmetrischen bzw. schiefsymmetrischen Teils der Matrix durchführen. Wir beweisen unter der Voraussetzung, daß die Eigenwerte der Matrix getrennt sind ($\delta > 0$), die asymptotisch quadratische Konvergenz dieser Verfahren gegen eine Endmatrix in Murnaghan-Form. Man beachte, daß die Voraussetzung $\delta > 0$ auch die Fälle identischer Real- oder Imaginärteile der Eigenwerte beinhaltet.

Wenn man sich im orthogonalen Teil der Verfahren auf die (Block-) Diagonalisierung entweder des symmetrischen oder des schiefsymmetrischen Teils der Matrix beschränkt, so läßt sich auch hier unter stärkeren Voraussetzungen die asymptotisch quadratische Konvergenz gegen Murnaghan-Form beweisen. Im einen Fall benötigt man die Trennung der Realteile, im anderen die Trennung der Imaginärteile der Eigenwerte (vgl.[56]). Die Beweisstrategien verlaufen dann analog zu denen von § 4.1, wobei in die Abschätzungen neben Lokalisierungsaussagen vom Gerschgorinschen Typ (s.[59],[33]) auch die Sätze von Wielandt und Hoffmann für symmetrische bzw. schiefsymmetrische Matrizen (s.[59], [20]) eingehen. Dadurch ergeben sich im Vergleich zu unseren universellen Verfahren wesentlich bessere Konstanten in den Abschätzungen, jedoch brechen die Beweise zusammen, falls gleiche Real- bzw. Imaginärteile der Eigenwerte auftreten.

In § 4.2 machen wir einige Anmerkungen zur globalen Konvergenz unserer Methoden.

4.1 ZUR ASYMPTOTISCH QUADRATISCHEN KONVERGENZ DER JACOBI-ÄHNLICHEN BLOCKVERFAHREN

Wir beschreiben im folgenden zwei Jacobi-ähnliche Blockverfahren zur Berechnung der Eigenwerte und Eigenvektoren beliebiger reeller J-symmetrischer Blockmatrizen, wobei die beiden Verfahren abgesehen von einer unterschiedlichen Normreduzierung im außerdiagonalen Teil der Matrix identisch sind. Sie bestehen aus der wiederholten Anwendung einer Sequenz von Transformationen, die durch den nachfolgenden Algorithmus charakterisiert wird. Wir haben diesen Algorithmus in den vorangegangenen Kapiteln bereits stückweise formuliert und resümieren ihn jetzt noch einmal als Ganzes:

4.1.1 Algorithmus:

Gegeben sei eine reelle J-symmetrische Blockmatrix A der Dimension $n = 2m$.

Schritt 1 (Normreduzierung):

Es sei $p = 1(1)m$.

Für jedes p berechne $c_{2p-1,2p}$, α, a mit Hilfe von (2.1.9),

$$\tanh x = - \frac{c_{2p-1,2p}}{16\alpha + 4a} ,$$

bestimme $\cosh x$, $\sinh x$ und führe die Transformation

$$A \to S^{-1}AS$$

gemäß (2.1.4) aus.

Schritt 2:

Die Pivotindizes (p,q) laufen in der Reihenfolge

$$\begin{matrix} (1,2),(1,3),\ldots,(1,m), \\ (2,3),\ldots,(2,m), \\ \ddots \quad \vdots \\ (m-1,m). \end{matrix}$$

Für ein festes Pivotpaar (p,q) führe die Schritte 2a und 2b aus:

Schritt 2a (Normreduzierung):

Berechne

$$(*) \quad \delta_+ - |\varepsilon_+| \,,\ \delta_- - |\varepsilon_-| \,,\ 4\alpha_1+\alpha_{11}-4|\beta_1|-|\beta_{11}| \,,\ 4\alpha_2+\alpha_{22}-4|\beta_2|-|\beta_{22}|$$

mit Hilfe von (2.2.10).

Setze $x_1 = x_2 = 0$.

(**) Berechne $\tilde{c}_{2p-1,2q}(x_1,x_2)$, $\tilde{c}_{2p,2q-1}(x_1,x_2)$ und $H(x_1,x_2)$ nach (2.2.12).

Falls mindestens drei der Terme (*) gleich null sind (im Sinne der Rechnergenauigkeit), so setze

$$x_1' = x_1 + \operatorname{artanh}\left(-\frac{\tilde{c}_{2p-1,2q}(x_1,x_2)}{16\tilde{\alpha}_1(x_1)+4\tilde{\alpha}_{11}(x_1)+4\tilde{\delta}_+(x_1,x_2)+4\tilde{\delta}_-(x_1,x_2)}\right),$$

$$x_2' = x_2$$

für $|\tilde{c}_{2p-1,2q}(x_1,x_2)| \geq |\tilde{c}_{2p,2q-1}(x_1,x_2)|$ bzw.

$$x_1' = x_1 \,,$$

$$x_2' = x_2 + \operatorname{artanh}\left(-\frac{\tilde{c}_{2p,2q-1}(x_1,x_2)}{16\tilde{\alpha}_2(x_2)+4\tilde{\alpha}_{22}(x_2)+4\tilde{\delta}_+(x_1,x_2)+4\tilde{\delta}_-(x_1,x_2)}\right)$$

für $|\tilde{c}_{2p-1,2q}(x_1,x_2)| < |\tilde{c}_{2p,2q-1}(x_1,x_2)|$.

Sonst berechne

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = H(x_1,x_2)^{-1} \begin{pmatrix} \tilde{c}_{2p-1,2q}(x_1,x_2) \\ \tilde{c}_{2p,2q-1}(x_1,x_2) \end{pmatrix}$$

und setze

$$\begin{pmatrix} x_1' \\ x_2' \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \gamma \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \,,$$

wobei $\gamma = 2^{-j}$, $j \in \mathbb{N}_0$ gemäß (2.2.28) so bestimmt wird, daß

$$F(x_1',x_2') - F(x_1,x_2) \leq -\frac{1}{3}\gamma \begin{pmatrix} \tilde{c}_{2p-1,2q}(x_1,x_2) \\ \tilde{c}_{2p,2q-1}(x_1,x_2) \end{pmatrix}^T H(x_1,x_2)^{-1} \begin{pmatrix} \tilde{c}_{2p-1,2q}(x_1,x_2) \\ \tilde{c}_{2p,2q-1}(x_1,x_2) \end{pmatrix}$$

gilt.

Setze $x_1 = x_1'$, $x_2 = x_2'$.

Führe die ab (**) aufgelistete Folge von Anweisungen nur einmal (Standard-Normreduzierung) bzw. solange aus, wie $|x_i| \leq 1$, $i = 1,2$ gilt und bis $\tilde{c}_{2p-1,2q}(x_1,x_2)$ und $\tilde{c}_{2p,2q-1}(x_1,x_2)$ hinreichend klein sind (Optimale Normreduzierung).

Bestimme $\cosh x_i$, $\sinh x_i$, $i = 1,2$ und führe die Transformation

$$A \to T^{-1}AT$$

gemäß (2.2.4) aus.

Schritt 2b (Diagonalisierung):

Falls

$$\|A_{pq}^{+}\| + \min\{|a_{2p-1,2p-1} - a_{2q-1,2q-1}|, |a_{2p-1,2p-1} - a_{2q,2q}|,$$
$$|a_{2p,2p} - a_{2q-1,2q-1}|, |a_{2p,2p} - a_{2q,2q}|\}$$
$$> \|A_{pq}^{-}\| + \big||a_{2p-1,2p}| - |a_{2q-1,2q}|\big|$$

gilt, so berechne

$$\tan 2y_1 = \frac{2a_{2p-1,2q-1}}{a_{2p-1,2p-1} - a_{2q-1,2q-1}} \quad , \quad y_1 \in (-\tfrac{\pi}{4},\tfrac{\pi}{4}] \quad ,$$

$$\tan 2y_2 = \frac{2a_{2p,2q}}{a_{2p,2p} - a_{2q,2q}} \quad , \quad y_2 \in (-\tfrac{\pi}{2},\tfrac{\pi}{2}]$$

unter Beachtung von Kriterium (3.3.9). Sonst berechne

$$\tan 2y_1 = -2\,\frac{a_{2p-1,2p}\,a_{2p,2q-1} - a_{2q-1,2q}\,a_{2p-1,2q}}{a_{2p-1,2p}^2 - a_{2q-1,2q}^2 + a_{2p-1,2q}^2 - a_{2p,2q-1}^2}\,, \quad y_1 \in (-\tfrac{\pi}{4},\tfrac{\pi}{4}]\,,$$

$$\tan y_2 = \begin{cases} \dfrac{a_{2p-1,2q}\cos y_1 + a_{2q-1,2q}\sin y_1}{a_{2p-1,2p}\cos y_1 - a_{2p,2q-1}\sin y_1}\,, \\[2ex] \dfrac{a_{2p,2q-1}\cos y_1 + a_{2p-1,2p}\sin y_1}{a_{2q-1,2q}\cos y_1 - a_{2p-1,2q}\sin y_1}\,, \end{cases} \qquad y_2 \in (-\tfrac{\pi}{2},\tfrac{\pi}{2}]$$

(wenn $|a_{2p-1,2p}\cos y_1 - a_{2p,2q-1}\sin y_1| > |a_{2q-1,2q}\cos y_1 - a_{2p-1,2q}\sin y_1|$, wähle die erste, sonst die zweite Formel).

Bestimme $\cos y_i$, $\sin y_i$, $i = 1,2$ und führe die Transformation

$$A \to U^{-1}AU$$

gemäß (3.3.4) aus.

Die Wahl der Normreduzierung (Standard- oder Optimale) im Schritt 2a ist so zu verstehen, daß man sich definitiv für eine entscheidet und diese einheitlich im Algorithmus anwendet. Insofern beschreibt 4.1.1 zwei verschiedene Verfahren.

Wir haben in der Formulierung des Algorithmus die Bestimmung der Eigenvektoren von A nicht berücksichtigt. Sollen sie auch berechnet werden, so ist eine zusätzliche Akkumulierung der Transformationsmatrizen erforderlich. Für die numerische Praxis bedarf der Algorithmus natürlich noch weiterer Spezifizierungen (Abbruchkriterien, Schwellen, stabile Berechnung der Transformationsparameter, Ausnutzung der J-Symmetrie in den Transformationen, etc.). Auf diese Probleme werden wir in § 5.1 eingehen.

Für den Beweis der asymptotisch quadratischen Konvergenz der beiden im Algorithmus beschriebenen Verfahren führen wir nun noch einige Bezeichnungen ein:

Die in 4.1.1 beschriebene Sequenz von Transformationen nennen wir einen *Zyklus*. $A^{(N)}$, $N \in \mathbb{N}_0$ bezeichne die Matrix A nach der Durchführung

des N-ten Zyklus und $A_k^{(N)} = (A_{ij}^{(k,N)})_{i,j=1(1)m}$, $k,N \in \mathbb{N}_0$ die Matrix nach der k-ten Transformation des (N+1)-ten Zyklus. Wir setzen für den ersten Teil eines Zyklus (Schritt 1 aus 4.1.1)

$$A^{(0)} = A \,, \qquad A_0^{(N)} = A^{(N)} \,,$$

$$A_k^{(N)} = S_k^{-1} A_{k-1}^{(N)} S_k \quad , \; k = 1(1)m$$

und für den zweiten Teil eines Zyklus (Schritt 2 aus 4.1.1)

$$A_k^{(N)} = U_k^{-1} T_k^{-1} A_{k-1}^{(N)} T_k U_k \quad , \; k = m+1(1)m+M \,,$$

$$A^{(N+1)} = A_{m+M}^{(N)}$$

mit $M = \frac{m(m-1)}{2}$.

Wenn wir nur Transformationen eines festen Zyklus betrachten, so wird aus Gründen der Übersichtlichkeit der obere Index N weggelassen. Bei Transformationen an einem festen Pivotblock lassen wir auch den Index k weg und schreiben wie in den vorbereitenden Lemmata aus den vorangegangenen Kapiteln A' und A'' für die Matrix nach der ersten bzw. zweiten Transformation.

Wir zeigen die asymptotisch quadratische Konvergenz für beide Verfahren aus 4.1.1 in einem gemeinsamen Beweis. Nach Lemma 1.3.5 wissen wir, daß das Paar $(S(A), \| C(D) \|)$ für J-symmetrische Matrizen A in gewissem Sinne ein Maß für die Abweichung von der Murnaghan-Form darstellt. Da wir nun die asymptotisch quadratische Konvergenz der Matrizenfolge $A^{(N)}$ gegen Murnaghan-Form nachweisen wollen, genügt es, die asymptotisch quadratische Konvergenz von $S(A^{(N)})$ und $\| C(D^{(N)}) \|$ gegen O zu beweisen. Dieser Beweis ist recht aufwendig und kompliziert. Daher geben wir hier zunächst eine Beweisskizze mit Landau-Symbolen:

Gegeben sei eine reelle J-symmetrische Blockmatrix A mit getrennten Eigenwerten ($\delta > 0$). Es gelte für ein $\varepsilon_N > 0$, welches hinreichend klein ist,

$$S(A^{(N)}) = O(\varepsilon_N) \,, \quad \| C(D^{(N)}) \| = O(\varepsilon_N) \,.$$

Wir haben dann zu zeigen:

$$S(A^{(N+1)}) = O(\varepsilon_N^2) \quad , \quad \| C(D^{(N+1)}) \| = O(\varepsilon_N^2) \; .$$

Zunächst erhalten wir durch die Normreduzierung auf den Diagonalblökken $\| C(A_{pp}') \| = O(\varepsilon_N^2)$ für jedes $p = 1(1)m$ (Lemma 3.1.2). Da die Diagonalblöcke, die nicht Pivotblock sind, sich bei der Transformation nicht ändern, können wir

$$\| C(D_m^{(N)}) \| = O(\varepsilon_N^2)$$

zeigen (Satz 4.1.3). Die Größenordnung des außerblockdiagonalen Teils von A bleibt nach dem ersten Teil des Zyklus erhalten:

$$S(A_m^{(N)}) = O(\varepsilon_N)$$

(Lemma 3.1.3, Satz 4.1.3). Im zweiten Teil des Zyklus gilt für ein festes Pivotpaar (p,q)

$$\hat{T}_{pq} = \begin{pmatrix} I + O(\varepsilon_N^2) & O(\varepsilon_N) \\ O(\varepsilon_N) & I + O(\varepsilon_N^2) \end{pmatrix} \quad , \quad \hat{U}_{pq} = \begin{pmatrix} I + O(\varepsilon_N^2) & O(\varepsilon_N) \\ O(\varepsilon_N) & I + O(\varepsilon_N^2) \end{pmatrix}$$

(Lemmata 3.2.8 und 3.3.3). Hieraus folgt zunächst $A_{ii}' = A_{ii} + O(\varepsilon_N^2)$, $A_{ii}'' = A_{ii}' + O(\varepsilon_N^2)$, $i = p,q$ (Lemmata 3.2.9 und 3.3.4) und damit

$$\| C(D'') \| = O(\varepsilon_N^2)$$

(Lemma 3.3.6). Es wird also die im ersten Teil des Zyklus für $\| C(D) \|$ erreichte Ordnung nicht wieder zerstört. Weiterhin folgt $S(A') = O(\varepsilon_N)$ (Lemma 3.2.10) und daraus

$$S(A'') = O(\varepsilon_N)$$

(Lemma 3.3.5). Nach Lemma 2.2.2 ist wegen $\delta > 0$ die Funktion F gleichmäßig konvex, so daß gewährleistet ist, daß die Normreduzierung in Schritt 2a gemäß der gedämpften Newton-Iteration (2.2.27), (2.2.28) erfolgt. Da nach der T-Transformation somit $\| C_{pq}' \| = O(\varepsilon_N^2)$

(Lemma 3.2.7) und damit $\|\hat{C}_{pq}'\| = O(\varepsilon_N^2)$ (Lemma 3.2.12) gilt, folgt mit Hilfe von $\|\hat{C}_{pp}'\| + \|\hat{C}_{qq}'\| = O(\varepsilon_N^2)$ (Lemma 3.2.12)

$$\|\hat{C}_{pq}''\| = O(\varepsilon_N^2)$$

(Lemma 3.3.7). Hieraus erhält man

$$\|A_{pq}''\| = O(\varepsilon_N^2)$$

(Lemma 3.3.8). Wir zeigen, daß für den zweiten Teil des Zyklus

$$S(A_k^{(N)}) = O(\varepsilon_N) \quad , \quad \|C(D_k^{(N)})\| = O(\varepsilon_N^2) \quad , \; k = m(1)m+M$$

(Satz 4.1.5) gilt, und beweisen dann, einer Idee von Ruhe ([41]) folgend,

$$S(A_{m+M}^{(N)}) = O(\varepsilon_N^2)$$

(Satz 4.1.6). ■

Wesentliche Teile dieses Beweises sind mit den Lemmata aus Kapitel 3 bereits abgedeckt. Wenn wir nun die neu eingeführte Notation auf die Voraussetzungen und Aussagen dieser Lemmata übertragen, so ist unter $\|A\|$ jeweils $\|A_k^{(N)}\|$ zu verstehen. Jedoch behalten die Aussagen wegen

$$\|A_k^{(N)}\| \leq \|A_0^{(0)}\| \quad , k,N \in \mathbb{N}_0$$

ihre Gültigkeit, wenn wir mit $\|A\|$ die Norm der Startmatrix $A = A_0^{(0)}$ bezeichnen.

