ÜBER DIE EIGENWERTBERECHNUNG
MITTELS
PRIMITIVER JACOBI-ÄHNLICHER VERFAHREN

DISSERTATION

zur

Erlangung des Grades eines Dr. rer. nat.
des Fachbereichs Mathematik und Informatik
der Fernuniversität - Gesamthochschule - Hagen

vorgelegt von
Wolfgang Zacharias
aus Dortmund

Dortmund 1989

Herrn Prof. Dr. K. Veselić möchte ich für seine stete Bereitschaft
zu anregenden Diskussionen als Betreuer dieser Arbeit meinen Dank aussprechen.

0. Einleitung

Diese Arbeit beschäftigt sich mit einer speziellen Klasse von Jacobi-ähnlichen Algorithmen zur Lösung des algebraischen Eigenwertproblems

$$A \cdot \mathrm{x} = \lambda \cdot \mathrm{x} \tag{0.1}$$

Basis dieser Verfahren ist das klassische Jacobi-Verfahren [12] zur Berechnung der Eigenwerte einer symmetrischen Matrix. Es beruht auf einer Folge von orthogonalen Ähnlichkeitstransformationen, die die Ausgangsmatrix auf Diagonalgestalt transformieren. Als Transformationsmatrizen dienen ebene Rotationen, deren Rotationswinkel so bestimmt werden, daß einzelne Außerdiagonalelemente der Matrix zu Null werden. Als Pivotelemente werden dabei sukzessive alle Außerdiagonalelemente der Matrix verwendet.

Das klassische Jacobi-Verfahren für reelle, symmetrische Matrizen zeichnet sich in zwei Punkten besonders aus. Indem die Symmetrie der Matrix genutzt wird, kann das Verfahren auf die Bearbeitung nur eines Dreiecks der Matrix beschränkt werden. Dadurch wird der Rechenaufwand halbiert und gleichzeitig benötigt der Algorithmus nur den halben Speicherplatz im Vergleich zur vollen Matrix A des Eigenwertproblems (0.1). Andererseits sind die Rotationswinkel der ebenen Rotationen, unabhängig von der Dimension der Ausgangsmatrix, aus nur 4 Matrixelementen berechenbar. Beide Umstände führen zu einem äußerst effektiven Algorithmus.

Verallgemeinerungen des klassischen Jacobi-Verfahrens konnten diese beiden Eigenschaften nur teilweise übernehmen. Bei Algorithmen für Matrizen ohne Symmetrieeigenschaften geht z.B. der Speicherplatzvorteil verloren. Einen Überblick über die einzelnen Modifikationen des klassischen Jacobi-Verfahrens findet man bei Hoppe [11]. Wir betrachten hier eine Klasse von Blockverallgemeinerungen des klassischen Jacobi-Verfahrens für nicht diagonalisierbare Matrizen, die als Endmatrix eine Blockdiagonalmatrix mit 2x2 - Blöcken auf der Diagonale anstreben. Aus diesen Blöcken kann man leicht die Eigenwerte berechnen. Gleichzeitig können so auch komplexe Eigenwerte berechnet werden.

Ein entscheidendes Problem bei Jacobi - ähnlichen Blockverfahren für nicht normale Matrizen liegt darin, daß man die Blockdiagonalgestalt nicht mehr allein durch unitäre Transformationen erzielen kann. Dies sieht man an folgender von Schur bewiesenen Aussage (vgl. [28]):

Seien $\lambda_1,\dots,\lambda_n$ die Eigenwerte der Matrix A. Dann gilt:

$$\| A \|_F^2 \geq \sum_{i=1}^{n} | \lambda_i |^2. \tag{0.2}$$

Hierbei tritt Gleichheit genau dann auf, wenn A normal ist.

Will man folglich Diagonalgestalt mit den Eigenwerten auf der Diagonale erreichen, so muß die Norm der Matrix minimiert werden. Daher kann man nicht mehr allein unitäre und somit normerhaltende Transformationen verwenden.

Eine weitreichende Konsequenz ergibt sich daraus, daß die Normveränderungen einzelner Ähnlichkeitstranformationen kontrolliert werden müssen. Dies bedingt, daß man bei Verwendung von Elementartransformationen die Veränderungen von $4 \cdot n$ Matrixelementen mit $n = \dim A$ berücksichtigen und die Parameter so berechnen muß, daß die Gesamtnorm der Matrix reduziert wird. Die meisten Jacobi - ähnlichen Algorithmen für nicht normale Matrizen bestehen deshalb aus einem Normreduzierungsschritt und einem Diagonalisierungsschritt. Dabei ist zu beachten, daß der Rechenaufwand der Parameterbestimmung des Normreduzierungsschrittes proportional zur Dimension der Matrix ist, während beim klassischen Jacobi - Verfahren nur jeweils 4 Matrixelemente benutzt wurden und die Zahl der arithmetischen Operationen zur Parameterbestimmung unabhängig von der Dimension der Matrix ist.

In dieser Arbeit wird gezeigt, daß man in bestimmten Situationen auch bei nicht normalen Matrizen auf den komplizierten Normreduktionsschritt verzichten kann und wie beim Jacobi-Verfahren lediglich die Elemente des Pivotblockes annulliert. Ein solches annullierendes Verfahren wollen wir primitiv nennen, in dem Sinne, daß lediglich einzelne Matrixelemente zu Null gesetzt werden im Gegensatz zur wesentlich komplexeren Parameterbestimmung bei normreduzierenden Verfahren. Bei der primitiven Winkelbestimmung für Blockverfahren sind pro Pivot (p,q) lediglich die Matrixelemente der (p,q)-Restriktion der Matrix, also 16 Matrixelemente, betroffen. Den ersten Hinweis auf diese Vorgehensweise findet man bei Durand [1].

Fast alle Jacobi-ähnliche Algorithmen benutzen zur Diagonalisierung der Ausgangsmatrix Elementarmatrizen. Ähnlichkeitstransformationen mit Elementarmatrizen zeichnen sich dadurch aus, daß jeweils nur ein Zeilen- bzw. ein Spaltenpaar der Ausgangsmatrix linear kombiniert werden. Dies ermöglicht eine effektive Implementierung solcher Algorithmen auf Paralellrechnern, die Linearkombinationen von Vektoren und Skalarprodukte in einem einzigen Maschinenzyklus berechnen können. Bei normreduzierenden Parameterbestimmungen kann man den Vorteil der schnellen Durchführung einer elementaren Ähnlichkeitstransformation nicht nutzen, da die Parameterbestimmung an sich als Berechnung der Normen von ganzen Zeilen und Spalten der transformierten Matrix nicht paralellisierbar ist. Hierin liegt die besondere Bedeutung der primitiven Parameterbestimmung. Diese benötigt nur die lokale Informationen des Pivotblocks und man kann die Parameter äußerst effektiv berechnen. Eventuell schlechtere Konvergenzeigenschaften von primitiven Algorithmen können in der Praxis bei Verwendung von Paralellrechnern in Zukunft von untergeordneter Bedeutung sein, da diese durch den Geschwindigkeitsvorteil der Paralellisierung wettgemacht wird.

Wir behandeln speziell die Klasse der J-symmetrischen Matrizen A, für die gilt

$$J \cdot A \cdot J = A^{\mathrm{T}} \quad \text{mit } J = J^{\mathrm{T}} = J^{-1}. \tag{0.3}$$

Dieser Matrixtyp entsteht bei der Diskretisierung von Differentialgleichungen für gedämpfte Schwingungen. Näheres ist im Kapitel 4 nachzulesen. Blockverfahren für J-symmetrische Matrizen besitzen wie beim klassischen Jacobi-Verfahren den Vorteil des geringen Speicherplatzbedarfes, falls J-orthogonale Transformationsmatrizen T mit

$$T^{-1} = J \cdot T^{\mathrm{T}} \cdot J \tag{0.4}$$

verwendet werden, die die J-Symmetrie der Ausgangsmatrix erhalten.
Damit besitzen primitive Jacobi-ähnliche Blockverfahren für J-symmetrische Matrizen beide Vorteile des klassischen Jacobi-Algorithmus.

Wir weisen für ein primitives Jacobi-ähnliches Blockverfahren, angewandt auf J-symmetrische Matrizen, die asymptotisch quadratische Konvergenz nach. Dies bedeutet, daß bei genügend kleiner Außerdiagonale der Startmatrix A das Verfahren die Außerdiagonale mit quadratischer Konvergenzordnung verkleinert, bis Blockdiagonalgestalt erreicht wird.

Nach der Einführung der grundlegenden Bezeichnungen und Definitionen im 1. Kapitel stellen wir anhand eines einfachen primitiven Algorithmus die Beweisidee der asymptotisch quadratischen Konvergenz vor. Das behandelte Verfahren diagonalisiert eine komplexe Matrix mit getrennten Eigenwerten. Innerhalb des Beweises müssen die Parameter der einzelnen Transformationen in Abhängigkeit von der Größenordnung der Außerdiagonalen und der Getrenntheit der Eigenwerte abgeschätzt werden.

Das Verfahren eignet sich deshalb besonders gut als Beispiel, weil die Parameter im Gegensatz zum später betrachteten Blockverfahren exakt berechenbar sind. Weiterhin wird in Abschnitt 2.4 eine neue Beweisidee vorgestellt, wie man das Außerdiagonalwachstum besonders scharf abschätzen kann.

Das dritte Kapitel beinhaltet den asymptotischen Konvergenzbeweis für einen primitiven Jacobi-ähnlichen Blockalgorithmus. Ausgangspunkt für dieses Verfahren ist ein unveröffentlichtes Manuskript von K. Veselić und E. Zakrajsek [25]. Das Verfahren annulliert durch die Kombination einer nichtunitären hyperbolischen Ähnlichkeitstransformation und einer Jacobi-Rotation die 4 Elemente des jeweiligen Pivotblockes. Die 4 Parameter sind Lösung eines nichtlinearen Gleichungssystems und können nicht exakt berechnet werden. Deswegen wird in 3.4 das Newton-Verfahren verwendet, um die Parameter näherunsweise zu bestimmen. Die mit Hilfe des Satzes von Newton-Kantorovitsch [17] gewonnene Abschätzung der Parameter wird wie im Kapitel 2 für den Beweis der asymptotisch quadratischen Konvergenz eingesetzt. Es ist dabei zu beachten, daß die Abschätzungen von der Norm der zu diagonalisierenden Matrix abhängen und daher relativ kompliziert werden. Im Abschnitt 3.17 wird dann nachgewiesen, daß trotz möglicher Normerhöhung im Laufe des Algorithmus das Verfahren dennoch quadratisch konvergiert.

Im 4. Kapitel wird aufgezeigt, wie der vorgestellte Algorithmus für J-symmetrische Matrizen zur Geschwindigkeitserhöhung in einen normreduzierenden Algorithmus von Hoppe [11] integriert werden kann. Es werden Möglichkeiten vorgestellt, wie aus dem quadratischen Eigenwertproblem der gedämpften Schwingung J-symmetrische Matrizen erzeugt werden können. Abschließend werden die numerischen Ergebnisse vorgestellt.

1. Grundlagen

In diesem Kapitel wollen wir die grundlegenden Bezeichnungen und Definitionen zur Verfügung stellen. Es ist als kurze Einführung in die benutzte Nomenklatur anzusehen. Die Bezeichnungen, die im 2. Kapitel auf einzelne Elemente angewendet werden, sind im 3. Kapitel auf die Blöcke von J-symmetrischen Matrizen anzuwenden.

1.1 Definitionen und Begriffsbildungen

Wir definieren zunächst die Blockeinteilung von Matrizen.

Definition 1.1

Sei $A \in \mathbb{R}^{2m,2m}$ für $m \geq 2$ gegeben. Dann definieren wir für A folgende Blockeinteilung

$$A = \left[A_{ij} \right]_{1 \leq i,j \leq m} \quad \text{mit}$$

$$A_{ij} = \begin{bmatrix} a_{ij} & a_{i,j+m} \\ a_{i+m,j} & a_{i+m,j+m} \end{bmatrix} \in \mathbb{R}^{2,2} \text{ für alle } 1 \leq i,j \leq m. \tag{1.1}$$

Für $1 \leq p < q \leq m$ bezeichnen wir die 4x4-Matrix

$$\hat{A}_{ij} = \begin{bmatrix} A_{ij} & A_{i,j+m} \\ A_{i+m,j} & A_{i+m,j+m} \end{bmatrix} \in \mathbb{R}^{4,4} \tag{1.2}$$

als Blockrestriktion der Matrix A.

Wir werden folgende Normen in den nachfolgenden Kapiteln verwenden:

Bezeichnung 1.2

Seien $x \in \mathbb{R}^n$ sowie $A = (a_{ij})_{1 \leq i,j \leq n} \in \mathbb{R}^{n,n}$ mit $n \geq 2$ gegeben. Dann sei

$\| x \| := \left[\sum_{i=1}^{n} | x_i |^2 \right]^{\frac{1}{2}}$ die euklidische Vektornorm,

$\| A \|_F := \left[\sum_{i,j=1}^{n} | a_{ij} |^2 \right]^{\frac{1}{2}}$ die Frobeniusnorm

$\| A \|_2 := \sqrt{\rho(A^* \cdot A)}$ die Spektralnorm,

wobei $\rho(A)$ der Spekralradius von A ist

$\| A \|_\infty := \max_{1 \leq i,j \leq n} | a_{ij} |$ die Maximumnorm.

Wir benötigen als eine der wesentlichen Größen ein Maß für die Größenordnung der Außerdiagonale sowie ein Maß für die Abweichung einer Matrix von der Normalität.

Definition 1.3

Sei $A \in \mathbb{R}^{2m,2m}$ mit $A = \left[A_{ij} \right]_{1 \leq i,j \leq m}$ für $m \geq 2$ gegeben. Dann sei die Außerdiagonalnorm der Blockmatrix A gegeben durch

$$S(A) := \left[\sum_{\substack{i,j=1 \\ i \neq j}}^{n} \| A_{ij} \|_F^2 \right]^{\frac{1}{2}} \tag{1.3}$$

Definition 1.4

Die Matrix $A \in \mathbb{R}^{n,n}$, $n \geq 2$, sei auf Schursche Dreiecksgestalt transformiert, d.h.

$$A = D + M$$

mit $D = \operatorname{diag}(\lambda_1,\dots,\lambda_n)$, wobei die λ_i gerade die n Eigenwerte der Matrix A sind, und M ist echte obere Dreiecksmatrix, d.h. $M = (m_{ij})_{1\leq i,j\leq n}$ mit $m_{ij} = 0$ für $i \geq j$.
Sei $\|\cdot\|$ eine Matrixnorm, dann ist das Henrici-Maß $\Delta(A)$ der Matrix A definiert als

$$\Delta(A) := \inf_{M} \| M \|. \tag{1.4}$$

Das Infimum wird dabei über alle Dreiecksmatrizen M gebildet, die in der Schurschen Dreiecksgestalt auftreten können.

Bemerkung:

Wir werden $\Delta_F(A)$ und $\Delta_2(A)$ verwenden. Dies seien die Henrici-Maße von A bezüglich der Frobenius- bzw. bezüglich der Spektralnorm gebildet. Es gilt speziell für diese Henrici-Maße:

$$\begin{aligned} \Delta_F(A) &= \left[\| A \|_F^2 - \| D \|_F^2 \right]^{\frac{1}{2}} \\ &= \left[\| A \|_F^2 - \sum_{i=1}^{n} | \lambda_i |^2 \right]^{\frac{1}{2}} \end{aligned} \tag{1.5}$$

Es ist stets $\Delta_2(A) \leq \Delta_F(A)$.

Definition 1.5:

Sei $E \in \mathbb{R}^{2m,2m}$ mit $E = \left[E_{ij} \right]_{1 \le i,j \le m} = \left[e_{ij} \right]_{1 \le i,j \le 2m}$ für $m \geq 2$ eine reelle Blockmatrix. Wir bezeichnen E als Elementarmatrix, falls für ein (p,q) mit $1 \leq p < q \leq m$ gilt

$$E_{ij} = I \cdot \delta_{ij} \text{ für alle Paare}$$
$$(i,j) \in \Big\{ (i,j) \;\big|\; 1 \le i,j\ m \text{ und } (i,j) \neq (p,q),(p,q+m),(p+m,q),(p+m,q+m) \Big\}. \tag{1.6}$$

Damit unterscheidet sich E nur in einer (p,q)-Restriktion von der Einheitsmatrix. Das Indexpaar (p,q) heißt Pivot und der zum Pivot gehörende Matrixblock E_{pq} heißt Pivotblock.

Bemerkung:

Durch die Durchführung von Ähnlichkeitstransformationen mit Elementarmatrizen werden nur zwei Zeilen und zwei Spalten der Matrix transformiert bzw. eine Blockzeile und -spalte.

Ein Algorithmus, der aus einer Reihe von Ähnlichkeitstransformationen mit Elementarmatrizen besteht, ist neben der Form der Transformationen und deren Wirkung auch durch die Auswahl der Reihenfolge der verwendeten Pivotpaare gekennzeichnet. Zur Beschreibung der Pivotauswahl benutzen wir folgende Bezeichnungen:

Definition 1.6:

Sei $A \in \mathbb{R}^{n,n}$ beliebig. Es sei $N := n \cdot (n-1)$ und $P_n := \{ (i,j) \mid 1 \le i,j \le n \text{ und } i \ne j \}$. Für ein elementweises Verfahren werden alle Nichtdiagonalelementen a_{pq} als Pivotelemente verwendet. Wir sagen, daß ein Zyklus des Verfahrens aus der Folge von Ähnlichkeitstransformationen besteht, die jedes Nichtdiagonalelement genau einmal als Pivotelement benutzen. Das Matrixelement a_{pq} für $p \ne q$ wird dabei als $\zeta(p,q)$-tes Pivotelement verwendet, wobei

$$\zeta : P_n \to \{1,\ldots,N\},\ \zeta(p,q) := q - p + \sum_{i=1}^{p-1} (n - i) \tag{1.7}$$

Diese Reihenfolge wird als zeilenzyklische Pivotstrategie bezeichnet.

Bezeichnung

Zur Beschreibung der jeweiligen Standes des Verfahrens führen wir folgende Nummerierung für die Folge der transformierten Matrix A ein.

$A^{l} = \left[A_{ij}^{l} \right]_{1 \le i,j \le m}$ sei die Matrix nach dem l-ten Zyklus

$A^{k,l} = \left[A_{ij}^{k,l} \right]_{1 \le i,j \le m}$ sei die Matrix nach k Ähnlilchkeitstransformationen im $(l+1)$-ten Zyklus

Definition 1.7:

Sei $A \in \mathbb{R}^{2m,2m}$ mit $A = \left[A_{ij} \right]_{1 \le i,j \le m}$, $m \ge 2$, eine reelle Blockmatrix. A heißt J-symmetrisch, falls eine orthogonale Matrix $J \in \mathbb{R}^{2m,2m}$ existiert mit

$$A^{T} = J \cdot A \cdot J \tag{1.8}$$

Bemerkung:

Falls A J-symmetrisch ist, ist das Produkt $J \cdot A$ symmetrisch. Wir betrachten in dieser Arbeit J-Symmetrie bzgl. der Matrix

$$J = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} \in \mathbb{R}^{2m,2m}, \text{ wobei I die Einheitsmatrix im } \mathbb{R}^{m,m} \text{ ist.}$$

Definition 1.8:

Sei $R \in \mathbb{R}^{2m,2m}$ mit $R = \left[R_{ij} \right]_{1 \le i,j \le m}$ mit $m \ge 2$ eine reelle Blockmatrix. R heißt J-orthogonal, falls eine orthogonale Matrix $J \in \mathbb{R}^{2m,2m}$ existiert mit

$$R^{-1} = J \cdot R^T \cdot J \tag{1.9}$$

Bemerkung:

Die Inverse J-orthogonaler Matrizen ist also sehr einfach zu berechnen. Verwendet man für Ähnlichkeitstransformationen J-orthogonale Elementarmatrizen, so ist die J-Symmetrie eine Invariante der Transformation. Dies ermöglicht die Ausnutzung der J-Symmetrie bei der Auswahl der Pivotelemente.

Für Blockverfahren, die in der Klasse der J-symmetrischen Blockmatrizen arbeiten und auf J-orthogonalen Elementarmatrizen basieren, verwenden wir folgende Pivotstrategie.

Definition 1.9:

Sei $A \in \mathbb{R}^{2m,2m}$ eine J-symmetrische, reelle Blockmatrix. Es sei $M := m \cdot (m-1)/2$ und es sei $Q_m := \{ (i,j) \mid 1 \le i < j \le m \}$. Für ein die J-Symmetrie ausnutzendes Blockverfahren werden alle Nichtdiagonalblöcke A_{pq} mit $p < q$ als Pivotblöcke verwendet. Der Block A_{pq} wird dabei als $\phi(p,q)$-ter Pivotblock benutzt, wobei

$$\phi : Q_m \to \{1,\ldots,M\},\ \phi(p,q) := q - p + \sum_{i=1}^{p-1} (m - i) \tag{1.10}$$

Diese Reihenfolge wird als zeilenzyklische Pivotstrategie bezeichnet.

Schließlich führen wir noch folgende Bezeichnung ein, die den gewünschten Endzustand der Blockdiagonale nach Durchführung des Verfahrens beschreibt.

Definition 1.10

Sei $A \in \mathbb{R}^{2m,2m}$ mit $A = \left[A_{ij} \right]_{1 \le i,j \le m}$ eine reelle Blockmatrix. Wir sagen die Matrix A besitzt Murnaghan-Form, falls A eine 2x2-Blockdiagonalmatrix ist mit normalen Diagonalblöcken, d.h.

$A = \mathrm{diag}(A_{11},\ldots,A_{mm})$, wobei die 2x2-Matrizen A_{ii} von der Form

$$A_{ii} = \begin{bmatrix} \lambda_i & 0 \\ 0 & \lambda_{i+m} \end{bmatrix} \text{ oder } A_{ii} = \begin{bmatrix} \mathrm{Re}\,\lambda_i & \mathrm{Im}\,\lambda_i \\ -\mathrm{Im}\,\lambda_i & \mathrm{Re}\,\lambda_i \end{bmatrix} \text{ sind.} \tag{1.11}$$

Hierbei sind λ_j für $j = 1,\ldots,2m$ die Eigenwerte der Matrix A.

2. Ein erstes Jacobi-ähnliches Annullierungsverfahren

Dieses Kapitel beschäftigt sich mit einem ersten primitiven Verfahren zur Berechnung der Eigenwerte einer Matrix $A \in \mathbb{C}^{n,n}$, die getrennte Eigenwerte besitzt. Es soll zur Veranschaulichung des aufwendigen Beweises der asymptotisch quadratischen Konvergenz von primitiven Verfahren dienen. Es gelte

$$\delta := \min_{i \neq j} | \lambda_i - \lambda_j | > 0 \tag{2.1}$$

Hierbei sei $\lambda_i \in \mathbb{C}$ der i-te Eigenwert von A für i = 1,...,n.

Dabei verstehen wir unter Primitivität, daß Ähnlichkeitstransformationen zur direkten Annlierung einzelner Außerdiagonalelemente eingesetzt werden wie es beim klassischen Jacobi-Verfahren durch orthogonale Drehungsmatrizen geschieht. Die Benutzung orthogonaler bzw. unitärer Ähnlichkeitstranformationen kann bei nichtnormalen Matrizen nicht zum Ziel führen, da diese zusätzlich in ihrer Norm reduziert werden müssen, um Diagonalgestalt bzw Blockdiagonalgestalt zu erlangen.
Deswegen arbeiten wir mit nichtunitären Elementarmatrizen, die in komplizierteren Algorithmen zur Normreduzierung eingesetzt werden. Bei einer Normreduzierungstaktik gestaltet sich allerdings die Berechnung der zugehörigen Parameter sehr kompliziert, da bei Transformationen mit Elementarmatrizen zwei Zeilen und zwei Spalten verändert werden, so daß die Veränderung von 4·n Matrixelementen bei der Normänderung berücksichtigt werden muß.
Einfacher ist es, direkt ein einzelnes Matrixelement auf Null zu transformieren. Die Parameterberechnung ist in diesem Fall elementar, allerdings ist für eine solche Transformation in der Regel nur im asymptotischen Grenzfall kleiner Außerdiagonalelemente der Matrix mit einer Normreduzierung zu rechnen, da die Linearkombination der Pivotzeilen bzw. -spalten insgesamt trotz Annllierung der Pivotelemente in der Summe normerhöhend wirken kann.

Durch zeilenzyklische Anwendung von Ähnlichkeitstransformationen wird in dem jetzt beschriebenen Algorithmus versucht, die Außerdiagonale der Matrix A zu annullieren. Dabei wird die einfachste mögliche Taktik eines primitiven Verfahrens verwandt, indem wir mit zwei Dreiecksmatrizen mit jeweils einem reellen Parameter die aktuellen Pivotelemente annullieren.

2.1 Die Transformationsmatrizen

Alle Jacobi-ähnlichen Verfahren gehen von möglichst einfachen Transformationsmatrizen aus. Sie unterscheiden sich nur in einem einzigen Matrixelement von der Einheitsmatrix, dem Pivotelement an der Stelle (p,q). Die (i,j)-Restriktionen S_{ij}, $T_{ij} \in \mathbb{C}^{n,n}$ der Transformationsmatrizen $\boldsymbol{S}$ bzw. $\boldsymbol{T}$ sind dabei gegeben durch

$$S_{pq} := \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \text{ und } T_{pq} := \begin{pmatrix} 1 & 0 \\ y & 1 \end{pmatrix} \text{ mit } x,y \in \mathbb{C} \tag{2.2}$$

Die Transformationsmatrizen sind nichtsinguläre Dreiecksmatrizen. Daher ergeben sich besonders einfache Formeln für die Matrixelemente der transformierten Matrix $\tilde{\boldsymbol{A}}$. Die Parameter x,y der beiden Matrizen werden so bestimmt, daß die Pivotelemente a_{pq} und a_{qp} verschwinden. Es gilt

$$\left.\begin{aligned} & a_{qp} = 0: \\ & x = \frac{a_{pq}}{a_{qq} - a_{pp}} \quad \text{für } a_{qq} \neq a_{pp} \\ & y = 0 \end{aligned}\right] \tag{2.3a}$$

$$\left.\begin{aligned} & a_{qp} \neq 0: \\ & x = \frac{1}{2\cdot a_{qp}} \cdot \left((a_{qq} - a_{pp}) - \sqrt{(a_{qq} - a_{pp})^2 - 4\cdot a_{qp}\cdot a_{pq}} \right) \\ & y = \frac{a_{qp}}{\sqrt{(a_{qq} - a_{pp})^2 - 4\cdot a_{qp}\cdot a_{pq}}} \quad \text{für } \sqrt{(a_{qq} - a_{pp})^2 - 4\cdot a_{qp}\cdot a_{pq}} \neq 0 \end{aligned}\right] \tag{2.3b}$$

Bemerkungen

Die in beiden Parametern x und y auftretende komplexe Quadratwurzel ist nur bis auf ihr Vorzeichen eindeutig bestimmt. Für eine beliebige komplexe Zahl $z := \alpha + i\cdot\beta \in \mathbb{C}$ gilt

$$\sqrt{z} := \pm \left[\sqrt{0.5\cdot\left[\sqrt{\alpha^2 + \beta^2} + \alpha\right]} + (\operatorname{sign}\beta)\cdot i\cdot \sqrt{0.5\cdot\left[\sqrt{\alpha^2 + \beta^2} - \alpha\right]} \right]$$

Für die Parameterwahl ist lediglich wichtig, daß bei x und y dieselbe Quadratwurzel zugrundegelegt wird. Wir verwenden in diesem Kapitel stets die positive komplexe Quadratwurzel.

Die Parameterbestimmung zur Annullierung der Pivotelemente ist nur dann wohldefiniert, falls die Elemente der Diagonale im Pivotblock ungleich sind. Dies führt auf die Forderung $a_{qq} \neq a_{pp}$ für die untransformierte Matrix bzw. auf die Forderung $\sqrt{(a_{qq} - a_{pp})^2 - 4 \cdot a_{qp} \cdot a_{pq}} \neq 0$ für die Matrix nach der Transformation mit S. Wir werden später eine Voraussetzung an die Größenordnung der Außerdiagonale der Matrix A stellen, die beide Forderungen garantiert. Im Fall $a_{qp} = 0$ wird nur eine Transformation durchgeführt. Daher wird T mit $y = 0$ zur Einheitsmatrix gesetzt. Damit ist die Wirkung der Ähnlichkeitstransformationen auf die Matrix A an der Pivotstelle (p,q) festgelegt. Es gilt

$$\tilde{A} = T^{-1} \cdot S^{-1} \cdot A \cdot S \cdot T \tag{2.4}$$

$$\tilde{a}_{ij} = a_{ij} \quad \text{für } (i,j) \notin \{p,q\} \quad \text{sowie} \quad \tilde{a}_{pq} = \tilde{a}_{qp} = 0 \tag{2.5a}$$

$$\left.\begin{aligned} \tilde{a}_{qq} &= a_{qq} + x \cdot a_{qp} \\ \tilde{a}_{pp} &= a_{pp} - x \cdot a_{qp} \end{aligned}\right] \tag{2.5b}$$

$$\left.\begin{aligned} \tilde{a}_{ip} &= (1 + x \cdot y) \cdot a_{ip} + y \cdot a_{iq} \\ \tilde{a}_{iq} &= a_{iq} + x \cdot a_{ip} \\ \tilde{a}_{qj} &= (1 + x \cdot y) \cdot a_{qj} - y \cdot a_{pj} \\ \tilde{a}_{pj} &= a_{pj} - x \cdot a_{qj} \end{aligned}\right] \quad \text{für } (i,j) \neq (p,q) \tag{2.5c}$$

Die Formeln (2.5) machen die Möglichkeit der effektiven Implementierung des Algorithmus auf Paralellrechnern deutlich, da die neuen Matrixelemente lediglich Linearkombinationen von Zeilen bzw. Spalten der alten Matrix sind. Gleichzeitig erkennt man die Stabilität des Algorithmus bei kleinen Parametern x und y.