Wir haben diese Lemmata unter den Generalvoraussetzungen (3.1.1) und (3.2.1) bewiesen. Daher bedarf es nun noch einer Untersuchung, wie klein $S(A^{(N)})$ und $\|C(D^{(N)})\|$ sein müssen, damit die Bedingungen (3.1.1),(3.2.1) für alle Iterierten $A_k^{(N)}$ während eines Zyklus gelten. Zunächst betrachten wir den ersten Teil eines Zyklus:

Lemma 4.1.2:
Es sei $A^{(N)}$, $N \in \mathbb{N}_0$ eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$ zu Beginn des (N+1)-ten Zyklus. Es gelte $\delta > 0$ und

$$S(A^{(N)}) \leq \|A\| \tilde{\varepsilon} , \quad \| C(D^{(N)}) \| \leq \|A\|^2 \tilde{\varepsilon} \tag{4.1.1}$$

mit

$$\tilde{\varepsilon} \leq \frac{1}{1.0017^m \sqrt{M}} \left(\frac{1}{114.044} \frac{\delta^3}{\|A\|^3} \right)^M \varepsilon , \quad \varepsilon = \frac{1}{1000 \sqrt{m-1}} \frac{\delta^4}{\|A\|^4} . \tag{4.1.2}$$

Dann folgt für $k = 0(1)m-1$

$$S(A_k^{(N)}) \leq \frac{\sqrt{2}}{500 \sqrt{m-1}} \delta ,$$

$$\| C(A_{ii}^{(k,N)}) \| \leq \frac{1}{500} \delta^2 , \quad i = 1(1)m . \tag{4.1.3}$$

Beweis:
Der Beweis wird per vollständiger Induktion über k geführt. Wegen $m \geq 2$, $M \geq 1$ und $\delta \leq \sqrt{2}\|A\|$ gilt dabei

$$\frac{1}{1.0017^m \sqrt{M}} < 1 , \quad \left(\frac{1}{114.044} \frac{\delta^3}{\|A\|^3} \right)^M < 1 . \tag{4.1.4}$$

Wir erhalten zunächst für $k = 0$ mit (4.1.1), (4.1.4) und $\delta \leq \sqrt{2}\|A\|$

$$S(A_0^{(N)}) \leq \|A\| \tilde{\varepsilon} \leq \|A\| \varepsilon \leq \frac{\sqrt{2}}{500 \sqrt{m-1}} \delta ,$$

$$\| C(A_{ii}^{(0,N)}) \| \leq \| C(D_0^{(N)}) \| \leq \|A\|^2 \tilde{\varepsilon} \leq \|A\|^2 \varepsilon \leq \frac{1}{500} \delta^2 , \; i = 1(1)m.$$

Es sei dann (4.1.3) für die Matrizen $A_0^{(N)}, \ldots, A_k^{(N)}$, $k < m-1$ erfüllt. Mit Lemma 3.1.3, sukzessive angewandt auf $A_k^{(N)}, \ldots, A_0^{(N)}$, folgt

$$\begin{aligned} S(A_{k+1}^{(N)}) &\leq 1.0017 \; S(A_k^{(N)}) \leq 1.0017^{k+1} \; S(A_0^{(N)}) \\ &\leq 1.0017^m \; S(A_0^{(N)}) \leq 1.0017^m \; \|A\| \tilde{\varepsilon} \end{aligned} \tag{4.1.5}$$

und damit nach (4.1.4) und $\delta \le \sqrt{2}\,\|A\|$

$$S(A_{k+1}^{(N)}) \le \frac{\sqrt{2}}{500\sqrt{m-1}}\,\delta \ .$$

Weiter folgt zunächst für $i = k+2(1)m$

$$\|\,C(A_{ii}^{(k+1,N)})\,\| \;=\; \|\,C(A_{ii}^{(0,N)})\,\| \;\le\; \|\,C(D_0^{(N)})\,\|$$

und dann für $i = 1(1)k+1$ mit Hilfe der Lemmata 3.1.2 und 3.1.3, wieder angewandt auf die Matrizen $A_k^{(N)},\ldots,A_0^{(N)}$

$$\begin{aligned}\|\,C(A_{ii}^{(k+1,N)})\,\| &= \|\,C(A_{ii}^{(i,N)})\,\| \\ &\le 2.533\,\frac{\|A\|^2}{\delta^4}\,\|\,C(A_{ii}^{(i-1,N)})\,\|^2 + 3.373\,\frac{\|A\|^2}{\delta^2}\,S^2(A_{i-1}^{(N)}) \\ &\le 2.533\,\frac{\|A\|^2}{\delta^4}\,\|\,C(D_0^{(N)})\,\|^2 + 3.373\cdot 1.0017^{2m}\,\frac{\|A\|^2}{\delta^2}\,S^2(A_0^{(N)}) \\ &\le 2.533\,\frac{\|A\|^6}{\delta^4}\,\tilde{\varepsilon}^{\,2} + 3.373\cdot 1.0017^{2m}\,\frac{\|A\|^4}{\delta^2}\,\tilde{\varepsilon}^{\,2} \ . \qquad (4.1.6)\end{aligned}$$

In beiden Fällen ergibt sich dann wiederum mit (4.1.4) und $\delta \le \sqrt{2}\,\|A\|$

$$\|\,C(A_{ii}^{(k+1,N)})\,\| \;\le\; \frac{1}{500}\,\delta^2 \ . \qquad \blacksquare$$

Wir können nun abschätzen, wie groß $S(A^{(N)})$ und $\|C(D^{(N)})\|$ nach dem kompletten ersten Teil des Zyklus sind.

Satz 4.1.3:
Unter den Voraussetzungen von Lemma 4.1.2 folgt

$$\begin{aligned}S(A_m^{(N)}) &\le 1.0017^m\,\|A\|\,\tilde{\varepsilon}\ , \\ \|\,C(D_m^{(N)})\,\| &\le 9.279\,m\;1.0017^{2m}\,\frac{\|A\|^6}{\delta^4}\,\tilde{\varepsilon}^{\,2}\ .\end{aligned} \qquad (4.1.7)$$

Beweis:
Wie in (4.1.5) ergibt sich zunächst wegen (4.1.3)

$$S(A_m^{(N)}) \le 1.0017^m \|A\| \tilde{\varepsilon}.$$

Alsdann folgt aus

$$\| C(D_m^{(N)}) \|^2 = \sum_{i=1}^{m} \| C(A_{ii}^{(m,N)}) \|^2 = \sum_{i=1}^{m} \| C(A_{ii}^{(i,N)}) \|^2$$

wie in (4.1.6) unter Beachtung von (4.1.3)

$$\| C(D_m^{(N)}) \| \le m \left(2.533 \frac{\|A\|^6}{\delta^4} \tilde{\varepsilon}^2 + 3.373 \cdot 1.0017^{2m} \frac{\|A\|^4}{\delta^2} \tilde{\varepsilon}^2 \right)$$

und damit wegen $1.0017^{2m} > 1$ und $\delta \le \sqrt{2}\|A\|$

$$\| C(D_m^{(N)}) \| \le 9.279\, m\, 1.0017^{2m} \frac{\|A\|^6}{\delta^4} \tilde{\varepsilon}^2 . \quad \blacksquare$$

Wir betrachten dann den zweiten Teil eines Zyklus:

Lemma 4.1.4:
Unter den Voraussetzungen von Lemma 4.1.2 gilt für $k = m(1)m+M-1$

$$S(A_k^{(N)}) \le \|A\| \varepsilon ,$$
$$\| C(D_k^{(N)}) \| \le \frac{1}{10000} \delta^2 . \qquad (4.1.8)$$

Beweis:
Wir beweisen die Behauptung wieder induktiv über k. Für $k = m$ gilt zunächst wegen (4.1.7) und (4.1.4)

$$S(A_m^{(N)}) \le 1.0017^m \|A\| \tilde{\varepsilon} \le \|A\| \varepsilon ,$$

$$\| C(D_m^{(N)}) \| \le 9.279\, m\, 1.0017^{2m} \frac{\|A\|^6}{\delta^4} \tilde{\varepsilon}^2 \le \frac{1}{10000} \delta^2 ,$$

wobei in die letzte Ungleichung noch $\frac{m}{m-1} \le 2$ und $\delta \le \sqrt{2}\|A\|$ eingeht.

Es gelte dann (4.1.8) für die Matrizen $A_m^{(N)},\dots,A_k^{(N)}$, $k < m+M-1$. Wenn wir Lemma 3.3.5 nacheinander auf $A_k^{(N)},\dots,A_m^{(N)}$ anwenden, so erhalten wir mit (4.1.7)

$$S(A_{k+1}^{(N)}) \le 114.044\,\frac{\|A\|^3}{\delta^3}\,S(A_k^{(N)}) \le \left(114.044\,\frac{\|A\|^3}{\delta^3}\right)^{k-m+1} S(A_m^{(N)})$$
$$\le \left(114.044\,\frac{\|A\|^3}{\delta^3}\right)^{M} 1.0017^m\,\|A\|\tilde{\varepsilon} \le \|A\|\varepsilon \quad . \tag{4.1.9}$$

Analog erhalten wir durch sukzessives Anwenden von Lemma 3.3.6 auf $A_k^{(N)},\dots,A_m^{(N)}$

$$\|C(D_{k+1}^{(N)})\| \le \sqrt{2}\,\|C(D_k^{(N)})\| + 36744\,\frac{\|A\|^6}{\delta^6}\,S^2(A_k^{(N)})$$
$$\le \sqrt{2}^{k-m+1}\,\|C(D_m^{(N)})\| + 36744\,\frac{\|A\|^6}{\delta^6}\sum_{l=m}^{k}\sqrt{2}^{k-l}\,S^2(A_l^{(N)}).$$

Mit Lemma 3.3.5 ergibt sich dann wie in (4.1.9)

$$\sum_{l=m}^{k}\sqrt{2}^{k-l}\,S^2(A_l^{(N)}) \le \sum_{l=m}^{k}\sqrt{2}^{k-l}\left(114.044\,\frac{\|A\|^3}{\delta^3}\right)^{2(l-m)} S^2(A_m^{(N)})$$
$$\le (k-m+1)\left(114.044\,\frac{\|A\|^3}{\delta^3}\right)^{2(k-m)} S^2(A_m^{(N)}) \quad .$$

Insgesamt erhalten wir mit (4.1.7)

$$\|C(D_{k+1}^{(N)})\| \le \sqrt{2}^{k-m+1}\,\|C(D_m^{(N)})\| + 2.826(k-m+1)\left(114.044\,\frac{\|A\|^3}{\delta^3}\right)^{2(k-m+1)} S^2(A_m^{(N)})$$
$$\le \sqrt{2}^{M}\|C(D_m^{(N)})\| + 2.826\,M\left(114.044\,\frac{\|A\|^3}{\delta^3}\right)^{2M} S^2(A_m^{(N)})$$
$$\le 9.279\,m\,1.0017^{2m}\sqrt{2}^{M}\,\frac{\|A\|^6}{\delta^4}\,\tilde{\varepsilon}^2 + 2.826\,M\,1.0017^{2m}\left(114.044\,\frac{\|A\|^3}{\delta^3}\right)^{2M}\|A\|^2\tilde{\varepsilon}^2.$$

Aus $m \le 2M$ und $\sqrt[4]{2} < 114.044\,\frac{\|A\|^3}{\delta^3}$ folgt schließlich mit $\delta \le \sqrt{2}\|A\|$

$$\| C(D_{k+1}^{(N)}) \| \leq 29.862\ M\ 1.0017^{2m} \left(114.044\ \frac{\|A\|^3}{\delta^3}\right)^{2M} \frac{\|A\|^6}{\delta^6}\ \tilde{\varepsilon}^2$$

$$\leq 29.862\ \frac{\|A\|^6}{\delta^4}\ \varepsilon^2 \leq \frac{1}{10000}\ \delta^2 . \qquad (4.1.10)$$

■

Als Konsequenz aus diesem Lemma ergibt sich der folgende

Satz 4.1.5:
Unter den Voraussetzungen von Lemma 4.1.2 folgt für $k = m(1)m+M$

$$S(A_k^{(N)}) \leq 1.0017^m \left(114.044\ \frac{\|A\|^3}{\delta^3}\right)^M \|A\|\ \tilde{\varepsilon}\ ,$$

$$\| C(D_k^{(N)}) \| \leq 29.862\ M\ 1.0017^{2m} \left(114.044\ \frac{\|A\|^3}{\delta^3}\right)^{2M} \frac{\|A\|^6}{\delta^4}\ \tilde{\varepsilon}^2\ . \qquad (4.1.11)$$

Beweis:
Die Aussagen folgen wie in (4.1.9) und (4.1.10) direkt aus Lemma 4.1.4. ■

Nach diesen Vorbereitungen beweisen wir nun die wichtige Aussage, daß unter den Voraussetzungen (4.1.1),(4.1.2) $S(A)$ nach einem vollständig durchgeführten Zyklus quadratisch klein wird.

Satz 4.1.6:
Unter den Voraussetzungen von Lemma 4.1.2 folgt

$$S(A_{m+M}^{(N)}) \leq 648.995\ M\ 1.00431^m \left(18398\ \frac{\|A\|^6}{\delta^6}\right)^M \frac{\|A\|^8}{\delta^7}\ \tilde{\varepsilon}^2\ . \qquad (4.1.12)$$

Beweis:
Der Beweis wird in Anlehnung an eine Beweisstrategie von Ruhe ([41]) geführt. Die Grundidee ist dabei die folgende:

Die Pivotblöcke des rechtsoberen Dreiecks von A werden zeilenzyklisch durchlaufen und jeweils quadratisch klein gemacht (vgl. Lemma 3.3.8). Hierbei bleiben die behandelten Blöcke quadratisch klein, wenn an den nächsten Pivotblöcken derselben Zeile transformiert wird (vgl.(3.3.45)). Wenn eine Blockzeile dann insgesamt quadratisch klein

ist, wird sie durch Transformationen in den folgenden Zeilen nicht mehr wesentlich geändert (vgl.(3.3.44)). Auf diese Weise wird sukzessive das gesamte rechtsobere Dreieck von A quadratisch klein gemacht.

Der Index k der Matrix $A_k = A_k^{(N)}$ sei nun für ein festes p mit $0 \le p \le m-1$ durch die Position (p,m) bestimmt, d.h. man betrachtet die Matrix, nachdem die Pivotindizes die komplette p-te Blockzeile durchlaufen haben. Die nächsten Pivotindizes sind (p+1,p+2).

Wir beweisen die Behauptung des Satzes, indem wir zunächst zeigen, daß für $p = 0(1)m-1$ Konstanten K_p existieren, so daß

$$\sum_{i=1}^{p} \sum_{j=i+1}^{m} \|A_{ij}^{(k)}\|^2 \le K_p \|A\|^2 \tilde{\varepsilon}^4 \qquad (4.1.13)$$

gilt, und berechnen dann K_{m-1} rekursiv aus $K_{m-2},\ldots,K_0$. Der Beweis von (4.1.13) wird per vollständiger Induktion über p geführt.

Induktionsanfang:
Für $p = 0$ ist die Summe in (4.1.13) leer und damit gleich 0. Also erfüllt $K_0 = 0$ die Ungleichung.

Induktionsvoraussetzung:
Es existiere eine Konstante K_{p-1}, die (4.1.13) am Ende der (p-1)-ten Zeile erfüllt, d.h. für die

$$\sum_{i=1}^{p-1} \sum_{j=i+1}^{m} \|A_{ij}^{(k-m+p)}\|^2 \le K_{p-1} \|A\|^2 \tilde{\varepsilon}^4$$

gilt.