2.2 Definition eines Zyklus des Verfahrens

Sei $A \in \mathbb{C}^{n,n}$ und $(p,q) \in P_n := (1,\dots,n\} \times \{1,\dots,n\} \setminus \{(i,i) | i \in 1,\dots,n$ und $i < j\}$ ein gegebenes Pivotpaar. Die Pivotelemente werden in zeilenzyklischer Reihenfolge gewählt:

$$
\begin{array}{llll}
(1,2)\ (1,3)\ \dots & (1, n-1) & (1 & ,n) \\
\quad\ \ (2,3)\ \dots & & (2 & ,n) \\
 & & (n-1 & ,n)
\end{array}
$$

Ein vollständiger Zyklus besteht folglich aus $N := n \cdot (n-1)/2$ Transformationen, so daß jedes Nichtdiagonalement einmal als Pivotelement benutzt wird. Setzen wir

$$\zeta : P_n \to \{1,\dots,N\} \quad \zeta(p,q) := q - p + \sum_{i=1}^{p-1} (n-i), \tag{2.6}$$

so werden die Matrixelemente a_{pq} und a_{qp} als $\zeta(p,q)$-te Pivotelemente annulliert. Sei A^i die Matrix nach dem i-ten Zyklus. Über den $(i+1)$-ten Zyklus wird die Matrix $A^i \in \mathbb{C}^{n,n}$ wie folgt transformiert:

$$
\begin{aligned}
A^0 &:= A \\
A^{k+1,i} &:= \left[T^{k,i}\right]^{-1} \cdot \left[S^{k,i}\right]^{-1} \cdot A^{k,i} \cdot S^{k,i} \cdot T^{k,i} \text{ für } k \in \{1,\dots N\},\ i \in \mathbb{N} \qquad (2.7) \\
A^{i+1} &:= A^{N,i}
\end{aligned}
$$

Hierbei ist das zu $S^{k,i}$ und $T^{k,i}$ gehörende Pivotelement über (2.6) durch die Zuordnung $k = \zeta(p,q)$ gegeben. Wir bezeichnen mit hochgestellten Indizes die laufende Nummer der Transformation und die aktuelle Nummer des Zyklus. Damit ist A^0 die vorgegebene Startmatrix und $A^{i+1} = A^{N,i}$ die Matrix nach dem i-ten Zyklus, da ein Zyklus aus N Ähnlichkeitstransformationen besteht.

2.3 Abschätzung der Winkel bei asymptotischer Betrachtung

Ein derart einfaches Verfahren, das lediglich durch sukzessives Nullsetzen der Pivotelemente a_{pq} und a_{qp} versucht, eine Matrix A auf Diagonalgestalt zu bringen, kann nicht global konvergent sein. Wir begnügen uns daher damit zu untersuchen, wie sich das Verfahren bei bereits kleiner Außerdiagonalnorm $S(A)$ verhält. Es sei folglich

$$S(A) := \left\{ \sum_{\substack{i,j=1 \\ i \neq j}}^{n} |a_{ij}|^2 \right\}^{\frac{1}{2}} \leq \frac{\delta}{100} \qquad \text{mit } \delta \text{ aus (2.1)} \tag{2.8}$$

Lemma 2.1:

Mit Voraussetzung (2.8) gilt für die Abstände zweier Diagonalelemente der Matrix A:

$$|a_{qq} - a_{pp}| \geq 0.98 \cdot \delta \quad \text{für } p \neq q. \tag{2.9}$$

Beweis:

Mit Hilfe der umgekehrten Dreiecksungleichung und des Satzes von Gerschgorin, in der Frobenius-Norm angewandt, gilt:

$$\begin{aligned} |a_{qq} - a_{pp}| &\geq |\lambda_q - \lambda_p| - |\lambda_q - a_{qq}| - |\lambda_p - a_{pp}| \\ &\geq \delta \qquad - 2 \cdot \left[\sum_{\substack{i,j=1 \\ i \neq j}}^{n} |a_{ij}|^2 \right]^{\frac{1}{2}} \\ &= \delta \qquad - 2 \cdot S(A) \\ &\geq \delta \qquad - \frac{2}{100} \cdot \delta = 0.98 \cdot \delta \end{aligned}$$

●

Bemerkung:

Der Hilfssatz zeigt, daß für beliebige Pivotelemente a_{pq} stets $a_{pp} \neq a_{qq}$ gilt. Ebenso folgt nach der ersten Transformation mit der elementaren Matrix T, daß

$$\sqrt{(a_{qq} - a_{pp})^2 - 4 \cdot a_{qp} \cdot a_{pq}} > \sqrt{0.98^2 \cdot \delta^2 - 4 \cdot S(A)} \geq \delta \cdot \sqrt{0.98^2 - 0.04} > 0$$

ist. Damit sind die in (2.3) definierten Parameter unter der Voraussetzung (2.8) stets wohldefiniert.

Lemma 2.2:

Unter der Voraussetzung (2.8) gilt:

$$|x| \leq 1.0205144 \cdot \frac{S(A)}{\delta} \qquad |y| \leq 1.0205144 \cdot \frac{S(A)}{\delta} \tag{2.10}$$

Beweis: Für $a_{qp} = 0$ ist $y = 0$ und es gilt

$$|x| \leq \frac{|a_{pq}|}{|a_{qq} - a_{pp}|} \leq \frac{S(A)}{0.98 \cdot \delta} = 1.02041 \cdot \frac{S(A)}{\delta}.$$

Für $a_{qp} \neq 0$ gilt:

$$|x| \leq \left|\frac{a_{pp} - a_{qq}}{2 \cdot a_{qp}}\right| \cdot \left| 1 - \sqrt{1 - \frac{4\, a_{pq}\, a_{qp}}{(a_{qq} - a_{pp})^2}} \right| \quad \text{und}$$

$$\left| \frac{4\, a_{pq}\, a_{qp}}{(a_{qq} - a_{pp})^2} \right| \leq \frac{4 \cdot |a_{pq}^2 + a_{qp}^2|}{|a_{qq} - a_{pp}|^2} \leq \frac{4\, S^2(A)}{0.98 \cdot \delta^2} \leq 2.0824656 \cdot 10^{-4}$$

Setzen wir $f(\xi) := |1 - \sqrt{1 - \xi}|$ für $\xi \in \mathbb{R}$, so folgt für $|\xi| \leq 2.084 \cdot 10^{-4}$

$$|f'(\xi)| = \frac{1}{2 \cdot \sqrt{1 - \xi}} \leq \frac{1}{2 \cdot (1 - 2.083 \cdot 10^{-4})^{\frac{1}{2}}} = 0.50005207 =: C \quad \text{und damit}$$

$$|x| \leq \left|\frac{a_{pp} - a_{qq}}{2 \cdot a_{qp}}\right| \cdot C \cdot \frac{4 \cdot |a_{pq}| \cdot |a_{qp}|}{|a_{qq} - a_{pp}|^2} \leq 2 \cdot C \cdot \frac{S(A)}{0.98 \cdot \delta} = 1.0205144 \cdot \frac{S(A)}{\delta}$$

$$|y| \leq \frac{|a_{qp}|}{|a_{pp} - a_{qq}|} \cdot \frac{1}{\sqrt{1 - \frac{4|a_{pq}|\,|a_{qp}|}{|a_{qq} - a_{pp}|^2}}}$$

$$\leq \frac{|a_{qp}|}{|a_{pp} - a_{qq}|} \cdot \frac{1}{(1 - 2.083 \cdot 10^{-4})^{\frac{1}{2}}} \leq 1.0205144 \cdot \frac{S(A)}{\delta}$$

●

Damit sind die Parameter bei kleiner Außerdiagonale ebenfalls klein.

2.4. Außerdiagonalwachstum pro Transformation

Im weiteren leiten wir eine Formel für den maximalen Zuwachs der Außerdiagonalnorm $S(A)$ unter den Voraussetzungen des Abschnittes 2.3 her. Wir formulieren zunächst die Aussage.

Satz: 2.3:

Es gelte (2.8) und sei weiter $(p,q) \in P_n$ ein Pivotpaar, dessen Annullierung zu einem Vergrößerung von $S(A)$ führt, d.h. für die transformierte Matrix $\tilde{A}$ gilt $S(\tilde{A}) > S(A)$. Dann gilt:

(1) Die nach Formel (2.3) zur Annulierung von a_{pq} und a_{qp} gewählten Parameter $x,y \in \mathbb{C}$ genügen den Abschätzungen:

$$|x| \leq 1.0205144 \cdot \gamma \cdot \left[\frac{S(A)}{\delta}\right]^2 \quad \text{und} \quad |y| \leq 1.0205144 \cdot \gamma \cdot \left[\frac{S(A)}{\delta}\right]^2. \tag{2.11}$$

(2) Für den maximalen Außerdiagonalnormzuwachs pro Transformation gilt

$$S^2(\tilde{A}) \leq \left[1 + \left[\gamma \cdot \frac{S(A)}{\delta}\right]^2\right] \cdot S^2(A). \tag{2.12}$$

Hierbei ist $\gamma = 2.073381$.

Beweis:

Wir betrachten die Differenz von $S(\tilde{A})$ zu $S(A)$ bei einer Ähnlichkeitstransformation, die zu einem Zuwachs in $S(A)$ führt. Es gilt:

$$S(\tilde{A}) > S(A) \Leftrightarrow$$

$$T := (|1 + xy|^2 + |x|^2 - 1) \cdot \sum_{i \neq p,q}^{n} (|a_{ip}|^2 + |a_{qi}|^2) + |y|^2 \cdot \sum_{i \neq p,q}^{n} (|a_{pi}|^2 + |a_{iq}|^2) +$$

$$2 \cdot \mathrm{Re}\left[(y + x + xy^2) \cdot \sum_{i \neq p,q}^{n} (a_{ip} \cdot a_{iq} - a_{qi} \cdot a_{pi})\right] > |a_{pq}|^2 + |a_{qp}|^2$$

$$\Leftrightarrow$$

$$T > |a_{pq}|^2 + |a_{qp}|^2$$

Wir betrachten im weiteren das Verhalten des Terms T. Es gilt:

$$|T| \le \left[\, |y|^2 + |1 + xy|^2 + |x|^2 - 1 + |\mathrm{Re}\,(y + x + xy^2)| \, \right] \cdot S^2(A)$$

Unter Verwendung der Abschätzungen für x und y aus Lemma 2.1 ergibt sich

$$|T| \le 2.073381 \cdot \left[\frac{S(A)}{\delta} \right] \cdot S^2(A) \tag{2.13}$$

Wir setzen zur Abkürzung $\gamma := 2.073381$. Wir zeigen nun mittels Induktion über i folgende Zwischenbehauptungen für alle $i \ge 2$:

$$\left.\begin{aligned} |x| &\le 1.0205144 \cdot \left[\frac{S(A)}{\delta}\right] \cdot \left[\frac{S(A)}{\delta} \cdot \gamma\right]^{\sum_{j=1}^{i-1} (\frac{1}{2})^j} \\ |y| &\le 1.0205144 \cdot \left[\frac{S(A)}{\delta}\right] \cdot \left[\frac{S(A)}{\delta} \cdot \gamma\right]^{\sum_{j=1}^{i-1} (\frac{1}{2})^j} \end{aligned}\right] \tag{2.14}$$

Induktionsanfang (i=2):

Für $|T|$ gilt die Abschätzung (2.13). Mit der Voraussetzung der Zunahme der Außerdiagonalnorm folgt wegen $|a_{pq}|^2 + |a_{qp}|^2 \le T$ für die Pivotelemente die Abschätzung:

$$|a_{pq}| \le \sqrt{\gamma} \cdot S(A) \cdot \left[\frac{S(A)}{\delta}\right]^{\frac{1}{2}} \quad \text{und} \quad |a_{qp}| \le \sqrt{\gamma} \cdot S(A) \cdot \left[\frac{S(A)}{\delta}\right]^{\frac{1}{2}}$$

Diese Abschätzung verwenden wir, um die Abschätzungen für die Parameter $x, y \in \mathbb{C}$ zu verbessern. Es folgt nach (2.3)

$$|x| \le 1.0205144 \cdot \left[\frac{S(A)}{\delta}\right] \cdot \left[\frac{S(A)}{\delta} \cdot \gamma\right]^{\frac{1}{2}}$$

$$|y| \le 1.0205144 \cdot \left[\frac{S(A)}{\delta}\right] \cdot \left[\frac{S(A)}{\delta} \cdot \gamma\right]^{\frac{1}{2}}$$

Damit ist der Induktionsanfang gezeigt.

Induktionsvoraussetzung: Die Formeln (2.14) gelten für ein $i \ge 2$.

Induktionsschritt ($i \to i + 1$):

Durch Einsetzen der Induktionsvoraussetzung in die Abschätzung von $|\mathrm{T}|$ folgt:

$$
\begin{aligned}
|\mathrm{T}| \leq \Bigg\{ & 2\cdot 1.0205144^2 \cdot \left[\frac{S(A)}{\delta}\right]^2 \cdot \left[\left[\frac{S(A)}{\delta}\cdot\gamma\right]^{\sum_{j=1}^{i-1}(\frac{1}{2})^j}\right]^2 \\
& + 1.0205144^4 \cdot \left[\frac{S(A)}{\delta}\right]^4 \cdot \left[\left[\frac{S(A)}{\delta}\cdot\gamma\right]^{\sum_{j=1}^{i-1}(\frac{1}{2})^j}\right]^4 \\
& + 1.0205144^2 \cdot \left[\frac{S(A)}{\delta}\right]^2 \cdot \left[\left[\frac{S(A)}{\delta}\cdot\gamma\right]^{\sum_{j=1}^{i-1}(\frac{1}{2})^j}\right]^2 \\
& + 2\cdot 1.0205144 \cdot \left[\frac{S(A)}{\delta}\right] \cdot \left[\frac{S(A)}{\delta}\cdot\gamma\right]^{\sum_{j=1}^{i-1}(\frac{1}{2})^j} \\
& + 1.0205144^3 \cdot \left[\frac{S(A)}{\delta}\right]^3 \cdot \left[\left[\frac{S(A)}{\delta}\cdot\gamma\right]^{\sum_{j=1}^{i-1}(\frac{1}{2})^j}\right]^3 \Bigg\} \cdot S^2(A) \\
= {} & S^2(A) \cdot \left[\frac{S(A)}{\delta}\right] \cdot \left[\left[\frac{S(A)}{\delta}\cdot\gamma\right]^{\sum_{j=1}^{i-1}(\frac{1}{2})^j}\right] \cdot \\
& \cdot \Bigg\{ 2\cdot 1.0205144^2 \cdot \left[\frac{S(A)}{\delta}\right] \cdot \left[\frac{S(A)}{\delta}\cdot\gamma\right]^{1\cdot\sum_{j=1}^{i-1}(\frac{1}{2})^j} \\
& + 1.0205144^4 \cdot \left[\frac{S(A)}{\delta}\right]^3 \cdot \left[\frac{S(A)}{\delta}\cdot\gamma\right]^{3\cdot\sum_{j=1}^{i-1}(\frac{1}{2})^j} \\
& + 1.0205144^4 \cdot \left[\frac{S(A)}{\delta}\right] \cdot \left[\frac{S(A)}{\delta}\cdot\gamma\right]^{1\cdot\sum_{j=1}^{i-1}(\frac{1}{2})^j} \\
& + 1.0205144^3 \cdot \left[\frac{S(A)}{\delta}\right]^2 \cdot \left[\frac{S(A)}{\delta}\cdot\gamma\right]^{2\cdot\sum_{j=1}^{i-1}(\frac{1}{2})^j} + 2\cdot 1.0205144 \Bigg\}
\end{aligned}
$$

Wegen $\frac{S(A)}{\delta}\cdot\gamma \leq 2.07034\cdot 0.01 \leq 1$ gilt $\left[\frac{S(A)}{\delta}\cdot\gamma\right]^{k\cdot\sum_{j=1}^{i-1}(\frac{1}{2})^j} \leq 1$ für $k \in \{1,2,3\}$ und man kann den Term innerhalb der großen geschweiften Klammer durch γ abschätzen. Es folgt insgesamt

$$
|\mathrm{T}| \leq S^2(A) \cdot \left[\frac{S(A)}{\delta}\right] \cdot \left[\frac{S(A)}{\delta}\cdot\gamma\right]^{\sum_{j=1}^{i-1}(\frac{1}{2})^j} \cdot \gamma
$$

Unter der Annahme einer Transformation mit Normzuwachs folgen hieraus verbesserte Abschätzungen für die Pivotelemente a_{pq} und a_{qp}:

$$|a_{pq}| \leq \sqrt{\gamma}\cdot S(A)\cdot\left[\frac{S(A)}{\delta}\right]^{\frac{1}{2}}\cdot\left[\frac{S(A)}{\delta}\cdot\gamma\right]^{\frac{1}{2}\cdot\sum_{j=1}^{i-1}(\frac{1}{2})^j}$$

$$|a_{qp}| \leq \sqrt{\gamma}\cdot S(A)\cdot\left[\frac{S(A)}{\delta}\right]^{\frac{1}{2}}\cdot\left[\frac{S(A)}{\delta}\cdot\gamma\right]^{\frac{1}{2}\cdot\sum_{j=1}^{i-1}(\frac{1}{2})^j}$$

Mit Hilfe dieser Verbesserungen erhalten wir schließlich neue Abschätzungen für die Parameter $x,y \in \mathbb{C}$. Es gilt:

$$(|x|,|y|) \leq 1.0205144\cdot\left[\frac{S(A)}{\delta}\right]\cdot\left[\frac{S(A)}{\delta}\cdot\gamma\right]^{\sum_{j=1}^{i}(\frac{1}{2})^j}$$

Die letzte Formel ergibt sich mit Hilfe der Gleichung $\frac{1}{2}\cdot\sum_{j=1}^{i-1}(\frac{1}{2})^j = \sum_{j=1}^{i}(\frac{1}{2})^j - \frac{1}{2}$. Damit ist der Induktionsbeweis geführt.

Die angegebene Abschätzung ergibt wegen $\lim_{i\to\infty}\sum_{j=1}^{i}(\frac{1}{2})^j = 1$ die Behauptung (1). Es ist

$$S^2(\tilde{A}) = S^2(A) + T - |a_{pq}|^2 - |a_{qp}|^2 \leq S^2(A) + |T|$$

Mit Behauptung (1) folgt wie im obigen Induktionsanfang die Abschätzung:

$$|T| \leq S^2(A)\cdot\left[\frac{S(A)}{\delta}\right]^2\cdot\gamma^2$$

Einsetzen liefert schließlich

$$S^2(\tilde{A}) \leq S^2(A) + S^2(A)\cdot\left[\frac{S(A)}{\delta}\right]^2\cdot\gamma^2 = S^2(A)\cdot\left[1 + \left[\frac{S(A)^2}{\delta}\right]\cdot\gamma^2\right]$$

●

Bemerkung:

Mit Hilfe von $\left[\frac{S(A)}{\delta}\right]^2 \leq 10^{-4}$ gemäß Voraussetzung (2.8) folgt eine absolute Abschätzung für den maximalen Außerdiagonalnormzuwachs. Es gilt

$$S^2(\tilde{A}) \leq (1 + \gamma^2\cdot 10^{-4})\cdot S^2(A) \quad \text{und damit} \quad S(\tilde{A}) \leq 1.000215\cdot S(A). \qquad (2.15)$$

2.5 Das Wachstum gekoppelter Matrixelemente

In diesem Abschnitt betrachten wir das Wachstum der Matrixelemente in den beiden durch die Pivotwahl gekoppelten Spalten p und q der Matrix A. Wir schaffen mit Hilfe der angegebenen Abschätzung eine technische Voraussetzung für den abschließenden Konvergenzbeweis im folgenden Abschnitt.

Lemma 2.4

Es gelte (2.8). Sei $\tilde{A}$ die durch die $\varphi(p,q)$-te Transformation erzeugt. Dann gilt für die in dieser Transformation erzielten Veränderung

$$|\tilde{a}_{pj}|^2 + |\tilde{a}_{qj}|^2 \leq (1 + 3.1246\cdot 10^{-4})\cdot(|a_{pj}|^2 + |a_{qj}|^2) \tag{2.16}$$

Beweis:

Laut Transformationsformeln (2.5) gilt:

$$\begin{aligned}
&|\tilde{a}_{pj}|^2 + |\tilde{a}_{qj}|^2 \\
&= |(1 + x\cdot y)\cdot a_{qj} - y\cdot a_{pj}|^2 + |a_{pj} - x\cdot a_{qj}|^2 \\
&= |1 + x\cdot y|^2\cdot|a_{qj}|^2 - 2\cdot\mathrm{Re}\Big[(y\cdot(1 + x\cdot y)\cdot a_{qj}\cdot a_{pj}\Big] + |y|^2\cdot|a_{pj}|^2 \\
&\quad + |a_{pj}|^2 - 2\cdot\mathrm{Re}\Big[x\cdot a_{qj}\cdot a_{pj}\Big] + |x|^2\cdot|a_{qj}|^2 \\
&= (|1+x\cdot y|^2 + |x|^2)\cdot|a_{qj}|^2 + (1 + |y|^2)\cdot|a_{pj}|^2 \\
&\quad - 2\cdot\mathrm{Re}\Big[(x + y\cdot(1+x\cdot y))\cdot a_{pj}\cdot a_{qj}\Big] \\
&\leq \Big[\max\{|1+x\cdot y|^2+|x|^2; 1 + |y|^2\} + \mathrm{Re}\,|x + y\cdot(1+x\cdot y)|\Big]\cdot(|a_{pj}|^2 + |a_{qj}|^2)
\end{aligned}$$

Nach Lemma 2.2 gilt $(|x|,|y|) \leq 1.0205144 \cdot \frac{S(A)}{\delta} \leq 1.0205144\cdot 10^{-2}$.

Es folgt:

$$\begin{aligned}
|\tilde{a}_{pj}|^2 + |\tilde{a}_{qj}|^2 &\leq (1 + 3\cdot|x|^2 + |x|^4)\cdot(|a_{pj}|^2 + |a_{qj}|^2) \\
&\leq (1 + 3.1246\cdot 10^{-4})\quad\cdot(|a_{pj}|^2 + |a_{qj}|^2)
\end{aligned}$$

●

2.6 Die Zugehörigkeit der Diagonale zu festen Eigenwerten

Bei genügend kleiner Außerdiagonale stellen die Diagonalelemente der Matrix A Näherungen für die Eigenwerte der Matrix dar. Im fortgeschrittenen Stadium des Verfahrens ist es nun nicht mehr möglich, daß ein Diagonalelement im Laufe der Transformation verschiedene Eigenwerte der Matrix A approximiert. Dies ist die Aussage des nachfolgenden Lemmas:

Lemma 2.5:

Es gelte die Voraussetzung (2.8). Sei λ_p ein zum Diagonalelement a_{pp} gehöriger Eigenwert. Dann gehört zum transformierten Block $\tilde{a}_{pp}$ ebenfalls der Eigenwert λ_p.

Beweis:

Laut Definition gilt für den Abstand eines zweiten Eigenwertes λ_q zu λ_p: $|\lambda_p - \lambda_q| \geq \delta$.
Die Transformationsformeln ergeben für die Veränderung der Diagonalelemente mit

$$\tilde{a}_{qq} = a_{qq} + x \cdot a_{qp} \quad \text{und} \quad \tilde{a}_{pp} = a_{pp} - x \cdot a_{qp}$$

und wegen $|x| \leq 1.0205144 \cdot \dfrac{S(A)}{\delta}$

$$|\tilde{a}_{ii} - a_{ii}| \leq |x| \cdot |a_{qp}| \leq 1.0205144 \cdot \frac{S^2(A)}{\delta} \leq 1.0205144 \cdot 10^{-4} \cdot \delta \quad \text{für } i=p,q$$

Wäre nun der Eigenwert λ_q zu $\tilde{a}_{pp}$ gehörig für ein beliebiges $q \neq p$,so folgte:

$$|\tilde{a}_{pp} - \lambda_q| \leq S(\tilde{A}) \quad \text{und} \quad |a_{pp} - \lambda_p| \leq S(A) \text{ also}$$

$$\begin{aligned} |\tilde{a}_{pp} - a_{pp}| &\geq |\lambda_p - \lambda_q| - |\tilde{a}_{pp} - \lambda_q| - |a_{pp} - \lambda_p| \\ &\geq \delta - S(\tilde{A}) - S(A) \\ &\geq \delta - (1 + 1.000215) \cdot S(A) \\ &\geq 0.97 \cdot \delta \end{aligned}$$

Damit folgte insgesamt der Widerspruch:

$$0.97 \cdot \delta \leq |\tilde{a}_{pp} - a_{pp}| \leq 1.0205144 \cdot 10^{-4} \cdot \delta$$

●

2.7. Nichtdiagonalelementwachstum

Als letzten Hilfssatz für den eigentlichen Konvergenzbeweis benötigen wir noch eine Abschätzung für den Zuwachs der Nichtdiagonalelemente.

Lemma 2.6:

Unter der Voraussetzung (2.8) gilt für die durch eine Transformation veränderten Nichtdiagonalelemente folgende Abschätzung:

$$\left.\begin{aligned} |\tilde{a}_{pj}| &\le (1 + 2\cdot 10^{-4})\cdot |a_{pj}| + 1.0205144\cdot\frac{S(A)}{\delta}\cdot |a_{qj}| \\ |\tilde{a}_{jp}| &\le (1 + 2\cdot 10^{-4})\cdot |a_{jp}| + 1.0205144\cdot\frac{S(A)}{\delta}\cdot |a_{jq}| \end{aligned}\right\} \quad \text{für } j \neq p,q \qquad (2.17)$$

Beweis:

Die Transformationsformeln lauten:

$$\tilde{a}_{jp} = (1 + x\cdot y)\cdot a_{jp} + y\cdot a_{jq} \quad \text{und} \quad \tilde{a}_{pj} = a_{pj} - x\cdot a_{qj}$$

Damit ergibt sich sofort

$$|\tilde{a}_{pj}| \le |a_{pj}| + 1.0205144\cdot\frac{S(A)}{\delta}\cdot |a_{qj}|$$

$$|\tilde{a}_{jp}| \le \left(1 + 1.0205144^2\cdot\frac{S^2(A)}{\delta^2}\right)\cdot |a_{jp}| + 1.0205144\cdot\frac{S(A)}{\delta}\cdot |a_{jq}|$$

Mit (2.8) erhalten wir aus der zweiten Formel bereits eine der behaupteten Abschätzungen. Die erste Abschätzung ergibt sich durch Vergröberung der ersten Formel.

●

2.8 Asymptotisch quadratische Konvergenz

Mit den Lemmata der vorangehenden Abschnitte sind wir nun in der Lage, die asymptotisch quadratische Konvergenz des Verfahrens zu zeigen.

Wir beweisen, daß bei genügend kleiner Außerdiagonalnorm $S(A^0)$ der Startmatrix $A^0 = A$ die Außerdiagonalnorm folgender Formel genügt:

$$S(A^{k+1}) \leq K(n,\delta)\cdot S^2(A^k) \quad \text{für alle } k \in \mathbb{N}_0 \text{ mit } K(n,\delta) \in \mathbb{R}^+. \tag{2.18}$$

Der Beweis beruht auf der Addition aller Außerdiagonalnormzuwächse der Transformationen einer Pivotzeile und wird induktiv über die Anzahl der Pivotzeilen geführt. Dazu definieren wir

$$s_{ij}^r := |a_{ij}^r|^2 + |a_{ji}^r|^2, \tag{2.19}$$

wobei der obere Index r das Matrixelement nach der r-ten Transformation bedeutet.

Satz 2.7:

Es sei

$$S(A^0) \leq 1.000215^{-N}\cdot\frac{\delta}{100} \quad \text{mit } N = n\cdot(n-1)/2. \tag{2.20}$$

(i) Dann existiert für alle $p \in \{0, \dots,(n-1)\}$ eine Konstante $C_p > 0$, sodaß gilt:

$$\sum_{\substack{i<j\leq n\\ 1\leq i\leq p}} s_{ij}^r \leq C_p\cdot\frac{S^4(A^0)}{\delta^2} \quad \text{,wobei } r=\varphi(p+1,p+2) \text{ ist.} \tag{2.21}$$

(ii) Das Verfahren bricht nach endlich vielen Schritten mit der Annullierung der Außerdiagonalen ab und die Konvergenzordnung ist quadratisch, d.h. es existiert ein $K(n,\delta) > 0$ mit

$$S(A^N) \leq K(n,\delta)\cdot S^2(A^0)$$

Beweis:

Die Voraussetzung (2.20) garantiert, daß alle Abschätzungen der vorangehenden Abschnitte über die gesamte Dauer eines Zyklus ihre Gültigkeit behalten, da der Außerdiagonalnormzuwachs gemäß (2.15) abschätzbar ist.

Der Beweis wird per Induktion über den Index p der Pivotzeile geführt:

Induktionsanfang (p=0):

Für p=0 ist die Aussage leer.

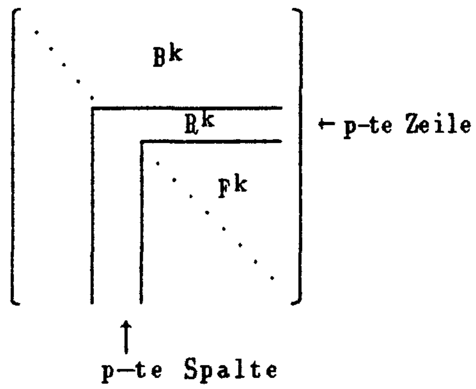
Induktionsvoraussetzung:

Für ein $p > 1$ existiere ein $C_{p-1} > 0$ mit

$$\sum_{\substack{i<j\leq n \\ 1\leq i\leq p-1}} s_{ij}^{r} \leq C_{p-1} \cdot \frac{S^4(A^0)}{\delta^2} \quad \text{mit } r = \varphi(p,p+1)$$

Induktionsschritt:

Es sei stets $r := \zeta(p,p+1)$. Für den eigentlichen Induktionsschritt unterteilen wir die Matrix A in drei Bereiche:

$A^k =$ 

für $k = r,\ldots,r+n-p$

Laut Induktionsvoraussetzung gilt bereits

$$\| B^r \|_E^2 = \sum_{\substack{i<j\leq n \\ 1\leq i\leq p-1}} s_{ij}^{r} \leq C_{p-1} \cdot \frac{S^4(A^0)}{\delta^2},$$

und es gilt natürlich $\| F^r \|_E^2 \leq \| E^r \|_E^2 \leq S^2(A^N)$.