Induktionsschritt:
Das Pivotpaar (p,q) durchlaufe nun die Folge (p,p+1),(p,p+2),...,(p,m). Zunächst schätzen wir den maximalen Normzuwachs der Zeilen 1 bis p-1 unter diesen Transformationen ab. Die Blöcke A_{ij} dieser Zeilen bleiben entweder unberührt, oder der Zuwachs kann nach (3.3.44) abgeschätzt werden. So gilt für das letzte Pivotpaar (p,m)

$$\|A_{ip}^{(k)}\|^2 + \|A_{im}^{(k)}\|^2 \le 2.001\,(\|A_{ip}^{(k-1)}\|^2 + \|A_{im}^{(k-1)}\|^2) \quad , \; i = 1(1)p-1$$

und damit

$$\sum_{i=1}^{p-1} \sum_{j=i+1}^{m} \|A_{ij}^{(k)}\|^2 \leq 2.001 \sum_{i=1}^{p-1} \sum_{j=i+1}^{m} \|A_{ij}^{(k-1)}\|^2 .$$

Es folgt sukzessive

$$\sum_{i=1}^{p-1} \sum_{j=i+1}^{m} \|A_{ij}^{(k)}\|^2 \leq 2.001^{m-p} \sum_{i=1}^{p-1} \sum_{j=i+1}^{m} \|A_{ij}^{(k-m+p)}\|^2$$

und damit nach Induktionsvoraussetzung

$$\sum_{i=1}^{p-1} \sum_{j=i+1}^{m} \|A_{ij}^{(k)}\|^2 \leq 2.001^{m-p} K_{p-1} \|A\|^2 \tilde{\varepsilon}^4 . \qquad (4.1.14)$$

Es bleibt die p-te Zeile abzuschätzen. Dazu verfolgen wir zunächst, wie ein als Pivot benutzter Block $A_{pi}^{(k+i-m)}$, $i = p+1(1)m$ sich verändert, wenn anschließend die weiteren Blöcke der p-ten Zeile als Pivot verwendet werden. Nach (3.3.45) gilt

$$\|A_{pi}^{(k)}\| \leq 1.000898^{m-i} \|A_{pi}^{(k-m+i)}\| + \sum_{j=i+1}^{m} 1.000898^{m-j} \, 55.497 \, \frac{\|A\|^2}{\delta^3} \, s(A_{k-m+j-1}) \, \|A_{ij}^{(k-m+j-1)}\|$$

für $i = p+1(1)m$. Hieraus erhalten wir mit Hilfe der Cauchy-Schwarzschen Ungleichung

$$\|A_{pi}^{(k)}\|^2 \leq 1.000898^{2m} \left\{ 2 \|A_{pi}^{(k-m+i)}\|^2 + 6160 \, \frac{\|A\|^4}{\delta^6} \left(\sum_{j=i+1}^{m} s(A_{k-m+j-1}) \, \|A_{ij}^{(k-m+j-1)}\| \right)^2 \right\} .$$

Nach Lemma 3.3.8 und Satz 4.1.5 ergibt sich mit (4.1.2), $\delta \leq \sqrt{2}\,\|A\|$ und $M \geq 1$

$$\|A_{pi}^{(k-m+i)}\| \leq 313.004 \, \frac{\|A\|^6}{\delta^7} \, s^2(A_{k-m+i-1}) + 79.440 \, \frac{\|A\|^2}{\delta^4} \, \|C(D_{k-m+i-1})\| \, s(A_{k-m+i-1})$$

$$\leq 313.004 \cdot 1.0017^{2m} \left(114.044 \frac{\|A\|^3}{\delta^3}\right)^{2M} \frac{\|A\|^8}{\delta^7} \tilde{\varepsilon}^2$$

$$+ 2.373 \sqrt{M}\, 1.0017^{2m} \left(114.044 \frac{\|A\|^3}{\delta^3}\right)^{2M} \frac{\|A\|^5}{\delta^4} \tilde{\varepsilon}^2$$

$$\leq 319.716 \sqrt{M}\, 1.0017^{2m} \left(114.044 \frac{\|A\|^3}{\delta^3}\right)^{2M} \frac{\|A\|^8}{\delta^7} \tilde{\varepsilon}^2 \quad , \; i = p+1(1)m \; .$$

Weiterhin gilt aufgrund der Cauchy-Schwarzschen Ungleichung und Satz 4.1.5

$$\left(\sum_{j=i+1}^{m} S(A_{k-m+j-1}) \, \|A_{ij}^{(k-m+j-1)}\|\right)^2 \leq \max^2\{S(A_k) \mid m \leq k \leq m+M\} (m-i) \sum_{j=i+1}^{m} \|A_{ij}^{(k-m+j-1)}\|^2$$

$$\leq \frac{1}{2} \max^2 \{S(A_k) \mid m \leq k \leq m+M\} (m-i) \; S^2(A_{k-m+j-1})$$

$$\leq \frac{1}{2} (m-i) \, 1.0017^{4m} \left(114.044 \frac{\|A\|^3}{\delta^3}\right)^{4M} \|A\|^4 \tilde{\varepsilon}^4$$

für $i = p+1(1)m$. Insgesamt haben wir damit

$$\|A_{pi}^{(k)}\|^2 \leq 1.000898^{2m} \left\{ 204437 \; M \; 1.0017^{4m} \left(114.044 \frac{\|A\|^3}{\delta^3}\right)^{4M} \frac{\|A\|^{16}}{\delta^{14}} \tilde{\varepsilon}^4 \right.$$

$$\left. + 3080 (m-i) 1.0017^{4m} \left(114.044 \frac{\|A\|^3}{\delta^3}\right)^{4M} \frac{\|A\|^8}{\delta^6} \tilde{\varepsilon}^4 \right\}$$

$$\leq (204437M + 12320(m-i)) \, 1.00215^{4m} \left(114.044 \frac{\|A\|^3}{\delta^3}\right)^{4M} \frac{\|A\|^{16}}{\delta^{14}} \tilde{\varepsilon}^4 \quad , i = p+1(1)m,$$

wobei die letzte Abschätzung aus $\delta \leq \sqrt{2}\, \|A\|$ folgt. Summation über die p-te Blockzeile liefert dann wegen $\sum_{i=p+1}^{m} (m-i) = \frac{1}{2}(m-p-1)(m-p)$

$$\sum_{i=p+1}^{m} \|A_{pi}^{(k)}\|^2 \leq Q_p \, \|A\|^2 \, \tilde{\varepsilon}^4 \quad ,$$

$$Q_p = (204437\, M(m-p) + 6160 (m-p-1)(m-p)) \, 1.00215^{4m} \left(114.044 \frac{\|A\|^3}{\delta^3}\right)^{4M} \frac{\|A\|^{14}}{\delta^{14}} \quad ,$$

und mit (4.1.14) folgt

$$\sum_{i=1}^{p} \sum_{j=i+1}^{m} \|A_{ij}^{(k)}\|^2 \leq K_p \|A\|^2 \tilde{\varepsilon}^4 \quad , \quad K_p = 2.001^{m-p} K_{p-1} + Q_p \quad ,$$

womit der Induktionsbeweis abgeschlossen ist.

Es gilt nun

$$s^2(A_{m+M}^{(N)}) = 2 \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \|A_{ij}^{(m+M)}\|^2 \leq 2 K_{m-1} \|A\|^2 \tilde{\varepsilon}^4 \quad .$$

Durch rekursives Einsetzen erhalten wir mit $K_0 = 0$

$$K_{m-1} = 2.001 K_{m-2} + Q_{m-1} = \sum_{p=1}^{m-1} 2.001^{\frac{(m-p-1)(m-p)}{2}} Q_p$$

und damit wegen $(m-p-1)(m-p) < (m-1)m = 2M$

$$K_{m-1} \leq 2.001^M \sum_{p=1}^{m-1} Q_p \quad .$$

Aus

$$\sum_{p=1}^{m-1} (m-p) = \frac{(m-1)m}{2} = M$$

und

$$\sum_{p=1}^{m-1} (m-p-1)(m-p) \leq \sum_{p=1}^{m-1} (m-p)^2 \leq \left(\sum_{p=1}^{m-1} (m-p) \right)^2 = M^2$$

folgt dann

$$K_{m-1} \leq 210597\, M^2\, 1.00215^{4m} \left(135.639 \frac{\|A\|^3}{\delta^3}\right)^{4M} \frac{\|A\|^{14}}{\delta^{14}}$$

und somit die Behauptung des Satzes. ■

Es folgt nun der eigentliche Hauptsatz, der alle bisherigen Ergebnisse zusammenfaßt:

Satz 4.1.7:
Es sei A eine reelle J-symmetrische Blockmatrix der Dimension $n = 2m$. Es gelte $\delta > 0$ und

$$s(A) \leq \|A\| \tilde{\tilde{\varepsilon}} \quad , \quad \|C(D)\| \leq \|A\|^2 \tilde{\tilde{\varepsilon}} \tag{4.1.15}$$

mit

$$\tilde{\tilde{\varepsilon}} < \frac{1}{648.995\, M\, 1.00431^m} \left(\frac{1}{18398} \frac{\delta^6}{\|A\|^6}\right)^M \frac{\delta^7}{\|A\|^7} \quad . \tag{4.1.16}$$

Dann konvergieren die durch 4.1.1 definierten Matrixfolgen $(A^{(N)})_{N\in\mathbb{N}_0}$ quadratisch gegen die Murnaghan-Form von A in dem Sinne, daß die Folgen $(s(A^{(N)})/\|A\|)_{N\in\mathbb{N}_0}$ und $(\|C(D^{(N)})\| / \|A\|^2)_{N\in\mathbb{N}_0}$ durch eine gemeinsame quadratisch konvergente Nullfolge majorisiert werden.

Beweis:
Es sei die Folge $(\varepsilon_N)_{N\in\mathbb{N}_0}$ definiert durch

$$\varepsilon_0 = \tilde{\tilde{\varepsilon}} \quad , \quad \varepsilon_{N+1} = K\, \varepsilon_N^2 \quad , N \in \mathbb{N}_0 \tag{4.1.17}$$

mit

$$K = 648.995\ M\ 1.00431^m \left(18398\, \frac{\|A\|^6}{\delta^6}\right)^M \frac{\|A\|^7}{\delta^7} \quad .$$

Wir beweisen nun zunächst per Induktion über N

$$s(A^{(N)})/\|A\| \leq \varepsilon_N \, , \, \|C(D^{(N)})\| / \|A\|^2 \leq \varepsilon_N \, , \, \varepsilon_N \leq \tilde{\tilde{\varepsilon}} \, , N\in\mathbb{N}_0 \, . \tag{4.1.18}$$

Für $N = 0$ haben wir aufgrund von (4.1.15)

$$s(A^{(0)})/\|A\| \leq \varepsilon_0 \, , \, \|C(D^{(0)})\| / \|A\|^2 \leq \varepsilon_0 \, .$$

Es seien dann die Ungleichungen (4.1.18) für ein $N \in \mathbb{N}_0$ erfüllt. Aus $m-1 \leq M$ und $\delta \leq \sqrt{2}\|A\|$ erhalten wir

$$\tilde{\tilde{\varepsilon}} \leq \frac{\sqrt{2}^3}{648.995\, \sqrt{m-1}\, \sqrt{M}\, 1.00431^m} \left(\frac{\sqrt{2}^3}{18398} \frac{\delta^3}{\|A\|^3}\right)^M \frac{\delta^4}{\|A\|^4}$$

$$\leq \frac{1}{1000\sqrt{m-1}\sqrt{M}\,1.0017^m}\left(\frac{1}{114.044}\frac{\delta^3}{\|A\|^3}\right)^M \frac{\delta^4}{\|A\|^4} \quad .$$

Damit sind für $\tilde{\varepsilon} = \varepsilon_N$ die Voraussetzungen (4.1.1),(4.1.2) erfüllt, und es folgt aus den Sätzen 4.1.5 und 4.1.6

$$S(A^{(N+1)}) \leq K\|A\|\varepsilon_N^2 \quad , \quad \|C(D^{(N+1)})\| \leq K\|A\|^2\varepsilon_N^2 \quad ,$$

wobei in die zweite Abschätzung nochmals $\delta \leq \sqrt{2}\|A\|$ eingeht. Es ergibt sich

$$S(A^{(N+1)})/\|A\| \leq \varepsilon_{N+1} \quad , \quad \|C(D^{(N+1)})\|/\|A\|^2 \leq \varepsilon_{N+1} \quad ,$$

und aus $\varepsilon_N \leq \tilde{\tilde{\varepsilon}}$ und $\tilde{\tilde{\varepsilon}} < \frac{1}{K}$ folgt

$$\varepsilon_{N+1} \leq \tilde{\tilde{\varepsilon}} \quad .$$

Die Folge $(\varepsilon_N)_{N\in\mathbb{N}_0}$ majorisiert also die Folgen

$$(S(A^{(N)})/\|A\|)_{N\in\mathbb{N}_0}$$

und

$$(\|C(D^{(N)})\|/\|A\|^2)_{N\in\mathbb{N}_0} \quad .$$

Nach (4.1.18) gilt $\varepsilon_N < \frac{1}{K}$, $N \in \mathbb{N}_0$ und somit

$$\varepsilon_{N+1} < \varepsilon_N \quad , \quad N \in \mathbb{N}_0 \quad .$$

Daher ist $(\varepsilon_N)_{N\in\mathbb{N}_0}$ streng monoton fallend und schließlich nach (4.1.17) eine quadratisch konvergente Nullfolge. ■

4.2 ZUR GLOBALEN KONVERGENZ DER JACOBI-ÄHNLICHEN BLOCKVERFAHREN

Das Problem der globalen Konvergenz der zwei in § 4.1 formulierten Verfahren ist nicht trivial. Die Praxis zeigt (vgl. auch § 5.1), daß die durch diese Verfahren erzeugten Matrizenfolgen für eine beliebige J-symmetrische Startmatrix gegen eine Endmatrix in Murnaghan-Form konvergieren. Man bedenke jedoch, daß für das ähnlich strukturierte komplexe Eberlein-Verfahren [10] kein Beweis der globalen Konvergenz selbst unter optimaler Pivotstrategie existiert. Zudem scheint ein entsprechender Beweis für unsere Verfahren durch die Blockarithmetik erheblich schwieriger und aufwendiger zu sein.

Wenn wir uns im orthogonalen Teil der Verfahren entweder generell für die Diagonalisierung des symmetrischen Teils oder generell für die Blockdiagonalisierung des schiefsymmetrischen Teils der Matrix entscheiden, so lassen sich bei optimaler Pivotstrategie Konvergenzsätze analog zu [48,Theorem 1] und [18,Theorem 1] beweisen. Die globale Konvergenz ist dann so zu verstehen, daß die Matrizenfolge gegen Normalität konvergiert (vgl.[38]) und der symmetrische Teil der iterierten Matrizen gegen eine feste Diagonalmatrix bzw. der schiefsymmetrische Teil gegen eine feste Matrix in Murnaghan-Form strebt. Im Falle getrennter Eigenwerte konvergiert die Matrizenfolge gegen die Murnaghan-Form der Ausgangsmatrix.

Weiterhin scheint eine Übertragung des Konvergenzsatzes von Hari ([19,Theorem 2.1]) auf die zeilenzyklische Version des erstgenannten Verfahrens möglich zu sein.

Die notwendigen Untersuchungen würden jedoch den Rahmen dieser Arbeit sprengen, so daß wir uns auf diese Anmerkungen beschränken wollen.

Hiermit sind die theoretischen Betrachtungen abgeschlossen, und wir kommen im nächsten Kapitel zur Diskussion der numerischen Resultate.

5. NUMERISCHE ANWENDUNGEN

Wir haben alle entwickelten Verfahren als ALGOL60-Prozeduren programmiert. Die numerischen Berechnungen wurden auf der Rechenanlage IBM 3031 des Rechenzentrums der Fernuniversität Hagen unter Benutzung des "Fast ALGOL60 Compiler of Delft, Release 1/2/1977" mit "Double-Precision"-Arithmetik (14-stellige hexadezimale bzw. 16-stellige dezimale Mantisse) durchgeführt. Aus Platzgründen haben wir darauf verzichtet, die ALGOL60-Programme in diese Arbeit aufzunehmen. Der interessierte Leser kann sie jedoch jederzeit bei dem Autor einsehen. Wir beabsichtigen, die Programme demnächst an anderer Stelle zu publizieren.

5.1 NUMERISCHE RESULTATE DER JACOBI-ÄHNLICHEN BLOCKVERFAHREN

Im Algorithmus 4.1.1 haben wir zwei Verfahren durch Beschreibung eines vollständigen Zyklus definiert. Für die folgenden numerischen Untersuchungen bezeichnen wir diese Verfahren als *Alg1* (Standard-Normreduzierung) und *Alg2* (Optimale Normreduzierung). Sie benötigen als Eingabe eine beliebige reelle J-symmetrische Matrix A und führen so viele Zyklen aus, bis eins der beiden folgenden Abbruchkriterien erfüllt ist:

(i) $$\max_{p<q} \max \{|a_{2p-1,2q-1}|,|a_{2p-1,2q}|,|a_{2p,2q-1}|,|a_{2p,2q}|\} \Big/ \max_{p=1(1)m} \max \{|a_{2p-1,2p-1}|,|a_{2p-1,2p}|,|a_{2p,2p}|\} \leq \varepsilon_s \quad ^{*)} \tag{5.1.1}$$

(*normales Ende*).

(ii) Es wurde keine Transformation in einem vollen Zyklus ausgeführt,

oder

es sind 50 Zyklen ausgeführt worden

(*abnormales Ende*).

Als Ausgabe liefern die Verfahren dann bei normalem Ende eine (im Sinne von $O(\varepsilon_s)$) blockdiagonale J-symmetrische Matrix $\bar{A}$ und die J-orthogonale Transformationsmatrix R, die die Startmatrix A durch eine Ähnlichkeitstransformation in die Endgestalt $\bar{A}$ überführt.

Die Endmatrix $\bar{A}$ muß nicht notwendig Murnaghan-Form besitzen, so daß die Näherungen der Eigenwerte direkt aus den Diagonalblöcken "abgelesen" werden könnten. Sie berechnen sich dann durch Lösen von m quadratischen Gleichungen als

$$\lambda_{2i-1,2i} = \frac{\bar{a}_{2i-1,2i-1} + \bar{a}_{2i,2i}}{2} \pm \sqrt{\frac{1}{4}(\bar{a}_{2i-1,2i-1} - \bar{a}_{2i,2i})^2 - \bar{a}_{2i-1,2i}^{\,2}}\,, \quad i = 1(1)m. \tag{5.1.2}$$

*) ε_s ist eine vorgegebene Schwelle, die in Abhängigkeit von der Maschinengenauigkeit gewählt wird.

Durch eine einfache Rechnung erhält man hieraus die Rechtseigenvektoren z_i von $\bar{A}$ (vgl.[26]), und die Rechts- und Linkseigenvektoren x_i bzw. y_i der ursprünglichen Matrix A ergeben sich als

$$x_i = R z_i \quad , \qquad y_i = J R z_i \quad , \quad i = 1(1)n \quad . \tag{5.1.3}$$

In Lemma 2.1.5 haben wir eine alternative Methode zur Bestimmung der Eigenwerte und Eigenvektoren aus der Endgestalt $\bar{A}$ aufgezeigt. Im Hinblick auf mögliche defektive oder fast-defektive 2x2-Diagonalblöcke von $\bar{A}$ erscheint jedoch die Berechnung nach (5.1.2) und (5.1.3) stabiler.

Wir wählen ε_S in (5.1.1) als $\varepsilon_S = \sqrt{\varepsilon_M}/100$, wobei ε_M die kleinste positive Zahl ist, für die das Ergebnis der Maschinenoperation $1 + \varepsilon_M$ größer als 1 ist. Bei der verwendeten Rechenanlage IBM 3031 ist $\varepsilon_M = 2^{-52} \approx 2.220446 \cdot 10^{-16}$.

Zur Speicherung der Matrix A wird aufgrund der J-Symmetrie nur das rechtsobere Dreieck (inklusive der Hauptdiagonalen) benötigt, während für die Akkumulierung der einzelnen Transformationsmatrizen eine volle nxn-Matrix bereitgestellt werden muß. Die Transformationen der Zeilen und Spalten von A werden (abweichend von der Formulierung für den Beweis der quadratischen Konvergenz) nicht blockweise, sondern elementweise ausgeführt. Dabei erfolgt die Berechnung der Elemente a_{ij}' gemäß der in [26] formulierten speicherplatzsparenden 3-Schritt-Strategie, d.h. für jedes Pivotpaar (p,q) in der Reihenfolge

(i) $(p,p),(p,q),(q,q),(p,q+1),(q,q+1),\ldots,(p,n),(q,n),$

(ii) $(1,p),(1,q),(2,p),(2,q),\ldots,(p,p),(p,q),(q,q),$

(iii) $(p,p+1),(p+1,q),(p,p+2),(p+2,q),\ldots,(p,q-1),(q-1,q),$

und die Transformationsparameter werden aus den definierenden Gleichungen (3.1.4),(3.2.16),(3.3.7),(3.3.8), etc. mit Hilfe der bekannten stabilen Formeln (vgl.[16]) bestimmt. Um eine zu zeitaufwendige Iteration zur Bestimmung des Infimums von $F(x_1,x_2)$ im Verfahren Alg2 zu vermeiden, begrenzen wir in der Praxis die Anzahl der Iterationsschritte für den Newton-Prozess (2.2.27) auf 5 und für den Eberlein-Prozess (2.2.29),(2.2.30) auf 10.

In § 2.2.2 haben wir die Normreduzierung mittels der Funktion $G(t_1,t_2)$ beschrieben. Obwohl wir die garantierte Normabnahme für den allgemeinen Fall nicht nachweisen konnten, formulieren wir hier ein Verfahren, welches den normreduzierenden Schritt für jeden Pivotblock A_{pq} auf diese Weise ausführt. Wir bezeichnen dieses Verfahren als *Alg3*. Es ist bis auf den Schritt 2a im Algorithmus 4.1.1 identisch mit den beiden Verfahren Alg1 und Alg2. Der Schritt 2a wird nun folgendermaßen ersetzt:

Schritt 2a*:

Berechne $c_{2p-1,2q}$, $c_{2p,2q-1}$, $H(0,0)$ mit Hilfe von (2.2.10).

Falls $\det H(0,0) > \varepsilon_M$ gilt, so setze

$$t_1 = - \frac{(16\alpha_2 + 4\alpha_{22} + 4\delta_+ + 4\delta_-)c_{2p-1,2q} - (4\delta_+ - 4\delta_-)c_{2p,2q-1}}{\det H(0,0)} ,$$

$$t_2 = - \frac{(16\alpha_1 + 4\alpha_{11} + 4\delta_+ + 4\delta_-)c_{2p,2q-1} - (4\delta_+ - 4\delta_-)c_{2p-1,2q}}{\det H(0,0)} .$$

Falls $\det H(0,0) \le \varepsilon_M$ oder $|t_i| > \frac{3}{4}$ für $i = 1$ oder $i = 2$ gilt, so setze für $|c_{2p-1,2q}| \ge |c_{2p,2q-1}|$

$$t_1 = - \frac{c_{2p-1,2q}}{16\alpha_1 + 4\alpha_{11} + 4\delta_+ + 4\delta_-} , \quad t_2 = 0$$

bzw. für $|c_{2p-1,2q}| < |c_{2p,2q-1}|$

$$t_1 = 0 , \quad t_2 = - \frac{c_{2p,2q-1}}{16\alpha_2 + 4\alpha_{22} + 4\delta_+ + 4\delta_-} .$$

Bestimme $\cosh x_i$, $\sinh x_i$, $i = 1,2$ aus $t_i = \tanh x_i$ und führe die Transformation

$$A \to T^{-1}AT$$

gemäß (2.2.4) aus.

Man beachte, daß durch diese Strategie die Parameter t_1, t_2 betragsmäßig durch $\frac{3}{4}$ beschränkt sind (vgl.[8]).

Für die Rechenpraxis haben wir alle drei Verfahren um einen zusätzlichen Programmschritt erweitert. Quasi als Vortransformation (oder als Schritt 1 im 1.Zyklus) wird die Eingabematrix A einer Ähnlichkeitstransformation gemäß Lemma 2.1.5 unterzogen, so daß nach diesem Schritt $\| C(D) \| = 0$ gilt. Die Parameter der Matrizen S werden dabei mit (2.1.29) und

$$\cosh x = \frac{1}{2}\left(\sqrt[8]{\frac{1+t}{1-t}} + \sqrt[8]{\frac{1-t}{1+t}}\right) \quad , \sinh x = \frac{1}{2}\left(\sqrt[8]{\frac{1+t}{1-t}} - \sqrt[8]{\frac{1-t}{1+t}}\right)$$

für $t = \tanh 4x$, $|t| < 1$ bestimmt. Da jedoch dieser Schritt nicht notwendig die Norm der Gesamtmatrix A reduziert (es existieren Beispiele, wo $\|A\|$ zunimmt), soll er nur für blockdiagonaldominante Matrizen ausgeführt werden. Wir fordern daher, daß der Quotient in (5.1.1) nicht größer als $5 \cdot 10^{-2}$ sein darf. Um das Auftreten großer Transformationsparameter in S zu vermeiden, fordern wir schließlich noch, daß

$$|\tanh 4x| \leq 0.999$$

und damit

$$|x| \leq 0.951 \quad \text{bzw.} \quad |\tanh x| \leq 0.740$$

gilt.

Zum Vergleich unserer drei Verfahren haben wir die folgenden J-symmetrischen Testmatrizen ausgewählt:

(1) 10 J-symmetrische Zufallsmatrizen *) der Dimension 20.

(2) 10 J-symmetrische Zufallsmatrizen der Dimension 40.

(3) Die Matrizen aus Beispiel (1) mit

(a) $A_{ij} := \frac{1}{100} A_{ij}$ für $i \neq j$, (b) $A_{ij} := \frac{1}{10000} A_{ij}$ für $i \neq j$.

*) Unter Zufallsmatrizen verstehen wir Matrizen, deren Elemente a_{ij} von einem Zufallszahlengenerator ("random") mit $a_{ij} \in (-1,1)$ erzeugt wurden.

(4) Die Matrizen aus Beispiel (2) mit

$$\text{(a)}\ A_{ij} := \frac{1}{100} A_{ij} \quad \text{für } i \neq j, \qquad \text{(b)}\ A_{ij} := \frac{1}{10000} A_{ij} \quad \text{für } i \neq j\ .$$

Die Eigenwerte dieser Matrizen treten reell und nicht-reell auf, sie sind in allen Beispielen gut getrennt ($\delta > 0$). Die Beispiele (3) und (4) stellen blockdiagonaldominante Matrizen dar, wie man sie in ähnlicher Form bei der Simultanen Iteration (vgl. § 5.3) und bei der Parameteranalyse (vgl. § 5.4) erhält.

(5)

$$A(\omega) = \frac{1}{2} \begin{pmatrix} \omega & -1 & & & & & & & & 1 & -\omega \\ 1 & 0 & \ddots & & & & & & & 0 & 1 \\ & \ddots & \ddots & \ddots & & & & & \iddots & & \\ & & \ddots & 0 & -1 & 1 & 0 & \ddots & & & \\ & & & 1 & -1 & -1 & 1 & & & & \\ & & & 1 & 1 & 1 & 1 & & & & \\ & & & 0 & 1 & -1 & 0 & \ddots & & & \\ & & \iddots & & & & & \ddots & \ddots & & \\ & & & & & & & \ddots & 0 & 1 \\ 1 & 0 & & & & & & & & -1 & -\omega \\ \omega & 1 & & & & & & & & & \end{pmatrix},$$

dim $A(\omega) = n = 2m$, m gerade, $\omega \geq 0$, alle freien Positionen sind gleich 0 (vgl.[53]).

Die Eigenwerte dieser Matrix sind die komplexen n-ten Wurzeln von ω. Für $\omega = 1$ ist $A(\omega)$ normal. Mit $\omega \to 0$ verschlechtert sich die Kondition der Matrix, und für $\omega = 0$ ist $\lambda = 0$ n-facher defektiver Eigenwert (mit nur einem Eigenvektor). Die Abweichung von der Normalität wird durch

$$\Delta(A) = \sqrt{\omega^2 - n\,\omega^{2/n} + n - 1}$$

bestimmt. Wir betrachten die Fälle

$$\text{(a)}\ \omega = 100, \quad \text{(b)}\ \omega = 20, \quad \text{(c)}\ \omega = 5, \quad \text{(d)}\ \omega = 1, \quad \text{(e)}\ \omega = 0.5,$$
$$\text{(f)}\ \omega = 0.01, \quad \text{(e)}\ \omega = 0$$

mit $n = 20$.

(6)
$$A = \begin{pmatrix} 0 & l_1 & & & & & \\ -l_1 & -d_1 & -r_1 & & & & \\ & r_1 & 0 & l_2 & & & \\ & & -l_2 & -d_2 & -r_2 & & \\ & & & r_2 & \ddots & \ddots & \\ & & & & \ddots & -r_{m-1} & \\ & & & & r_{m-1} & 0 & l_m \\ & & & & & -l_m & -d_m \end{pmatrix}$$

mit

$$l_j = (m+1)\sqrt{\frac{m-j+2}{m-j+1}} \quad , \quad r_j = -(m+1)\sqrt{\frac{m-j}{m-j+1}}$$

(vgl.[26],[53]).

Diese Matrix ergibt sich aus der Linearisierung eines quadratischen Eigenwertproblems (vgl. § 1.3), welches die gedämpften Schwingungen eines Systems von m Massen beschreibt. Wenn

$$d_j = d \quad , \quad j = 1(1)m$$

gilt, so ist die Dämpfung proportional, und die Eigenwerte sind exakt bekannt. Sie berechnen sich als

$$\lambda_{2j-1,2j} = \frac{1}{2}\left(-d \pm \sqrt{d^2 - 16(m+1)^2 \sin^2 \frac{j\pi}{2(m+1)}}\right), \quad j = 1(1)m.$$

Wir untersuchen für $n = 2m = 20$ die Fälle

(a) $d_j = d = 1$, (b) $d_j = d = 44 \sin\frac{\pi}{22}$,

(c) $d_j = d = 30$, (d) $d_j = d = 50$

und für $n = 2m = 40$

(e) $d_j = 0.1\,j$, (f) $d_j = d = 1.$

Dabei sind in (a) alle Eigenwerte nicht-reell, in (b) ebenfalls bis

auf einen doppelten reellen defektiven Eigenwert ("kritische" Dämpfung), in (c) treten 10 reelle und 10 nicht-reelle Eigenwerte auf, und in (d) sind alle reell. In (e) und (f) sind wiederum alle Eigenwerte nicht-reell. Die Abweichung von der Normalität berechnet sich hier für $d_j = d$ als

$$\Delta(A) = \sqrt{(m-m_1)d^2 + 8m_1(m+1)^2}\ ,$$

wobei $2m_1$ mit $0 \le m_1 \le m$ die Anzahl der reellen Eigenwerte von A bestimmt.

(7)
$$A(\omega) = \begin{pmatrix} 0 & 1 & 0 & 1 & \cdots & 0 & 1 \\ -1 & -\omega & -1 & 0 & & -1 & 0 \\ 0 & 1 & 0 & 1 & & & \\ -1 & 0 & -1 & -\omega & & & \\ \vdots & & & & \ddots & & \\ & & & & & 0 & 1 \\ & & & & & -1 & 0 \\ 0 & 1 & & & 0 & 1 & 0 & 1 \\ -1 & 0 & \cdots & & -1 & 0 & -1 & -\omega \end{pmatrix},$$

dim $A(\omega) = n = 2m$, $\omega \ge 0$.

Die Eigenwerte dieser Matrix sind $\lambda_{1,2} = \frac{1}{2}(-\omega \pm \sqrt{\omega^2 - n^2})$, m-1 mal $\lambda = 0$ und m-1 mal $\lambda = -\omega$. Dabei sind die (m-1)-fachen Eigenwerte 0 und $-\omega$ nicht-defektiv, während für $\omega = n$ der doppelte Eigenwert $\lambda_{1,2} = -\frac{1}{2}\omega$ defektiv ist. Die Abweichung von der Normalität wird hier durch

$$\Delta(A) = \begin{cases} \omega & \text{für } \omega \le n \\ n & \text{für } \omega > n \end{cases}$$

bestimmt. Wir betrachten

(a) $\omega = 1$, (b) $\omega = 10$, (c) $\omega = 20$, (d) $\omega = 30$

für $n = 20$. Hierbei ist das Paar $\lambda_{1,2}$ für (a) und (b) nicht-reell, für (c) doppelt reell defektiv und für (d) reell und getrennt.

(8)
$$A(\omega) = \begin{pmatrix} 0 & 1-m & 0 & 1 & \cdots & 0 & 1 \\ -1+m & -\omega & -1 & 0 & & -1 & 0 \\ 0 & 1 & 0 & 1-m & & & \vdots \\ -1 & 0 & -1+m & -\omega & & & \vdots \\ \vdots & & & & \ddots & & \vdots \\ \vdots & & & & & 0 & 1 \\ \vdots & & & & & -1 & 0 \\ 0 & 1 & \cdots & & 0 & 1 & 0 \quad 1-m \\ -1 & 0 & & & -1 & 0 & -1+m \quad -\omega \end{pmatrix},$$

$\dim A(\omega) = n = 2m$, $\omega \geq 0$.

Hier sind die Eigenwerte $\lambda = 0$ und $\lambda = -\omega$ (jeweils einfach) und m-1 Paare $\lambda_{1,2} = \frac{1}{2}(-\omega \pm \sqrt{\omega^2 - n^2})$. Die (m-1)-fachen Eigenwerte λ_1 bzw. λ_2 sind jeweils nicht-defektiv, unabhängig davon, ob sie reell oder nicht-reell sind, nur für $\omega = n$ ist der (n-2)-fache Eigenwert $\lambda = -\frac{1}{2}\omega$ defektiv. Hier gilt

$$\Delta(A) = \begin{cases} \omega \sqrt{m-1} & \text{für } \omega \leq n \\ n \sqrt{m-1} & \text{für } \omega > n \end{cases}.$$

Wir untersuchen

(a) $\omega = 1$, (b) $\omega = 10$, (c) $\omega = 20$, (d) $\omega = 30$

für $n = 20$. Die Aussagen über die Eigenwertpaare $\lambda_{1,2}$ gelten hier analog zum Beispiel (7).

(9) Die J-symmetrischen Matrizen der Dimension 6 aus den Untersuchungen von § 2.2.1

(a) (A6) mit den nicht-defektiven Eigenwerten

$$\lambda_{1,2} = 2, \quad \lambda_{3,4} = 3, \quad \lambda_{5,6} = \frac{9}{2} \pm \frac{\sqrt{3}}{2} i,$$

(b) (A15) mit den nicht-defektiven Eigenwerten

$$\lambda_{1,2} = 1 + \sqrt{3} i, \quad \lambda_{3,4} = 1 - \sqrt{3} i, \quad \lambda_{5,6} = 1,$$

(c) (A7) mit dem defektiven Eigenwert $\lambda_{1,2,3,4} = 1$ und $\lambda_5 = 0$, $\lambda_6 = 4$.

(10)
$$A = \begin{pmatrix} 0 & \sqrt{1/2} & 0 & 0 \\ -\sqrt{1/2} & \sqrt{2} & \sqrt{3/2} & 0 \\ 0 & -\sqrt{3/2} & 0 & \sqrt{2} \\ 0 & 0 & -\sqrt{2} & -\sqrt{2} \end{pmatrix}$$

(vgl.[26],[53]).

Diese Matrix beschreibt eine Schwingung von zwei Massen, und die Eigenwerte sind $\lambda_{1,2} = i$ und $\lambda_{3,4} = -i$, jeweils defektiv.

(11) Zwei Beispiele aus der Praxis, die bei der Lösung von Schwingungsproblemen auftreten. Wir betrachten zwei J-symmetrische Matrizen der Dimension

(a) 90 , (b) 180 ,

die aus quadratischen Eigenwertproblemen (1.3.2) mit symmetrischen Matrizen M,D und K resultieren. Diese Probleme beschreiben jeweils die gedämpften Schwingungen eines dynamischen Systems einer mit Federn und Dämpfern gelagerten Maschinenkonstruktion. Hierbei sind die Matrizen M und D jeweils Diagonalmatrizen, und die Matrix K weist in beiden Fällen eine Bandstruktur mit "Löchern" auf, so daß die J-symmetrischen Matrizen A dünn besetzt sind und ebenfalls eine Bandstruktur mit "Löchern" besitzen (vgl. § 1.3). Die Eigenwerte sind für beide Matrizen ausschließlich nicht-reell, wobei sowohl die Real- als auch die Imaginärteile getrennt sind. Für (a) liegen die Realteile zwischen -2.142 und $-8.650 \cdot 10^{-6}$, die positiven Imaginärteile zwischen 8.870 und 755.890, und für (b) die Realteile zwischen -25.488 und $-6.180 \cdot 10^{-15}$ und die positiven Imaginärteile zwischen 16.830 und 1565.333 .