Gesucht ist eine positive Konstante $C_p > 0$ mit

$$\| B^{r+n-p} + R^{r+n-p} \|_E^2 = \| B^{\varphi(p+1,p+2)} + R^{\varphi(p+1,p+2)} \|_E^2 \leq C_p \cdot \frac{S^4(A^0)}{\delta^2}.$$

Mit Hilfe des Lemmas (2.6) lassen sich alle Matrixelemente in der p-ten Zeile nach Verwendung der gesamten Zeile als Pivotelemente abschätzen. Es gilt:

$$|a_{pj}^{r+n-p}| \leq (1 + 2\cdot 10^{-4})^{n-j} \cdot |a_{pj}^{r+j-p}| + \sum_{i=j+1}^{n} 1.0205144^{n-i}\, 1.0205144 \cdot \frac{S(A^i)}{\delta} \cdot |a_{ij}^{r+i-p-1}|$$

$$|a_{jp}^{r+n-p}| \leq (1 + 2\cdot 10^{-4})^{n-j} \cdot |a_{jp}^{r+j-p}| + \sum_{i=j+1}^{n} 1.0205144^{n-i}\, 1.0205144 \cdot \frac{S(A^i)}{\delta} \cdot |a_{ji}^{r+i-p-1}|$$

Quadrieren der Formel mit anschließender Anwendung der Cauchy-Schwarzschen Ungleichung ergibt:

$$\begin{aligned}
|a_{pj}^{r+n-p}|^2 &\leq 2\cdot(1 + 2\cdot 10^{-4})^{2(n-j)} \cdot |a_{pj}^{r+j-p}|^2 \\
&\quad + 2\cdot\left[\sum_{i=j+1}^{n} 1.0205144^{n-i}\, 1.0205144 \cdot \frac{S(A^i)}{\delta} \cdot |a_{ij}^{r+i-p-1}| \right]^2 \\
&\leq 2\cdot(1 + 2\cdot 10^{-4})^{2(n-j)} \cdot |a_{pj}^{r+j-p}|^2 \\
&\quad + 2\cdot\left[\sum_{i=j+1}^{n} 1.0205144^{2(n-i)} \cdot \left[1.0205144 \cdot \frac{S(A^i)}{\delta} \right]^2 \cdot \sum_{i=j+1}^{n} |a_{ij}^{r+i-p-1}|^2 \right]^2 \\
&\leq 2\cdot 1.0002^{2(n-j)} \cdot \left[|a_{pj}^{r+j-p}|^2 + \sum_{i=j+1}^{n} \left[1.0205144 \cdot \frac{S(A^i)}{\delta} \right]^2 \cdot \sum_{i=j+1}^{n} |a_{ij}^{r+i-p-1}|^2 \right]
\end{aligned}$$

Eine analoge Formel gilt für $|a_{jp}^{r+n-p}|^2$. Nun können wir die Quadrate aufsummieren:

$$\sum_{j=p+1}^{n} s_{pj}^{r+n-p} \le$$

$$2\cdot\left[\sum_{j=p+1}^{n} 1.0002^{2(n-j)}\cdot s_{pj}^{r+j-p} + \sum_{j=p+1}^{n} 1.0002^{2(n-j)}\cdot \sum_{i=j+1}^{n}\sum_{j=p+1}^{n} s_{ij}^{r+i-p-1}\cdot\left[1.0205144\cdot\frac{S(A^{i})}{\delta}\right]^{2}\right]$$

$$\le$$

$$2\cdot\left[\sum_{j=p+1}^{n} 1.0002^{2(n-j)}\cdot s_{pj}^{r+j-p} + \left\{\sum_{i=p+2}^{n}\sum_{j=p+1}^{i-1} 1.0002^{2(n-j)}\cdot s_{ij}^{r+i-p-1}\right\}\cdot \sum_{i=p+2}^{n}\left[1.0205144\cdot\frac{S(A^{i})}{\delta}\right]^{2}\right]$$

Es gilt

$$\sum_{j=p+1}^{i-1} s_{ij}^{r+i-p-1} \le \| F^{r}\|_{E}^{2} \le S^{2}(A^{N})$$

Damit folgt

$$\sum_{j=p+1}^{n} s_{pj}^{r+n-p} \le 2\cdot 1.0002^{2(n-p)}\cdot$$

$$\cdot\left[\sum_{j=p+1}^{n} \cdot s_{pj}^{r+j-p} + (n-p-1)\cdot 1.0205144^{2}\cdot\left[\frac{S(A^{N})}{\delta}\right]^{2}\cdot\left\{\sum_{i=p+2}^{n}\sum_{j=p+1}^{i-1} s_{ij}^{r+i-p-1}\right\}\right]$$

Die Parameter x,y waren so gewählt, daß die Pivotelemente verschwinden, also

$$a_{pj}^{r+j-p} = a_{jp}^{r+j-p} = 0 \quad \text{und} \quad \sum_{j=p+1}^{n} s_{pj}^{r+j-p} = 0$$

Insgesamt folgt

$$\sum_{j=p+1}^{n} s_{pj}^{r+n-p} \le 2\cdot 1.0002^{2(n-p)}\cdot(n-p-1)\cdot 1.0205144^{2}\cdot\left[\frac{S(A^{N})}{\delta}\right]^{2}\cdot\sum_{i=p+2}^{n} S^{2}(A^{N})$$

$$\le 2\cdot 1.0002^{2(n-p)}\cdot(n-p-1)\cdot 1.0205144^{2}\cdot\left[\frac{S(A^{N})}{\delta}\right]^{2}\cdot(n-p-1)\cdot S^{2}(A^{N})$$

$$\underbrace{\leq 2\cdot(n-p-1)^2\cdot 1.0205144^{2(n-p)}}_{C(n\ ,p)}\cdot\frac{S^4(A^N)}{\delta^2}$$

$$\cdot\frac{S^4(A^N)}{\delta^2}$$

Damit folgt für die p-te Zeile und Spalte:

$$\| R^{r+n-p} \|_E^2 \leq C(n,p)\cdot\frac{S^4(A^N)}{\delta^2}$$

Für die Matrixelemente im bereits transformierten Bereich B gilt, daß sie bei Transformationen der p-ten Zeile entweder überhaupt nicht verändert werden oder gerade zwei Zeilen bzw. zwei Spalten gekoppelt transformiert werden. Daher können wir zur Abschätzung von $\| B^{r+n-p} \|_E^2$ das Lemma 2.5 verwenden. Es folgt:

$$\begin{aligned}\| B^{r+n-p} \|_E^2 &\leq (1 + 3.1246\cdot 10^{-4})\cdot\| B^{r+n-p-1} \|_E^2\\ &\leq \ldots \leq (1 + 3.1246\cdot 10^{-4})^{n-p}\cdot\| B^{r} \|_E^2\\ &\leq (1 + 3.1246\cdot 10^{-4})^{n-p}\cdot C_{p-1}\cdot\frac{S^4(A^N)}{\delta^2}\end{aligned}$$

Damit erhalten wir

$$\begin{aligned}\| B^{r+n-p} + R^{r+n-p} \|_E^2 &\leq (1 + 3.1246\cdot 10^{-4})^{n-p}\cdot C_{p-1}\cdot\frac{S^4(A^N)}{\delta^2} + C(n,p)\cdot\frac{S^4(A^N)}{\delta^2}\\ &\leq \underbrace{\left[(1 + 3.1246\cdot 10^{-4})^{n-p}\cdot C_{p-1} + C(n,p)\right]}_{C_p}\cdot\frac{S^4(A^N)}{\delta^2}\\ &:= \qquad C_p \cdot\frac{S^4(A^N)}{\delta^2}\end{aligned}$$

Damit ist der Induktionsbeweis geführt.

Durch sukzessive Benutzung aller Zeilen der Matrix als Pivotzeilen erhalten wir schließlich nach $N=n(n-1)$ Transformationen:

$$\| E^N \|_E^2 = \| B^N + R^N \|_E^2 \leq C_N\cdot\frac{S^4(A^N)}{\delta^2}$$

Mit $S(A^N) \leq 1.000215^N\cdot S(A^0)$ folgt insgesamt

$$\| E^N \|_E = S(A^N) \leq \sqrt{C_N}\cdot\frac{S^2(A^N)}{\delta} \leq 1.000215^N\cdot\sqrt{C_N}\cdot\frac{S^2(A^0)}{\delta}$$

oder

$$S(A^N) \leq K(n,\delta) \cdot S^2(A^0) \text{ mit } K(n,\delta) := 1.000215^N \cdot \sqrt{C_N} \cdot \frac{1}{\delta}; \quad N = n(n-1)/2$$

Damit ist die asymptotisch quadratische Konvergenz gezeigt.

•

Bemerkung:

Die im Satz 2.8 verwendete Konstante C_p ist Lösung folgender linearer Differenzengleichung

$$C_p = (1 + 3.1246 \cdot 10^{-4})^{n-p} \cdot C_{p-1} + 2(n-p-1)^2 \cdot 1.0205144^{2(n-p)} \quad \text{mit } C_0 = 0.$$

Die Lösung läßt sich abschätzen durch die Lösung der folgenden linearen Differenzengleichung mit konstanten Koeffizienten:

$$\tilde{C}_p = \underbrace{(1 + 3.1246 \cdot 10^{-4})^n}_{b} \cdot \tilde{C}_{p-1} + \underbrace{2 \cdot (n-1)^2 \cdot 1.0205144^{2n}}_{a} \text{mit } \tilde{C}_0 = 0.$$

$$:= b \cdot \tilde{C}_{p-1} + a$$

Die Lösung lautet $\tilde{C}_p = a \cdot \sum_{i=0}^{p-1} b^i$

Wegen a, b > 1 folgt schließlich die Abschätzung

$$C_N \leq \tilde{C}_N = a \cdot \sum_{i=0}^{N-1} b^i \leq a \cdot N \cdot b^N \text{ mit } N = n \cdot (n-1)/2.$$

Folglich erhält man $K(n,\delta) \leq 1.000215^N \cdot \sqrt{a \cdot N \cdot b^N} \cdot \frac{1}{\delta}$

Im nachfolgenden Kapitel werden die einzelnen Beweisschritte des asymptotischen Konvergenznachweises auf Blockmatrizen verallgemeinert und so auf ein primitives Verfahren angewendet, das für J-symmetrische Matrizen konstruiert ist. Der bis jetzt beschriebene Konvergenzbeweis ist relativ einfach ausgefallen, weil wegen der Forderung der einfachen Eigenwerte mit Mindestabstand $\delta > 0$ eine wesentliche Schwierigkeit des eigentlichen primitiven Verfahrens zur Diagonalisierung von J-symmetrischen Matrizen nicht auftritt. Es handelt sich hierbei um die Abhängigkeit aller Abschätzungskonstanten von der Norm der aktuellen Iterierten.

3. Ein primitives Verfahren für J-symmetrische Matrizen

Wir beschäftigen uns in diesem Kapitel mit einem primitiven Algorithmus zur Berechnung der Eigenwerte einer J-symmetrischen Matrix. Das Verfahren transformiert durch Annullierung von Außerdiagonalblöcken der Startmatrix A die Matrix auf Blockdiagonalgestalt. Dabei bleibt die J-Symmetrie der Ausgangsmatrix erhalten. Die Ähnlichkeitstransformation besteht aus zwei hintereinander auf jedem Pivotblock auszuführenden Transformationen, einer hyperbolischen nichtunitären Transformation und einer anschließenden orthogonalen Transformation mit jeweils zwei Parametern. Die Parameter werden so bestimmt, daß die 4 Matrixelemente des Pivotblockes näherungsweise zu Null gesetzt werden, da die exakten Gleichungen auf ein nicht explizit lösbares nichtlineares Nullstellenproblem führen.

Ziel ist es, für dieses Verfahren die asymptotisch quadratische Konvergenz nachzuweisen, d.h. bei genügend kleiner Außerdiagonalnorm $S(A) \le \epsilon$ zu zeigen, daß ein Zyklus die Größenordnung von $S(A)$ von $O(\epsilon)$ auf $O(\epsilon^2)$ reduziert. Die Pivotwahl soll dabei zeilenzyklisch durchgeführt werden. Die einzige Voraussetzung an die Ausgangsmatrix ist die Getrenntheit der Eigenräume der Diagonalblöcke. Es soll gelten

$$A \in \mathbb{R}^{2m,2m},\ m \ge 2,\ A \text{ sei J-symmetrisch bezüglich } J = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}, \tag{3.1}$$

wobei $I \in \mathbb{R}^{m,m}$ die Einheitsmatrix ist und

$$\delta := \min_{i \neq j} \operatorname{dist} \{ \mathcal{E}_i , \mathcal{E}_j \} > 0. \tag{3.2}$$

Hierbei sei $\mathcal{E}_k$ für $k = 1,\dots,m$ die Zweipunktmenge, die das Paar von Eigenwerten der Matrix A, das zum Diagonalblock A_{kk} gehört, beinhaltet. Wir gehen dabei von genügend kleiner Außerdiagonale aus, sodaß die Eigenwerte der Diagonalblöckke A_{kk} als Näherungswerte für zwei bestimmte Eigenwerte von A aufgefaßt werden können.

3.1 Die Transformationsmatrizen und ihre Eigenschaften

Die hyperbolische Transformation T ist folgendermaßen gegeben. Es sei A eine reelle J-symmetrische Blockmatrix mit dim $A := n = 2\cdot m$. Es sei weiter (p,q) für $1 \le p < q \le m$ ein fest vorgegebenes Pivotpaar.

Dann ist die Transformationsmatrix $T^{\varphi(p,q)}$ mit $\varphi(p,q)$ gemäß (3.15) für das Pivotpaar (p,q), die den Pivotblock A_{pq} annulliert, gegeben durch die Restriktion $\hat{T}_{pq}$:

$$T_{pp} := \begin{bmatrix} \cosh y_1 & 0 \\ 0 & \cosh y_2 \end{bmatrix}, \quad T_{pq} := \begin{bmatrix} 0 & \sinh y_1 \\ \sinh y_2 & 0 \end{bmatrix}$$

$$\text{mit } y_1, y_2 \in \mathbb{R} \tag{3.3}$$

$$T_{qp} := \begin{bmatrix} 0 & \sinh y_2 \\ \sinh y_1 & 0 \end{bmatrix}, \quad T_{qq} := \begin{bmatrix} \cosh y_2 & 0 \\ 0 & \cosh y_1 \end{bmatrix}$$

$$T_{ij} := I \cdot \delta_{ij} \qquad \text{für } (i,j) \neq (p,p),(p,q),(q,p),(q,q)$$

Die so definierte zweiparametrige Elementarmatrix $T(y_1,y_2)$ ist J-orthogonal, aber nicht orthogonal, so daß unter Ähnlichkeitstransformationen mit $T(y_1,y_2)$ die Matrix A zwar J-symmetrisch bleibt, aber die Norm der Matrix nicht invariant ist.
Falls es nicht von Interesse ist, auf welchem Pivotblock die Transformation T wirkt, lassen wir die Nummer $\varphi(p,q)$ der Transformation weg.

Die Matrix T ist stets regulär, und sie hat als Funktion der Parameter $y_i, \tilde{y}_i \in \mathbb{R}$ folgende Eigenschaften

$$\begin{aligned} T(0,0) &= I & T(\tilde{y}_1 + y_1, \tilde{y}_2 + y_2) &= T(\tilde{y}_1,\tilde{y}_2) + T(y_1,y_2) \\ T^{-1}(y_1,y_2) &= T(-y_1,-y_2) & T(y_1,y_2)^{T} &= T(y_1,y_2) \end{aligned} \tag{3.4}$$

Für feste Parameter $y_1, y_2 \in \mathbb{R}$ und ein festes Pivotpaar (p,q) sei die Ähnlichkeitstransformation auf A beschrieben durch

$$A' := (T^{\varphi(p,q)})^{-1}(y_1,y_2) \cdot A \cdot (T^{\varphi(p,q)})(y_1,y_2) \tag{3.5}$$

Dann ist A' wieder J-symmetrisch, und es gelten folgende Transformationsvorschriften für die einzelnen (i,j)-Blöcke der Matrix A:

$$\left.\begin{aligned} A'_{pi} &= T_{pp} \cdot A_{pi} - T_{pq} \cdot A_{qi} \\ A'_{qi} &= -T_{qp} \cdot A_{pi} + T_{qq} \cdot A_{qi} \end{aligned}\right\} \quad \text{für } i = 1, \ldots m \text{ und } i \neq p,q \tag{3.6}$$

$$\begin{aligned} A'_{pp} &= T_{pp} \cdot A_{pp} \cdot T_{pp} - T_{pq} \cdot A_{qp} \cdot T_{pp} + T_{pp} \cdot A_{pq} \cdot T_{qp} - T_{pq} \cdot A_{qq} \cdot T_{qp} \\ A'_{pq} &= T_{pp} \cdot A_{pp} \cdot T_{pq} - T_{pq} \cdot A_{qp} \cdot T_{pq} + T_{pp} \cdot A_{pq} \cdot T_{qq} - T_{pq} \cdot A_{qq} \cdot T_{qq} \\ A'_{qq} &= -T_{qp} \cdot A_{pp} \cdot T_{pq} + T_{qq} \cdot A_{qp} \cdot T_{pq} - T_{qp} \cdot A_{pq} \cdot T_{qq} + T_{qq} \cdot A_{qq} \cdot T_{qq} \end{aligned} \tag{3.7}$$

Für die p-te und q-te Blockspalte und A'_{qp} ergeben sich wegen der J-Symmetrie entsprechende Formeln. Für spätere Betrachtungen benötigen wir ebenfalls die Transformationsformeln einzelner Elemente der Matrix $A = \left[a_{ij} \right]_{1 \le i,j \le n}$. Bei der expliziten Berechnung transformierter Elemente verwendet man die J-Symmetrie der Matrix und transformiert nur die obere Dreieckshälfte der Matrix A. Deshalb geben wir hier Formeln an, die auch nur solche Matrixelemente verwenden, die in der oberen Dreieckshälfte liegen. Es gilt

$$\begin{aligned}
&\left.\begin{aligned} a'_{ip} &= a_{ip}\cdot ch_1 + a_{i,q+m}\cdot sh_1\\ a'_{i,q+m} &= a_{i,q+m}\cdot ch_1 + a_{ip}\cdot sh_1\end{aligned}\right\}\quad \text{für } i = 1,\dots,p-1\\
&\left.\begin{aligned} a'_{pi} &= a_{pi}\cdot ch_1 + a_{i,q+m}\cdot sh_1\\ a'_{i,q+m} &= a_{i,q+m}\cdot ch_1 + a_{pi}\cdot sh_1\end{aligned}\right\}\quad \text{für } i = p+1,\dots,m \text{ und } i \neq q\\
&\left.\begin{aligned} a'_{pi} &= a_{pi}\cdot ch_1 - a_{i,q+m}\cdot sh_1\\ a'_{i,q+m} &= a_{i,q+m}\cdot ch_1 - a_{pi}\cdot sh_1\end{aligned}\right\}\quad \text{für } i = m+1,\dots,q+m-1 \text{ und } i \neq p+m\\
&\left.\begin{aligned} a'_{pi} &= a_{pi}\cdot ch_1 - a_{q+m,i}\cdot sh_1\\ a'_{q+m,i} &= a_{q+m,i}\cdot ch_1 - a_{pi}\cdot sh_1\end{aligned}\right\}\quad \text{für } i = q+m-1,\dots,n
\end{aligned}\tag{3.8a}$$

$$\begin{aligned}
&\left.\begin{aligned} a'_{iq} &= a_{iq}\cdot ch_2 + a_{i,p+m}\cdot sh_2\\ a'_{i,p+m} &= a_{i,p+m}\cdot ch_2 + a_{iq}\cdot sh_2\end{aligned}\right\}\quad \text{für } i = 1,\dots,q-1 \text{ und } i \neq p\\
&\left.\begin{aligned} a'_{qi} &= a_{qi}\cdot ch_2 + a_{i,p+m}\cdot sh_2\\ a'_{i,p+m} &= a_{i,p+m}\cdot ch_2 + a_{qi}\cdot sh_2\end{aligned}\right\}\quad \text{für } i = q+1,\dots,m\\
&\left.\begin{aligned} a'_{qi} &= a_{qi}\cdot ch_2 - a_{i,p+m}\cdot sh_2\\ a'_{i,p+m} &= a_{i,p+m}\cdot ch_2 - a_{qi}\cdot sh_2\end{aligned}\right\}\quad \text{für } i = m+1,\dots,p+m-1\\
&\left.\begin{aligned} a'_{qi} &= a_{qi}\cdot ch_2 - a_{p+m,i}\cdot sh_2\\ a'_{p+m,i} &= a_{p+m,i}\cdot ch_2 - a_{qi}\cdot sh_2\end{aligned}\right\}\quad \text{für } i = p+m-1,\dots,n \text{ und } i \neq q+m
\end{aligned}\tag{3.8b}$$

und weiter gilt für die Elemente der Pivotblöcke

$$
\begin{aligned}
a'_{pp} &= a_{pp}\cdot ch_1^2 + 2\cdot a_{p,q+m}\cdot ch_1\cdot sh_1 - a_{q+m,q+m}\cdot sh_1^2\\
a'_{p,p+m} &= a_{p,p+m}\cdot ch_1\cdot ch_2 + a_{pq}\cdot ch_1\cdot sh_2 - a_{p+m,q+m}\cdot ch_2\cdot sh_1 + a_{q,q+m}\cdot sh_1\cdot sh_2\\
a'_{p+m,p+m} &= a_{p+m,p+m}\cdot ch_2^2 - 2\cdot a_{q,p+m}\cdot ch_2\cdot sh_2 - a_{q,q}\cdot sh_2^2
\end{aligned}
$$

$$
\begin{aligned}
a'_{pq} &= a_{pq}\cdot ch_1\cdot ch_2 + a_{p,p+m}\cdot ch_1\cdot sh_2 + a_{q,q+m}\cdot ch_2\cdot sh_1 - a_{p+m,q+m}\cdot sh_1\cdot sh_2\\
a'_{p,q+m} &= (a_{pp} - a_{q+m,q+m})\cdot ch_1\cdot sh_1 + a_{p,q+m}\cdot(ch_1^2 + sh_1^2)\\
a'_{q,p+m} &= (a_{qq} - a_{p+m,p+m})\cdot ch_2\cdot sh_2 + a_{q,p+m}\cdot(ch_2^2 + sh_2^2) \qquad (3.8c)\\
a'_{p+m,q+m} &= a_{p+m,q+m}\cdot ch_1\cdot ch_2 - a_{q,q+m}\cdot ch_1\cdot sh_2 - a_{p,p+m}\cdot ch_2\cdot sh_1 - a_{pq}\cdot sh_1\cdot sh_2
\end{aligned}
$$

$$
\begin{aligned}
a'_{q,q} &= a_{qq}\cdot ch_2^2 + 2\cdot a_{q,p+m}\cdot ch_2\cdot sh_2 - a_{p+m,p+m}\, sh_2^2\\
a'_{q,q+m} &= a_{q,q+m}\cdot ch_1\cdot ch_2 - a_{p+m,q+m}\cdot ch_1\cdot sh_2 + a_{p,q}\cdot ch_2\cdot sh_1 + a_{p,p+m}\cdot sh_1\cdot sh_2\\
a'_{q+m,q+m} &= a_{q+m,q+m}\cdot ch_1^2 - 2\cdot a_{p,q+m}\cdot ch_1\cdot sh_1 - a_{p,p}\cdot sh_1^2
\end{aligned}
$$

Hierbei bedeutet $ch_i = \cosh y_i$ und $sh_i = \sinh y_i$ für $i = 1,2$.

Auf die hyperbolische Transformation folgt unmittelbar bei jedem Pivotpaar (p,q) eine orthogonale Ähnlichkeitstransformation mit der Elementarmatrix $U^{\varphi(p,q)}$, die folgendermaßen gegeben ist:

$$
U_{pp} := \begin{bmatrix}\cos x_1 & 0\\ 0 & \cos x_2\end{bmatrix},\quad U_{pq} := \begin{bmatrix}-\sin x_1 & 0\\ 0 & -\sin x_2\end{bmatrix}
$$

$$
\text{mit } x_1, x_2 \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right] \qquad (3.9)
$$

$$
U_{qp} := \begin{bmatrix}\sin x_1 & 0\\ 0 & \sin x_2\end{bmatrix},\quad U_{qq} := \begin{bmatrix}\cos x_1 & 0\\ 0 & \cos x_2\end{bmatrix}
$$

$$
U_{ij} := I\cdot\delta_{ij} \qquad \text{für } (i,j) \neq (p,p),(p,q),(q,p),(q,q)
$$

Die zweiparametrige Elementarmatrix $U(x_1,x_2)$ ist J-orthogonal und orthogonal, so daß unter Ähnlichkeitstransformationen mit $U(x_1,x_2)$ die Matrix A ihre J-Symmetrie beibehält und gleichzeitig die Norm der Matrix nicht verändert wird.
Für feste Parameter $x_1,x_2 \in (-\frac{\pi}{2}, \frac{\pi}{2}]$ und ein festes Pivotpaar (p,q) sei die Ähnlichkeitstransformation auf A' beschrieben durch

$$A'' := (U^{\varphi(p,q)})^{-1}(x_1,x_2) \cdot A' \cdot (U^{\varphi(p,q)})(x_1,x_2)$$

Wie oben lassen wir die oberen Indizes $\varphi(p,q)$, die nur bei der Beschreibung eines ganzen Zyklus des Verfahrens von Bedeutung sind, hier zunächst weg. Dann transformiert sich die Matrix A' in Blockschreibweise wie folgt:

$$\left.\begin{aligned} A''_{pi} &= U_{pp} \cdot A'_{pi} - U_{pq} \cdot A'_{qi} \\ A''_{qi} &= -U_{qp} \cdot A'_{pi} + U_{qq} \cdot A'_{qi}\,, \end{aligned}\right\} \quad \text{für } i = 1, \dots m \text{ und } i \neq p,q \tag{3.10}$$

$$\begin{aligned} A''_{pp} &= U_{pp} \cdot A'_{pp} \cdot U_{pp} - U_{pq} \cdot A'_{qp} \cdot U_{pp} + U_{pp} \cdot A'_{pq} \cdot U_{qp} - U_{pq} \cdot A'_{qq} \cdot U_{qp} \\ A''_{pq} &= U_{pp} \cdot A'_{pp} \cdot U_{pq} - U_{pq} \cdot A'_{qp} \cdot U_{pq} + U_{pp} \cdot A'_{pq} \cdot U_{qq} - U_{pq} \cdot A'_{qq} \cdot U_{qq} \\ A''_{qq} &= -U_{qp} \cdot A'_{pp} \cdot U_{pq} + U_{qq} \cdot A'_{qp} \cdot U_{pq} - U_{qp} \cdot A'_{pq} \cdot U_{qq} + U_{qq} \cdot A'_{qq} \cdot U_{qq} \end{aligned}$$

Für spätere Betrachtungen sind folgende zwei Eigenschaften der Transformation von Bedeutung, die sich aufgrund der Orthogonalität von U ergeben:

$$\begin{aligned} \| A''_{pi} \|_F^2 + \| A''_{qi} \|_F^2 &= \| A'_{pi} \|_F^2 + \| A'_{qi} \|_F^2 \\ \| A''_{ip} \|_F^2 + \| A''_{iq} \|_F^2 &= \| A'_{ip} \|_F^2 + \| A'_{iq} \|_F^2 \end{aligned} \quad \text{für } i = 1,\dots m \text{ und } i \neq p,q. \tag{3.11}$$

Wegen der J-Symmetrie von A folgt:

$$S^2(A'') - S^2(A') = 2 \cdot \| A''_{pq} \|_F^2 - 2 \cdot \| A'_{pq} \|_F^2 \tag{3.12}$$

Schließlich benötigen wir ebenfalls die Transformationsformeln einzelner Elemente der Matrix $A' = \left[a'_{ij} \right]_{1 \le i,j \le n}$. Wir geben die Formeln bezüglich der oberen Dreieckshälfte der Matrix A an. Es gilt

$$\left.\begin{aligned} a''_{ip} &= a'_{ip} \cdot c_1 - a'_{iq} \cdot s_1 \\ a''_{iq} &= a'_{iq} \cdot c_1 + a'_{ip} \cdot s_1 \end{aligned}\right\} \text{für } i = 1,\ldots,p-1$$

$$\left.\begin{aligned} a''_{pi} &= a'_{pi} \cdot c_1 - a'_{iq} \cdot s_1 \\ a''_{iq} &= a'_{iq} \cdot c_1 + a'_{pi} \cdot s_1 \end{aligned}\right\} \text{für } i = p+1,\ldots,q-1$$

$$\left.\begin{aligned} a''_{pi} &= a'_{pi} \cdot c_1 - a'_{qi} \cdot s_1 \\ a''_{qi} &= a'_{qi} \cdot c_1 + a'_{pi} \cdot s_1 \end{aligned}\right\} \text{für } i = q+1,\ldots,m \qquad (3.13a)$$

$$\left.\begin{aligned} a''_{p,i+m} &= a'_{p,i+m} \cdot c_1 - a'_{q,i+m} \cdot s_1 \\ a''_{q,i+m} &= a'_{q,i+m} \cdot c_1 + a'_{p,i+m} \cdot s_1 \end{aligned}\right\} \text{für } i = 1,\ldots,m \text{ und } i \neq p,q$$

$$\left.\begin{aligned} a''_{i,p+m} &= a'_{i,p+m} \cdot c_2 - a'_{i,q+m} \cdot s_2 \\ a''_{i,q+m} &= a'_{i,q+m} \cdot c_2 + a'_{i,p+m} \cdot s_2 \end{aligned}\right\} \text{für } i = m+1,\ldots,p+m-1$$

$$\left.\begin{aligned} a''_{p+m,i} &= a'_{p+m,i} \cdot c_2 - a'_{i,q+m} \cdot s_2 \\ a''_{i,q+m} &= a'_{i,q+m} \cdot c_2 + a'_{p+m,i} \cdot s_2 \end{aligned}\right\} \text{für } i = p+m+1,\ldots,q+m-1$$

$$\left.\begin{aligned} a''_{p+m,i} &= a'_{p+m,i} \cdot c_2 - a'_{q+m,i} \cdot s_2 \\ a''_{q+m,i} &= a'_{q+m,i} \cdot c_2 + a'_{p+m,i} \cdot s_2 \end{aligned}\right\} \text{für } i = q+m+1,\ldots,n \qquad (3.13b)$$

$$\left.\begin{aligned} a''_{i,p+m} &= a'_{i,p+m} \cdot c_2 - a'_{i,q+m} \cdot s_2 \\ a''_{i,q+m} &= a'_{i,q+m} \cdot c_2 + a'_{i,p+m} \cdot s_2 \end{aligned}\right\} \text{für } i = 1,\ldots,m \text{ und } i \neq p,q$$

und für die Elemente der Pivotblöcke gilt weiter

$$\begin{aligned}
a''_{p,p} &= a'_{p,p}\cdot c_1^2 - 2\cdot a'_{p,q}\cdot c_1\cdot s_1 + a_{q,q}\cdot s_1^2 \\
a''_{p,p+m} &= a'_{p,p+m}\cdot c_1\cdot c_2 - a'_{q,p+m}\cdot c_2\cdot s_1 - a'_{p,q+m}\cdot c_1\cdot s_2 + a'_{q,q+m}\cdot s_1\cdot s_2 \\
a''_{p+m,p+m} &= a'_{p+m,p+m}\cdot c_2^2 - 2\cdot a'_{p+m,q+m}\cdot c_2\cdot s_2 + a'_{q+m,q+m}\cdot s_2^2
\end{aligned}$$

$$\begin{aligned}
a''_{p,q} &= a'_{p,q}\cdot(c_1^2 - s_1^2) - a'_{q,q}\cdot c_1\cdot s_1 + a'_{p,p}\cdot c_1\cdot s_1 \\
a''_{p,q+m} &= a'_{p,p+m}\cdot c_1\cdot s_2 - a'_{q,p+m}\cdot s_1\cdot s_2 + a'_{p,q+m}\cdot c_1\cdot c_2 - a'_{q,q+m}\cdot s_1\cdot c_2 \\
a''_{q,p+m} &= a'_{q,p+m}\cdot c_1\cdot c_2 - a'_{q,q+m}\cdot c_1\cdot s_2 + a'_{p,p+m}\cdot c_2\cdot s_1 - a'_{p,q+m}\cdot s_1\cdot s_2 \\
a''_{p+m,q+m} &= a'_{p+m,p+m}\cdot c_2\cdot s_2 + a'_{p+m,q+m}\cdot(c_2^2 - s_2^2) - a'_{q+m,q+m}\cdot c_2\cdot s_2
\end{aligned} \tag{3.13c}$$

$$\begin{aligned}
a''_{q,q} &= a'_{q,q}\cdot c_1^2 + 2\cdot a'_{p,q}\cdot c_1\cdot s_1 + a'_{p,p}\cdot s_1^2 \\
a''_{q,q+m} &= a'_{q,q+m}\cdot c_1\cdot c_2 + a'_{q,p+m}\cdot c_1\cdot s_2 + a'_{p,q+m}\cdot c_2\cdot s_1 + a'_{p,p+m}\cdot s_1\cdot s_2 \\
a''_{q+m,q+m} &= a'_{q+m,q+m}\cdot c_2^2 + 2\cdot a'_{p+m,q+m}\cdot c_2\cdot s_2 + a'_{p+m,p+m}\cdot s_2^2
\end{aligned}$$

Hierbei bedeutet $c_i = \cos x_i$ und $s_i = \sin x_i$ für $i = 1{,}2$.