Bevor wir die Ergebnisse der drei Algorithmen im Detail beschreiben, wollen wir einige Bemerkungen zu unseren allgemeinen Rechenerfahrungen machen.

Für alle oben beschriebenen Beispiele liefern unsere Algorithmen eine Endmatrix in ($O(\varepsilon_s)$-) Blockdiagonalform. Dies ist ein Indiz dafür, daß unsere Verfahren generell global konvergieren, obwohl ein Beweis dieser Tatsache sehr schwierig sein dürfte (vgl. § 4.2). Ein gewisser Nachteil Jacobi-ähnlicher Verfahren ist bekanntlich der, daß sie keinen Nutzen aus einer Bandgestalt oder einer ähnlichen schwach besetzten Struktur der Startmatrix ziehen können. Dieses Phänomen konnten wir auch beobachten. So wird in den Beispielen (5),(6) und (11) jeweils schon im 1.Zyklus die Struktur der Matrix zerstört, so daß die Verfahren vom 2.Zyklus an auf vollbesetzten Matrizen arbeiten müssen.

Im normreduzierenden Schritt für die Außerdiagonalblöcke konnten wir für Alg1 und Alg2 registrieren, daß das Kriterium (2.2.28) immer schon für $\gamma_k = 1$ erfüllt ist, und daß für Alg3 die Parameterwahl gemäß (2.2.35) stets eine Normabnahme bewirkt (vgl. dazu die Anmerkungen in § 2.2.2). Im direkten Vergleich der Normabnahme in Alg1 und Alg3 zeigte sich, daß die Normreduzierung mittels $G(t_1,t_2)$ nie wesentlich schlechter als die mittels $F(x_1,x_2)$ (höchstens in der 6. oder 7. Dezimalstelle), stattdessen aber in den meisten Fällen deutlich besser als diese ist. Bei der optimalen Normreduzierung in Alg2 registrierten wir, daß im Durchschnitt 2 - 3 Iterationsschritte zur Minimierung von $F(x_1,x_2)$ ausgeführt werden, wobei in den ersten beiden Zyklen in der Regel 3 - 5 Schritte und in den letzten Zyklen nur 1 Schritt benötigt wird. Bei allen drei Verfahren war zu beobachten, daß die Normreduzierung fast ausschließlich mittels $F(x_1,x_2)$ bzw. $G(t_1,t_2)$ durchgeführt wird. Die alternative Methode gemäß der Eberlein-Strategie wird in unseren Beispielen lediglich für die (konstruierte) Matrix (9c) angewandt (man beachte hierbei die verschiedenen Kriterien zum Aufrufen der Alternativ-Strategie in Alg1 bzw. Alg2 und Alg3).

Wir fassen nun unsere numerischen Resultate in den folgenden Tabellen zusammen. Die Bezeichnungen sind dabei

$\Delta(A)$ Abweichung von der Normalität für die Startmatrix A (wenn die Eigenwerte nicht explizit bekannt sind, wird dieser Wert aus einem Programmprotokoll berechnet),

CPU Rechenzeit in Sekunden (inklusive der Akkumulation der Transformationsmatrizen, ohne Ein- und Ausgabe der Matrix),

Zyklen Anzahl der Iterationszyklen,

Kond Konditionszahl der Transformationsmatrix R, definiert als $K(R) = \|R\|_1 \|R^{-1}\|_1 = \|R\|_1 \|R\|_\infty$ (aufgrund der J-Orthogonalität von R),

$S(\bar{A})$ Abweichung von der Blockdiagonalgestalt für die Endmatrix $\bar{A}$.

Ergebnisse der Beispiele (1) - (4):

	CPU			Zyklen			Kond		
	Alg1	Alg2	Alg3	Alg1	Alg2	Alg3	Alg1	Alg2	Alg3
(1)	5.16	5.83	4.94	7.5	7.2	7.7	57.01	56.92	57.43
(2)	44.33	46.21	42.64	9.4	8.8	9.5	154.65	149.96	157.92
(3a)	2.65	2.73	2.47	3.9	3.7	3.9	13.47	13.25	13.47
(3b)	1.49	1.56	1.40	2.1	2.1	2.1	8.46	8.46	8.46
(4a)	20.83	21.85	20.15	4.7	4.5	4.7	38.37	38.56	38.47
(4b)	11.34	11.71	10.93	2.3	2.3	2.3	15.74	15.74	15.74

Die aufgelisteten Werte sind Mittelwerte. $S(\bar{A})$ ist in allen Beispielen höchstens 10^{-10}. Wir konnten hier unser theoretisches Resultat aus Kapitel 4 bestätigen, alle drei Verfahren sind für alle Beispielmatrizen asymptotisch quadratisch konvergent. Es fällt auf, daß die Methode Alg2 mit Ausnahme der stark blockdiagonaldominanten Matrizen (3b) und (4b) weniger Zyklen als Alg1 benötigt, jedoch in der CPU-Zeit aufgrund der aufwendigeren Parameterbestimmung nicht besser liegt. Die Kondition der Transformationsmatrizen ist für Alg2 nur bei den Matrizen (4a) ein wenig schlechter als die für Alg1. Dies ist eine Bestätigung der sinnvollen Parameterbeschränkung für diese Methode (vgl. § 2.2.2). Als schnellstes Verfahren erweist sich Alg3. Es erfordert in etwa die gleiche Zyklenzahl wie Alg1 und liefert vergleichbare Konditionszahlen, benötigt aber deutlich weniger CPU-Zeit als Alg1.

Die oben beschriebene Vortransformation, die im Schritt 1 des

1. Zyklus die Blockdiagonale D normalisiert, erweist sich als sehr wirkungsvoll. Sie wird bei sämtlichen blockdiagonaldominanten Matrizen (3a),(3b),(4a) und (4b) aufgerufen und bewirkt dort eine sofortige quadratische Konvergenz von $S(A^{(N)})$ und $\| C(D^{(N)}) \|$ gegen 0. Dabei wird die Norm der Gesamtmatrix A in keinem Beispiel vergrößert. Im Vergleich dazu erweisen sich zusätzlich getestete Programmvarianten, die diesen Schritt nach der üblichen Eberlein-Strategie ausführen, als deutlich langsamer. Das Problem ist hierbei, daß für die Startmatrix zwar $S(A)$ schon klein, aber $\| C(D) \|$ noch groß ist und somit die quadratische Konvergenz erst später einsetzt.

Ergebnisse Beispiel (5):

	ω	$\Delta(A)$	CPU Alg1	CPU Alg2	CPU Alg3	Zyklen Alg1	Zyklen Alg2	Zyklen Alg3	Kond Alg1	Kond Alg2	Kond Alg3	$S(\bar{A})$ Alg1	$S(\bar{A})$ Alg2	$S(\bar{A})$ Alg3
(a)	100	99.94	8.79	7.34	7.75	13	9	12	324.02	271.96	323.91	$1\sigma^{15}$	$1\sigma^{12}$	$1\sigma^{11}$
(b)	20	19.80	6.22	6.37	6.18	9	8	10	102.68	82.14	103.58	$1\sigma^{10}$	$1\sigma^{15}$	$1\sigma^{16}$
(c)	5	4.53	4.55	4.76	4.23	7	6	7	30.68	30.36	30.67	$1\sigma^{15}$	$1\sigma^{10}$	$1\sigma^{16}$
(d)	1	0	2.55	2.77	2.12	6	6	5	9.28	9.28	9.28	$1\sigma^{15}$	$1\sigma^{15}$	$1\sigma^{10}$
(e)	0.5	0.77	4.09	4.58	3.76	6	6	6	14.86	14.84	14.86	$1\sigma^{14}$	$1\sigma^{14}$	$1\sigma^{14}$
(f)	0.01	2.53	6.97	8.04	6.44	10	10	10	322.48	232.18	322.48	$1\sigma^{11}$	$1\sigma^{15}$	$1\sigma^{12}$
(g)	0	4.36	33.80	39.83	32.46	49	48	50	$6\cdot10^{13}$	$6\cdot10^{14}$	$5\cdot10^{13}$	$1\sigma^{12}$	$1\sigma^{15}$	$1\sigma^{11}$

Ergebnisse Beispiel (6):

	d_j	$\Delta(A)$	CPU Alg1	CPU Alg2	CPU Alg3	Zyklen Alg1	Zyklen Alg2	Zyklen Alg3	Kond Alg1	Kond Alg2	Kond Alg3	$S(\bar{A})$ Alg1	$S(\bar{A})$ Alg2	$S(\bar{A})$ Alg3
(a)	1	3.16	3.55	3.76	3.37	5	5	5	9.74	9.74	9.74	$1\sigma^{12}$	$1\sigma^{12}$	$1\sigma^{12}$
(b)	6.26	36.34	3.98	4.29	3.66	6	6	6	51.95	52.08	51.99	$1\sigma^{11}$	$1\sigma^{11}$	$1\sigma^{11}$
(c)	30	96.64	4.20	4.72	4.01	6	6	6	39.85	37.98	39.85	$1\sigma^{10}$	$1\sigma^{10}$	$1\sigma^{10}$
(d)	50	98.39	4.15	4.55	3.96	6	6	6	25.28	25.28	25.28	$1\sigma^{12}$	$1\sigma^{14}$	$1\sigma^{12}$
(e)	0.1j	5.95	34.85	34.36	32.01	8	7	8	53.26	53.26	53.26	$1\sigma^{11}$	$1\sigma^{14}$	$1\sigma^{11}$
(f)	1	4.47	29.33	30.77	28.98	6	6	6	18.71	18.71	18.71	$1\sigma^{10}$	$1\sigma^{10}$	$1\sigma^{10}$

Auch hier ist in allen Beispielen mit getrennten Eigenwerten die Konvergenz von $S(A^{(N)})$ und $\| C(D^{(N)}) \|$ gegen 0 asymptotisch quadratisch. Für die Matrix (5g) mit dem 20-fachen Eigenwert $\lambda = 0$ ist die Konvergenz nur linear, während für das Beispiel (6b) mit einem doppelten reellen defektiven Eigenwert zwar $\| C(D^{(N)}) \|$ im Laufe der Iteration groß bleibt, aber $S(A^{(N)})$ asymptotisch quadratisch gegen 0 konvergiert. Die Methode Alg2 benötigt im Beispiel (5) bei starker Abweichung von der Normalität wesentlich weniger Zyklen als Alg1, jedoch ist sie auch hier nur im Beispiel (5a) schneller als Alg1. Im Beispiel (6) hingegen erzielt Alg2 gegenüber Alg1 trotz starker Abweichung von der Normalität wie in (b),(c) und (d) keine Verbesserung der Zyklenzahl und bewirkt damit eine deutlich schlechtere CPU-Zeit. Die Aussagen über Alg2, die wir bei den Beispielen (1) - (4) hinsichtlich der Kondition von R gemacht haben, bestätigen sich auch hier. Generell sind die Konditionszahlen zufriedenstellend, und es fällt auf, daß wir im Beispiel (6b) trotz des doppelten defektiven Eigenwertes keine großen Konditionszahlen erhalten (man beachte, daß unsere Methoden den defektiven 2x2-Block nicht zu normalisieren versuchen). Lediglich im Beispiel (5g) wächst $K(R)$ katastrophal an, und der Eigenwert 0 wird nur auf 2 Dezimalstellen genau berechnet. Schließlich ist das Verfahren Alg3 wieder bezüglich der Zyklenzahl und der Kondition mit Alg1 nahezu identisch, jedoch in allen Fällen deutlich schneller.

Wir können an diesen Beispielen zwei interessante Phänomene beobachten. So zeigt das Beispiel (5), daß unsere Methoden bei starker Abweichung von der Normalität bzw. bei schlechter Kondition der Matrix aufwendiger und langsamer werden. Andrerseits vergrößert sich die Zyklenzahl aber auch bei vergleichbaren Matrizen (s.(6a) und (6f)) mit wachsender Dimension.

Ergebnisse Beispiel (7):

	ω	$\Delta(A)$	CPU			Zyklen			Kond			$S(\bar{A})$		
			Alg1	Alg2	Alg3	Alg1	Alg2	Alg3	Alg1	Alg2	Alg3	Alg1	Alg2	Alg3
(a)	1	1	1.88	1.99	1.72	3	3	3	8.61	8.61	8.61	10^{-13}	10^{-13}	10^{-13}
(b)	10	10	3.00	3.21	2.29	5	5	4	11.79	12.38	10.94	10^{-11}	10^{-12}	10^{-10}
(c)	20	20	3.58	3.75	3.27	6	6	6	83.13	84.68	79.56	10^{-10}	10^{-10}	10^{-11}
(d)	30	20	1.61	1.75	1.48	3	3	3	13.21	13.21	13.21	10^{-13}	10^{-13}	10^{-13}

Ergebnisse Beispiel (8):

	ω	$\Delta(A)$	CPU Alg1	CPU Alg2	CPU Alg3	Zyklen Alg1	Zyklen Alg2	Zyklen Alg3	Kond Alg1	Kond Alg2	Kond Alg3	$S(\bar{A})$ Alg1	$S(\bar{A})$ Alg2	$S(\bar{A})$ Alg3
(a)	1	3	1.60	1.21	1.55	3	2	3	8.43	8.53	8.90	$1\sigma^{12}$	$1\sigma^{10}$	$1\sigma^{12}$
(b)	10	30	1.63	1.78	1.50	3	3	3	12.02	11.03	10.66	$1\sigma^{9}$	$1\sigma^{10}$	$1\sigma^{10}$
(c)	20	60	4.20	4.49	3.97	6	6	6	141.40	145.18	141.52	$1\sigma^{10}$	$1\sigma^{10}$	$1\sigma^{10}$
(d)	30	60	1.65	1.82	1.49	3	3	3	11.98	13.88	11.98	$1\sigma^{10}$	$1\sigma^{9}$	$1\sigma^{10}$

Ergebnisse Beispiele (9) und (10):

	$\Delta(A)$	CPU Alg1	CPU Alg2	CPU Alg3	Zyklen Alg1	Zyklen Alg2	Zyklen Alg3	Kond Alg1	Kond Alg2	Kond Alg3	$S(\bar{A})$ Alg1	$S(\bar{A})$ Alg2	$S(\bar{A})$ Alg3
(9a)	10.05	0.21	0.17	0.19	6	4	6	11.62	12.91	11.59	$1\sigma^{10}$	$1\sigma^{14}$	$1\sigma^{11}$
(9b)	8	0.21	0.17	0.18	6	4	6	15.37	17.83	16.16	$1\sigma^{15}$	$1\sigma^{14}$	$1\sigma^{16}$
(9c)	2.83	0.63	0.23	0.60	37	12	37	$6\cdot10^{7}$	$2\cdot10^{8}$	$5\cdot10^{7}$	$1\sigma^{8}$	$1\sigma^{8}$	$1\sigma^{8}$
(10)	2.83	0.64	0.29	0.55	39	15	36	$1\cdot10^{8}$	$2\cdot10^{8}$	$1\cdot10^{8}$	$1\sigma^{11}$	$1\sigma^{10}$	$1\sigma^{11}$

Alle Matrizen in diesen Beispielen besitzen mehrfache Eigenwerte. Wir konnten hier die interessante Tatsache beobachten, daß unsere Methoden für die Matrizen mit mehrfachen nicht-defektiven Eigenwerten ((7a),(7b),(7d),(8a),(8b),(8d),(9a),(9b)) eine asymptotisch quadratische Konvergenz von $S(A^{(N)})$ und $\| C(D^{(N)}) \|$ gegen 0 (im Sinne des Satzes 4.1.7) aufweisen, selbst wenn die algebraische Vielfachheit der Eigenwerte größer als 2 ist. In den übrigen Beispielen (es tritt jeweils mindestens ein defektiver Eigenwert auf) ist die Konvergenz linear. Dabei ist bemerkenswert, daß im Beispiel (7c) schon ein doppelter reeller defektiver Eigenwert die quadratische Konvergenz verhindert. Hier bleibt $\| C(D^{(N)}) \|$ groß, während $S(A^{(N)})$ zwar schnell, aber nur linear gegen 0 konvergiert (vgl. dazu Beispiel (6b)).

Die Konditionszahlen der Transformationsmatrizen sind bei den nicht-defektiven Problemen erwartungsgemäß gut. Bei den defektiven Matrizen (9c) und (10) wird $K(R)$ katastrophal groß, wogegen die Kondition bei den defektiven Beispielen (7c) und (8c) akzeptabel ist. Hier stecken die reellen defektiven Eigenwerte in 2x2-Diagonalblöcken. Ansonsten werden unsere bisherigen Resultate hinsichtlich der internen Vergleiche zwischen den drei Verfahren durch die obigen Zahlen bestätigt.

Ergebnisse Beispiel (11):

	$\Delta(A)$	CPU			Zyklen			Kond			$s(\bar{A})$		
		Alg1	Alg2	Alg3	Alg1	Alg2	Alg3	Alg1	Alg2	Alg3	Alg1	Alg2	Alg3
(a)	19.93	224.60	223.64	222.82	9	8	10	19.27	19.27	19.27	10^{-8}	10^{-9}	10^{-13}
(b)	265.45	1521.95	1556.07	1452.71	9	9	9	54.20	54.20	54.20	10^{-10}	10^{-10}	10^{-12}

Unsere bisherigen Beobachtungen und Schlußfolgerungen werden durch diese Untersuchung von großen Matrizen aus der Praxis unterstrichen. Insbesondere sind alle drei Methoden wieder asymptotisch quadratisch konvergent, da die Eigenwerte getrennt sind. Auffällig ist hier jedoch die Anzahl der (orthogonalen bzw. normreduzierenden) Außerdiagonaltransformationen. So werden von allen drei Verfahren bei der Lösung des 90x90-Problems (a) vom 2. Zyklus an statt der möglichen 990 Transformationen nur höchstens 540 und bei der Behandlung des 180x180-Problems (b) vom 1. Zyklus an statt 4005 nur höchstens 1980 Transformationen ausgeführt. Dieses Phänomen erklärt sich aus der speziellen Struktur der Matrizen. Wie wir erwähnten, sind beide Matrizen schwach besetzt, und offenbar werden einige Null-Blöcke durch Transformationen an anderen Pivots nur so schwach gestört, daß für sie selbst keine Transformationen erforderlich sind.

Allgemein ist die Genauigkeit der berechneten Eigenwerte zufriedenstellend, der relative Fehler liegt bei den nicht-defektiven Problemen in der Größenordnung $K(R)\,\varepsilon_M$, d.h. die Eigenwerte sind auf 13 - 14 Stellen genau. Natürlich verschlechtert sich die Genauigkeit bei defektiven Eigenwerten. Tritt ein doppelter reeller defektiver Eigenwert auf, so wird er nur auf 7 - 8 Stellen exakt bestimmt, und in den Beispielen (5g),(9c) und (10), für die wir sehr große Konditionszahlen erhielten, werden die Eigenwerte auch wieder nur mit einem Fehler der Größenordnung $K(R)\varepsilon_M$ berechnet.

Als Fazit unserer praktischen Untersuchungen können wir festhalten, daß alle drei betrachteten Algorithmen zuverlässige global konvergente Verfahren zur Berechnung der Eigenwerte und Eigenvektoren beliebiger J-symmetrischer Matrizen darstellen. Für nicht-defektive Probleme erweist sich die Konvergenz als asymptotisch quadratisch (was hoffen läßt, daß sich unser Konvergenzsatz 4.1.7 auch auf den

Fall mehrfacher nicht-defektiver Eigenwerte erweitern läßt, zumindest wenn die Normreduzierung für die Außerdiagonalblöcke ohne Eberlein-Schritte durchgeführt wird), wogegen sie für defektive Matrizen (wie bei allen bekannten Jacobi-ähnlichen Prozessen) nur linear ist. Aufgrund des 2x2-Block-Charakters unserer Methoden konvergieren die Größen $S(A^{(N)})$ auch dann noch asymptotisch quadratisch gegen 0, wenn ein doppelter reeller defektiver Eigenwert auftritt, jedoch nur, wenn die Matrix keine weiteren mehrfachen Eigenwerte besitzt (vgl. (6b) und (7c)).

Im internen Vergleich der drei Algorithmen hat sich Alg2 gegenüber Alg1 nicht so gut bewährt. Alg2 ist nur bei extrem starker Abweichung von der Normalität schneller als Alg1, selbst wenn die Zyklenzahl in vielen Fällen günstiger ist. Somit ist in der Praxis Alg1 vorzuziehen. Als schnellste Methode erwies sich jedoch Alg3. Da sie zuverlässig arbeitet, auch wenn ihre normreduzierenden Eigenschaften nicht vollständig bewiesen sind, werden wir uns in den Untersuchungen in den folgenden Paragraphen stets auf diese Methode beziehen.

Wir beschließen diesen Abschnitt mit zwei Resultaten negativen Charakters. Zum einen haben wir Untersuchungen mit modifizierten Verfahren, die im Schritt 2 in 4.1.1 zunächst die Diagonalisierung und dann die Normreduzierung vornehmen, durchgeführt. Nach dem diagonalisierenden Schritt gilt entweder $\varepsilon_+ = \varepsilon_- = 0$ (Jacobi-Transformation) oder $\beta_1 = \beta_2 = 0$ (Paardekooper-Transformation), so daß sich die Berechnung der normreduzierenden Funktion $F(x_1,x_2)$ vereinfacht. Jedoch erweist sich diese Strategie als ineffizient. Die Verfahren werden im Schnitt um 2 Zyklen langsamer.

Des weiteren haben wir Tests mit einer anderen Pivotstrategie angestellt. Da viele der von uns betrachteten Matrizen eine Bandstruktur aufweisen, liegt es auf der Hand, bei diesen Matrizen die Pivotindizes in zyklischer Diagonal-Reihenfolge laufen zu lassen in der Hoffnung, daß die Bandgestalt nicht sofort vollständig zerstört wird. Jedoch bringt auch diese Strategie keine Vorteile gegenüber der üblichen zeilenweisen Pivotreihenfolge, im Durchschnitt werden genauso viele Transformationen benötigt. Zudem scheinen die Konvergenzbeweise bei dieser Strategie ungleich schwerer zu sein.

5.2 NUMERISCHER VERGLEICH MIT STANDARD-ALGORITHMEN ZUR EIGENWERTBERECHNUNG

In diesem Paragraphen vergleichen wir unsere Methode Alg3 mit den in der numerischen Praxis gebräuchlichsten Verfahren zur Lösung des vollständigen J-symmetrischen bzw. allgemeinen algebraischen Eigenwertproblems, mit dem Verfahren von Veselić ([53]), dem komplexen Eberlein-Verfahren ([8]) und dem QR-Verfahren von Francis ([13]).

Die Bezeichnungen sind dabei

Alg3 wie in § 5.1 definiert,

Ves das Veselić-Verfahren (optimale Variante) als ALGOL-Prozedur, wie in [53] beschrieben, inklusive Eigenvektorberechnung,

Eber das komplexe Eberlein-Verfahren, als ALGOL-Prozedur "comeig" in [10] beschrieben, inklusive Eigenvektorberechnung,

QR das QR-Verfahren, zusammengesetzt aus den ALGOL-Prozeduren "orthes" ([31]), "ortrans" ([40]) und "hqr2" ([40]), inklusive Eigenvektorberechnung.

Zum Vergleich der Verfahren wählen wir exakt die Testmatrizen aus [53], d.h. mit den Bezeichnungen aus § 5.1 die Matrizen (1),(2),(3a),(3b),(4a),(4b), hier aber jeweils 20 statt 10 Matrizen, und (5d),(5e),(5f),(6a),(6b),(6c),(6d),(6e),(10).

In den nachfolgenden Tabellen haben wir die Resultate für Ves, Eber und QR aus [53] übernommen. *Audiag* bezeichnet dabei für Alg3, Ves und Eber das betragsgrößte Außer(block)diagonalelement, und *CPU* und *Zyklen* sind wie in § 5.1 definiert.

	CPU				Zyklen			Audiag		
	Alg3	Ves	Eber	QR	Alg3	Ves	Eber	Alg3	Ves	Eber
(1)	5.08	6.67	29.66	2.36	7.90	7.25	9.85			
(2)	42.49	49.35	256.65	16.20	9.35	8.65	11.25			
(3a)	2.37	2.92	23.23	2.20	3.70	3.55	7.60	Für alle Matrizen höchstens 10^{-10}		
(3b)	1.38	1.79	21.53	2.17	2.05	2.15	7.10			
(4a)	20.08	21.64	196.31	14.58	4.75	4.30	8.70			
(4b)	10.85	11.97	178.00	14.05	2.30	2.45	7.95			

	CPU				Zyklen			Audiag		
	Alg3	Ves	Eber	QR	Alg3	Ves	Eber	Alg3	Ves	Eber
(5d)	2.12	2.25	16.51	2.17	5	5	6	10^{-11}	10^{-12}	10^{-15}
(5e)	3.76	4.97	24.82	2.54	6	6	8	10^{-14}	10^{-13}	10^{-14}
(5f)	6.44	9.40	40.21	3.40	10	11	13	10^{-12}	10^{-14}	10^{-13}
(6a)	3.37	4.52	18.48	2.66	5	5	7	10^{-12}	10^{-12}	10^{-14}
(6b)	3.66	4.93	48.48	2.50	6	6	35*)	10^{-11}	10^{-12}	10^{-8}
(6c)	4.01	5.20	26.40	2.61	6	6	9	10^{-10}	10^{-12}	10^{-12}
(6d)	3.96	5.37	14.27	2.38	6	6	6	10^{-12}	10^{-12}	10^{-12}
(6e)	32.01	37.29	202.44	13.89	8	7	11	10^{-11}	10^{-12}	10^{-14}
(10)	0.55	0.66	0.82	0.15	36	32	35*)	10^{-11}	10^{-7}	10^{-9}

*) Die maximale Zyklenzahl für Eber ist 35.

Hinsichtlich der Genauigkeit erweisen sich die vier Methoden als adäquat. In allen nicht-defektiven Beispielen werden die Eigenwerte bis auf die letzten 2 - 3 Stellen exakt bestimmt, und die defektiven Eigenwerte in (6b) und (10) berechnen sich auf 7 - 8 Stellen genau.

Fazit der Resultate für (1) - (4):
Die tabellierten Werte für die Beispiele (1) - (4) sind wieder Mittelwerte. Unser Verfahren Alg3 ist hier gegenüber der Methode von Veselić zwischen 7% und 24% schneller, obwohl die Zyklenzahl teilweise leicht zugenommen hat. Im Vergleich zur Eberlein-Methode ist Alg3 bei den "vollen" Matrizen ungefähr um einen Faktor 6 schneller, bei den schwach blockdiagonaldominanten um ca. einen Faktor 10 und bei den stark blockdiagonaldominanten um ca. einen Faktor 16. Die Tatsache, daß die Eberlein-Methode bei den Matrizen (3) und (4) im Vergleich zu (1) und (2) nur unwesentlich schneller wird, liegt darin begründet, daß diese Methode echte Diagonaldominanz zur schnelleren Konvergenz benötigt und kaum Nutzen aus einer Blockdiagonaldominanz ziehen kann. Jedoch treten gerade diese Matrizen, wie schon in § 5.1 erwähnt, in Unterproblemen bei Verfahren der Simultanen Iteration auf (vgl. § 5.3). Gegenüber dem QR-Verfahren ist Alg3 bei den "vollen" Problemen (1) und (2) im Mittel um einen Faktor 2.2 bzw. 2.6 langsamer. Jedoch kann der QR-Algorithmus nahezu gar nicht von der Blockdiagonaldominanz der Matrizen profitieren, so daß Alg3 bei 10^{-2}-Blockdiagonaldominanz fast gleich schnell wird und bei 10^{-4}-Blockdiagonaldominanz die CPU-Zeit des QR-Verfahrens deutlich unterbietet. Interessant ist hier, daß Alg3 bei größerer Dimension eine stärkere Blockdiagonaldominanz benötigt, um das QR-Verfahren zu schlagen.

Fazit der Resultate für (5),(6) und (10):
Bei diesen Beispielen beträgt die Verbesserung der Rechenzeit von Alg3 gegenüber der Methode von Veselić zwischen 14% und 32% mit Ausnahme der normalen Matrix (5d), wo der Zeitgewinn nur ca. 6% ausmacht. Der Vergleich von Alg3 mit dem komplexen Eberlein-Verfahren bringt hier wieder, abgesehen von dem kleinen 4x4-Beispiel (10), einen Vorteil zugunsten Alg3 um einen Faktor zwischen 3.5 und 13.5. Dabei ist die Differenz in Beispiel (6b) so deutlich, weil die Eberlein-Methode den defektiven 2x2-Block der Matrix zu diagonalisieren versucht. Das QR-Verfahren erhält bei den Matrizen (6) und (10) einen zusätzlichen Vorteil, da hier aufgrund der Tridiagonalität die Prozeduren "orthes" und "ortrans" nicht aufgerufen werden müssen. Unsere Methode Alg3 ist im Vergleich zum QR-Verfahren bei der normalen Matrix (5d) ein wenig schneller, ansonsten mit Ausnahme von (10) um einen Faktor zwischen 1.2 und 1.9 langsamer, bei (10) beträgt der Faktor ungefähr 2.3 .

Als Gesamtfazit können wir festhalten, daß unsere Methode Alg3 gegenüber dem gleichartigen Verfahren von Veselić Verbesserungen hinsichtlich der Rechenzeit zwischen 7% und 32% erzielt, obwohl sich die Zyklenzahl in manchen Fällen leicht vergrößert hat. Der Vergleich gegenüber der komplexen Eberlein-Methode sieht unsere Methode um einen Faktor 3.5 bis 16 schneller, wobei der Unterschied bei blockdiagonaldominanten Matrizen besonders drastisch wird. Ein weiterer Nachteil der Eberlein-Methode ist der, daß sie die J-Symmetrie der Matrizen zerstört und zudem mit komplexer Arithmetik arbeiten muß. Somit hat sie einen ungefähr 4 mal größeren Speicherplatzbedarf als Alg3. Im Vergleich mit dem QR-Algorithmus ist unser Verfahren bei den physikalischen Problemen um einen Faktor zwischen 1.2 und 1.9 und bei den "vollen" Zufallsmatrizen um einen Faktor 2.2 bzw. 2.6 langsamer. Jedoch gleicht sich die CPU-Zeit bei schwacher Blockdiagonaldominanz der Matrizen an und sieht unser Verfahren um so deutlicher im Vorteil, wie die Blockdiagonaldominanz stärker wird (vgl. dazu auch § 5.4). Somit kann unser Verfahren zumindest dann mit dem QR-Algorithmus konkurrieren, wenn es darum geht, blockdiagonaldominante Probleme (wie sie in der Praxis häufig auftreten) zu lösen (vgl. dazu auch [16,S.300]). Dabei ist zu beachten, daß Alg3 gegenüber dem QR-Verfahren nur in etwa die Hälfte des Speicherplatzes benötigt, da das QR-Verfahren ebenfalls die Symmetrie des Problems zerstört. In [53] wird

auf einen weiteren, nicht unwesentlichen Aspekt im Vergleich der beiden Methoden hingewiesen. So ist es bei der Behandlung sehr großer Matrizen oft zwingend erforderlich, eine geeignete Blockteilung vorzunehmen, um dann nur gewisse Blöcke im Zentralspeicher des Rechners zu bearbeiten. Hierbei hat sich jedoch gezeigt, daß Jacobi-ähnliche Prozesse gegenüber QR-ähnlichen Verfahren weitaus weniger Zeit für die entsprechenden Ein- und Ausgabetransfers benötigen (vgl.[4]).

5.3 KOMBINATION DER JACOBI-ÄHNLICHEN BLOCKVERFAHREN MIT DEM VERFAHREN DER SIMULTANEN ITERATION

In diesem Paragraphen beschreiben wir eine wichtige Anwendung unserer Verfahren bei der Lösung eines Unterproblems innerhalb der sogenannten *Simultanen Iteration*. Dazu betrachten wir das verallgemeinerte reelle lineare nxn-Eigenwertproblem

$$Ax = \lambda Bx \ . \tag{5.3.1}$$

In der Praxis ist es häufig nicht erforderlich, alle n Eigenwerte des Problems zu berechnen, sondern nur eine gewisse vorgegebene Teilmenge. Sollen beispielsweise nur die r betragsmäßig größten bestimmt werden, so wird durch die Iterationsvorschrift

$$BX_k = AX_{k-1} \quad , \ k = 1,2,3,... \tag{5.3.2}$$

eine Folge von reellen nxr-Matrizen $X_0, X_1, X_2, ...$ generiert, deren Spaltenvektoren unter gewissen Bedingungen gegen den Vektorraum der zu diesen Eigenwerten gehörigen Eigenvektoren konvergieren, d.h. im Konvergenzfall spannen die r Iterationsvektoren den Eigenvektorraum über $\mathbb{C}$ auf, womit dann natürlich auch die gesuchten Eigenwerte bestimmt sind.

Die Grundidee zu dieser Iteration (auch *Subspace-Iteration* oder *Treppeniteration* genannt) stammt von Bauer ([2]). Natürlich muß die grobe Iterationsvorschrift (5.3.2) in der praktischen Anwendung weiter verfeinert werden, um eine stabile Konvergenz der Matrizen X_k zu gewährleisten. So muß unter anderem dafür gesorgt werden, daß die Spaltenvektoren von X_k linear unabhängig bleiben. Dazu wird in geeignet gewählten Abständen das rxr-Eigenwertproblem

$$A_k y = \mu B_k y \tag{5.3.3}$$

mit

$$A_k = X_k^T A X_k \ , \quad B_k = X_k^T B X_k$$

vollständig gelöst, und die Matrix X_k wird durch Rechtsmultiplikation

mit der erhaltenen Eigenvektormatrix Y "verbessert".

Wenn nun das Problem (5.3.1) aus einem quadratischen Eigenwertproblem (1.3.2) mit symmetrischen positiv definiten Matrizen M,D und K resultiert, d.h. wenn A symmetrisch und B symmetrisch, aber indefinit mit einer Signatur $(\frac{n}{2},\frac{n}{2})$ ist (vgl.(1.3.3)), so überträgt sich diese Struktur unter gewissen Voraussetzungen im Laufe der Iteration auf das kleinere Problem (5.3.3) (vgl.[26]). Aus § 1.3 ist aber bekannt, daß ein solches Problem äquivalent zu einem J-symmetrischen gewöhnlichen Eigenwertproblem ist. Somit können wir dieses Unterproblem mit unseren Jacobi-ähnlichen Blockverfahren lösen, wobei diese entscheidend davon profitieren, daß die J-symmetrischen Matrizen mit fortschreitender (simultaner) Iteration immer stärker blockdiagonaldominant werden.