3.2 Definition eines Zyklus des Verfahrens

Sei $A \in \mathbb{R}^{2m,2m}$ und $(p,q) \in Q_m := \{(i,j) | \; 1 \le i < j \le m\}$ ein gegebenes Pivotpaar. Die Pivotelemente werden in zeilenzyklischer Reihenfolge gewählt:

$$\begin{array}{ccccc} (1,2) & (1,3) \ldots & (1,m-1) & (1 & ,m) \\ & (2,3) \ldots & (2,m-1) & (2 & ,m) \\ & & \ldots & & \\ & & & (m-1 & ,m) \end{array} \tag{3.14}$$

Ein vollständiger Zyklus besteht folglich aus $M := m \cdot (m-1)/2$ Transformationen, sodaß alle Nichtdiagonalblöcke einmal als Pivotblöcke benutzt werden. Setzen wir

$$\varphi\colon Q_m \to \{1,\ldots,M\} \quad \varphi(p,q) := q - p + \sum_{i=1}^{p-1} (m-i), \tag{3.15}$$

so werden die Matrixblöcke A_{pq} und A_{qp} als $\varphi(p,q)$-te Pivotelemente annulliert. Sei A^i die Matrix nach dem i-ten Zyklus. Über den (i+1)-ten Zyklus wird die Matrix $A^i \in \mathbb{R}^{2m,2m}$ wie folgt transformiert:

$$\begin{aligned} A^0 &:= A \\ A^{k+1,i} &:= \left[U^{k,i}\right]^{-1} \cdot \left[T^{k,i}\right]^{-1} \cdot A^{k,i} \cdot T^{k,i} \cdot U^{k,i} \text{ für } k \in \{1,..,M\} \, ; \, i \in \mathbb{N} \\ A^{i+1} &:= A^{M,i} \end{aligned} \tag{3.16}$$

Hierbei sind $T^{k,i}$ und $U^{k,i}$ die Transformationsmatrizen der k-ten Transformation im i-ten Zyklus. Der zugehörige Pivotblock ist durch $k = \varphi(p,q)$ gekennzeichnet. Damit ist A^i die Startmatrix des i-ten Zyklus und $A^{i+1} = A^{M,i}$ die Matrix nach dem i-ten Zyklus, da ein Zyklus aus M Ähnlichkeitstransformationen besteht.

3.3 Die Eigenwerte der Blockdiagonale als Näherungen für die exakten Eigenwerte

Die beiden Ähnlichkeitstransformationen werden dazu benutzt, die Blockaußerdiagonale der Matrix A zu annullieren. Dies wird dadurch erreicht, daß die 4 Parameter der beiden Transformationen so gewählt werden, daß der Pivotblock näherungsweise annulliert wird. Im Laufe der Iteration soll damit die Reduktion der Startmatrix A auf Blockdiagonalgestalt erreicht werden. Wie im einfachen Fall der Diagonalisierung einer Matrix braucht man ein Kriterium, in welchem Zusammenhang die Eigenwerte der 2x2-Matrizen der Blockdiagonale zu den exakten Eigenwerten der Matrix A stehen.

In diesem Abschnitt beschreiben wir das für das Annullierungsverfahren grundlegende Theorem, das den Abstand der Eigenwerte der Blockdiagonale zu den exakten Eigenwerten der Matrix A abschätzt. Neben der Größe der Außerdiagonalelemente $S(A)$ geht zusätzlich das Henrici-Maß $\Delta_F(A_{jj})$ der Diagonalblöcke A_{jj} in die Abschätzung ein. Das Theorem stellt eine Blockverallgemeinerung der Eigenwerteinschließungsaussagen von Gerschgorin dar und stammt von Meyer-Veselić [13].

Theorem 3.1

Das Spektrum einer Blockmatrix $A \in \mathbb{R}^{2m,2m}$ der Form $A = \left[A_{ij} \right]_{1 \le i,j \le m}$ ist enthalten in der Menge H:

$$H = \bigcup_{j=1}^{m} H_j \quad \text{mit}$$

$$H_j = \bigcup_{i=1}^{2} \left\{ z \in \mathbb{C} \mid |\mu_i^{(j)} - z| \le \frac{y_j}{\left(\sqrt{y_j + \frac{1}{4}} - \frac{1}{2}\right)} \cdot \sum_{\substack{k=1 \\ k \ne j}}^{m} \| A_{jk} \|_F \right\} \quad \text{bzw.} \tag{3.17a}$$

$$H_j = \bigcup_{i=1}^{2} \left\{ z \in \mathbb{C} \mid |\mu_i^{(j)} - z| \le \left[1 + \left(\sqrt{y_j + \frac{1}{4}} - \frac{1}{2}\right) \right] \cdot \sum_{\substack{k=1 \\ k \ne j}}^{m} \| A_{jk} \|_F \right\}. \tag{3.17b}$$

Hierbei ist $\mu_i^{(j)}$ der i - te Eigenwert des Diagonalblockes A_{jj} für $i = 1,2$, und es ist

$$y_j = \frac{\Delta_F(A_{jj})}{\sum\limits_{\substack{k=1 \\ k \neq j}}^{m} \| A_{jk} \|_F}. \tag{3.18}$$

Die Norm $\| \ \|_F$ ist dabei die Frobenius - Norm und Δ_F das in der Frobenius - Norm berechnete Henrici - Maß.

<u>Beweis:</u>

Der Beweis ist bei Meyer [14] nachzulesen.

<u>Bemerkung:</u>

Die Blockverallgemeinerung der Einschließungsaussagen nach Gerschgorin lassen sich auf beliebige Blockeinteilungen der Matrix A anwenden. In unserem Spezialfall haben alle Matrixblöcke die Dimension 2. Im Falle normaler Diagonalblöcke A_{jj} ist das Henrici-Maß $\Delta_F(A_{jj}) = 0$ und damit $y_j = 0$, so daß die Einschließungsmengen H_j sich zu direkten Blockverallgemeinerungen der Gerschgorin - Kreisen reduzieren und lediglich die Norm der Außerdiagonalblöcke einer Zeile eine Rolle spielt.

<u>Bemerkung:</u>

Für nichtnormale Diagonalblöcke A_{jj} genügt es, die Größenordnung des Produktes

$$\Delta_F(A_{jj}) \cdot \sum_{\substack{k=1 \\ k \neq j}}^{m} \| A_{jk} \|_F$$

zu kontrollieren. Ein iteratives Verfahren, daß eine Folge von Matrizen $(A^i)_{i \in \mathbb{N}}$ erzeugt, für die

$$\lim_{i\to\infty} \Delta_F(A_{jj}^i) \cdot \sum_{\substack{k=1\\k\neq j}}^{m} \| A_{jk}^i \|_F = 0 \quad \text{für alle } j \in 1,\dots m \tag{3.19}$$

gilt, transformiert eine Startmatrix A auf Blockdiagonalgestalt und die Eigenwerte der Blockdiagonale approximieren die exakten Eigenwerte der Matrix A. Die Genauigkeit der Näherungen für die Eigenwerte ist durch die rechte Seite der Ungleichung in der Definition der H_j gegeben. Sie besagt, daß einfache Eigenwerte $\mu_i^{(j)}$ einen Fehler der Größenordnung $S(A)$ im Vergleich zu den exakten Eigenwerten λ_i haben. Bei doppelten, defektiven Eigenwerten $\mu_i^{(j)}$ sinkt die Genauigkeit auf die Größenordung $\sqrt{S(A)}$. Dazu schauen wir uns das Henrici-Maß $\Delta_F(A)$ genauer an.

J-symmetrische 2x2-Matrizen sind leicht zu charakterisieren. Es gilt

<u>Hilfssatz:</u>

Sei $A = \begin{bmatrix} a & b \\ -b & d \end{bmatrix}$ eine reelle, J-symmetrische Matrix.

A ist genau dann normal, falls $a = d$ oder $b = 0$ gilt.

A ist genau dann defektiv, falls $| a - d | = 2\cdot|b| \neq 0$.

Das Henrici-Maß als Maß für die Abweichung von der Normalität ist gegeben durch

$$\Delta_F(A) = \min\left\{ | a - d | , 2\cdot|b|\right\} \leq \| A \|_F \tag{3.20}$$

<u>Beweis:</u>

Für eine reelle, normale Matrix gilt $A^T\cdot A = A\cdot A^T$. Es ist

$$A^T\cdot A = \begin{bmatrix} (a^2+b^2) & b(a-d) \\ b(a-d) & (b^2+d^2) \end{bmatrix} \text{ und } A\cdot A^T = \begin{bmatrix} (a^2+b^2) & b(d-a) \\ b(d-a) & (b^2+d^2) \end{bmatrix}.$$

Es folgt

$$A^T\cdot A = A\cdot A^T \Leftrightarrow b\cdot(a-d) = 0 \Leftrightarrow b = 0 \text{ oder } (a-d) = 0$$

Eine defektive 2x2-Matrix muß einen doppelten Eigenwert λ besitzen und es muß $b \neq 0$ gelten. Es folgt

$$(a - \lambda)^2 - b^2 = 0 \quad (d - \lambda)^2 - b^2 = 0 \text{ und } b\cdot(a + d - 2\cdot\lambda) = 0.$$

Es ergibt sich somit $\lambda = (a + d)/2$ und damit

$$|\, a - d \,| = 2\cdot|b| \neq 0.$$

Das Henrici-Maß $\Delta_F(A)$ in der Frobenius-Norm ist gegeben durch

$$\Delta_F(A) = \left\{ \|A\|_F^2 - \sum_{i=1}^{2} \mu_i^2 \right\}^{\frac{1}{2}} \leq \|A\|_F,$$

wobei μ_1 und μ_2 die beiden Eigenwerte von A sind mit

$$\mu_{1,2} = \frac{a + d}{2} \pm \frac{1}{2}\cdot\left[(a - d)^2 - 4\cdot b^2 \right]^{\frac{1}{2}}$$

und $\|A\|_F = \left[a^2 + 2\cdot b^2 + d^2 \right]^{\frac{1}{2}}$ gilt.

Für $|\, a - d \,| \geq 2\cdot|b|$ sind die Eigenwerte reell, und es gilt

$$\sum_{i=1}^{2} \mu_i^2 = a^2 - 2\cdot b^2 + d^2$$

Für $|\, a - d \,| < 2\cdot|b|$ sind die Eigenwerte komplex konjugiert, und es gilt

$$\sum_{i=1}^{2} |\mu_i|^2 = 2\cdot(\mathrm{Re}\,\mu_1)^2 + 2\cdot(\mathrm{Im}\,\mu_1)^2 = 2\cdot a\cdot d + 2\cdot b^2$$

Es folgt

$$\|A\|_F^2 - \sum_{i=1}^{2} |\mu_i| = \begin{cases} (a-d)^2 & \text{für } |\, a - d \,| < 2\cdot|b| \\ 4\cdot b^2 & \text{für } |\, a - d \,| \geq 2\cdot|b| \end{cases}$$

Damit gilt

$$\Delta_F(A) = \min\left\{ |\, a - d \,|, 2\cdot|b| \right\}$$

Bemerkung:

Die Zeilensumme einer Außerdiagonale ist durch die Außerdiagonalnorm S(A) abschätzbar. Es gilt

$$\sum_{\substack{k=1\\k\neq j}}^{m} \| A_{jk} \|_F \leq \sqrt{m-1} \cdot \left\{ \sum_{\substack{k=1\\k\neq j}}^{m} \| A_{jk} \|_F^2 \right]^{\frac{1}{2}} \leq \frac{\sqrt{m-1}}{\sqrt{2}} \cdot S(A)$$

Daher genügt es, innerhalb des Verfahrens die Produkte $\Delta_F(A_{jj}) \cdot S(A)$ für $j = 1,\ldots,m$ zu kontrollieren. Der asymptotische Konvergenzbeweis beruht im wesentlichen darauf, daß bei genügend kleiner Außerdiagonale S(A) die Henrici-Maße der Blockdiagonale beschränkt werden können. In diesem Fall stimmt die Genauigkeit der Näherungswerte für die Eigenwerte und die Konvergenzgeschwindigkeit des iterativen Verfahrens mit der Größenordnung von S(A) überein.

Bemerkung:

Theorem 3.1 erlaubt, Näherungen für die exakten Eigenwerte auszurechnen, obwohl die Norm der Matrix während der Iteration anwächst. Die bekannten Jacobiähnlichen Verfahren benutzen entweder unitäre und damit nicht normverändernde Transformationen, oder im Falle nichtunitärer Transformationen werden die Parameter so bestimmt, daß eine Normreduzierung garantiert ist. Der Satz von Meyer-Veselić zeigt, daß ein beschränktes Normwachstum durchaus möglich ist und trotzdem Approximationen der Eigenwerte berechnet werden können. Diesen Aspekt werden wir ausnutzen, um einfachere Parameterbestimmungen verwenden zu können, die die Rechengeschwindigkeit des Verfahrens positiv beeinflussen, aber ansonsten keine Normreduzierung garantieren.

3.4 Die näherungsweise Bestimmung der Parameter zur Annullierung des Pivotblockes

In diesem Abschnitt wollen wir beschreiben, wie die 4 Parameter y_i und x_i der hyperbolischen Transformation *T* bzw. der orthogonalen Transformation *U* zusammenwirken müssen, damit die vier Matrixelemente des Pivotblockes A_{pq} annulliert werden. Die Transformationsformeln (3.8) und (3.13) ergeben für die Elemente nach beiden Transformationen unter Verwendung der Abkürzung $s_i = \sin x_i$, $c_i = \cos x_i$, $sh_i = \sinh y_i$ sowie $ch_i = \cosh y_i$ für $i = 1,2$ folgende Formeln:

$$\begin{aligned} a''_{pq} &= \cos(2x_1)\cdot a'_{pq} - 0.5\cdot\sin(2x_1)\cdot(a'_{qq} - a'_{pp}) \\ &= \cos(2x_1)\cdot\Big[ch_1\cdot ch_2\cdot a_{pq} + ch_2\cdot sh_1\cdot a_{q,q+m} + ch_1\cdot sh_2\cdot a_{p,p+m} - sh_1\cdot sh_2\cdot a_{p+m,q+m}\Big] \\ &\quad - \frac{1}{2}\cdot\sin(2x_1)\cdot\Big(ch_2^2\cdot a_{qq} + 2\cdot ch_2\cdot sh_2\cdot a_{q,p+m} - sh_2^2\cdot a_{p+m,p+m} \\ &\qquad - ch_1^2\cdot a_{pp} - 2\cdot ch_1\cdot sh_1\cdot a_{p,q+m} + sh_1^2\cdot a_{q+m,q+m}\Big) =: F_1(x_1,y_1,y_2,x_2) \end{aligned} \tag{3.21a}$$

$$\begin{aligned} a''_{p,q+m} &= a'_{p,p+m}\cdot\cos(x_1)\cdot\sin(x_2) - a'_{q,p+m}\cdot\sin(x_1)\cdot\sin(x_2) + \\ &\qquad a'_{p,q+m}\cdot\cos(x_1)\cdot\cos(x_2) - a'_{q,q+m}\cdot\sin(x_1)\cdot\cos(x_2) \\ &= c_1\cdot s_2\cdot\Big(a_{p,p+m}\cdot ch_1\cdot ch_2 + a_{p,q}\cdot ch_1\cdot sh_2 - a_{p+m,q+m}\cdot ch_2\cdot sh_1 + a_{q,q+m}\cdot sh_1\cdot sh_2\Big) \\ &+ s_1\cdot s_2\cdot\Big((a_{p+m,p+m} - a_{q,q})\cdot ch_2\cdot sh_2 - a_{q,p+m}\cdot(ch_2^2 + sh_2^2)\Big) \\ &+ c_1\cdot c_2\cdot\Big((a_{p,p} - a_{q+m,q+m})\cdot ch_1\cdot sh_1 + a_{p,q+m}\cdot(ch_1^2 + sh_1^2)\Big) \\ &- s_1\cdot c_2\cdot\Big(a_{q,q+m}\cdot ch_1\cdot ch_2 - a_{p+m,q+m}\cdot ch_1\cdot sh_2 + a_{p,q}\cdot ch_2\cdot sh_1 + a_{p,p+m}\cdot sh_1\cdot sh_2\Big) \\ &=: F_2(x_1,y_1,y_2,x_2) \end{aligned} \tag{3.21b}$$

$$\begin{aligned} a''_{p+m,q} = -a''_{q,p+m} &= a'_{q,q+m}\cdot\cos(x_1)\cdot\sin(x_2) + a'_{p,q+m}\cdot\sin(x_1)\cdot\sin(x_2) - \\ &\qquad a'_{q,p+m}\cdot\cos(x_1)\cdot\cos(x_2) - a'_{p,p+m}\cdot\sin(x_1)\cdot\cos(x_2) \\ &= c_1\cdot s_2\cdot\Big(a_{q,q+m}\cdot ch_1\cdot ch_2 - a_{p+m,q+m}\cdot ch_1\cdot sh_2 + a_{p,q}\cdot ch_2\cdot sh_1 + a_{p,p+m}\cdot sh_1\cdot sh_2\Big) \\ &+ s_1\cdot s_2\cdot\Big((a_{p,p} - a_{q+m,q+m})\cdot ch_1\cdot sh_1 + a_{p,q+m}\cdot(ch_1^2 + sh_1^2)\Big) \\ &- c_1\cdot c_2\cdot\Big((a_{q,q} - a_{p+m,p+m})\cdot ch_2\cdot sh_2 + a_{q,p+m}\cdot(ch_2^2 + sh_2^2)\Big) \\ &- s_1\cdot c_2\cdot\Big(a_{p,p+m}\cdot ch_1\cdot ch_2 + a_{p,q}\cdot ch_1\cdot sh_2 - a_{p+m,q+m}\cdot ch_2\cdot sh_1 + a_{q,q+m}\cdot sh_1\cdot sh_2\Big) \\ &=: F_3(x_1,y_1,y_2,x_2) \end{aligned} \tag{3.21c}$$

$$
\begin{aligned}
a''_{p+m,q+m} &= \cos(2x_2)\cdot a'_{p+m,q+m} - 0.5\cdot\sin(2x_2)\cdot(a'_{q+m,q+m} - a'_{p+m,p+m}) \\
&= \cos(2x_2)\cdot\Big[ch_1\cdot ch_2\cdot a_{p+m,q+m} - ch_2\cdot sh_1\cdot a_{p,p+m} - ch_1\cdot sh_2\cdot a_{q,q+m} - sh_1\cdot sh_2\cdot a_{p,q} \Big] \\
&\quad -\frac{1}{2}\cdot\sin(2x_2)\cdot\Big(ch_1^2\cdot a_{q+m,q+m} - 2\cdot ch_1\cdot sh_1\cdot a_{p,q+m} - sh_1^2\cdot a_{p,p} \\
&\qquad - ch_2^2\cdot a_{p+m,p+m} + 2\cdot ch_2\cdot sh_2\cdot a_{q,p+m} + sh_2^2\cdot a_{q,q} \Big) =: F_4(x_1,y_1,y_2,x_2)
\end{aligned}
\tag{3.21d}
$$

Diese vier Transformationsformeln bilden ein nichtlineares Gleichungssystem in den vier Parametern der beiden Transformationen.

$$
\begin{aligned}
F : \mathbb{R}^4 &\to \mathbb{R}^4 ; \\
\boldsymbol{x} := (x_1,y_1,y_2,x_1) &\to (F_1,F_2,F_3,F_4)^T_{(\boldsymbol{x})} = \boldsymbol{0}
\end{aligned}
\tag{3.22}
$$

Eine exakte Lösung ist nicht möglich. Innerhalb des Verfahrens müssen wir uns daher mit einer näherungsweisen Bestimmung der Parameter begnügen. Da die vier Funktionen beliebig oft differenzierbar sind, bietet sich das Newton - Verfahren zur Approximation der exakten Lösung an, da wir an einem quadratisch konvergenten Verfahren interessiert sind.

Da nichts Näheres über die Lage der Nullstelle im $\mathbb{R}^4$ bekannt ist, starten wir das Verfahren im Nullpunkt und es kommt im wesentlichen auf das Verhalten der partiellen Ableitungen in Null an. Ein schnelles Verfahren wird auf eine sehr genaue Bestimmung der Nullstelle verzichten müssen und wird lediglich einen Schritt des Newton - Verfahrens durchführen. Eine genauere Berechnung der Nullstelle ist ohnehin nicht von Bedeutung, da selbst exakt erzeugte Nullen im Pivotblock bereits in der nachfolgenden Transformation in der Größenordnung der Außerdiagonalnorm quadratisch gestört werden und sofort verschwinden. Deshalb benötigen wir lediglich eine quadratische Verkleinerung des Pivotblockes. Eine genauere Berechnung der Nullstelle ist nicht nötig.

Die Komponentenfunktionen F_1 und F_4 haben ebenso wie F_2 und F_3 eine ähnliche Struktur. Deswegen geben wir an dieser Stelle lediglich für F_1 und F_2 die partiellen Ableitungen an, die für das Newton - Verfahren benötigt werden. Wir verwenden dabei abkürzend $s_i = \sin x_i$, $c_i = \cos x_i$, $sh_i = \sinh y_i$ sowie $ch_i = \cosh y_i$. Es gilt:

$$\frac{\partial F_1}{\partial x_1}(x_1,y_1,y_2,x_2) =$$

$$-2\cdot\sin(2\cdot x_1)\cdot\left(a_{p,q}\cdot ch_1\cdot ch_2 + a_{p,p+m}\cdot ch_1\cdot sh_2 + a_{q,q+m}\cdot ch_2\cdot sh_1 - a_{p+m,q+m}\cdot sh_1\cdot sh_2 \right)$$

$$- \cos(2\cdot x_1)\cdot\Big(a_{q,q}\cdot ch_2^2 + 2\cdot a_{q,p+m}\cdot ch_2\cdot sh_2 - a_{p+m,p+m} sh_2^2 \qquad (3.23a)$$
$$- a_{p,p}\cdot ch_1^2 - 2\cdot a_{p,q+m}\cdot ch_1\cdot sh_1 + a_{q+m,q+m}\cdot sh_1^2 \Big)$$

und $\dfrac{\partial F_1}{\partial x_1}(0) = (a_{p,p} - a_{q,q})$

$$\frac{\partial F_1}{\partial y_1}(x_1,y_1,y_2,x_2) =$$

$$= \cos(2x_1)\cdot \left(sh_1\cdot ch_2\cdot a_{p,q} + ch_2\cdot ch_1\cdot a_{q,q+m} + sh_1\cdot sh_2\cdot a_{p,p+m} - ch_1\cdot sh_2\cdot a_{p+m,q+m}\right)$$

$$- \frac{1}{2}\cdot\sin(2x_1)\cdot\left(\sinh(2y_1)\cdot(a_{q+m,q+m} - a_{p,p}) - 2\cdot\cosh(2y_1)\cdot a_{p,q+m}\right) \qquad (3.23b)$$

und $\dfrac{\partial F_1}{\partial y_1}(0) = a_{q,q+m}$

$$\frac{\partial F_1}{\partial y_2}(x_1,y_1,y_2,x_2) =$$

$$= \cos(2x_1)\cdot \left(ch_1\cdot sh_2\cdot a_{p,q} + sh_2\cdot sh_1\cdot a_{q,q+m} + ch_1\cdot ch_2\cdot a_{p,p+m} - sh_1\cdot ch_2\cdot a_{p+m,q+m}\right)$$

$$- \frac{1}{2}\cdot\sin(2x_1)\cdot\left(\sinh(2y_2)\cdot(a_{q,q} - a_{p+m,p+m}) - 2\cdot\cosh(2y_2)\cdot a_{q,p+m}\right) \qquad (3.23c)$$

und $\dfrac{\partial F_1}{\partial y_2}(0) = a_{p,p+m}$

$$\frac{\partial F_1}{\partial x_2}(x_1,y_1,y_2,x_2) = 0 \tag{3.23d}$$

$$\frac{\partial F_2}{\partial x_1}(x_1,y_1,y_2,x_2) =$$

$$\begin{aligned}
= -\, s_1 \cdot s_2 \cdot &\Big(a_{p,p+m} \cdot ch_1 \cdot ch_2 + a_{p,q} \cdot ch_1 \cdot sh_2 - a_{p+m,q+m} \cdot ch_2 \cdot sh_1 + a_{q,q+m} \cdot sh_1 \cdot sh_2 \Big) \\
+\, c_1 \cdot s_2 \cdot &\Big((a_{p+m,p+m} - a_{q,q}) \cdot ch_2 \cdot sh_2 - a_{q,p+m} \cdot (ch_2^2 + sh_2^2) \Big) \\
-\, s_1 \cdot c_2 \cdot &\Big((a_{p,p} - a_{q+m,q+m}) \cdot ch_1 \cdot sh_1 + a_{p,q+m} \cdot (ch_1^2 + sh_1^2) \Big) \\
-\, c_1 \cdot c_2 \cdot &\Big(a_{q,q+m} \cdot ch_1 \cdot ch_2 - a_{p+m,q+m} \cdot ch_1 \cdot sh_2 + a_{p,q} \cdot ch_2 \cdot sh_1 + a_{p,p+m} \cdot sh_1 \cdot sh_2 \Big)
\end{aligned} \tag{3.24a}$$