In [26] haben wir in Weiterentwicklung der Methode von Borri und Mantegazza ([3]) ein Verfahren der Simultanen Iteration für das oben beschriebene Problem formuliert. Dieses Verfahren haben wir hier mit unserer Methode Alg3 kombiniert. Hinsichtlich der technischen Details sei der interessierte Leser auf [26] verwiesen. Wir zeigen im folgenden einige typische Resultate auf.

Dazu betrachten wir die quadratischen Eigenwertprobleme der Dimension 45 bzw. 90, die wir in § 5.1, Beispiel (11a) und (11b) beschrieben haben. Die Bezeichnungen bedeuten

Sim Anzahl der Iterationsschritte für die Simultane Iteration,

Jac Anzahl der Zyklen für Alg3.

CPU ist wie in § 5.1 erklärt, und r,k sind wie in diesem Paragraphen definiert.

Ergebnisse Beispiel (11a):

r	CPU	Sim	Jac
12	135.38	36	20
18	194.89	26	16
28	587.55	41	15

r = 12:

k	11	16	21	26	31	36
Jac	6	3	4	3	2	2

r = 18:

k	11	16	21	26
Jac	6	4	3	3

r = 28:

k	26	31	36	41
Jac	6	3	3	3

Ergebnisse Beispiel (11b):

r	CPU	Sim	Jac
12	140.24	36	21
18	267.84	36	19
28	877.03	66	38

r = 12:

k	16	21	26	31	36
Jac	6	3	5	4	3

r = 18:

k	11	16	21	26	31	36
Jac	7	4	3	2	2	1

r = 28:

k	16	21	26	31	36	41	46	51	56	61	66
Jac	7	5	4	5	4	3	2	2	2	2	2

Beide Beispiele demonstrieren deutlich, wie sich im Laufe der Simultanen Iteration die Zyklenzahl von Alg3 verringert. Hinsichtlich der absoluten CPU-Zeit sehen wir hier ein Kriterium von Rutishauser ([43]) bestätigt, welches besagt, daß das Verfahren der Simultanen Iteration nur dann angewandt werden soll, wenn $r < \frac{n}{5}$ gilt (vgl. dazu die Ergebnisse aus § 5.1).

5.4 NUMERISCHE ANWENDUNGEN IN PHYSIKALISCHEN PROBLEMSTELLUNGEN (PARAMETERANALYSE)

In der physikalischen Praxis geschieht es häufig, daß das mathematische Modell, welches das gegebene physikalische Problem beschreibt, parameterabhängig formuliert wird. So ist es möglich, durch Änderungen der Parameter verschiedene physikalische Bedingungen zu simulieren. Natürlich hofft der Mathematiker, daß sich durch eine kleine Parameteränderung die mathematische Lösung nicht wesentlich ändert und daß sich durch eine einmal gewonnene Lösung die Berechnung der "benachbarten" Probleme stark vereinfacht. Wir beschreiben im folgenden eine typische Problemstellung dieser Art und demonstrieren die Leistungsfähigkeit unserer Verfahren bei ihrer numerischen Behandlung.

Dazu betrachten wir das reelle quadratische Eigenwertproblem

$$(\lambda^2 M + \lambda D(\tau) + K)x = 0 \qquad (5.4.1)$$

mit

$$M = \mathrm{diag}(m_1,\ldots,m_m) \ , \ D(\tau) = \tau\,\mathrm{diag}(d_1,\ldots,d_m)$$

und einer symmetrischen Bandmatrix K (vgl.(1.3.2)). Es beschreibt die gedämpften Schwingungen eines linearen dynamischen Systems von m Massen, wobei die Dämpfung (linear) variabel gehalten wird. Wenn wir dieses Problem nach der Strategie (1.3.8),(1.3.9) mit anschließender Zeilen- und Spaltenpermutation in ein gewöhnliches J-symmetrisches Eigenwertproblem doppelter Dimension transformieren, so erhalten wir eine Matrix der Form

$$A(\tau) = A_0 + \tau\,\mathrm{diag}(0,-\frac{d_1}{m_1},\ldots,0,-\frac{d_m}{m_m}) \ , \qquad (5.4.2)$$

d.h. eine Änderung des Parameters τ bewirkt hier lediglich eine Änderung der Hauptdiagonalen von $A(\tau)$. Nun wird die Dämpfung variiert, indem τ sukzessive die vorgegebenen Werte $0 = \tau_0,\tau_1,\ldots,\tau_s$ annimmt. Somit müssen s+1 J-symmetrische Eigenwertprobleme für die Matrizen

$$A_k = A(\tau_k) \quad , \ k = 0(1)s$$

gelöst werden. Für die Praxis ist dann interessant, wie sich die Eigenwerte der A_k und damit die Frequenzen der Schwingungen für die verschiedenen Werte von τ verhalten. Man startet mit dem ungedämpften Problem ($\tau = 0$) und verfolgt die "Wanderung" der Eigenwerte mit langsam zunehmender Dämpfung.

Natürlich kann man bei dieser sogenannten *Parameteranalyse* jedes einzelne Problem unabhängig von den anderen lösen, jedoch arbeitet man sehr viel rationeller, wenn man statt der Matrizen A_k die "vorbehandelten" Matrizen

$$\overset{\circ}{A}_k = R_{k-1}^{-1} A_k R_{k-1} \quad , \; k = 1(1)s \tag{5.4.3}$$

betrachtet, wobei R_{k-1} die J-orthogonale Matrix bezeichnet, die A_{k-1} per Ähnlichkeitstransformation auf Blockdiagonalgestalt bringt. Denn aus Stetigkeitsgründen wird $\overset{\circ}{A}_k$ nicht stark von der Blockdiagonalgestalt abweichen, wenn τ_k nahe bei τ_{k-1} liegt. Ist dann $\overset{\circ}{R}_k$ die resultierende J-orthogonale Transformationsmatrix, die $\overset{\circ}{A}_k$ blockdiagonalisiert, so erhalten wir R_k als

$$R_k = R_{k-1} \overset{\circ}{R}_k \; .$$

Es gilt nun, die Berechnung von $\overset{\circ}{A}_k$ mit geringst möglichem Aufwand auszuführen. Aus (5.4.2) folgt

$$A_k = A_{k-1} + (\tau_k - \tau_{k-1}) \operatorname{diag}(0, -\frac{d_1}{m_1}, \ldots, 0, -\frac{d_m}{m_m}), \; k = 1(1)s$$

und damit aus (5.4.3)

$$\overset{\circ}{A}_k = R_{k-1}^{-1} A_{k-1} R_{k-1} + (\tau_k - \tau_{k-1}) R_{k-1}^{-1} \operatorname{diag}(0, -\frac{d_1}{m_1}, \ldots, 0, -\frac{d_m}{m_m}) R_{k-1} \; . \tag{5.4.4}$$

In dieser Darstellung ist der erste Summand aus der Blockdiagonalisierung von A_{k-1} bzw. $\overset{\circ}{A}_{k-1}$ bekannt, und der zweite Summand ergibt sich aufgrund der J-Orthogonalität von R_{k-1} nach einigen elementaren Umformungen als

$$(\tau_k - \tau_{k-1}) \sum_{i=1}^{m} \frac{d_i}{m_i} J (e_{2i}^T R_{k-1})^T (e_{2i}^T R_{k-1}) \quad . \tag{5.4.5}$$

Natürlich ist die Matrix (5.4.5) wieder J-symmetrisch, so daß bei der Berechnung von $\overset{\circ}{A}_k$ nur das rechtsobere Dreieck in Betracht gezogen werden muß.

Diese geschickte Realisierung der Parameteranalyse ist aufgrund der Blockdiagonaldominanz der $\overset{\circ}{A}_k$ speziell auf unsere Jacobi-ähnlichen Verfahren abgestimmt. Selbstverständlich ist es auch möglich, eine solche Analyse mit Hilfe anderer Algorithmen durchzuführen. Wir betrachten zum Vergleich das gebräuchlichste Verfahren zur Eigenwertberechnung, das QR-Verfahren ([13]). Bei diesem Verfahren hat $R_{k-1}^{-1}A_{k-1}R_{k-1}$ in (5.4.4) jeweils rechtsobere Quasidreiecksform, und da die Transformationen bei einer stabilen Implementierung ausschließlich orthogonal durchgeführt werden, ist der zweite Summand in (5.4.4) symmetrisch und berechnet sich als

$$-(\tau_k-\tau_{k-1})\sum_{i=1}^{m}\frac{d_i}{m_i}(e_{2i}^T R_{k-1})^T(e_{2i}^T R_{k-1}) \quad . \tag{5.4.6}$$

Jedoch weisen die Matrizen $\overset{\circ}{A}_k$ keine Symmetrieeigenschaften auf, so daß das QR-Verfahren auf den vollen nxn-Matrizen arbeiten muß. Hier besitzen die $\overset{\circ}{A}_k$ aus Stetigkeitsgründen eine Fast-Quasidreiecksform, wenn die Parameter τ_k und τ_{k-1} nahe beieinander liegen.

Wir haben beide beschriebenen Parameteranalysen jeweils für zwei Probleme der Form (5.4.1) durchgeführt. Die erstgenannte haben wir mit Hilfe unseres Verfahrens Alg3 realisiert, und für die zweite haben wir den QR-Algorithmus aus den ALGOL-Prozeduren "orthes" ([31]), "ortrans" ([40]) und "hqr2" ([40]) kombiniert, dabei jedoch die eigentliche Eigenvektorberechnung in "hqr2" unterdrückt, so daß nur die Transformationsmatrizen akkumuliert werden. Als Beispiele haben wir wieder die in § 5.1, (11a) und (11b) beschriebenen quadratischen Eigenwertprobleme der Dimension 45 bzw. 90 gewählt. Diese sind von der Form (5.4.1), wenn man die Dämpfungsmatrix D durch den Faktor τ parametrisiert, wobei für beide nur jeweils 12 der 45 bzw. 90 Dämpfungskonstanten d_i ungleich 0 sind, so daß sich die Berechnung der Matrizen (5.4.5) und (5.4.6) weiter vereinfacht.

Die Parameter τ_k seien äquidistant aus dem Intervall [0,2] gewählt, d.h. es gelte

$$\tau_k = \frac{2k}{s} \quad , \; k = 0(1)s \quad .$$

Wir untersuchen für beide Probleme jeweils die Fälle $s = 10$, $s = 20$ und $s = 40$.

Resultate:
Für $\tau_0 = 0$ braucht Alg3 je 8 Zyklen für (11a) und (11b) bei einer CPU-Zeit von 122.55 bzw. 801.85 Sekunden, während das QR-Verfahren 110.98 Sekunden für (11a) und 847.35 Sekunden für (11b) benötigt. Man beachte hierbei, daß die Matrix A_0 rein schiefsymmetrisch ist, so daß das Verfahren Alg3 ausschließlich mit Paardekooper-Transformationen arbeitet.

Die Generierung der Matrizen $\mathring{A}_k$ dauert dann (für alle Werte von s) für Alg3 im Durchschnitt 4.60 Sekunden ((11a)) bzw. 18.13 Sekunden ((11b)) und für das QR-Verfahren 7.48 Sekunden bzw. 31.59 Sekunden.

Für die Blockdiagonalisierung der $\mathring{A}_k$ bzw. für deren Transformation auf Quasidreiecksform ergeben sich die folgenden durchschnittlichen CPU-Zeiten und Zyklenzahlen:

(11a):

s	CPU Alg3	CPU QR	Zyklen (Alg3)
10	55.02	94.91	3.50
20	50.24	94.60	3.15
40	44.36	94.61	2.55

(11b):

s	CPU Alg3	CPU QR	Zyklen (Alg3)
10	502.19	740.31	3.60
20	455.69	739.85	3.40
40	411.66	739.73	3.00

Schließlich erhalten wir als totale CPU-Zeiten für die vollständig durchgeführten Parameteranalysen (Generierung und Behandlung aller Matrizen $A_0, \mathring{A}_1, \ldots, \mathring{A}_s$):

(11a):

s	Alg3	QR
10	718.83	1134.83
20	1219.07	2152.69
40	2080.73	4195.07

(11b):

s	Alg3	QR
10	6000.31	8566.29
20	10278.89	16276.25
40	17996.70	31700.18

Es sei hier am Rande erwähnt, daß in der Praxis die Abbruchkriterien der beiden Methoden oftmals abgeschwächt werden können. So ist es beispielsweise für unser Verfahren Alg3 nicht unbedingt erforderlich, die Matrizen $\mathring{A}_k$ auf eine 10^{-10}-Blockdiagonalgestalt zu bringen. Es langt vielfach schon eine 10^{-5}-Blockdiagonalität, wodurch der Aufwand um ca. 1 Zyklus verringert wird.

Als Fazit können wir festhalten, daß unsere Methode Alg3 wesentlich stärker von der "Vorbehandlung" profitiert als das QR-Verfahren. Während das QR-Verfahren für die Fast-Quasidreiecksmatrizen $\mathring{A}_k$ nur zwischen 12% und 15% weniger CPU-Zeit als für die "volle" Matrix A_0 benötigt (und zwar unabhängig von der Schrittweite der τ_k), wird Alg3 auf den fast-blockdiagonalen Matrizen $\mathring{A}_k$ im Vergleich zur ersten Matrix A_0 wesentlich schneller, was sich in der Zyklenzahl und in der Rechenzeit ausdrückt, und dieser Effekt verstärkt sich noch mit einer kleineren Schrittweite für die Parameter τ_k. Hinsichtlich der totalen CPU-Zeit für die komplette Parameteranalyse erweist sich unsere Methode Alg3 gegenüber dem QR-Verfahren um einen (gerundeten) Faktor 1.4 bzw. 1.6 schneller, wenn die Analyse mit einem groben Raster ($s = 10$) durchgeführt wird, und bei feineren Rastern beträgt der Faktor 1.6 bzw. 1.8 für $s = 20$ und 1.8 bzw. 2.0 für $s = 40$. Diese Ergebnisse unterstreichen noch einmal deutlich die Effektivität unserer Methoden auf blockdiagonaldominanten Problemen.

In beiden beschriebenen Parameteranalysen werden die Transformationsmatrizen R_k mitgeführt. Damit können die Eigenvektoren der Matrizen A_k jederzeit ohne großen Aufwand berechnet werden. Dies wiederum ermöglicht es, in jedem beliebigen Schritt Rückschlüsse auf die vollständige Lösung des linearen Systems ziehen zu können. Des weiteren ist es möglich, mit Hilfe der Kondition der R_k die Schrittweite der Parameter τ_k zu steuern, wenn diese nicht äquidistant sondern variabel gewählt werden. Falls jedoch keine Notwendigkeit zur Akkumulation der Transformationsmatrizen besteht, d.h. wenn beispielsweise bei fester Schrittweite ausschließlich die Eigenwerte der A_k von Interesse sind, so läßt sich die QR-Parameteranalyse effektiver durchführen. In diesem Fall wird das QR-Verfahren (ohne Eigenvektorteil) direkt auf die ursprünglichen Matrizen A_k angewandt. Wir haben dazu den QR-Algorithmus aus den ALGOL-Prozeduren "orthes" ([31]) und "hqr" ([32]) zusammengestellt und wiederum auf die be-

schriebenen Beispiele angewandt. Die folgenden Daten sind die absoluten CPU-Zeiten für die Behandlung aller Matrizen $A_0, A_1, \ldots, A_s$. Zum Vergleich haben wir noch einmal unsere obigen Ergebnisse für die vollständige Alg3-Parameteranalyse mit aufgeführt:

(11a):

s	Alg3	QR
10	718.83	536.47
20	1219.07	1015.02
40	2080.73	1972.12

(11b):

s	Alg3	QR
10	6000.31	4034.82
20	10278.89	7628.84
40	17996.70	14816.89

Hier ergeben sich also Gesamtrechenzeiten, die bis zu einem Faktor 1.5 zugunsten des QR-Verfahrens ausfallen. Jedoch sei noch einmal betont, daß diese QR-Parameteranalyse keine Informationen über die Eigenvektoren der A_k liefert. Falls diese für gewisse Parameterwerte benötigt werden, so müssen sie in einer zusätzlichen Rechnung bestimmt werden. Somit ist die QR-Parameteranalyse im Vergleich zur Alg3-Analyse nur dann schneller, wenn die Eigenvektorberechnung nicht zu oft erforderlich wird.

LITERATURVERZEICHNIS

[1] Armijo,L.: Minimization of functions having Lipschitz-continuous first partial derivatives, Pacif. J. Math. 16, 1 - 3 (1966).

[2] Bauer,F.L.: Das Verfahren der Treppeniteration und verwandte Verfahren zur Lösung algebraischer Eigenwertprobleme, Z. Angew. Math. Physik 8, 214 - 235 (1957).

[3] Borri,M.,Mantegazza,P.: Efficient solution of quadratic eigenproblems arising in dynamic analysis of structures, Comp. Meths. Appl. Mech. Engrg. 12, 19 - 31 (1977).

[4] Braun,K.A.,Johnson,Th.L.: Hypermatrix generalization of the Jacobi and Eberlein method for computing eigenvalues and eigenvectors of hermitian and non-hermitian matrices, Comp. Meths. Appl. Mech. Engrg. 4, 1 - 18 (1974).

[5] Brebner,M.A.,Grad,J.: Eigenvalues of $Ax = \lambda Bx$ for real symmetric matrices A and B computed by reduction to a pseudosymmetric form and the HR process, Lin. Alg. and its Appl. 43, 99 - 118 (1982).

[6] Bunse-Gerstner,A.: Der HR-Algorithmus zur numerischen Bestimmung der Eigenwerte einer Matrix, Dissertation, Univ. Bielefeld, 1978.

[7] Dunford,N.,Schwartz,J.T.: Linear operators Part I, Interscience Publishers, New York - London - Sydney, 1957.

[8] Eberlein,P.J.: A Jacobi-like method for the automatic computation of eigenvalues and eigenvectors of an arbitrary matrix, SIAM J. 10, 74 - 88 (1962).

[9] Eberlein,P.J.,Boothroyd,J.: Solution to the eigenproblem by a norm-reducing Jacobi-type method, Numer. Math. 11, 1 - 12 (1968).

[10] Eberlein,P.J.: Solution to the complex eigenproblem by a norm-reducing Jacobi-type method, Numer. Math. 14, 232 - 245 (1970).

[11] Elkin,R.: Convergence theorems for Gauss-Seidel and other minimization algorithms, Dissertation, Univ. of Maryland, College Park, Maryland, 1968.

[12] Forsythe,G.E.,Henrici,P.: The cyclic Jacobi method for computing the principal values of a complex matrix, Trans. Amer. Math. Soc. 94, 1 - 23 (1960).

[13] Francis,J.G.F.: The QR transformation, Parts I and II, Comp. J. 4, 265 - 271, 332 - 345 (1961/62).

[14] Gohberg,I.,Lancaster,P.,Rodman,L.: Matrices and indefinite scalar products, Birkhäuser-Verlag, Basel - Boston - Stuttgart, 1983.

[15] Goldstein,A.: Constructive real analysis, Harper & Row, New York, 1967.

[16] Golub,G.H.,Van Loan,C.F.: Matrix computations, North Oxford Academic Publishing, Oxford, 1983.

[17] Gregory,R.T.: Computing eigenvalues and eigenvectors of a symmetric matrix on the ILLIAC, Math. Tab. and Other Aids to Comp. 7, 215 - 220 (1953).

[18] Hari,V.: A Jacobi-like eigenvalue algorithm for general real matrices, Glasnik Mat. 11 (31), 367 - 378 (1976).

[19] Hari,V.: On the global convergence of the Eberlein method for real matrices, Numer. Math. 39, 361 - 369 (1982).

[20] Hari,V.: On the quadratic convergence of the Paardekooper method I, Glasnik Mat. 17 (37), 183 - 195 (1982).

[21] Hari,V.: On the quadratic convergence of the Paardekooper method II, Preprint, erscheint demnächst in Glasnik Matematički.

[22] Hari,V.: On the quadratic convergence of the Paardekooper method III, Preprint, erscheint demnächst in Glasnik Matematički.

[23] Hari,V.: On convergence of certain cyclic Jacobi-like norm-reducing processes for real matrices, unveröffentlichtes Manuskript.

[24] Henrici,P.: On the speed of convergence of cyclic and quasicyclic Jacobi methods for computing eigenvalues of hermitian matrices, SIAM J. 6, 144 - 162 (1958).

[25] Henrici,P.: Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices, Numer. Math. 4, 24 - 41 (1962).

[26] Hoppe,P.: Numerische Behandlung quadratischer Eigenwertprobleme der Form $(\lambda^2 M + \lambda D + K)x = 0$, Diplomarbeit, Univ. Dortmund, 1979.

[27] Hoppe,P.,Veselić,K.: A Jacobi-like method for the symmetric indefinite eigenproblem $Sx = \lambda Tx$ with complex eigenvalues, Glasnik Mat. 19 (39), 383 - 390 (1983).

[28] Jacobi,C.G.J.: Über ein leichtes Verfahren die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen, J. Reine Angew. Math. 30, 51 - 94 (1846).

[29] Lancaster,P.: Lambda-matrices and vibrating systems, Pergamon Press, Oxford, 1966.

[30] Mal'cev,A.I.: Foundations of linear algebra, Freeman and Company, San Francisco - London, 1963.

[31] Martin,R.S.,Wilkinson,J.H.: Similarity reduction of a general matrix to Hessenberg form, Numer. Math. 12, 349 - 368 (1968).

[32] Martin,R.S.,Peters,G.,Wilkinson,J.H.: The QR algorithm for real Hessenberg matrices, Numer. Math. 14, 219 - 231 (1970).

[33] Meyer,D.,Veselić,K.: On some new inclusion theorems for the eigenvalues of partitioned matrices, Numer. Math. 34, 431 - 437 (1980).

[34] Mirsky,L.: On the minimization of matrix norms, Amer. Math. Monthly 65, 106 - 107 (1958).

[35] Müller,P.C.: Stabilität und Matrizen, Springer-Verlag, Berlin - Heidelberg - New York, 1977.

[36] Murnaghan,F.D.,Wintner,A.: A canonical form for real matrices under orthogonal transformations, Proc. Nat. Acad. Sci. 17, 417 - 420 (1931).

[37] Ortega,J.M.,Rheinboldt,W.C.: Iterative solution of nonlinear equations in several variables, Academic Press, New York - San Francisco - London, 1970.

[38] Paardekooper,M.H.C.: An eigenvalue algorithm based on norm-reducing transformation, Dissertation, Techn. Hoogesch. Eindhoven, 1969.

[39] Paardekooper,M.H.C.: An eigenvalue algorithm for skew-symmetric matrices, Numer. Math. 17, 189 - 202 (1971).

[40] Peters,G.,Wilkinson,J.H.: Eigenvectors of real and complex matrices by LR and QR triangularizations, Numer. Math. 16, 181 - 204 (1970).

[41] Ruhe,A.: On the quadratic convergence of a generalization of the Jacobi method to arbitrary matrices, BIT 8, 210 - 231 (1968).

[42] Ruhe,A.: The norm of a matrix after a similarity transformation, BIT 9, 53 - 58 (1969).

[43] Rutishauser,H.: Computational aspects of F.L. Bauer's simultaneous iteration method, Numer. Math. 13, 4 - 13 (1969).

[44] Sacks-Davis,R.: Solution to the eigenproblem by Jacobi-type methods, Dissertation, School of Math. Sciences, Univ. of Melbourne, 1974.

[45] Sacks-Davis,R.: A real norm-reducing eigenvalue algorithm, Austral. Comp. J. 7, 65 - 69 (1975).

[46] Schönhage,A.: Zur Konvergenz des Jacobi-Verfahrens, Numer. Math. 3, 374 - 380 (1961).

[47] Schönhage,A.: Zur quadratischen Konvergenz des Jacobi-Verfahrens, Numer. Math. 6, 410 - 412 (1964).

[48] Veselić,K.: A convergent Jacobi method for solving the eigenproblem of arbitrary real matrices, Numer. Math. 25, 179 - 184 (1976).

[49] Veselić,K.: Some convergent Jacobi-like procedures for diagonalising J-symmetric matrices, Numer. Math. 27, 67 - 75 (1976).

[50] Veselić,K.: On a class of Jacobi-like procedures for diagonalising arbitrary real matrices, Numer. Math. 33, 157 - 172 (1979).

[51] Veselić,K.: On a new class of elementary matrices, Numer. Math. 33, 173 - 180 (1979).

[52] Veselić,K.,Wenzel,H.J.: A quadratically convergent Jacobi-like method for real matrices with complex eigenvalues, Numer. Math. 33, 425 - 435 (1979).

[53] Veselić,K.: A global Jacobi method for a symmetric indefinite problem $Sx = \lambda Tx$, Comp. Meths. Appl. Mech. Engrg. 38, 273 - 290 (1983).

[54] Voevodin,V.V.: Čislennye metody lineinoj algebry, Nauka, Moskau, 1966.

[55] Waller,H.,Krings,W.: Matrizenmethoden in der Maschinen- und Bauwerksdynamik, B.I.-Wissenschaftsverlag, Mannheim - Wien - Zürich, 1975.

[56] Wenzel,H.J.: Über quadratisch konvergente Jacobi-ähnliche Block-

verfahren für beliebige reelle Matrizen mit komplexen Eigenwerten, Dissertation, Fernuniv. Hagen, 1983.

[57] Werner,H.: Praktische Mathematik I, Springer-Verlag, Berlin - Heidelberg - New York, 1970.

[58] Wilkinson,J.H.: Note on the quadratic convergence of the cyclic Jacobi process, Numer. Math. 4, 296 - 300 (1962).

[59] Wilkinson,J.H.: The algebraic eigenvalue problem, Oxford University Press, Oxford, 1965.

[60] Wilkinson,J.H.: Modern error analysis, SIAM Review 13, 548 - 568 (1971).

[61] Wilkinson,J.H.,Reinsch,C.: Handbook of automatic computation II, Linear algebra, Springer-Verlag, Berlin - Heidelberg - New York, 1971.

[62] Zurmühl,R.: Matrizen und ihre technischen Anwendungen, Springer-Verlag, Berlin - Göttingen - Heidelberg, 1964.

ANHANG

$$A(\omega) = \left(\begin{array}{cccc|cc} 1 & 0 & 0 & 0 & \omega & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -\omega & 0 \\ \hline \omega & 0 & 0 & \omega & 1 & 1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{array}\right), \quad \omega \in \mathbb{R}, \tag{A1}$$

$\lambda_{1,2} = 1,\ \lambda_{3,4} = 2,\ \lambda_{5,6} = 1 \pm i.$

$$A = \left(\begin{array}{cccc|cc} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 2 & 0 & -1 & 0 & 0 \\ 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & -1 & -1 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{array}\right), \tag{A2}$$

$\lambda_{1,2} = 1,\ \lambda_{3,4} = 2,\ \lambda_{5,6} = 1 \pm i.$

$$A = \left(\begin{array}{cccc|cc} 1 & 0 & 0 & -1 & -1 & 0 \\ 0 & -1 & 1 & 0 & 0 & 1 \\ 0 & -1 & -3 & 0 & 0 & -1 \\ 1 & 0 & 0 & 3 & 1 & 0 \\ \hline -1 & 0 & 0 & -1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & -1 \end{array}\right), \tag{A3}$$

$\lambda_{1,2} = 2,\ \lambda_{3,4} = -2,\ \lambda_5 = 1,\ \lambda_6 = -1.$

$$A = \left(\begin{array}{cccc|cc} 3 & -3 & -3 & 1 & 1 & -1 \\ 3 & 2 & -2 & 3 & 1 & -2 \\ -3 & 2 & 6 & -3 & -1 & 2 \\ -1 & 3 & 3 & 1 & -1 & 1 \\ \hline 1 & -1 & -1 & 1 & 3 & -1 \\ 1 & -2 & -2 & 1 & 1 & 2 \end{array}\right) , \qquad \text{(A4)}$$

$\lambda_{1,2} = 2,\ \lambda_{3,4} = 4,\ \lambda_{5,6} = \frac{5}{2} \pm \frac{\sqrt{3}}{2} i$.

$$A = \left(\begin{array}{cccc|cc} 3 & -1 & 0 & 1 & 0 & -2 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 1 & 0 & 2 \\ \hline 0 & 0 & 0 & 0 & 2 & -2 \\ 2 & 0 & 0 & 2 & 2 & 2 \end{array}\right) , \qquad \text{(A5)}$$

$\lambda_{1,2} = 2,\ \lambda_{3,4} = 1,\ \lambda_{5,6} = 2 \pm 2i$.

$$A = \left(\begin{array}{cccc|cc} 3 & 2 & 1 & 1 & 3 & 1 \\ -2 & 5 & 2 & -2 & 1 & 2 \\ 1 & -2 & 1 & 1 & -1 & -2 \\ -1 & -2 & -1 & 1 & -3 & -1 \\ \hline 3 & -1 & -1 & 3 & 4 & -1 \\ -1 & 2 & 2 & -1 & 1 & 5 \end{array}\right) , \qquad \text{(A6)}$$

$\lambda_{1,2} = 2,\ \lambda_{3,4} = 3,\ \lambda_{5,6} = \frac{9}{2} \pm \frac{\sqrt{3}}{2} i$.

$$A = \left(\begin{array}{cccc|cc} 2 & 0 & 0 & -1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 4 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \end{array}\right) , \tag{A7}$$

$\lambda_{1,2,3,4} = 1,\ \lambda_5 = 0,\ \lambda_6 = 4.$

$$A = \left(\begin{array}{cccc|cc} 2 & 0 & 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ -2 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{array}\right) , \tag{A8}$$

$\lambda_{1,2,3,4} = 1,\ \lambda_{5,6} = 1 \pm \sqrt{3}\,i\,.$

$$A = \left(\begin{array}{cccc|cc} -3 & 5 & 4 & 0 & 0 & 0 \\ -5 & 3 & 0 & -4 & 0 & 0 \\ 4 & 0 & 3 & 5 & 0 & 0 \\ 0 & -4 & -5 & -3 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{array}\right) , \tag{A9}$$

$\lambda_{1,2,3,4} = 0,\ \lambda_{5,6} = 1 \pm i\,.$

$$A = \left(\begin{array}{cccc|cc} 3 & 0 & 0 & 1 & -1 & 0 \\ 0 & 3 & -1 & 0 & -1 & 0 \\ 0 & 1 & 5 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 & -1 & 0 \\ \hline -1 & 1 & 1 & 1 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 \end{array}\right) , \qquad \text{(A10)}$$

$\lambda_{1,2} = 4,\ \lambda_{3,4} = 1.70284 \pm 1.20563\,i\ ,\ \lambda_5 = 4.59431,\ \lambda_6 = 3.$

$$A = \left(\begin{array}{cccc|cc} 3 & 1 & 1 & -2 & 1 & 0 \\ -1 & 2 & 0 & -1 & 0 & 0 \\ 1 & 0 & 2 & 1 & 0 & 0 \\ 2 & -1 & -1 & 1 & -1 & 0 \\ \hline 1 & 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) , \qquad \text{(A11)}$$

$\lambda_{1,2} = 2,\ \lambda_{3,4} = 1.35605 \pm 2.06011\,i\ ,\ \lambda_5 = 3.28791,\ \lambda_6 = 1.$

$$A = \left(\begin{array}{cccc|cc} 5 & 1 & 2 & 1 & -2 & -2 \\ -1 & 2 & 0 & -2 & 0 & 0 \\ 2 & 0 & 2 & 1 & 0 & 0 \\ -1 & -2 & -1 & 3 & 2 & 2 \\ \hline -2 & 0 & 0 & -2 & 9 & 4 \\ 2 & 0 & 0 & 2 & -4 & -3 \end{array}\right) , \qquad \text{(A12)}$$

$\lambda_{1,2} = 5,\ \lambda_{3,4} = 1,\ \lambda_{5,6} = 3 \pm \sqrt{20}\,.$

$$A = \left(\begin{array}{cccc|cc} 2 & 6 & -4 & 1 & -2 & -2 \\ -6 & 1 & 0 & 2 & 0 & 0 \\ -4 & 0 & 1 & 4 & 0 & 0 \\ -1 & 2 & -4 & 0 & 2 & 2 \\ \hline -2 & 0 & 0 & -2 & 3 & 6 \\ 2 & 0 & 0 & 2 & -6 & 7 \end{array}\right) , \qquad \text{(A13)}$$

$\lambda_{1,2} = 1 + 4i$, $\lambda_{3,4} = 1 - 4i$, $\lambda_{5,6} = 5 \pm \sqrt{32}\, i$.

$$A = \left(\begin{array}{cccc|cc} 4 & 0 & 0 & 2 & 2 & 0 \\ 0 & 3 & 1 & 4 & 1 & 0 \\ 0 & -1 & 1 & -4 & -1 & 0 \\ -2 & 4 & 4 & 0 & 2 & 0 \\ \hline 2 & -1 & -1 & -2 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) , \qquad \text{(A14)}$$

$\lambda_{1,2} = 2$, $\lambda_{3,4} = 0.72722 \pm 2.72129\, i$, $\lambda_5 = 5.54555$, $\lambda_6 = 1$.

$$A = \left(\begin{array}{cccc|cc} 4 & -3 & 1 & 2 & -1 & -1 \\ 3 & -2 & 2 & 1 & 1 & 1 \\ 1 & -2 & 2 & -1 & 1 & 1 \\ -2 & 1 & 1 & 0 & 1 & 1 \\ \hline -1 & -1 & 1 & -1 & 1 & 0 \\ 1 & 1 & -1 & 1 & 0 & 1 \end{array}\right) , \qquad \text{(A15)}$$

$\lambda_{1,2} = 1 + \sqrt{3}\, i$; $\lambda_{3,4} = 1 - \sqrt{3}\, i$, $\lambda_{5,6} = 1$.

LEBENSLAUF

29.09.1952	Geboren in Hamm (Westf.) Vater: Heinz Hoppe Mutter: Margret Hoppe, geb. Peters Beruf des Vaters: Bundesbahnamtmann Staatsangehörigkeit: Deutsch Konfession: Römisch-katholisch
1958 - 1962	Volksschule in Hamm
1962 - 1971	Neusprachliches Gymnasium in Hamm
13.05.1971	Abitur
Oktober 1971 - Juni 1973	Soldat auf Zeit bei der Bundeswehr
Oktober 1973 - Juli 1980	Studium der Mathematik und Informatik an der Universität Dortmund
18.07.1980	Diplom in Mathematik
16.08.1980	Heirat Ehefrau: Birgitta Hoppe, geb. Usdrowski
15.09.1980 - 30.11.1980	Wissenschaftliche Hilfskraft an der Fernuniversität - Gesamthochschule - Hagen
seit 01.12.1980	Wissenschaftlicher Angestellter an der Fernuniversität - Gesamthochschule - Hagen

Hagen, den 27.07.1984 Peter Hoppe