und $\dfrac{\partial F_2}{\partial x_1}(\mathit{0}) = -a_{q,q+m}$

$$\frac{\partial F_2}{\partial y_1}(x_1,y_1,y_2,x_2) =$$

$$\begin{aligned}
= c_1 \cdot s_2 \cdot &\Big(a_{p,p+m} \cdot sh_1 \cdot ch_2 + a_{p,q} \cdot sh_1 \cdot sh_2 - a_{p+m,q+m} \cdot ch_2 \cdot ch_1 + a_{q,q+m} \cdot ch_1 \cdot sh_2 \Big) \\
+\, c_1 \cdot c_2 \cdot &\Big((a_{p,p} - a_{q+m,q+m}) \cdot \cosh(2y_1) + a_{p,q+m} \cdot 2 \cdot \sinh(2y_1) \Big) \\
-\, s_1 \cdot c_2 \cdot &\Big(a_{q,q+m} \cdot sh_1 \cdot ch_2 - a_{p+m,q+m} \cdot sh_1 \cdot sh_2 + a_{p,q} \cdot ch_2 \cdot ch_1 + a_{p,p+m} \cdot ch_1 \cdot sh_2 \Big)
\end{aligned} \tag{3.24b}$$

und $\dfrac{\partial F_2}{\partial y_1}(\mathit{0}) = (a_{p,p} - a_{q+m,q+m})$

$$\frac{\partial F_2}{\partial y_2}(x_1,y_1,y_2,x_2) =$$

$$\begin{aligned}
= \quad c_1 \cdot s_2 \cdot &\Big(a_{p,p+m} \cdot ch_1 \cdot sh_2 + a_{p,q} \cdot ch_1 \cdot ch_2 - a_{p+m,q+m} \cdot sh_2 \cdot sh_1 + a_{q,q+m} \cdot sh_1 \cdot ch_2 \Big) \\
-\, s_1 \cdot s_2 \cdot &\Big((a_{q,q} - a_{p+m,p+m}) \cdot \cosh(2y_2) + a_{q,p+m} \cdot 2 \cdot \sinh(2y_2) \Big) \\
-\, s_1 \cdot c_2 \cdot &\Big(a_{q,q+m} \cdot ch_1 \cdot sh_2 - a_{p+m,q+m} \cdot ch_1 \cdot ch_2 + a_{p,q} \cdot sh_2 \cdot sh_1 + a_{p,p+m} \cdot sh_1 \cdot ch_2 \Big)
\end{aligned} \tag{3.24c}$$

und $\dfrac{\partial F_2}{\partial y_2}(\mathit{0}) = 0$

$$
\begin{aligned}
&\frac{\partial F_2}{\partial x_2}(x_1,y_1,y_2,x_2) = \\
&= c_1 \cdot c_2 \cdot \Big(a_{p,p+m} \cdot ch_1 \cdot ch_2 + a_{p,q} \cdot ch_1 \cdot sh_2 - a_{p+m,q+m} \cdot ch_2 \cdot sh_1 + a_{q,q+m} \cdot sh_1 \cdot sh_2 \Big) \\
&+ s_1 \cdot c_2 \cdot \Big((a_{p+m,p+m} - a_{q,q}) \cdot ch_2 \cdot sh_2 - a_{q,p+m} \cdot (ch_2^2 + sh_2^2) \Big) \\
&- c_1 \cdot s_2 \cdot \Big((a_{p,p} - a_{q+m,q+m}) \cdot ch_1 \cdot sh_1 + a_{p,q+m} \cdot (ch_1^2 + sh_1^2) \Big) \\
&+ s_1 \cdot s_2 \cdot \Big(a_{q,q+m} \cdot ch_1 \cdot ch_2 - a_{p+m,q+m} \cdot ch_1 \cdot sh_2 + a_{p,q} \cdot ch_2 \cdot sh_1 + a_{p,p+m} \cdot sh_1 \cdot sh_2 \Big)
\end{aligned}
\tag{3.24d}
$$

$$
\text{und} \quad \frac{\partial F_2}{\partial x_2}(\mathit{0}) = a_{p,p+m}
$$

Für die Komponenten F_3 und F_4 ergeben sich analoge Formeln. Wir verzichten auf die Angabe der zweiten partiellen Ableitungen der Funktion F, obwohl diese später benötigt werden und geben lediglich die Jacobi-Matrix $J(\boldsymbol{x})$ der partiellen Ableitungen im Punkt $\mathit{0} \in \mathbb{R}^4$ an, die für den ersten Schritt des Newton-Verfahrens benötigt wird:

$$
J(\mathit{0}) = \left.\frac{\partial F}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=0}
$$

$$
= \begin{bmatrix}
(a_{pp} - a_{qq}) & a_{q,q+m} & a_{p,p+m} & 0 \\
-a_{q,q+m} & (a_{pp} - a_{q+m,q+m}) & 0 & a_{p,p+m} \\
-a_{p,p+m} & 0 & (a_{p+m,p+m} - a_{qq}) & a_{q,q+m} \\
0 & -a_{p,p+m} & -a_{q,q+m} & (a_{p+m,p+m} - a_{q+m,q+m})
\end{bmatrix}
\tag{3.25}
$$

Die Näherungen $\boldsymbol{x}^{(k)}$ des Newton-Verfahrens werden durch folgende Iteration erzeugt:

$$
\begin{aligned}
\boldsymbol{x}^{(0)} &= \mathit{0} \\
\boldsymbol{x}^{(k+1)} &= \boldsymbol{x}^{(k)} - J^{-1}(\boldsymbol{x}^{(k)}) \cdot F(\boldsymbol{x}^{(k)}) \qquad \text{für } k \geq 0
\end{aligned}
\tag{3.26}
$$

Im Verfahren selbst werden wir uns mit dem ersten Schritt des Verfahrens begnügen, d.h. wir berechnen lediglich

$$
\boldsymbol{x}^{(1)} = - J^{-1}(\mathit{0}) \cdot F(\mathit{0}) \tag{3.27}
$$

Im nächsten Abschnitt werden wir die quadratische Konvergenz des Newton-Verfahrens bei genügend kleiner Außerdiagonale nachweisen. An dieser Stelle untersuchen wir lediglich die Lösbarkeit des linearen Gleichungssystems für die 1. Iterierte $x^{(1)}$. Dazu benötigen wir als Voraussetzung ebenfalls eine genügend kleine Außerdiagonale. Wir werden die Größenordnung der Außerdiagonale $S(A)$ so beschränken, wie es insgesamt für die asymptotisch quadratische Konvergenz des gesamten Verfahrens nötig ist. Hier wird die Beschränkung der Außerdiagonale lediglich zur Trennung der Eigenwerte $\mu_i^{(p)}$ und $\mu_i^{(q)}$ unterschiedlicher Diagonalblöcke A_{pp} und A_{qq} benötigt.

Generalvoraussetzung an die Außerdiagonale:

Es sei δ durch (3.2) gegeben und es gelte

$$S(A) := \left\{ \sum_{\substack{i,j=1 \\ i \neq j}}^{m} \| A_{ij} \|_F^2 \right\}^{\frac{1}{2}} \leq \frac{1}{6.4 \cdot 10^4 \cdot \sqrt{m-1}} \cdot \frac{\delta^9}{\| A \|_F^8} \tag{3.28}$$

Unter dieser Generalvoraussetzung werden wir alle weiteren Betrachtungen durchführen. Mit ihr zeigen wir zunächst, daß der erste Schritt des Newton-Verfahrens durchführbar ist, d.h. die Matrix $J(0)$ ist invertierbar.

Hilfssatz 3.3

Unter der Generalvoraussetzung gilt für die Eigenwerte $\mu_i^{(j)}$ der Blockdiagonale

$$\delta(A) := \min_{i \neq j} \operatorname{dist}\{ \tilde{\mathcal{E}}_i , \tilde{\mathcal{E}}_j \} \geq 0.976 \cdot \delta ,. \text{ wobei} \tag{3.29}$$

$\tilde{\mathcal{E}}_k = \{\mu_1^{(k)}, \mu_2^{(k)}\}$ für $k = 1,\dots,m$ ist

Beweis:

Laut Theorem (3.1) liegen die exakten Eigenwerte λ_i der Matrix A in der Vereinigung aller Mengen

$$H_j = \bigcup_{i=1}^{2} \left\{ z \in \mathbb{C} \mid | \mu_i^{(j)} - z | \leq \left[1 + \left(\sqrt{y_j + \tfrac{1}{4}} - \tfrac{1}{2} \right) \right] \cdot \sum_{\substack{k=1 \\ k \neq j}}^{m} \| A_{jk} \|_F \right\}$$

für j = 1,...,m mit (3.18)

$$y_j = \frac{\Delta_F(A_{jj})}{\sum_{\substack{k=1 \\ k \neq j}}^{m} \| A_{jk} \|_F}.$$

Die exakten Eigenwerte haben den Mindestabstand $\delta > 0$. Setzen wir

$$R_j := \left[1 + \left(\sqrt{y_j + \tfrac{1}{4}} - \tfrac{1}{2} \right) \right] \cdot \sum_{\substack{k=1 \\ k \neq j}}^{m} \| A_{jk} \|_F,$$

so folgt wegen $\Delta_F(A_{jj}) \leq \| A \|_F$ und mit $\delta \leq \sqrt{2} \cdot \| A \|_F$ die Abschätzung

$$\sum_{\substack{k=1 \\ k \neq j}}^{m} \| A_{jk} \|_F \leq \frac{\sqrt{m-1}}{\sqrt{2}} \cdot S(A) \leq \frac{1}{6.4 \cdot 10^4 \cdot \sqrt{2}} \cdot \frac{\delta^9}{\| A \|_F^8} \leq \frac{1}{8 \cdot 10^3} \cdot \frac{\delta^2}{\| A \|_F} \leq \frac{\sqrt{2}}{8 \cdot 10^3} \cdot \delta$$

und damit ergibt sich für den maximalen Radius $R := \max_{j=1}^{m} R_j$ der Kreise H_j:

$$\begin{aligned} R &\leq \frac{\sqrt{2}}{16 \cdot 10^3} \cdot \delta + \sqrt{\left[\frac{\sqrt{2}}{16 \cdot 10^3} \cdot \delta \right]^2 + \frac{1}{8 \cdot 10^3} \cdot \frac{\delta^2}{\| A \|_F} \cdot \| A \|_F} \\ &\leq 0.012 \cdot \delta. \end{aligned}$$

Da die Mengen $\mathcal{E}_i$ und $\mathcal{E}_j$ für $i \neq j$ gemäß (3.2) einen Mindestabstand δ haben und gleichzeitig in den Kreisen H_j mit maximalen Radius $0.012 \cdot \delta$ liegen, müssen die Mengen $\tilde{\mathcal{E}}_i$ und $\tilde{\mathcal{E}}_j$ ebenfalls getrennt liegen für $i \neq j$, und für zwei beliebige Eigenwerte der Diagonalblöcke, die in verschiedenen $\tilde{\mathcal{E}}_k$ liegen, folgt:

$$| \mu_k^{(i)} - \mu_l^{(j)} | \geq | \lambda_i - \lambda_j | - | \mu_k^{(i)} - \lambda_i | - | \mu_l^{(j)} - \lambda_j |$$
$$\geq \delta - 2 \cdot 0.012 \cdot \delta = 0.976 \cdot \delta.$$

Hierbei sind λ_i bzw. λ_j die zu den beiden Eigenwerten $\mu_k^{(i)}$ und $\mu_l^{(j)}$ gehörenden exakten Eigenwerte der Matrix A.

●

Die Getrenntheit der Mengen der exakten Eigenwerte überträgt sich damit bei kleiner Außerdiagonale auf die Mengen der Eigenwerte der Diagonalblöcke. Dies ist die Garantie für die Regularität der Matrix $J(0)$, denn die Linearisierung im Ursprung besitzt eine ganz spezielle Struktur, sie kann als Kroneckerprodukt mit den Diagonalblöcken A_{pp} und A_{qq} geschrieben werden:

Definition 3.4

Unter dem Kronecker-Produkt zweier Matrizen $A \in \mathbb{R}^{k,k}$ und $B \in \mathbb{R}^{n,n}$, in Zeichen $B \otimes A$, verstehen wir die Matrix $C \in \mathbb{R}^{nk,nk}$ gegeben durch:

$$C = B \otimes A := \begin{bmatrix} a_{11}B & a_{12}B \dots & a_{1k}B \\ a_{21}B & a_{22}B \dots & a_{2k}B \\ \vdots & \vdots & \vdots \\ a_{k1}B & a_{k2}B \dots & a_{kk}B \end{bmatrix} \qquad \text{mit } a_{ij}B \in \mathbb{R}^{n,n} \tag{3.30}$$

Bemerkung:

Die Matrix $J(0)$ der Linearisierung in Null kann geschrieben werden als

$$J(0) = A_{pp} \otimes I_2 - I_2 \otimes A_{qq} \qquad \text{mit } I_2 \text{ als Einheitsmatrix im } \mathbb{R}^{2,2}.$$

Die Eigenwerte eines Kroneckerproduktes lassen sich aus den Eigenwerten der zugrundeliegenden Matrizen A_{pp} und A_{qq} berechnen, wie man aus den folgenden beiden Sätzen aus Friedman [6] sieht:

Satz 3.5

Seien $A_i, B_i \in \mathbb{R}^{d,d}$ für $i = 1,\dots,p$ und $C = A_1 \otimes B_1 + A_2 \otimes B_2 + \dots + A_p \otimes B_p$ .Die Menge der Matrizen $B_1,\dots, B_p$ seien simultan reduzierbar, d.h. es existiert eine nichtsinguläre Matrix $P \in \mathbb{R}^{d,d}$ mit

$$P \cdot B_k \cdot P^{-1} = \begin{bmatrix} B_{k1} & & & \\ & B_{k2} & & 0 \\ & & \ddots & \\ & 0 & & B_{kl} \end{bmatrix} \text{ für } k=1,\dots,p \text{ und } B_{ki} \in \mathbb{R}^{n,n} \text{ mit } n \cdot p = d. \quad (3.31)$$

Dann ist jeder Eigenwert der Matrix

$$M_j = A_1 \otimes B_{1j} + A_2 \otimes B_{2j} + \dots + A_p \otimes B_{pj} \quad (3.32)$$

Eigenwert der Matrix C und umgekehrt jeder Eigenwert der Matrix C ein Eigenwert einer der Matrizen M_j für $j=1,\dots,p$.

Folgerung:

Ist y ein Eigenvektor aller Matrizen $B_1,\dots,B_p$, so daß

$$B_k \cdot y = \mu_k \cdot y \quad \text{für } k=1,\dots,p$$

gilt und ist x ein Vektor mit

$$(\mu_1 \cdot A_1 + \dots + \mu_p \cdot A_p) \cdot x = \lambda \cdot x,$$

dann ist λ ein Eigenwert von C, und es gilt

$$C \cdot (x \otimes y) = \lambda \cdot (x \otimes y)$$

Bemerkung:

Eine besonders einfache Darstellung der Eigenwerte von C ergibt sich, falls p = 2 und $A_2 = I_2$ bzw. $B_1 = I_2$ mit I_2 als Identität im $\mathbb{R}^{2,2}$ ist, d.h.

$$C = A \otimes I_2 + I_2 \otimes B \tag{3.33}$$

Seien μ_j die Eigenwerte von B und λ_j die Eigenwerte von A für j = 1,2. Dann sind die vier Eigenwerte der Matrix C gegeben durch $(\mu_i + \lambda_j)$ für i,j = 1,2.

Mit obiger Bemerkung kann man folglich die Eigenwerte der Linearisierung $J(\mathit{0})$ angeben:

Satz 3.6

Seien μ_i^p die Eigenwerte von A_{pp} und μ_i^q die Eigenwerte von A_{qq} für i = 1,2. Diese sind unmittelbar als Lösung eines quadratischen Gleichungssystems berechenbar. Es ist

$$\mu_{1,2}^p = \frac{a_{pp} + a_{p+m,p+m}}{2} \pm \sqrt{\left[\frac{a_{pp} - a_{p+m,p+m}}{2}\right]^2 - a_{p,p+m}^2} \quad \text{und}$$

$$\mu_{1,2}^q = \frac{a_{qq} + a_{q+m,q+m}}{2} \pm \sqrt{\left[\frac{a_{qq} - a_{q+m,q+m}}{2}\right]^2 - a_{q,q+m}^2}$$

Die Eigenwerte von $J(\mathit{0})$ sind dann gegeben durch die vier Differenzen

$$(\mu_1^q - \mu_1^p),\ (\mu_1^q - \mu_2^p),\ (\mu_2^q - \mu_1^p) \text{ und } (\mu_2^q - \mu_2^p).$$

Weiterhin folgt sofort mit Hilfssatz (3.3), da die Eigenwerte der Diagonalblöcke A_{pp} und A_{qq} den Mindestabstand $0.976 \cdot \delta$ haben, daß die Matrix $J(\mathit{0})$ regulär ist, da die Differenzen von Null wegbeschränkt sind.

Bemerkung:

Damit ist gezeigt, daß für genügend kleine Außerdiagonalnorm der erste Newton-Schritt stets durchführbar ist, und gleichzeitig haben wir einen wesentlichen Beitrag für den Beweis der quadratischen Konvergenz des Newton - Verfahrens erhalten. Da die Eigenwerte von H von Null wegbeschränkt sind, ist der Spektralradius der Inversen $J^{-1}(\mathit{0})$ beschränkt. Dies verwenden wir, um die Spektralnorm von J^{-1} abzuschätzen.

Wir kommen nun im nächsten Abschnitt zum Beweis der quadratischen Konvergenz des Newton - Verfahrens, das zur näherungsweisen Bestimmung der vier Parameter der Ähnlichkeitstransformationen dient.

3.5 Der Konvergenzbeweis des Parameterbestimmung

In diesem Kapitel wollen wir nachweisen, daß die nach Newton gebildete Iterationsvorschrift (3.26) bzw. (3.27) zur näherungsweisen Bestimmung der Transformationsparameter quadratisch konvergiert. Dazu verwenden wir den Satz von Newton - Kantorovitsch. Der Beweis benötigt im wesentlichen 3 Voraussetzungen, die kombiniert die quadratische Konvergenz bzw. die lokale Eindeutigkeit des Nullstellenproblems garantieren:

1) Die Beschränktheit der Spektralnorm der Inversen $J^{-1}(\boldsymbol{0})$ der Linearisierung
2) Die Beschränktheit der 1. Iterierten $\boldsymbol{x}^{(1)}$ des Newton - Verfahrens
3) Eine Lipschitz - Bedingung für die Linearisierung $J(\boldsymbol{x})$ bezüglich $\boldsymbol{x} \in \mathbb{R}^4$

Die erste Voraussetzung erhalten wir aus einem Satz von Henrici [8], der es ermöglicht, die Spektralnorm einer beliebigen, auch nichtnormalen Matrix durch ihren Spektralradius abzuschätzen. Der Spektralradius von $J^{-1}(\boldsymbol{0})$ ist bereits durch die Darstellung der Eigenwerte von $J(\boldsymbol{0})$ als Kehrwert der kleinsten Differenz der Eigenwerte der Diagonalblöcke A_{pp} und A_{qq} abgeschätzt. Es gilt:

Satz 3.7

Sei $A \in \mathbb{R}^{n,n}$ eine nichtnormale, nichtsinguläre Matrix. Dann gilt:

$$\| A^{-1} \|_2 \leq \frac{f(y)}{y} \cdot \rho(A^{-1}) \text{ mit } y := \rho(A^{-1}) \cdot \Delta_2(A) \text{ und } f(y) := \sum_{i=1}^{n} y^i. \tag{3.34}$$

$\rho(A)$ sei hierbei der Spektralradius von A und $\Delta_2(A)$ das Henrici - Maß gemessen in der Spektralnorm.

Um den obigen Satz aus $J(\boldsymbol{0})$ anwenden zu können, benötigen wir zunächst eine Abschätzung von $\Delta_2(J(\boldsymbol{0}))$. Der Beweis des nachfolgenden Hilfssatzes verdeutlicht gleichzeitig die Struktur der Dreiecksgestalt, auf die $J(\boldsymbol{0})$ unitär transformiert werden kann.

<u>Hilfssatz 3.8</u>

Das Henrici-Maß $\Delta_2(J(\mathit{0}))$ der Matrix $J(\mathit{0})$ der Linearisierung ist beschränkt. Es gilt

$$\Delta_2(J(\mathit{0})) \leq 2 \cdot \| A \|_F. \tag{3.35}$$

<u>Beweis:</u>

Die Matrix $J(\mathit{0})$ ist darstellbar in der Form $J(\mathit{0}) = A_{pp} \otimes I_2 - I_2 \otimes A_{qq}$ mit I_2 als Einheitsmatrix im $\mathbb{R}^{2,2}$. Nach Schur existieren unitäre Ähnlichkeitstransformationen U und V, die die Diagonalblöcke A_{pp} und A_{qq} auf Dreiecksgestalt transformieren,nämlich:

$$U^* \cdot A_{pp} \cdot U := D_{pp} + M_{pp} := \begin{bmatrix} \mu_1^p & \Delta_F(A_{pp}) \\ 0 & \mu_2^p \end{bmatrix} \quad \text{und}$$

$$V^* \cdot A_{qq} \cdot V := D_{qq} + M_{qq} := \begin{bmatrix} \mu_1^q & \Delta_F(A_{qq}) \\ 0 & \mu_2^q \end{bmatrix}.$$

mit D_{ii} als Diagonalanteil der Dreiecksmatrix und M_{ii} als echte obere Dreiecksmatrix für i = p,q. Dann folgt mit den Rechenregeln für Kroneckerprodukte:

$$\begin{aligned}
&(U^* \otimes V^*) \cdot J(\mathit{0}) \cdot (U \otimes V) \\
&= (U^* \otimes V^*) \cdot \left[((A_{pp} \otimes I_2) \cdot (U \otimes V) - (I_2 \otimes A_{qq}) \cdot (U \otimes V) \right] \\
&= (U^* \cdot A_{pp} \cdot U) \otimes (V^* \cdot I_2 \cdot V) \qquad - (U^* \cdot I_2 \cdot U) \otimes (V^* \cdot A_{qq} \cdot V) \\
&= (D_{pp} + M_{pp}) \otimes I_2 \qquad - I_2 \otimes (D_{qq} + M_{qq}) \\
&= (D_{pp} \otimes I_2 - I_2 \otimes D_{qq}) \qquad + (M_{pp} \otimes I_2 - I_2 \otimes M_{qq}) \\
&= \begin{bmatrix} (\mu_1^p - \mu_1^q) & & & \mathit{0} \\ & (\mu_1^p - \mu_2^q) & & \\ & & (\mu_2^p - \mu_1^q) & \\ \mathit{0} & & & (\mu_2^p - \mu_2^q) \end{bmatrix} + \underbrace{\begin{bmatrix} 0 & \Delta_F(A_{pp}) & \Delta_F(A_{qq}) & 0 \\ 0 & 0 & 0 & \Delta_F(A_{pp}) \\ 0 & 0 & 0 & \Delta_F(A_{pp}) \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{:= \mathbf{M}(J(0))}.
\end{aligned}$$

Damit folgt wegen $\Delta_F(A_{ii}) \leq \| A \|_F$ für i = p,q

$$\Delta_2(J(\mathit{0})) \leq \Delta_F(J(\mathit{0})) = \| M(J(\mathit{0})) \|_F = \sqrt{2\cdot\Delta_F(A_{pp})^2 + 2\cdot\Delta_F(A_{qq})^2}$$

$$\leq 2\cdot\| A \|_F$$

●

Somit ist die Spektralnorm der Linearisierung abschätzbar. Es gilt:

Satz 3.9

Für die Spektralnorm der Linearisierung $J(\mathit{0})$ gilt

$$\| J(\mathit{0})^{-1}\|_2 \leq 2.05\cdot\frac{1}{\delta}\cdot\sum_{i=0}^{3}\left[2.05\cdot\| A \|_F\cdot\frac{1}{\delta}\right]^i \tag{3.36}$$

Beweis:

Es gilt $\| J(\mathit{0})^{-1} \|_2 \leq (1 + y + y^2 + y^3)\cdot\rho(J(\mathit{0})^{-1})$ mit $y:=\rho(J(\mathit{0})^{-1})\cdot\Delta_2(J(\mathit{0}))$.

Die vorhergehenden Betrachtungen ergeben:

$$0 < y < 2.05\cdot\| A \|_F\cdot\frac{1}{\delta}$$

Einsetzen der Abschätzung für y ergibt die Behauptung.

●

Damit haben wir eine Abschätzung für $\| J(\mathit{0})^{-1} \|_2$ gefunden. Zwei weitere Voraussetzungen müssen noch erfüllt werden. Dazu verwenden wir für den Punkt 2 eine Abschätzung für die Lösung eines linearen Gleichungssystems, dessen Matrix ein Kroneckerprodukt ist. Diese Abschätzung stammt von Meyer [15].

Satz 3.10

Gegeben sei ein lineares System der Form

$$J(\theta)\, x = (A_{pp} \otimes I_2 - I_2 \otimes A_{qq}) \cdot x = C \text{ mit } C \in \mathbb{R}^4.$$

Die Lösung $x \in \mathbb{R}^4$ dieses Systems genügt der Abschätzung

$$\| x \| \leq G_4(y) \cdot \frac{\| C \|}{\min\limits_{i,j=1}^{2} | \mu_i^p - \mu_j^q |} \quad \text{mit} \tag{3.37a}$$

$$G_4(y) := 1 + y + y^2 + y^3 + y^4 \text{ und } y := \frac{\Delta_F(A_{pp}) + \Delta_F(A_{qq})}{\min\limits_{i,j=1}^{2} | \mu_i^p - \mu_j^q |}. \tag{3.37b}$$

Der Beweis ist bei Meyer [15] nachzulesen. Mit Hilfe dieses Satzes erhalten wir das

Lemma 3.11

Unter der Generalvoraussetzung (3.28) genügt die 1. Iterierte $x^{(1)}$ des Newton-Verfahrens (3.26) zur Bestimmung der 4 Transformationsparameter der Abschätzung:

$$\| x^{(1)} \| \leq G_4\left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right] \cdot \frac{1}{0.976 \cdot \delta} \cdot \frac{\sqrt{2}}{2} \cdot \frac{1}{6.4 \cdot 10^4 \cdot \sqrt{m-1}} \cdot \frac{\delta^9}{\| A \|_F^8} \tag{3.38}$$

Beweis:

$$\| x^{(1)} \| \leq G_4(y) \cdot \frac{\| F(\theta) \|}{\min\limits_{i,j=1}^{2} | \mu_i^p - \mu_j^q |}$$

Es ist nach Satz (3.8) $0 < y < 2.05 \cdot \| A \|_F \cdot \frac{1}{\delta}$. Weiterhin gilt nach Hilfssatz (3.3) für den Nenner $\min\limits_{i,j=1}^{2} | \mu_i^p - \mu_j^q | \geq 0.976 \cdot \delta$. Schließlich ist die rechte Seite $F(\theta)$ gerade gegeben durch den Vektor $F(\theta) = (a_{pq}, a_{p,q+m}, a_{p+m,q}, a_{p+m,q+m})^T$. Damit folgt

$$\| F(\theta) \| \leq \frac{\sqrt{2}}{2} \cdot S(A)$$

Insgesamt folgt so

$$\| x^{(1)} \| \le G_4 \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right] \cdot \frac{1}{0.976 \cdot \delta} \cdot \frac{\sqrt{2}}{2} \cdot S(A)$$

Einsetzen der Generalvoraussetzung für $S(A)$ liefert die Behauptung.

●

<u>Bemerkung:</u>

Aus der obigen Abschätzung für $\| x^{(1)} \|$ lassen sich durch Vergröberung der Abschätzung unter Zuhilfenahme der Formel $\delta \le \sqrt{2} \cdot \| A \|_F$ mehrere abgestufte Abschätzungen für $\| x^{(1)} \|$ zeigen, deren Gültigkeit wir weiter unten in vielen Abschätzungen verwenden werden. Es gilt:

$$\| x^{(1)} \| \le \sum_{i=0}^{4} \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \cdot \frac{1}{0.976 \cdot \delta} \cdot \frac{\sqrt{2}}{2} \cdot \frac{1}{6.4 \cdot 10^4 \cdot \sqrt{m-1}} \cdot \frac{\delta^9}{\| A \|_F^8}$$

$$\le \frac{1}{0.976} \cdot \frac{\sqrt{2}}{2} \cdot \frac{1}{6.4 \cdot 10^4} \cdot \sum_{i=0}^{4} 2.05^i \cdot \left[\| A \|_F \cdot \frac{1}{\delta} \right]^i \cdot \frac{\delta^8}{\| A \|_F^8}$$

$$\le \frac{1}{0.976} \cdot \frac{\sqrt{2}}{2} \cdot \frac{1}{6.4 \cdot 10^4} \cdot$$

$$\left[\frac{\delta^4}{\| A \|_F^4} + 2.05 \cdot \frac{\delta^3}{\| A \|_F^3} + 2.05^2 \cdot \frac{\delta^2}{\| A \|_F^2} + 2.05^3 \cdot \frac{\delta}{\| A \|_F} + 2.05^4 \right] \cdot \frac{\delta^4}{\| A \|_F^4}$$

$$\le \frac{1}{0.976} \cdot \frac{\sqrt{2}}{2} \cdot \frac{1}{6.4 \cdot 10^4} \cdot$$

$$\left[4 + 2.05 \cdot 2 \cdot \sqrt{2} + 2.05^2 \cdot 2 + 2.05^3 \cdot \sqrt{2} + 2.05^4 \right] \cdot \frac{\delta^4}{\| A \|_F^4}$$

$$=: C \cdot \frac{\delta^4}{\| A \|_F^4} \le \sqrt{2} \cdot C \cdot \frac{\delta^3}{\| A \|_F^3} \le 2 \cdot C \cdot \frac{\delta^2}{\| A \|_F^2} \le 2 \cdot \sqrt{2} \cdot C \cdot \frac{\delta}{\| A \|_F} \le 4 \cdot C \qquad (3.39)$$

mit $C = \frac{1}{0.976} \cdot \frac{\sqrt{2}}{2} \cdot \frac{1}{6.4 \cdot 10^4} \cdot \left[4 + 2.05 \cdot 2 \cdot \sqrt{2} + 2.05^2 \cdot 2 + 2.05^3 \cdot \sqrt{2} + 2.05^4 \right]$

$\approx 5.4391332 \cdot 10^{-4} < 5.44 \cdot 10^{-4}$

Damit folgt als gröbste Abschätzung: $\| x^{(1)} \| \le 2.176 \cdot 10^{-3}$

Schließlich benötigen wir im Konvergenzbeweis des Newton-Verfahrens eine Lipschitzbedingung für die Funktion $J(\boldsymbol{x})$. Eine solche Abschätzung ist bei Abschätzung aller auftretenden Matrixelemente durch die Norm der Matrix lokal um $\mathit{0} \in \mathbb{R}^4$ leicht zu gewinnen, da J eine C^∞- Funktion ist und die Komponenten der Hessematrix H der zweiten partiellen Ableitungen auf jedem Kompaktum, also zum Beispiel auf einer abgeschlossenen Kugel um Null, ihr Maximum annehmen. Die Hesse-Matrix H hat auch bei Berücksichtigung der Symmetrie immer noch 32 verschiedene Komponenten. Es wäre an dieser Stelle äußerst mühsam, die Abschätzung dieser 32 partiellen 2. Ableitungen konkret durchzuführen. Ich verzichte auf die explizite Angabe von $H(\boldsymbol{x})$. Die Rechnungen sind trivial, aber sehr langwierig.

Für den Konvergenzbeweis benutzen wir den Satz von Newton-Kantorovitsch:

<u>Theorem 3.12</u> (Newton-Kantorovitsch)

Sei $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ Fréchet-differenzierbar und $D_0 \subset D$ konvex. Wir betrachten

$$\boldsymbol{x}^{(k+1)} := \boldsymbol{x}^{(k)} - F'^{-1}(\boldsymbol{x}^{(k)}) \cdot F(\boldsymbol{x}^{(k)}) \quad \text{für } k \in \mathbb{N}_0. \tag{3.40}$$

Für die in der Iterationsvorschrift (3.40) auftretende Ableitung existiere ein $\gamma > 0$ mit

$$\| F'(\boldsymbol{x}) - F'(\boldsymbol{y}) \|_2 \leq \gamma \cdot \| \boldsymbol{x} - \boldsymbol{y} \| \quad \text{für alle } \boldsymbol{x}, \boldsymbol{y} \in D_0. \tag{3.41}$$

Sei nun ein $\boldsymbol{x}^{(0)} \in D_0$ gegeben mit $\| F'^{-1}(\boldsymbol{x}^{(0)}) \|_2 \leq \beta$ und $\| F'^{-1}(\boldsymbol{x}^{(0)}) \cdot F(\boldsymbol{x}^{(0)}) \| \leq \eta$ gegeben. Gilt dann für

$$t^* := \frac{1}{\beta \cdot \gamma} \cdot \left[1 - \sqrt{1 - 2 \cdot \alpha} \right] \quad \text{die Beziehung}$$

$$S(\boldsymbol{x}^{(0)}, t^*) := \left\{ \boldsymbol{x} \in \mathbb{R}^4 \mid \| \boldsymbol{x} - \boldsymbol{x}^{(0)} \| \leq t^* \right\} \subset D_0 \tag{3.42}$$

und ist

$$\alpha := \beta \cdot \gamma \cdot \eta \leq \frac{1}{2}, \tag{3.43}$$

dann konvergiert die Folge $(\boldsymbol{x}^{(k)})_{k \in \mathbb{N}_0}$ aus (3.40) gegen ein $\boldsymbol{x}^* \in \mathbb{R}$ mit $F(\boldsymbol{x}^*) = 0$ und

$$\| \boldsymbol{x}^* - \boldsymbol{x}^{(k)} \|_F \leq \frac{1}{\beta \cdot \gamma \cdot 2^k} \cdot (2 \cdot \alpha)^{2^k} \text{ für } k \in \mathbb{N}_0. \tag{3.44}$$

Die Lösung ist eindeutig in $S(\boldsymbol{x}^{(0)}, t^{**})$ mit

$$t^{**} := \frac{1}{\beta \cdot \gamma} \cdot \left[1 + \sqrt{1 - 2 \cdot \alpha} \right] \tag{3.45}$$

Der Beweis findet sich in Ortega - Rheinbold [17].

Mit Hilfe dieses Satzes erhalten wir unter Berücksichtigung der bereits gezeigten Voraussetzungen:

Satz 3.13

Die gemäß (3.26) gebildete Newton - Iteration zur näherungsweisen Bestimmung der 4 Parameter der Ähnlichkeitstransformationen

$$\begin{aligned} \boldsymbol{x}^{(0)} &= \boldsymbol{0} \\ \boldsymbol{x}^{(k+1)} &= \boldsymbol{x}^{(k)} - J^{-1}(\boldsymbol{x}^{(k)}) \cdot F(\boldsymbol{x}^{(k)}) \qquad \text{für } k \geq 0 \end{aligned}$$

konvergiert für alle Startwerte $\boldsymbol{x}^{(0)} \in S(\boldsymbol{0}, R) \subset \mathbb{R}^4$ mit $R \leq 4.352 \cdot 10^{-3}$ gegen die exakte Lösung $\boldsymbol{x}^*$ des zugehörigen nichtlinearen Nullstellenproblems $F(\boldsymbol{x}) = 0$. Die Lösung ist lokal eindeutig in $S(\boldsymbol{0}, R)$. Es gilt die Fehlerabschätzung

$$\| \boldsymbol{x}^* - \boldsymbol{x}^{(k)} \| \leq \frac{1}{\beta \cdot \gamma \cdot 2^k} \cdot (2 \cdot \alpha)^{2^k} \tag{3.46}$$

mit

$$\begin{aligned} \beta &\leq 2.05 \cdot \frac{1}{\delta} \cdot \sum_{i=0}^{3} \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \\ \gamma &\leq 3.4424 \cdot 10^{-1} \cdot \| A \|_F \\ \eta &\leq \sum_{i=0}^{4} \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \cdot \frac{1}{0.976 \cdot \delta} \cdot \frac{\sqrt{2}}{2} \cdot \frac{1}{6.4 \cdot 10^4 \cdot \sqrt{m-1}} \cdot \frac{\delta^9}{\| A \|_F^8} \end{aligned} \tag{3.47}$$

und $\alpha := \beta \cdot \gamma \cdot \eta$. Insbesondere gelten die folgenden Abschätzungen für die 1. Iterierte:

$$\| x^* - x^{(1)} \| \leq 9.152 \cdot 10^{-6} \cdot \frac{\delta^4}{\| A \|_F^4} \leq \ldots \leq 3.6592 \cdot 10^{-5} \tag{3.48}$$

$$\| F(x^{(1)}) \| \leq 1.46993 \cdot 10^5 \cdot \left[\max_{\xi \in S(0,R)} \| J(\xi) \|_\infty \right] \cdot \frac{\| A \|_F^{12}}{\delta^{14}} \cdot S^2(A) \tag{3.49}$$

$$\leq 2.99138 \cdot 10^5 \cdot \frac{\| A \|_F^{13}}{\delta^{14}} \cdot S^2(A)$$

<u>Beweis:</u>

Es gilt bei Verwendung des Mittelwertsatzes und der Verträglichkeit von $\|\cdot\|_2$ mit $\|\cdot\|_F$

$$\| J(x) - J(y) \|_2 \leq \| J(x) - J(y) \|_F$$

$$\leq 8 \cdot \left[\max_{\xi \in \{x + \omega(y - x) \mid 0 \leq \omega \leq 1\}} \| J'(\xi) \|_\infty \right] \cdot \| x - y \|$$

Eine längliche Rechnung ergibt, daß das Maximum der 32 verschiedenen Komponenten der 2. partiellen Ableitungen von $F(x)$ gerade bei den ungemischten zweiten Ableitungen angenommen wird. Das Maximum der 2. partiellen Ableitungen von F_i wird angenommen von $\frac{\partial^2 F_i}{\partial^2 x_i}$ für i = 1,2,3,4. Wir erhalten für $x, y \in S(0,R) \subset \mathbb{R}^4$ mit $R \leq 4.352 \cdot 10^{-3}$ die Abschätzung

$$\left[\max_{\xi \in \{x + \omega(y - x) \mid 0 \leq \omega \leq 1\}} \| J'(\xi) \|_\infty \right] \leq \left[\max_{\xi \in S(0,R)} \| J'(\xi) \|_\infty \right] \leq 4.303 \cdot 10^{-2} \cdot \| A \|_F$$

Wir erhalten für die Lipschitzkonstante

$$\gamma \leq 3.4424 \cdot 10^{-1} \cdot \| A \|_F$$

Es ist weiter $x^{(0)} = 0$ und damit folgt nach Satz (3.9)

$$\| F'^{-1}(x^{(0)}) \|_2 = \| J^{-1}(0) \|_2 \leq 2.05 \cdot \frac{1}{\delta} \cdot \sum_{i=0}^{3} \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i := \beta$$

Schließlich erhalten wir als Abschätzung der 1. Iterierten nach Theorem (3.12)

$$\| F'^{-1}(x^{(0)})\cdot F(x^{(0)}) \| = \| J^{-1}(0)\cdot F(0) \| = \| x^{(1)} \| \le$$

$$\sum_{i=0}^{4}\left[2.05\cdot\| A \|_F\cdot\frac{1}{\delta}\right]^i \cdot \frac{1}{0.976\cdot\delta}\cdot\frac{\sqrt{2}}{2}\cdot\frac{1}{6.4\cdot 10^4\cdot\sqrt{m-1}}\cdot\frac{\delta^9}{\| A \|_F^8} := \eta$$

Es folgt

$$\alpha = \beta\cdot\gamma\cdot\eta \le \left\{2.05\cdot\frac{1}{\delta}\cdot\sum_{i=0}^{3}\left[2.05\cdot\| A \|_F\cdot\frac{1}{\delta}\right]^i\right\}\cdot\left\{3.4424\cdot 10^{-1}\cdot\| A \|_F\right\}\cdot$$
$$\left\{\sum_{i=0}^{4}\left[2.05\cdot\| A \|_F\cdot\frac{1}{\delta}\right]^i\cdot\frac{1}{0.976\cdot\delta}\cdot\frac{\sqrt{2}}{2}\cdot\frac{1}{6.4\cdot 10^4\cdot\sqrt{m-1}}\cdot\frac{\delta^9}{\| A \|_F^8}\right\}$$

$$\le\frac{\delta^7}{\| A \|_F^7}\cdot\left\{2.05\cdot 3.4424\cdot 10^{-1}\cdot\frac{1}{0.976\cdot}\cdot\frac{\sqrt{2}}{2}\cdot\frac{1}{6.4\cdot 10^4}\right\}$$
$$\cdot\left\{1 + 2\cdot C + 3\cdot C^2 + 4\cdot(C^3 + C^4) + 3\cdot C^5 + 2\cdot C^6 + C^7\right\}$$

mit $C := \left[2.05\cdot\| A \|_F\cdot\frac{1}{\delta}\right]$. Wegen $\delta \le \sqrt{2}\cdot\| A \|_F$ folgt

$$\alpha \le \left\{2.05\cdot 3.4424\cdot 10^{-1}\cdot\frac{1}{0.976\cdot}\cdot\frac{\sqrt{2}}{2}\cdot\frac{1}{6.4\cdot 10^4}\right\}\cdot$$
$$\left\{8\cdot\sqrt{2} + 2.05\cdot 2\cdot 8 + 3\cdot 2.05^2\cdot 4\cdot\sqrt{2} + 4\cdot(4\cdot 2.05^3 + 2\cdot\sqrt{2}\cdot 2.05^4)\right.$$
$$\left. + 3\cdot 2\cdot 2.05^5 + 2\cdot\sqrt{2}\cdot 2.05^6 + 2.05^7\right\}$$

und schließlich $\alpha \le 8.408\cdot 10^{-3} \le 0.5$

Es gilt

$$t^* := \frac{1}{\beta\cdot\gamma}\cdot\left[1 - \sqrt{1 - 2\cdot\alpha}\right] \le 2\cdot\eta < 4.352\cdot 10^{-3}$$

und deswegen ist $S(0, t^*) \subset D_0$. Damit sind die Voraussetzungen des Satzes von Newton-Kantorovitsch erfüllt, und es folgt die quadratische Konvergenz des Newton-Verfahrens gegen die Nullstelle x^* des Problems $F(x) = 0$. Für die 1. Iterierte erhalten wir die Abschätzung

$$\| x^* - x^{(1)} \| \leq 2 \cdot \frac{1}{\beta \cdot \gamma} \cdot \alpha^2 = 2 \cdot \alpha \cdot \eta \leq 2 \cdot 5.44 \cdot 10^{-4} \cdot 8.408 \cdot 10^{-3} \cdot \frac{\delta^4}{\| A \|_F^4}$$

$$\leq 9.152 \cdot 10^{-6} \cdot \frac{\delta^4}{\| A \|_F^4} \leq \ldots \leq 3.6592 \cdot 10^{-5}$$

Unter den starken Voraussetzungen an die Außerdiagonale ist die 1. Iterierte eine sehr gute Näherung für die exakte Nullstelle des Problems. Für die Größenordnung des Pivotblockes ergibt sich nach der näherungsweisen Annullierung mit den Ähnlichkeitstransformationen mit $x^{(1)}$:

Es gilt nach dem Mittelwertsatz

$$\| F(x^{(1)}) \| = \| F(x^{(1)}) - F(x^*) \|$$

$$\leq 4 \cdot \left[\max_{\xi \in \{ x^{(1)} + \omega(x^{(1)} - x^*) \,|\, 0 \leq \omega \leq 1\}} \| J(\xi) \|_\infty \right] \cdot \| x^{(1)} - x^* \|$$

Die Verbindungsgerade zwischen $x^{(1)}$ und x^* liegt wegen $\| x^{(1)} \| \leq 2.176 \cdot 10^{-3}$ und $\| x^{(1)} - x^* \| \leq 4.574 \cdot 10^{-6}$ noch ganz in $S(\mathit{0}, R) \subset \mathbb{R}^4$ mit $R = 4.352 \cdot 10^{-3}$. Es folgt

$$\begin{aligned} \| F(x^{(1)}) \| &\leq 4 \cdot \left[\max_{\xi \in S(\mathit{0},R)} \| J(\xi) \|_\infty \right] \cdot \| x^{(1)} - x^* \| \\ &\leq 8 \cdot \left[\max_{\xi \in S(\mathit{0},R)} \| J(\xi) \|_\infty \right] \cdot \frac{1}{\beta \cdot \gamma} \cdot \alpha^2 \\ &= 8 \cdot \left[\max_{\xi \in S(\mathit{0},R)} \| J(\xi) \|_\infty \right] \cdot \beta \cdot \gamma \cdot \eta^2 \end{aligned}$$

Nun gilt

$$\begin{aligned} \beta &\leq 2.05 \cdot \frac{1}{\delta} \cdot \sum_{i=0}^{3} \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \\ \gamma &\leq 3.4424 \cdot 10^{-1} \cdot \| A \|_F \\ \eta &\leq \sum_{i=0}^{4} \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \cdot \frac{1}{0.976 \cdot \delta} \cdot \frac{\sqrt{2}}{2} \cdot S(A) \end{aligned}$$

Es folgt

$$
\begin{aligned}
&\| F(\boldsymbol{x}^{(1)})\| \\
&\le 8 \cdot \left[\max_{\xi \in S(\mathit{0},R)} \| J(\xi) \|_\infty \right] \cdot 2.05 \cdot \frac{1}{\delta} \cdot \sum_{i=0}^{3} \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \cdot \\
&\quad \cdot 3.4424 \cdot 10^{-1} \cdot \| A \|_F \cdot \left[\sum_{i=0}^{4} \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \cdot \frac{1}{0.976 \cdot \delta} \cdot \frac{\sqrt{2}}{2} \cdot S(A) \right]^2 \\
&\le \left[8.2 \cdot 0.34424 \cdot \frac{1}{0.976^2} \right] \cdot \left[\max_{\xi \in S(\mathit{0},R)} \| J(\xi) \|_\infty \right] \cdot \cdot \frac{\| A \|_F^{12}}{\delta^{14}} \cdot S^2(A) \\
&\quad \cdot \left[\frac{\delta^3}{\| A \|_F^3} \cdot \sum_{i=0}^{3} \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \right] \cdot \left[\frac{\delta^4}{\| A \|_F^4} \cdot \sum_{i=0}^{4} \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \right]^2 \\
&\le 2.9633 \cdot \left[2.05^3 + 2.05^2 \cdot \sqrt{2} + 2.05 \cdot 2 + 2 \cdot \sqrt{2} \right] \cdot \\
&\qquad \cdot \left[2.05^4 + 2.05^3 \cdot \sqrt{2} + 2.05^2 \cdot 2 + 2.05 \cdot 2 \cdot \sqrt{2} + 4 \right]^2 \cdot \\
&\qquad \cdot \left[\max_{\xi \in S(\mathit{0},R)} \| J(\xi) \|_\infty \right] \cdot \frac{\| A \|_F^{12}}{\delta^{14}} \cdot S^2(A) \\
&\le 1.46993 \cdot 10^5 \cdot \left[\max_{\xi \in S(\mathit{0},R)} \| J(\xi) \|_\infty \right] \cdot \frac{\| A \|_F^{12}}{\delta^{14}} \cdot S^2(A)
\end{aligned}
$$

Das Maximum der Jacobi-Matrix $J(\boldsymbol{x})$ in der Sphäre $S(\mathit{0},R)$ läßt sich leicht abschätzen. Es gilt

$$
\max_{\xi \in S(\mathit{0},R)} \| J(\xi) \|_\infty \le 2.035044 \cdot \| A \|_F .
$$

●

Damit ist die quadratische Konvergenz des Newton-Verfahres nachgewiesen. Gleichzeitig kennen wir die Größenordnung des Parametervektors $\boldsymbol{x} \in \mathbb{R}^4$, und wir haben gezeigt, daß der Pivotblock A_{pq} nach der Transformation von der Ordnung $O(S(A))$ nach der Transformation die Größenordnung $O(S^2(A))$ besitzt.

In den nächsten Abschnitten beschäftigen wir uns mit den Veränderungen der Matrixblöcke A_{ij} durch die einzelnen Transformationen unter Verwendung der nun bekannten Größenordnung der Transformationsparameter. Wir können damit alle Wirkungen über einen gesamten Zyklus des Verfahrens in ihrer Gesamtheit berechnen.

3.6 Abschätzung der Transformationsmatrix der hyperbolischen Transformation

In diesem Abschnitt wollen wir die Pivotblöcke der hyperbolischen Transformation in ihrer Größenordnung unter der Generalvoraussetzung abschätzen. Wir benötigen solche Abschätzungen sowohl für die hyperbolische als auch für die orthogonale Transformation, um die Veränderungen der Matrix A pro Transformationsschritt abschätzen zu können. Hieraus werden wir eine Abschätzung für den maximalen Zuwachs der Außerdiagonalnorm $S(A)$ gewinnen.

Satz 3.14

Es gelte die Generalvoraussetzung (3.28). Die Parameter y_1 und y_2 seien durch das Newton-Verfahren (3.26) für das Pivotpaar $(p,q) \in P_m$ bestimmt. Dann folgt für die Pivotblöcke der hyperbolischen Transformation $\mathcal{T}(y_1,y_2)$

$$\begin{aligned} \| T_{pp} \|_2 &= \| T_{qq} \|_2 \leq (1 + 2.3675 \cdot 10^{-6}) \\ \| T_{pq} \|_2 &= \| T_{qp} \|_2 \leq (1 + 2.3675 \cdot 10^{-6}) \cdot \| x^{(1)} \| \end{aligned} \tag{3.50}$$

mit (3.39)

$$\begin{aligned} \| x^{(1)} \| &\leq \sum_{i=0}^{4} \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \cdot \frac{1}{0.976 \cdot \delta} \cdot \frac{\sqrt{2}}{2} \cdot S(A) \\ &\leq 5.44 \cdot 10^{-4} \cdot \frac{\delta^4}{\| A \|_F^4} \leq \ldots \leq 2.176 \cdot 10^{-3} \end{aligned}$$

Weiter gilt für den Abstand der Diagonalblöcke T_{ii} für $i=p,q$ zur Einheitsmatrix $I_2 \in \mathbb{R}^{2,2}$ mit der Bezeichnung $\check{T}_{pp} = T_{pp} - I_2$ bzw. $\check{T}_{qq} = T_{qq} - I_2$

$$\begin{aligned} \| \check{T}_{pp} \|_2 &= \| \check{T}_{qq} \|_2 \\ \| \check{T}_{pp} \|_2^2 + 2 \cdot \| \check{T}_{pp} \|_2 &= \| T_{pq} \|_2^2. \end{aligned} \tag{3.51}$$

Beweis:

Der Diagonalblock T_{pp} ist gegeben durch

$$T_{pp} := \begin{bmatrix} \cosh y_1 & 0 \\ 0 & \cosh y_2 \end{bmatrix}$$

Damit sind die beiden Eigenwerte von T_{pp} bekannt, und der Spektralradius lautet

$$\rho(T_{pp}) = \max \{ \cosh(y_1), \cosh(y_2) \}$$

Wegen

$$\cosh(y_i) = \frac{1}{\sqrt{1 - \tanh^2(y_i)}} \leq \frac{1}{\sqrt{1 - y_i^2}} \leq \frac{1}{\sqrt{1 - 4.735 \cdot 10^{-6}}} \leq 1 + 2.3675 \cdot 10^{-6}$$

folgt

$$\| T_{pp} \|_2 = \| T_{qq} \|_2 \leq (1 + 2.3675 \cdot 10^{-6})$$

Der Block T_{pq} ist im Gegensatz zu den Diagonalblöcken nichtsymmetrisch und damit gilt

$$\| T_{pq} \|_2 = \sqrt{\rho(T_{pq}^T \cdot T_{pq})} \quad \text{mit} \quad T_{pq}^T \cdot T_{pq} = \begin{bmatrix} \sinh^2 y_1 & 0 \\ 0 & \sinh^2 y_2 \end{bmatrix}$$

Hiermit folgt sofort

$$\| T_{pq} \|_2 = \max \{ |\sinh(y_1)|, |\sinh(y_2)| \}.$$

Mit

$$\begin{aligned} \sinh(y_i) = \frac{\tanh(y_i)}{\sqrt{1 - \tanh^2(y_i)}} & \leq (1 + 2.3675 \cdot 10^{-6}) \cdot \tanh(y_i) \\ \leq (1 + 2.3675 \cdot 10^{-6}) \cdot |y_i| & \leq (1 + 2.3675 \cdot 10^{-6}) \cdot \| x^{(1)} \| \end{aligned}$$

folgt die Abschätzung für $\| T_{pq} \|_2$. Es gilt schließlich

$$\breve{T}_{pp} = T_{pp} - I_2 = \begin{bmatrix} (\cosh y_1 - 1) & 0 \\ 0 & (\cosh y_2 - 1) \end{bmatrix}$$

Damit gilt

$$\rho(\check{T}_{pp}) = \rho(\check{T}_{qq}) = \max \{ \cosh(y_1) - 1, \cosh(y_2) - 1 \}$$

Sei o.B.d.A. $\rho(\check{T}_{pp}) = (\cosh(y_1) - 1)$. Dann gilt

$$\| \check{T}_{pp} \|_2^2 + 2 \cdot \| \check{T}_{pp} \|_2 = (\cosh(y_1) - 1)^2 + 2 \cdot (\cosh(y_1) - 1)$$
$$= \cosh^2(y_1) - 1 = \sinh^2(y_1) = \| \check{T}_{pq} \|_2,$$

denn mit $\rho(\check{T}_{pp}) = (\cosh(y_1) - 1)$ folgt $\sinh(y_1) \geq \sinh(y_2)$.

●

<u>Bemerkung:</u>

Aus der Darstellung von $\check{T}_{pp}$ ergibt sich für deren Norm

$$\| \check{T}_{pp} \|_2 = \frac{\| T_{pq} \|_2^2}{2 + \| \check{T}_{pp} \|_2} = \frac{\| T_{pq} \|_2^2}{2 + \max(ch_1 - 1, ch_2 - 1)} = \frac{\| T_{pq} \|_2^2}{1 + \max(ch_1, ch_2)}$$
$$\leq \frac{1}{2} \cdot \| T_{pq} \|_2^2 \leq \frac{1}{2} \cdot (1 + 2.3675 \cdot 10^{-6})^2 \cdot \| x^{(1)} \|^2 = O(S(A)^2)$$

mit $ch_i = \cosh(y_i)$ für $i = 1,2$. Damit folgt insgesamt

$$T_{pp} = I_2 + O(S(A)^2) \quad \text{und } T_{pq} = O(S(A)) \tag{3.52}$$

Also ist die hyperbolische Transformationsmatrix T in den Diagonalblöcken nur quadratisch in der Größenordnung der Außerdiagonalnorm von der Identität entfernt und in den von Null verschiedenen Außerdiagonalblöcken in der Größenordnung der Außerdiagonalnorm von Null verschieden. Für gegen Null konvergierende Außerdiagonalnorm $S(A)$ konvergiert damit T gegen die Einheitsmatrix in $\mathbb{R}^{n,n}$.

Die Blöcke der hyperbolischen Transformation $T(y_1, y_2)$ sind damit abgeschätzt und wir können nun berechnen, wie sich die Matrix A duch eine einzelne hyperbolische Transformation verändert.

3.7 Änderung der Matrixblöcke durch die hyperbolische Transformation

Durch eine hyperbolische Ähnlichkeitstransformation mit der Matrix $T(y_1,y_2)$ beim Pivot $(p,q) \in P_m$ verändern sich die Blöcke A_{pp}, A_{qq} sowie A_{pq} der Matrix A. Der nachfolgende Satz gibt Abschätzungen für die Größenordnung dieser Veränderung an.

Satz 3.15

Es gelte die Generalvoraussetzung (3.28). Die Parameter y_1 und y_2 seien durch das Newton-Verfahren (3.26) für das Pivotpaar $(p,q) \in P_m$ bestimmt. Dann folgt für die Pivotblöcke A'_{ij} der Matrix A' nach der hyperbolischen Transformation $T(y_1,y_2)$

$$\begin{aligned} A'_{ii} &= A_{ii} + W_{ii} \quad \text{mit} \\ \| W_{ii} \|_F &\le 1992.2 \cdot \frac{\| A \|_F^9}{\delta^{10}} \cdot S^2(A) \quad \text{für } i = p,q \end{aligned} \tag{3.53}$$

$$\begin{aligned} A'_{pq} &= A_{pq} + W_{pq} \quad \text{mit} \\ \| W_{pq} \|_F &\le 49.23 \cdot \frac{\| A \|_F^5}{\delta^5} \cdot S(A) \end{aligned} \tag{3.54}$$

Insbesondere gilt $\| A'_{pq} \|_F \le 2.35244 \cdot 10^{-3} \cdot \delta$ (3.55)

Beweis:

Unter Verwendung der Formeln (3.7) für die blockweise Veränderung der Matrix A unter der Ähnlichkeitstransformation mit $T(y_1,y_2)$ gilt

$$\begin{aligned} &A'_{pp} \\ &= T_{pp} \cdot A_{pp} \cdot T_{pp} + T_{pp} \cdot A_{pq} \cdot T_{qp} - T_{pq} \cdot A_{qp} \cdot T_{pp} - T_{pq} \cdot A_{qq} \cdot T_{qp} \\ &= A_{pp} - \left[A_{pp} - T_{pp} \cdot A_{pp} \cdot T_{pp} \right] + T_{pp} \cdot A_{pq} \cdot T_{qp} - T_{pq} \cdot A_{qp} \cdot T_{pp} - T_{pq} \cdot A_{qq} \cdot T_{qp} \\ &= A_{pp} + \left[\breve{T}_{pp} \cdot A_{pp} + A_{pp} \cdot \breve{T}_{pp} + \breve{T}_{pp} \cdot A_{pp} \cdot \breve{T}_{pp} \right] \\ &\qquad + T_{pp} \cdot A_{pq} \cdot T_{qp} - T_{pq} \cdot A_{qp} \cdot T_{pp} - T_{pq} \cdot A_{qq} \cdot T_{qp} \end{aligned}$$

Somit erhalten wir als Darstellung der Veränderung von A_{pp}

$$\begin{aligned} W_{pp} = & \left[\check{T}_{pp} \cdot A_{pp} + A_{pp} \cdot \check{T}_{pp} + \check{T}_{pp} \cdot A_{pp} \cdot \check{T}_{pp} \right] \\ & + T_{pp} \cdot A_{pq} \cdot T_{qp} - T_{pq} \cdot A_{qp} \cdot T_{pp} - T_{pq} \cdot A_{qq} \cdot T_{qp} , \end{aligned}$$

und es folgt

$$\begin{aligned} \| W_{pp} \| \le \; & 2 \cdot \| \check{T}_{pp} \|_2 \cdot \| A_{pp} \| + \| \check{T}_{pp} \|_2^2 \cdot \| A_{pp} \| + \| A_{pq} \| \cdot \| T_{pp} \|_2 \cdot \| T_{qp} \|_2 \\ & + \| A_{qp} \| \cdot \| T_{pp} \|_2 \cdot \| T_{pq} \|_2 + \| A_{qq} \| \cdot \| T_{pq} \|_2 \cdot \| T_{qp} \|_2 \\ = \; & \| A_{pp} \| \cdot \left[2 \cdot \| \check{T}_{pp} \|_2 + \| \check{T}_{pp} \|_2^2 \right] \\ & + 2 \cdot \| A_{pq} \| \cdot \| T_{pp} \|_2 \cdot \| T_{pq} \|_2 + \| A_{qq} \| \cdot \| T_{pq} \|_2^2 \\ \le \; & \| T_{pq} \|_2^2 \cdot \left[\| A_{pp} \| + \| A_{qq} \| \right] + 2 \cdot \| A_{pq} \| \cdot \| T_{pp} \|_2 \cdot \| T_{pq} \|_2 . \end{aligned}$$

Die letzte Abschätzung folgt wegen $2 \cdot \| \check{T}_{pp} \|_2 + \| \check{T}_{pp} \|_2^2 = \| T_{pq} \|_2^2$ nach Satz (3.13) und wegen $\| \check{T}_{pp} \|_\infty = \| \check{T}_{pp} \|_2$ bzw. $\| T_{ij} \|_\infty = \| T_{ij} \|_2$ für $(i,j) \in \{ (p,q),(p,p) \}$.

Wegen $\| A_{pp} \| + \| A_{qq} \| \le \sqrt{2} \cdot \| A \|_F$ und $\| A_{pq} \| \le \frac{\sqrt{2}}{2} \cdot S(A)$ ergibt sich

$$\begin{aligned} \| W_{pp} \| \le \; & \| T_{pq} \|_2^2 \cdot \sqrt{2} \cdot \| A \|_F + 2 \cdot \frac{\sqrt{2}}{2} \cdot S(A) \cdot \| T_{pp} \|_2 \cdot \| T_{pq} \|_2 \\ \le \; & \sqrt{2} \cdot (1 + 2.3675 \cdot 10^{-6})^2 \cdot \\ & \cdot \left[\| x^{(1)} \|^2 \cdot \| A \|_F + S(A) \cdot \| x^{(1)} \| \right] \end{aligned}$$

Mit $\| x^{(1)} \| \le \sum_{i=0}^{4} \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \cdot \frac{1}{0.976 \cdot \delta} \cdot \frac{\sqrt{2}}{2} \cdot S(A)$ folgt

$$\| W_{pp} \| \le \sqrt{2} \cdot (1 + 2.3675 \cdot 10^{-6})^2 \cdot S^2(A) \cdot$$

$$\left[\left[\sum_{i=0}^{4} \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \cdot \frac{1}{0.976 \cdot \delta} \cdot \frac{\sqrt{2}}{2} \right]^2 \cdot \| A \|_F + \sum_{i=0}^{4} \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \cdot \frac{1}{0.976 \cdot \delta} \cdot \frac{\sqrt{2}}{2} \right]$$

$$\leq \sqrt{2}\cdot(1 + 2.3675\cdot 10^{-6})^2\cdot S^2(A)\cdot$$

$$\left[\left[\sum_{i=0}^{4} 2.05^{i}\cdot\left[\frac{\delta}{\|A\|_F}\right]^{4-i}\cdot\frac{1}{0.976}\cdot\frac{\sqrt{2}}{2}\right]^2\cdot\frac{\|A\|_F^9}{\delta^{10}}\right.$$

$$\left.+\frac{\|A\|_F^4}{\delta^5}\cdot\sum_{i=0}^{4} 2.05^{i}\cdot\left[\frac{\delta}{\|A\|_F}\right]^{4-i}\cdot\frac{1}{0.976\cdot\delta}\cdot\frac{\sqrt{2}}{2}\right]$$

$$\leq \sqrt{2}\cdot(1 + 2.3675\cdot 10^{-6})^2\cdot S^2(A)\cdot\frac{\|A\|_F^9}{\delta^{10}}\cdot$$

$$\left[\left[\sum_{i=0}^{4} 2.05^{i}\cdot\left[\frac{\delta}{\|A\|_F}\right]^{4-i}\cdot\frac{1}{0.976}\cdot\frac{\sqrt{2}}{2}\right]^2\right.$$

$$\left.+\frac{\delta^5}{\|A\|_F^5}\cdot\sum_{i=0}^{4} 2.05^{i}\cdot\left[\frac{\delta}{\|A\|_F}\right]^{4-i}\cdot\frac{1}{0.976\cdot\delta}\cdot\frac{\sqrt{2}}{2}\right]$$

Wegen $\delta \leq \sqrt{2}\cdot\|A\|_F$ folgt insgesamt

$$\|W_{pp}\| \leq C\cdot\frac{\|A\|_F^9}{\delta^{10}}\cdot S^2(A) \text{ mit}$$

$$C = \sqrt{2}\cdot(1+2.3675\cdot 10^{-6})^2\cdot\left[\left[\frac{\sqrt{2}}{0.976\cdot 2}\cdot\sum_{i=0}^{4} 2.05^{i}\cdot\sqrt{2}^{4-i}\right]^2 + \frac{4}{0.976}\cdot\sum_{i=0}^{4} 2.05^{i}\cdot\sqrt{2}^{4-i}\right]$$

$$\approx 1992.19126$$

Der Beweis für den Block A_{qq} ist identisch. Für den Nichtdiagonalblock A'_{pq} erhalten wir in Blockschreibweise

$$\begin{aligned}
&A'_{pq}\\
&= T_{pp}\cdot A_{pp}\cdot T_{pq} + T_{pp}\cdot A_{pq}\cdot T_{qq} - T_{pq}\cdot A_{qp}\cdot T_{pq} - T_{pq}\cdot A_{qq}\cdot T_{qq}\\
&= A_{pq} - \left[A_{pq} - T_{pp}\cdot A_{pq}\cdot T_{qq}\right] + T_{pp}\cdot A_{pp}\cdot T_{pq} - T_{pq}\cdot A_{qp}\cdot T_{pq} - T_{pq}\cdot A_{qq}\cdot T_{qq}\\
&= A_{pq} + \left[\breve{T}_{pp}\cdot A_{pq} + A_{pq}\cdot\breve{T}_{qq} + \breve{T}_{pp}\cdot A_{pq}\cdot\breve{T}_{qq}\right]\\
&\quad + T_{pp}\cdot A_{pp}\cdot T_{pq} - T_{pq}\cdot A_{qp}\cdot T_{pq} - T_{pq}\cdot A_{qq}\cdot T_{qq}
\end{aligned}$$

Somit erhalten wir als Darstellung der Veränderung von A_{pp}

$$\begin{aligned}
W_{pq} = &\left[\breve{T}_{pp}\cdot A_{pq} + A_{pq}\cdot\breve{T}_{qq} + \breve{T}_{pp}\cdot A_{pq}\cdot\breve{T}_{qq}\right]\\
&+ T_{pp}\cdot A_{pp}\cdot T_{pq} - T_{pq}\cdot A_{qp}\cdot T_{pq} - T_{pq}\cdot A_{qq}\cdot T_{qq},
\end{aligned}$$

und es folgt

$$
\begin{aligned}
\| W_{pq} \| \le\ & 2\cdot\| \breve{T}_{pp} \|_2 \cdot \| A_{pq} \| + \| \breve{T}_{pp} \|_2^2 \cdot \| A_{pq} \| + \| A_{pp} \| \cdot \| T_{pp} \|_2 \cdot \| T_{pq} \|_2 \\
& + \| A_{pq} \| \cdot \| T_{pq} \|_2^2 + \| A_{qq} \| \cdot \| T_{pq} \|_2 \cdot \| T_{qq} \|_2 \\
=\ & \| A_{pq} \| \cdot \left[2\cdot \| \breve{T}_{pp} \|_2 + \| \breve{T}_{pp} \|_2^2 \right] \\
& + \| A_{pq} \| \cdot \| T_{pq} \|_2^2 + \left[\| A_{pp} \| + \| A_{qq} \| \right] \cdot \| T_{pq} \|_2 \cdot \| T_{pp} \|_2 \\
\le\ & \| T_{pq} \|_2 \cdot \| T_{pp} \|_2 \cdot \left[\| A_{pp} \| + \| A_{qq} \| \right] + 2\cdot \| A_{pq} \| \cdot \| T_{pq} \|_2^2
\end{aligned}
$$

Es folgt wie oben

$$
\begin{aligned}
\| W_{pq} \| \le\ & 2\cdot(1 + 2.3675\cdot 10^{-6})\cdot \| x^{(1)} \|^2 \cdot \frac{\sqrt{2}}{2} \cdot S(A) \\
& +(1 + 2.3675\cdot 10^{-6})\cdot \| x^{(1)} \| \cdot \sqrt{2} \cdot \| A \|_F
\end{aligned}
$$

Mit der obigen Abschätzung für $\| x^{(1)} \|$ und mit der gröberen Abschätzung $\| x^{(1)} \| \le 2.176\cdot 10^{-3}$ folgt

$$
\begin{aligned}
\| W_{pq} \| \le\ & (1 + 2.3675\cdot 10^{-6})\cdot \sqrt{2} \cdot (2.176\cdot 10^{-3})^2 \cdot S(A) \\
& +(1 + 2.3675\cdot 10^{-6})\cdot \sum_{i=0}^{4} \left[2.05\cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \cdot \frac{\| A \|_F}{0.976\cdot \delta} \cdot S(A) \\
\le\ & \Bigg\{ (1 + 2.3675\cdot 10^{-6})\cdot \sqrt{2} \cdot (2.176\cdot 10^{-3})^2 \cdot \frac{\delta^5}{\| A \|_F^5} \\
& +(1 + 2.3675\cdot 10)\cdot \sum_{i=0}^{4} 2.05^i \cdot \left[\frac{\delta}{\| A \|_F} \right]^{4-i} \cdot \frac{1}{0.976} \Bigg\} \cdot \frac{\| A \|_F^5}{\delta^5} \cdot S(A),
\end{aligned}
$$

also insgesamt

$$
\| W_{pq} \| \le C \cdot \frac{\| A \|_F^5}{\delta^5} \cdot S(A)
$$

$$
\text{mit } C = \left\{ (1 + 2.3675\cdot 10^{-6})\cdot \left\{ (2.176\cdot 10^{-3})^2 \cdot 8 + \frac{1}{0.976} \cdot \sum_{i=0}^{4} 2.05^i \cdot \left[\frac{\delta}{\| A \|_F} \right] \right\} \right\}
$$

$$
\approx 49.22957
$$

Mit der Generalvoraussetzung (3.28) erhalten wir aus der letzten Abschätzung

$$\| A'_{pq} \| \le \| A_{pq} \| + \| W_{pq} \|$$

$$\le \frac{\sqrt{2}}{2} \cdot S(A) + 49.22957 \frac{\| A \|_F^5}{\delta^5} \cdot S(A)$$

$$\le \left\{ \frac{\sqrt{2}}{2} \cdot \frac{\delta^5}{\| A \|_F^5} + 49.22957 \right\} \frac{\| A \|_F^5}{\delta^5} \cdot \frac{1}{6.4 \cdot 10^4 \cdot \sqrt{m-1}} \cdot \frac{\delta^9}{\| A \|_F^8}$$

$$\le \left\{ 4 + 49.22957 \right\} \cdot \frac{2 \cdot \sqrt{2}}{6.4 \cdot 10^4} \cdot \delta \le 2.35244 \cdot 10^{-3} \cdot \delta$$

●

3.8 Wachstum der Außerdiagonale und der Norm

Mit Hilfe der Abschätzungen der Normveränderung unter der hyperbolischen Transformation T läßt sich der maximale Außerdiagonalnormzuwachs sowie der maximale Normzuwachs für den hyperbolischen Teil einer Ähnlichkeitstransformation abschätzen. Dies ist die Aussage der folgenden beiden Sätze.

Satz 3.16

Es gelte die Generalvoraussetzung (3.28). Die Parameter y_1 und y_2 seien durch das Newton-Verfahren (3.26) für das Pivotpaar $(p,q) \in P_m$ bestimmt. Dann folgt für den Normzuwachs der Außerdiagonalnorm nach der hyperbolischen Transformation

$$S(A') \leq 53.976 \cdot \frac{\| A \|_F^5}{\delta^5} \cdot S(A) \tag{3.56}$$

Beweis:

Für die Änderung der Außerdiagonalnorm ergibt sich

$$\begin{aligned} S^2(A') - S^2(A) &= \sum_{\substack{i,j=1 \\ i \neq j}}^{m} \| A'_{ij} \|_F^2 - \sum_{\substack{i,j=1 \\ i \neq j}}^{m} \| A_{ij} \|_F^2 = \\ &= 2 \cdot \sum_{i \neq p,q}^{m} \left[\| A'_{pi} \|_F^2 + \| A'_{qi} \|_F^2 \right] - 2 \cdot \sum_{i \neq p,q}^{m} \left[\| A_{pi} \|_F^2 + \| A_{qi} \|_F^2 \right] \\ &\quad + 2 \cdot \left[\| A'_{pq} \|_F^2 - \| A_{pq} \|_F^2 \right], \end{aligned}$$

denn es werden lediglich die p-te und q-te Blockzeile bzw -spalte durch die zum Pivotpaar (p,q) gehörende Transformation geändert. Die Blockdarstellung (3.6) der Veränderung liefert

$$\begin{aligned} \| A'_{pi} \|_F &\leq \| T_{pp} \|_2 \cdot \| A_{pi} \|_F + \| T_{pq} \|_2 \cdot \| A_{qi} \|_F \\ \| A'_{qi} \|_F &\leq \| T_{qp} \|_2 \cdot \| A_{pi} \|_F + \| T_{qq} \|_2 \cdot \| A_{qi} \|_F \end{aligned}$$

Quadrieren der Formeln und anschließende Anwendung der Cauchy - Schwarzschen Ungleichung ergibt

$$\begin{aligned}\| A'_{pi} \|_F^2 &\le \| T_{pp} \|_2^2 \cdot \| A_{pi} \|_F^2 + \| T_{pq} \|_2^2 \cdot \| A_{qi} \|_F^2 \\ &\quad + 2 \cdot \| T_{pp} \|_2 \cdot \| A_{pi} \|_F \cdot \| T_{pq} \|_2 \cdot \| A_{qi} \|_F \\ &\le 2 \cdot \left[\| T_{pp} \|_2^2 \cdot \| A_{pi} \|_F^2 + \| T_{pq} \|_2^2 \cdot \| A_{qi} \|_F^2 \right].\end{aligned}$$

Analog folgt

$$\| A'_{qi} \|^2 \le 2 \cdot \left[\| T_{qp} \|_2^2 \cdot \| A_{pi} \|_F^2 + \| T_{qq} \|_2^2 \cdot \| A_{qi} \|_F^2 \right].$$

Mit Hilfe der Abschätzung (3.50) für $\| T_{ij} \|_2$ mit $(i,j) \in \{ (p,q),(q,p),(p,p),(q,q) \}$ folgt

$$\begin{aligned}\| A'_{pi} \|_F^2 &\le 2 \cdot \Big[(1 + 2.3675 \cdot 10^{-6})^2 \cdot 2.176^2 \cdot 10^{-6} \cdot \| A_{pi} \|_F^2 \\ &\quad + (1 + 2.3675 \cdot 10^{-6})^2 \cdot \| A_{qi} \|_F^2 \Big].\end{aligned}$$

$$\begin{aligned}\| A'_{qi} \|_F^2 &\le 2 \cdot \Big[(1 + 2.3675 \cdot 10^{-6})^2 \cdot 2.176^2 \cdot 10^{-6} \cdot \| A_{pi} \|_F^2 \\ &\quad + (1 + 2.3675 \cdot 10^{-6})^2 \cdot \| A_{qi} \|_F^2 \Big].\end{aligned}$$

und damit

$$\begin{aligned}\| A'_{pi} \|_F^2 &\le 9.5 \cdot 10^{-6} \cdot \| A_{qi} \|_F^2 + (2 + 9.5 \cdot 10^{-6}) \cdot \| A_{pi} \|_F^2 \\ \| A'_{qi} \|_F^2 &\le 9.5 \cdot 10^{-6} \cdot \| A_{pi} \|_F^2 + (2 + 9.5 \cdot 10^{-6}) \cdot \| A_{qi} \|_F^2 .\end{aligned} \tag{3.57}$$

Durch Summation der Formeln über i erhält man

$$\begin{aligned}&2 \cdot \sum_{i \ne p,q}^{m} \left[\| A'_{pi} \|_F^2 + \| A'_{qi} \|_F^2 \right] - 2 \cdot \sum_{i \ne p,q}^{m} \left[\| A_{pi} \|_F^2 + \| A_{qi} \|_F^2 \right] = \\ &\le 2 \cdot \sum_{i \ne p,q}^{m} \left[9.5 \cdot 10^{-6} \cdot \| A_{qi} \|_F^2 + (1 + 9.5 \cdot 10^{-6}) \cdot \| A_{pi} \|_F^2 \right] \\ &\quad + 2 \cdot \sum_{i \ne p,q}^{m} \left[9.5 \cdot 10^{-6} \cdot \| A_{pi} \|_F^2 + (1 + 9.5 \cdot 10^{-6}) \cdot \| A_{qi} \|_F^2 \right]\end{aligned}$$

$$= 2 \cdot \sum_{i \neq p,q}^{m} \left[(1 + 2 \cdot 9.5 \cdot 10^{-6}) \cdot \| A_{qi} \|_F^2 + (1 + 2 \cdot 9.5 \cdot 10^{-6}) \cdot \| A_{pi} \|_F^2 \right]$$

$$= 2 \cdot (1 + 2 \cdot 9.5 \cdot 10^{-6}) \cdot \sum_{i \neq p,q}^{m} \left[\| A_{qi} \|_F^2 + \| A_{pi} \|_F^2 \right]$$

$$= (2 + 1.9 \cdot 10^{-5}) \cdot \sum_{i \neq p,q}^{m} \left[\| A_{qi} \|_F^2 + \| A_{pi} \|_F^2 \right] \leq (2 + 1.9 \cdot 10^{-5}) \cdot S^2(A)$$

Aus der Darstellung von A'_{pq} nach (3.54) ergibt sich

$$\| A'_{pq} \|_F^2 - \| A_{pq} \|_F^2$$

$$\leq 2 \cdot \| A_{pq} \|_F \cdot \| W_{pq} \|_F + \| W_{pq} \|_F^2$$

$$\leq \sqrt{2} \cdot 49.22957 \frac{\| A \|_F^5}{\delta^5} \cdot S^2(A) + 49.22957^2 \cdot \frac{\| A \|_F^{10}}{\delta^{10}} \cdot S^2(A)$$

$$\leq \left[\sqrt{2} \cdot 49.22957 \cdot \frac{\delta^5}{\| A \|_F^5} + 49.22957^2 \right] \cdot \frac{\| A \|_F^{10}}{\delta^{10}} \cdot S^2(A)$$

$$\leq \left[8 \cdot 49.22957 + 49.22957^2 \right] \cdot \frac{\| A \|_F^{10}}{\delta^{10}} \cdot S^2(A) = 2817.387 \cdot \frac{\| A \|_F^{10}}{\delta^{10}} \cdot S^2(A)$$

Man addiert jetzt beide Abschätzungen und bekommt für den Normzuwachs

$$S^2(A') = S^2(A) + (2 + 1.9 \cdot 10^{-5}) \cdot S^2(A) + 2817.387 \cdot \frac{\| A \|_F^{10}}{\delta^{10}} \cdot S^2(A)$$

$$\leq \left[(3 + 1.9 \cdot 10^{-5}) \cdot \frac{\delta^{10}}{\| A \|_F^{10}} + 2817.387 \right] \cdot \frac{\| A \|_F^{10}}{\delta^{10}} \cdot S^2(A)$$

$$\leq 2913.387 \cdot \frac{\| A \|_F^{10}}{\delta^{10}} \cdot S^2(A)$$

oder

$$S(A') \leq 53.976 \cdot \frac{\| A \|_F^5}{\delta^5} \cdot S(A)$$

●

Wir kennen nun den Außerdiagonalnormzuwachs und sind damit in der Lage, den maximalen Normzuwachs unter der hyperbolischen Transformation abzuschätzen.

Satz 3.17

Es gelte die Generalvoraussetzung (3.28). Die Parameter y_1 und y_2 seien durch das Newton-Verfahren (3.26) für das Pivotpaar $(p,q) \in P_m$ bestimmt. Dann folgt für den maximalen Normzuwachs der Matrix A durch die hyperbolische Transformation:

$$\| A' \|_F \leq (1 + 2.1225 \cdot 10^{-5}) \cdot \| A \|_F \tag{3.58}$$

Beweis:

Die Veränderung der Norm der Matrix ist gegeben durch

$$\| A' \|_F^2 - \| A \|_F^2 = \sum_{\substack{i,j=1 \\ i \neq j}}^{m} \| A'_{ij} \|_F^2 + \sum_{i=1}^{m} \| A'_{ii} \|_F^2 - \sum_{\substack{i,j=1 \\ i \neq j}}^{m} \| A_{ij} \|_F^2 - \sum_{i=1}^{m} \| A_{ii} \|_F^2$$

$$= S^2(A') - S^2(A) + \| A'_{pp} \|_F^2 - \| A_{pp} \|_F^2 + \| A'_{qq} \|_F^2 - \| A_{qq} \|_F^2$$

Die Veränderungen der Außerdiagonalnorm und der Diagonalblöcke sind bereits abgeschätzt. Aus dem Satz 3.16 folgt also

$$S^2(A') - S^2(A) \leq \left[2929.8729 \cdot \frac{\| A \|_F^{10}}{\delta^{10}} - 1 \right] \cdot S^2(A)$$

Nach (3.53) gilt weiter

$$\| A'_{ii} \|_F^2 - \| A_{ii} \|_F^2 \leq 2 \cdot \| A_{ii} \|_F \cdot \| W_{ii} \|_F + \| W_{ii} \|_F^2$$

$$\| W_{ii} \|_F \leq 1992.2 \cdot \frac{\| A \|_F^9}{\delta^{10}} \cdot S^2(A) \qquad \text{für } i = p,q$$

und damit

$$\| A'_{ii} \|_F^2 - \| A_{ii} \|_F^2 \leq \left[2 \cdot 1992.2 + 1992.2^2 \cdot \frac{\| A \|_F^8}{\delta^{10}} \cdot S^2(A) \right] \cdot \frac{\| A \|_F^{10}}{\delta^{10}} \cdot S^2(A)$$

$$\leq \left[2 \cdot 1992.2 + 1992.2^2 \cdot \frac{16}{4.096 \cdot 10^9} \right] \cdot \frac{\| A \|_F^{10}}{\delta^{10}} \cdot S^2(A)$$

$$\leq 3984.4155 \cdot \frac{\| A \|_F^{10}}{\delta^{10}} \cdot S^2(A)$$

Damit folgt mit der Generalvoraussetzung (3.28)

$$\| A' \|_F^2 - \| A \|_F^2$$

$$\leq \left[2929.8729 \cdot \frac{\| A \|_F^{10}}{\delta^{10}} - 1 \right] \cdot S^2(A) + 2 \cdot 3984.4155 \cdot \frac{\| A \|_F^{10}}{\delta^{10}} \cdot S^2(A)$$

$$\leq \left[10898.704 - \frac{\delta^{10}}{\| A \|_F^{10}} \right] \cdot \frac{\| A \|_F^{10}}{\delta^{10}} \cdot S^2(A)$$

$$\leq 10866.704 \cdot \frac{1}{4.096 \cdot 10^9} \cdot \frac{\delta^8}{\| A \|_F^8} \cdot \| A \|_F^2$$

$$\leq 4.245 \cdot 10^{-5} \cdot \| A \|_F^2.$$

Wurzelziehen ergibt die Behauptung.

●

Mit diesen beiden Sätzen sind alle wesentlichen Veränderungen, die die hyperbolische Transformation an der Matrix A vornimmt, charakterisiert.

3.9 Abschätzung der Transformationsmatrix der orthogonalen Transformation

Es ist das Ziel der nächsten beiden Abschnitte, die Normveränderungen unter der orthogonalen Transformation U abzuschätzen, um auch bei der zweiten Ähnlichkeitstransformation den Außerdiagonalnormzuwachs zu kontrollieren. Dazu betrachten wir zunächst die einzelnen (i,j) - Restriktionen von U:

Satz 3.18

Es gelte die Generalvoraussetzung (3.28). Die Parameter x_1 und x_2 seien durch das Newton - Verfahren (3.26) für das Pivotpaar $(p,q) \in P_m$ bestimmt. Dann folgt für die Pivotblöcke der orthogonalen Transformation $U(x_1,x_2)$

$$\begin{aligned} \| U_{pp} \|_2 &= \| U_{qq} \|_2 \le 1 \\ \| U_{pq} \|_2 &= \| U_{qp} \|_2 \le \| x^{(1)} \|, \end{aligned} \tag{3.59}$$

wobei

$$\begin{aligned} \| x^{(1)} \| &\le \sum_{i=0}^{4} \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \cdot \frac{1}{0.976 \cdot \delta} \cdot \frac{\sqrt{2}}{2} \cdot S(A) \\ &\le 5.44 \cdot 10^{-4} \cdot \frac{\delta^4}{\| A \|_F^4} \le \ldots \le 2.176 \cdot 10^{-3} \quad \text{ist.} \end{aligned}$$

Weiter gilt für den Abstand der Diagonalblöcke U_{ii} für i=p,q zur Einheitsmatrix $I_2 \in \mathbb{R}^{2,2}$ mit der Bezeichnung $\breve{U}_{pp} = U_{pp} - I_2$ bzw. $\breve{U}_{qq} = U_{qq} - I_2$

$$\begin{aligned} \| \breve{U}_{pp} \|_2 &= \| \breve{U}_{qq} \|_2 \\ \| \breve{U}_{pp} \|_2^2 + 2 \cdot \| \breve{U}_{pp} \|_2 &\le (1 + 7.1 \cdot 10^{-6}) \cdot \| U_{pq} \|_2^2. \end{aligned} \tag{3.60}$$

Beweis:

Der Diagonalblock U_{pp} ist gegeben durch

$$U_{pp} := \begin{bmatrix} \cos x_1 & 0 \\ 0 & \cos x_2 \end{bmatrix}$$

Damit sind die Eigenwerte der Matrix U_{pp} bekannt, und es gilt

$$\| U_{pp} \|_2 = \rho(U_{pp}) = \max \{ \cos(x_1), \cos(x_2) \} \leq 1$$

Wegen $U_{pq} = -U_{qp} = \begin{bmatrix} \sin x_1 & 0 \\ 0 & \sin x_2 \end{bmatrix}$ ist

$$\| U_{pq} \|_2 = \| U_{qp} \|_2 = \rho(U_{pq}) = \max \{ |\sin(x_1)|, |\sin(x_2)| \}$$

$$\| U_{pq} \|_2 = \| U_{qp} \|_2 \leq \max \{ |x_1|, |x_2| \} \leq \| x^{(1)} \|$$

mit $\quad \| x^{(1)} \| \leq \sum_{i=0}^{4} \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \cdot \frac{1}{0.976 \cdot \delta} \cdot \frac{\sqrt{2}}{2} \cdot S(A)$

Für $\check{U}_{pp}$ erhalten wir somit

$$\check{U}_{pp} = U_{pp} - I_2 = \begin{bmatrix} (\cos(x_1)-1) & 0 \\ 0 & (\cos(x_2)-1) \end{bmatrix}$$

$$\| \check{U}_{pp} \|_2 = \| \check{U}_{qq} \|_2 = \max \{ 1-\cos(x_1), 1-\cos(x_2) \}.$$

Sei o.B.d.A. $\| \check{U}_{pp} \|_2 = 1 - \cos(x_1)$. Dann folgt

$$\begin{aligned} \| \check{U}_{pp} \|_2^2 + 2 \cdot \| \check{U}_{pp} \|_2 &= (1-\cos(x_1))^2 + 2 \cdot (1-\cos(x_1)) \\ &\leq \frac{2 \cdot \sin^2(x_1)}{\sqrt{1 - \sin^2(x_1)}} - \sin^2(x_1) \\ &\leq \left[\frac{2}{\sqrt{1 - \sin^2(x_1)}} - \sqrt{1 - \sin^2(x_1)} \right] \cdot \sin^2(x_1) \\ &\leq \left[\frac{2}{\sqrt{1 - \|x^{(1)}\|^2}} - \sqrt{1 - \|x^{(1)}\|^2} \right] \cdot \sin^2(x_1) \end{aligned}$$

Mit $1 - \cos(x_1) = \max \{ 1-\cos(x_1), 1-\cos(x_2) \}$ folgt $\sin(x_1) = \max \{ |\sin(x_1)|, |\sin(x_2)| \}$ und damit insgesamt wegen $\| x^{(1)} \| \leq 2.176 \cdot 10^{-3}$

$$\begin{aligned} \| \check{U}_{pp} \|_2^2 + 2 \cdot \| \check{U}_{pp} \|_2 &\leq \left[\frac{2}{\sqrt{1 - 2.176^2 \cdot 10^{-6}}} - \sqrt{1 - 2.176^2 \cdot 10^{-6}} \right] \cdot \| U_{pq} \|_2^2 \\ &\leq \left[\frac{2}{\sqrt{1 - 2.176^2 \cdot 10^{-6}}} - \sqrt{1 - 2.176^2 \cdot 10^{-6}} \right] \cdot \| U_{pq} \|_2^2 \\ &\leq (1 + 7.1 \cdot 10^{-6}) \cdot \| U_{pq} \|_2^2 \end{aligned}$$

●

3.10 Änderung der Matrixblöcke durch die orthogonale Transformation

Nach der Abschätzung der Blöcke der Transformationsmatrix U können wir nun die blockweise Veränderung der Matrix A' abschätzen.

Satz 3.19

Es gelte die Generalvoraussetzung (3.28). Die Parameter x_1 und x_2 seien durch das Newton-Verfahren (3.26) für das Pivotpaar $(p,q) \in P_m$ bestimmt. Dann folgt für die Pivotblöcke A''_{ij} der Matrix A'' nach beiden Ähnlichkeitstransformationen

$$\begin{aligned} A''_{ii} &= A'_{ii} + W'_{ii} \quad \text{mit} \\ \| W'_{ii} \|_F &\leq 3960.73 \cdot \frac{\| A \|_F^9}{\delta^{10}} \cdot S^2(A) \quad \text{für } i = p,q \end{aligned} \tag{3.61}$$

$$\begin{aligned} A''_{pq} &= A'_{pq} + W'_{pq} \quad \text{mit} \\ \| W'_{pq} \|_F &\leq 49.23097 \cdot \frac{\| A \|_F^5}{\delta^5} \cdot S(A) \end{aligned} \tag{3.62}$$

Beweis:

Mit der blockweisen Darstellung der orthogonalen Transformation nach (3.10) ergibt sich für die Änderung der Diagonalblöcke

$$A''_{pp} = A'_{pp} + W'_{pp} \quad \text{mit}$$

$$W'_{pp} = -A'_{pp} + U_{pp} \cdot A'_{pp} \cdot U_{pp} - U_{pq} \cdot A'_{qp} \cdot U_{pp} + U_{pp} \cdot A'_{pq} \cdot U_{qp} - U_{pq} \cdot A'_{qq} \cdot U_{qp}$$

Mit

$$\breve{U}_{pp} \cdot A'_{pp} + A'_{pp} \cdot \breve{U}_{pp} + \breve{U}_{pp} \cdot A'_{pp} \cdot \breve{U}_{pp} = U_{pp} \cdot A'_{pp} \cdot U_{pp} - A'_{pp}$$

erhalten wir für die Veränderung von A'_{pp}

$$\begin{aligned} W'_{pp} = & \left[\breve{U}_{pp} \cdot A'_{pp} + A'_{pp} \cdot \breve{U}_{pp} + \breve{U}_{pp} \cdot A'_{pp} \cdot \breve{U}_{pp} \right] \\ & - U_{pq} \cdot A'_{qp} \cdot U_{pp} + U_{pp} \cdot A'_{pq} \cdot U_{qp} - U_{pq} \cdot A'_{qq} \cdot U_{qp} \end{aligned}$$

und

$$\begin{aligned}
&\| W'_{pp} \|_F \\
&\le 2\cdot\| \breve{U}_{pp} \|_2\cdot\| A'_{pp} \|_F + \| \breve{U}_{pp} \|_2^2\cdot\| A'_{pp} \|_F + \| A'_{qp} \|_F\cdot\| U_{pp} \|_2\cdot\| U_{pq} \|_2 \\
&\quad + \| A'_{pq} \|_F\cdot\| U_{pp} \|_2\cdot\| U_{qp} \|_2 + \| A'_{qq} \|_F\cdot\| U_{pq} \|_2\cdot\| U_{qp} \|_2 \\
&= \| A'_{pp} \|_F\cdot\left[2\cdot\| \breve{U}_{pp} \|_2 + \| \breve{U}_{pp} \|_2^2\right] \\
&\quad + 2\cdot\| A'_{pq} \|_F\cdot\| U_{pp} \|_2\cdot\| U_{pq} \|_2 + \| A'_{qq} \|_F\cdot\| U_{pq} \|_2^2 \\
&\le (1+7.1\cdot10^{-6})\cdot\| U_{pq} \|_2^2\cdot\left[\| A'_{pp} \|_F + \| A'_{qq} \|_F\right] + 2\cdot\| A'_{pq} \|_F\cdot\| U_{pp} \|_2\cdot\| U_{pq} \|_2.
\end{aligned}$$

Die letzte Abschätzung folgt wegen (3.59) und wegen $\| \breve{U}_{pp} \|_\infty = \| \breve{U}_{pp} \|_2$ bzw. $\| U_{ij} \|_\infty = \| U_{ij} \|_2$ für $(i,j) \in \{ (p,q),(q,p) \}$.

Wegen $\| A'_{pp} \|_F + \| A'_{qq} \|_F \le \sqrt{2}\cdot\| A' \|_F \le \sqrt{2}\cdot(1 + 2.1225\cdot10^{-5})\cdot\| A \|_F$ und $\| A'_{pq} \|_F \le \frac{\sqrt{2}}{2}\cdot S(A') \le 32.2745\cdot\frac{\| A \|_F^5}{\delta^5}\cdot S(A)$ ergibt sich

$$\begin{aligned}
\| W'_{pp} \|_F &\le (1+7.1\cdot10^{-6})\cdot\sqrt{2}\cdot(1 + 2.1225\cdot10^{-5})\cdot\| U_{pq} \|_2^2\cdot\| A \|_F \\
&\quad + 2\cdot32.2745\cdot\frac{\| A \|_F^5}{\delta^5}\cdot S(A)\cdot\| U_{pp} \|_2\cdot\| U_{pq} \|_2 \\
&\le (1+7.1\cdot10^{-6})\cdot\sqrt{2}\cdot(1 + 2.1225\cdot10^{-5})\cdot\| x^{(1)}\|^2\cdot\| A \|_F \\
&\quad + 2\cdot32.2745\cdot\frac{\| A \|_F^5}{\delta^5}\cdot S(A)\cdot\| x^{(1)}\|
\end{aligned}$$

Mit $\| x^{(1)} \| \le \sum_{i=0}^{4}\left[2.05\cdot\| A \|_F\cdot\frac{1}{\delta}\right]^i\cdot\frac{1}{0.976\cdot\delta}\cdot\frac{\sqrt{2}}{2}\cdot S(A)$ folgt

$$\begin{aligned}
\| W'_{pp} \|_F &\le (1+2.833\cdot10^{-5})\cdot\left[\sum_{i=0}^{4}\left[2.05\cdot\| A \|_F\cdot\frac{1}{\delta}\right]^i\right]^2\cdot\frac{\sqrt{2}}{2\cdot0.976^2\cdot\delta^2}\cdot S^2(A)\cdot\| A \|_F \\
&\quad + \sqrt{2}\cdot32.2745\cdot\frac{\| A \|_F^5}{\delta^5}\cdot\sum_{i=0}^{4}\left[2.05\cdot\| A \|_F\cdot\frac{1}{\delta}\right]^i\cdot\frac{1}{0.976\cdot\delta}\cdot S^2(A)
\end{aligned}$$

$$\leq \left\{ (1+2.833\cdot10^{-5})\cdot \frac{\sqrt{2}}{2\cdot0.976^2} \cdot \left[\sum_{i=0}^{4} 2.05^i \cdot \left[\frac{\delta}{\| A \|_F}\right]^{4-i} \right]^2 \right.$$

$$\left. + \sqrt{2}\cdot 32.2745 \cdot \frac{1}{0.976} \cdot \sum_{i=0}^{4} 2.05^i \cdot \left[\frac{\delta}{\| A \|_F}\right]^{4-i} \right\} \cdot \frac{\| A \|_F^9}{\delta^{10}} \cdot S^2(A)$$

Wegen $\delta \leq \sqrt{2}\cdot\| A \|_F$ folgt insgesamt

$$\| W_{pp} \|_F \leq C \cdot \frac{\| A \|_F^9}{\delta^{10}} \cdot S^2(A) \text{ mit}$$

$$C = \frac{(1+2.833\cdot10^{-5})\cdot\sqrt{2}}{2\cdot0.976^2} \cdot \left[\sum_{i=0}^{4} 2.05^i \cdot \sqrt{2}^{4-i} \right]^2 + \frac{\sqrt{2}\cdot 32.2745}{0.976} \cdot \sum_{i=0}^{4} 2.05^i \cdot \sqrt{2}^{4-i}$$

$$\approx 3960.73 \leq 3961$$

Der Beweis für den Block A'_{qq} ist identisch. Für den Nichtdiagonalblock A''_{pq} erhalten wir in Blockschreibweise

$$\begin{aligned} A''_{pq} &= U_{pp}\cdot A'_{pp}\cdot U_{pq} - U_{pq}\cdot A'_{qp}\cdot U_{pq} + U_{pp}\cdot A'_{pq}\cdot U_{qq} - U_{pq}\cdot A'_{qq}\cdot U_{qq} \\ &= A'_{pq} + \left[\breve{U}_{pp}\cdot A'_{pq} + A'_{pq}\cdot \breve{U}_{qq} + \breve{U}_{pp}\cdot A'_{pq}\cdot \breve{U}_{qq} \right] \\ &\quad + U_{pp}\cdot A'_{pp}\cdot U_{pq} - U_{pq}\cdot A'_{qp}\cdot U_{pq} - U_{pq}\cdot A'_{qq}\cdot U_{qq} \end{aligned}$$

Somit erhalten wir als Darstellung der Veränderung von A'_{pq}

$$\begin{aligned} W'_{pq} &= \left[\breve{U}_{pp}\cdot A'_{pq} + A'_{pq}\cdot \breve{U}_{qq} + \breve{U}_{pp}\cdot A'_{pq}\cdot \breve{U}_{qq} \right] \\ &\quad + U_{pp}\cdot A'_{pp}\cdot U_{pq} - U_{pq}\cdot A'_{qp}\cdot U_{pq} - U_{pq}\cdot A'_{qq}\cdot U_{qq}, \end{aligned}$$

und es folgt

$$\begin{aligned} \| W'_{pq} \|_F \leq \; & \\ & 2\cdot\| \breve{U}_{pp} \|_2\cdot\| A'_{pq} \|_F + \| \breve{U}_{pp} \|_2^2\cdot\| A'_{pq} \|_F + \| A'_{pp} \|_F\cdot\| U_{pp} \|_2\cdot\| U_{pq} \|_2 \\ & + \| A'_{qp} \|_F\cdot\| U_{pq} \|_2^2 + \| A'_{qq} \|_F\cdot\| U_{pq} \|_2\cdot\| U_{qq} \|_2 \\ = \; & \| A'_{pq} \|_F\cdot\left[2\cdot\| \breve{U}_{pp} \|_2 + \| \breve{U}_{pp} \|_2^2\right] \\ & + \| A'_{pq} \|_F\cdot\| U_{pq} \|_2^2 + \left[\| A'_{pp} \|_F + \| A'_{qq} \|_F\right]\cdot\| U_{pq} \|_2\cdot\| U_{pp} \|_2 \\ \leq \; & \| U_{pq} \|_2\cdot\| U_{pp} \|_2\cdot\left[\| A'_{pp} \|_F + \| A'_{qq} \|_F\right] \\ & + 2\cdot(1+7.1\cdot10^{-6})\cdot\| A'_{pq} \|_F\cdot\| U_{pq} \|_2^2 \end{aligned}$$

Wie oben ergibt sich

$$
\begin{aligned}
\| W_{pq} \|_F \le & \| \boldsymbol{x}^{(1)} \| \cdot \sqrt{2} \cdot (1 + 2.833 \cdot 10^{-5}) \cdot \| A \|_F \\
& + 2 \cdot (1 + 7.1 \cdot 10^{-6}) \cdot 32.2745 \cdot \frac{\| A \|_F^5}{\delta^5} \cdot S(A) \cdot \| \boldsymbol{x}^{(1)} \|^2 \\
& \le (1 + 2.833 \cdot 10^{-5}) \cdot \frac{1}{0.976 \cdot \delta} \sum_{i=0}^{4} \left[2.05 \cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \cdot S(A) \cdot \| A \|_F \\
& + 2 \cdot (1 + 7.1 \cdot 10^{-6}) \cdot 32.2745 \cdot 2.176^2 \cdot 10^{-6} \cdot \frac{\| A \|_F^5}{\delta^5} \cdot S(A) \\
& \le \left\{ 1.02462 \cdot \sum_{i=0}^{4} 2.05^i \cdot \left[\frac{\delta}{\| A \|_F} \right]^{4-i} + 3.057 \cdot 10^{-4} \right\} \cdot \frac{\| A \|_F^5}{\delta^5} \cdot S(A) \\
& \le 49.23097 \cdot \frac{\| A \|_F^5}{\delta^5} \cdot S(A)
\end{aligned}
$$

●

3.11 Gesamtänderung der Außerdiagonale durch beide Transformationen

In diesem Abschnitt wollen wir die Änderung der Außerdiagonalnorm nach Ausführung beider Transformationen zu einem festen Pivotpaar (p,q) schätzen.

Satz 3.20

Es gelte die Generalvoraussetzung (3.28). Die Parameter beider Transformationen seien durch das Newton-Verfahren (3.26) für das Pivotpaar $(p,q) \in P_m$ bestimmt. Dann folgt für den Normzuwachs der Außerdiagonalnorm nach beiden Transformationen:

$$S(A'') \leq 118.813 \cdot \frac{\| A \|_F^5}{\delta^5} \cdot S(A) \tag{3.63}$$

Beweis:

Wegen der Orthogonalität von U und der J-Symmetrie von A gilt nach (3.12)

$$S^2(A'') = S^2(A') + 2 \cdot \left[\| A''_{pq} \|_F^2 - \| A'_{pq} \|_F^2 \right],$$

Mit (3.61) erhalten wir

$$\| A''_{pq} \|_F^2 - \| A'_{pq} \|_F^2 \leq 2 \cdot \| A'_{pq} \|_F \cdot \| W'_{pq} \|_F + \| W'_{pq} \|_F^2$$

Wir wissen bereits

$$\| W'_{pq} \|_F \leq 49.23097 \cdot \frac{\| A \|_F^5}{\delta^5} \cdot S(A)$$

$$\| A'_{pq} \|_F \leq 32.2745 \cdot \frac{\| A \|_F^5}{\delta^5} \cdot S(A) \qquad \text{und damit}$$

$$\| A''_{pq} \|_F^2 - \| A'_{pq} \|_F^2 \leq \left[2 \cdot 32.2745 \cdot 49.23097 + 49.23097^2 \right] \cdot \frac{\| A \|_F^{10}}{\delta^{10}} \cdot S^2(A)$$

$$\leq 5601.4981 \cdot \frac{\| A \|_F^{10}}{\delta^{10}} \cdot S^2(A)$$

Mit Formel (3.56) folgt insgesamt

$$S^2(A'') \le \left[53.967^2 + 2 \cdot 5601.4981 \right] \cdot \frac{\| A \|_F^{10}}{\delta^{10}} \cdot S^2(A)$$

$$= 14116.384 \cdot \frac{\| A \|_F^{10}}{\delta^{10}} \cdot S^2(A) \text{ oder}$$

$$S(A'') \le 118.813 \cdot \frac{\| A \|_F^{5}}{\delta^{5}} \cdot S(A)$$

3.12 Trennung der Eigenwerte der Blockdiagonale nach beiden Transformationen

Unter der Generalvoraussetzung, die eine genügend kleine Außerdiagonale der Matrix A fordert, wissen wir, daß die Eigenwerte $\mu_i^{(p)}$ und $\mu_i^{(q)}$ der Diagonalblöcke A_{pp} und A_{qq} getrennt liegen. Laut (3.29) haben sie einen Mindestabstand von

$$\delta(A) := \min \left\{ \; | \mu_k^{(j)} - \mu_l^{(j)} | \; | \; k,l \in \{1,2\} \text{ und } 1 \leq i < j \leq m \right\} \geq 0.976 \cdot \delta$$

Wir zeigen in diesem Abschnitt, daß unter der Generalvoraussetzung die Getrenntheit dieser Eigenwerte auch nach Ausführung der beiden Ähnlichkeitstransformationen zu einem einzelnen Pivot erhalten bleibt.

Satz 3.21

Es gelte die Generalvoraussetzung (3.28). Die Matrix A werde für ein Pivotpaar $(p,q) \in P_m$ sowohl der hyperbolischen als auch der orthogonalen Ähnlichkeitstransformation unterworfen. Dann sind auch nach beiden Transformationen die Eigenwerte der Diagonalblöcke getrennt, genauer gilt:

$$\delta(A'') := \min \left\{ \; | \mu_k^{(j)} - \mu_l^{(j)} | \; | \; k,l \in \{1,2\} \text{ und } 1 \leq i < j \leq m \right\} \geq 0.8937 \cdot \delta \qquad (3.64)$$

Beweis:

Unter der Generalvoraussetzung wissen wir, daß für die nicht transformierte Matrix A gilt

$$\delta(A) := \min \left\{ \; | \mu_k^{(j)} - \mu_l^{(j)} | \; | \; k,l \in \{1,2\} \text{ und } 1 \leq i < j \leq m \right\} \geq 0.976 \cdot \delta$$

Für die Außerdiagonalnorm erhielten wir nach Satz (3.20)

$$S(A'') \leq 118.813 \cdot \frac{\| A \|_F^5}{\delta^5} \cdot S(A)$$

und

$$S(A) \leq \frac{1}{6.4 \cdot 10^4 \cdot \sqrt{m-1}} \cdot \frac{\delta^9}{\| A \|_F^8}.$$

Schließlich gilt für das Henrici-Maß der Diagonalblöcke A''_{jj}

$$\Delta_F(A''_{jj}) \leq \| A''_{jj} \|_F \leq \| A_{jj} \|_F + \| A'_{jj} - A_{jj} \|_F + \| A''_{jj} - A'_{jj} \|_F$$
$$\leq \| A_{jj} \|_F + \| W_{jj} \|_F + \| W'_{jj} \|_F$$

Für W_{jj} und W'_{jj} kennen wir nach (3.53) und (3.61) die Abschätzungen

$$\| W_{ii} \|_F \leq 1992.2 \cdot \frac{\| A \|_F^9}{\delta^{10}} \cdot S^2(A) \qquad \text{für } i = p,q$$

$$\| W'_{ii} \|_F \leq 3960.73 \cdot \frac{\| A \|_F^9}{\delta^{10}} \cdot S^2(A) \qquad \text{für } i = p,q$$

Es folgt

$$\Delta_F(A''_{jj}) \leq \| A_{jj} \|_F + 5952.93 \cdot \frac{\| A \|_F^9}{\delta^{10}} \cdot S^2(A)$$

Laut Theorem (3.1) liegen die exakten Eigenwerte λ_i der Matrix A in der Vereinigung aller Mengen

$$H_j = \bigcup_{i=1}^{2} \left\{ z \in \mathbb{C} \mid | \mu_i^{(j)} - z | \leq \left[1 + (\sqrt{y_j + \tfrac{1}{4}} - \tfrac{1}{2}) \right] \cdot \sum_{\substack{k=1 \\ k \neq j}}^{m} \| A_{jk} \|_F \right\}$$

für j = 1,...,m mit (3.18)

$$y_j = \frac{\Delta_F(A_{jj})}{\sum\limits_{\substack{k=1 \\ k \neq j}}^{m} \| A_{jk} \|_F}.$$

Die exakten Eigenwerte haben den Mindestabstand $\delta > 0$. Wir setzen

$$R_j := \left[1 + \left(\sqrt{y_j + \tfrac{1}{4}} - \tfrac{1}{2}\right)\right] \cdot \sum_{\substack{k=1\\k\neq j}}^{m} \| A''_{jk} \|_F .$$

Mit den beiden folgenden Abschätzungen

$$\sum_{\substack{k=1\\k\neq j}}^{m} \| A''_{jk} \|_F \le \frac{\sqrt{m-1}}{\sqrt{2}} \cdot S(A'') \le \frac{118.813}{6.4\cdot 10^4 \cdot \sqrt{2}} \cdot \frac{\delta^4}{\| A \|_F^3} \le \frac{118.813}{3.2\cdot 10^4 \cdot \sqrt{2}} \cdot \frac{\delta^2}{\| A \|_F}$$

$$\le \frac{118.813}{3.2\cdot 10^4} \cdot \delta \approx 3.716\cdot 10^{-3} \cdot \delta$$

und

$$\Delta_F(A''_{jj}) \cdot \sum_{\substack{k=1\\k\neq j}}^{m} \| A''_{jk} \|_F \le$$

$$\le \left[\| A \|_F + 5952.93 \cdot \frac{\| A \|_F^9}{\delta^{10}} \cdot S^2(A) \right] \cdot \frac{118.813}{3.2\cdot 10^4 \cdot \sqrt{2}} \cdot \frac{\delta^2}{\| A \|_F}$$

$$\le \left[\| A \|_F + 5952.93 \cdot \frac{1}{4.096\cdot 10^9} \cdot \frac{\delta^8}{\| A \|_F^7} \right] \cdot \frac{118.813}{3.2\cdot 10^4 \cdot \sqrt{2}} \cdot \frac{\delta^2}{\| A \|_F}$$

$$\le \left[1 + 5952.93 \cdot \frac{16}{4.096\cdot 10^9} \right] \cdot \frac{118.813}{3.2\cdot 10^4 \cdot \sqrt{2}} \cdot \delta^2$$

$$\le 2.627\cdot 10^{-3} \cdot \delta^2$$

erhalten wir für den maximalen Radius $R := \max_{j=1}^{m} R_j$ der Kreise H_j:

$$R \le 1.8576\cdot 10^{-3} \cdot \delta + \sqrt{3.4504\cdot 10^{-6} \cdot \delta^2 + 2.627\cdot 10^{-3} \cdot \delta^2}$$

$$\le 5.3145\cdot 10^{-2} \cdot \delta.$$

Mit den gleichen Argumenten wie in Hilfssatz (3.3) folgt

$$\begin{aligned} | \mu_k^{(i)} - \mu_l^{(j)} | &\geq |\lambda_i - \lambda_j| - | \mu_k^{(i)} - \lambda_i| - | \mu_l^{(j)} - \lambda_j | \\ &\geq \delta - 2 \cdot 5.3145 \cdot 10^{-2} \cdot \delta = 0.8937 \cdot \delta. \end{aligned}$$

Hierbei sind λ_i bzw. λ_j die zu den beiden Eigenwerten $\mu_k^{(i)}$ und $\mu_l^{(j)}$ gehörenden exakten Eigenwerte der Matrix A.

●

Damit bleiben die Eigenwerte der Diagonalblöcke auch nach einer vollen Transformation zu einem einzelnen Pivotpaar (p,q) getrennt. Mit Hilfe dieser Eigenschaft läßt sich nun zeigen, daß bei genügend kleiner Außerdiagonale die Eigenwerte der Diagonalblöcke gegen einen festen Eigenwert der Startmatrix A konvergieren müssen und nicht im Laufe der Iteration ihre Zugehörigkeit zu einem bestimmten Eigenwert der Matrix A wechseln.

3.13 Zugehörigkeit der Näherungen zu festen Eigenwerten der Matrix

Der vorangegangene Abschnitt hat gezeigt, daß die Gesamttransformation auf einem Pivotpaar (p,q) ϵ P_m die Getrenntheit der Eigenwerte $\mu_k^{(i)}$ der Diagonalblöcke A_{ii} für k = 1,2 und i = 1,...,m nicht beeinflußt. Die Eigenwerte der Diagonalblöcke stellen im Sinne von Theorem (3.1) Näherungen für die exakten Eigenwerte der Matrix A dar. Dieser Abschnitt soll zeigen, daß bei genügend fortgeschrittener Annullierung der Außerdiagonale die Näherungen $\mu_k^{(i)}$ ihre Zugehörigkeit zu den exakten Eigenwerten nicht mehr ändern können. Diese Aussage verhindert, daß in der Schlußphase des Algorithmus die Außerdiagonale bereits annulliert ist und lediglich Veränderungen der Diagonalblöcke stattfinden, die zum Tausch der Eigenwertzuordnungen führen. Wir benötigen zunächst eine Aussage, wie sich die Eigenwerte einer Matrix bei Störung der Matrix verändern. Das zitierte Resultat stammt von Wilkinson - Ostrowski [28].

Satz 3.22

Seien $A,B \in \mathbb{R}^{n,n}$ zwei Matrizen mit $\| A \|_F , \| B \|_F < 1$. Sei weiter $\epsilon > 0$ und $A' = A + \epsilon \cdot B$. Dann gibt es zu jedem Eigenwert λ von A einen Eigenwert λ' der Matrix A' mit

$$|\lambda - \lambda_i| < (n+2) \cdot \left[n^2 \cdot \epsilon\right]^{\frac{1}{n}} \tag{3.65}$$

Das Ergebnis ist nur für kleine Dimensionen sinnvoll anwendbar. In größeren Dimensionen geben simple Normabschätzungen der Matrizen bessere Schranken. In unserem Fall ist dies ohne Bedeutung, da wir nur 2x2 - Blöcke betrachten.

Wir wollen die Änderung der Eigenwerte der Diagonalblöcke nach beiden Ähnlichkeitstransformationen abschätzen.

Satz 3.23

Es gelte die Generalvoraussetzung (3.28). Die Matrix A werde für ein Pivotpaar $(p,q) \in P_m$ sowohl der hyperbolischen als auch der orthogonalen Ähnlichkeitstransformation unterworfen. Dann approximieren die Eigenwerte $\mu_k^{(i)}$ der Diagonalblöcke A_{pp} bzw. A_{qq} und die Eigenwerte $\hat{\mu}_k^{(i)}$ der transformierten Blöcke A''_{pp} bzw. A''_{qq} für $k = 1,2$ und $i = p,q$ dieselben vier Eigenwerte der Matrix A.

Beweis:

Wir beschreiben zunächst die Ausgangssituation. Laut Theorem (3.1) liegen die exakten Eigenwerte der Matrix A in der Vereinigung aller Kreise

$$H_i^{(j)} = \left\{ z \in \mathbb{C} \;\middle|\; |\mu_i^{(j)} - z| \le \left[1 + \left(\sqrt{y_j + \tfrac{1}{4}} - \tfrac{1}{2}\right)\right] \cdot \sum_{\substack{k=1 \\ k \ne j}}^{m} \| A_{jk} \|_F \right\}$$

für $j = 1,\dots,m$ und $i = 1,2$.

Da aufgrund der Generalvoraussetzung alle Radien der Kreise $H_i^{(j)}$ kleiner als der Mindestabstand δ der exakten Eigenwerte sind, kann in jedem Kreis $H_i^{(p)}$ bzw. $H_i^{(q)}$ für $i = 1,2$ nur ein exakter Eigenwert der Matrix A liegen. Wir bezeichnen diesen exakten Eigenwert mit $\lambda_i^{(p)}$ bzw. $\lambda_i^{(q)}$. Die gleiche Situation ist auch nach Durchführung der beiden Ähnlichkeitstransformationen auf dem Pivotpaar (p,q) gegeben. Wir betrachten jetzt o.B.d.A. den Block A_{pp}. Nach beiden Transformationen gilt mit (3.60) und (3.53)

$$\begin{aligned} A''_{pp} &= A_{pp} + W_{pp} + W'_{pp} \\ &= (1+\alpha)\cdot \| A \|_F \cdot \left[\frac{A_{pp}}{(1+\alpha)\| A \|_F} + \frac{\| W_{pp} + W'_{pp} \|_F}{\| A \|_F} \cdot \frac{W_{pp} + W'_{pp}}{(1+\alpha)\| W_{pp} + W'_{pp} \|_F} \right] \end{aligned}$$

für $\alpha > 0$ beliebig. Die Eigenwerte von A''_{pp} und der geklammerten Matrix unterscheiden sich nur um den Faktor $(1+\alpha)\cdot\| A \|_F$. Auf die Matrix in der Klammer können wir den Satz von Wilkinson - Ostrowski anwenden. Wir setzen

$$A := \frac{A_{pp}}{(1+\alpha)\| A \|_F}, \quad B := \frac{W_{pp} + W'_{pp}}{(1+\alpha)\| W_{pp} + W'_{pp} \|_F} \quad \text{und } \epsilon := \frac{\| W_{pp} + W'_{pp} \|_F}{\| A \|_F}.$$

Dann gilt für den maximalen Abstand zweier zueinander gehörender Eigenwerte der Matrizen A und $A' = (A + \epsilon B)$ wegen dim $A = 2$ die Abschätzung

$$| \lambda - \lambda' | < 8 \cdot \sqrt{\epsilon}$$

mit

$$\epsilon = \frac{\| W_{pp} + W'_{pp} \|_F}{\| A \|_F} \leq \frac{1}{\| A \|_F} \cdot \left[\| W_{pp} \| + \| W'_{pp} \|_F \right]$$

Wegen

$$\| W_{pp} \|_F \leq 1992.2 \cdot \frac{\| A \|_F^9}{\delta^{10}} \cdot S^2(A) \quad \text{und}$$

$$\| W'_{pp} \|_F \leq 3960.73 \cdot \frac{\| A \|_F^9}{\delta^{10}} \cdot S^2(A) \quad \text{folgt}$$

$$\epsilon \leq \frac{1}{\| A \|_F} \cdot 5952.93 \cdot \frac{\| A \|_F^9}{\delta^{10}} \cdot S^2(A)$$

Es ergibt sich

$$| \lambda - \lambda' | < 617.2419 \cdot \frac{\| A \|_F^4}{\delta^5} \cdot S(A)$$

Unter der Generalvoraussetzung (3.28) folgt schließlich

$$| \lambda - \lambda' | < 617.2419 \cdot \frac{1}{6.4 \cdot 10^4 \cdot \sqrt{m-1}} \cdot \frac{\delta^4}{\| A \|_F^4}$$

Dies ist der Abstand der Eigenwerte der normierten Matrix. Die Eigenwerte von A''_{pp} und A_{pp} unterscheiden sich gerade um den Faktor $(1+\alpha) \cdot \| A \|_F$ von den oben genannten, d.h. es gilt

$$| \hat{\mu}_i^{(p)} - \mu_i^{(p)} | < 617.2419 \cdot (1+\alpha) \cdot \frac{1}{6.4 \cdot 10^4 \cdot \sqrt{m-1}} \cdot \frac{\delta^4}{\| A \|_F^3}$$

$$\leq 617.2419 \cdot (1+\alpha) \cdot \frac{2 \cdot \sqrt{2}}{6.4 \cdot 10^4} \cdot \delta$$

$$\leq (1+\alpha) \cdot 0.0273 \cdot \delta \quad \text{für } i = 1,2 \text{ und } \alpha > 0$$

Wegen $\alpha > 0$ beliebig folgt somit

$$| \hat{\mu}_i^{(p)} - \mu_i^{(p)} | \leq 0.0273 \cdot \delta \quad \text{für } i = 1,2$$

Damit ist die Veränderung der Eigenwerte der Diagonalblöcke durch die beiden Ähnlichkeitstransformationen sehr gering. Wir erhalten für den Abstand von $\hat{\mu}_i^{(p)}$ und dem zu $\mu_i^{(p)}$ gehörenden exakten Eigenwert $\lambda_i^{(p)}$

$$\begin{aligned} | \hat{\mu}_i^{(p)} - \lambda_i^{(p)} | &\leq | \hat{\mu}_i^{(p)} - \mu_i^{(p)} | + | \mu_i^{(p)} - \lambda_i^{(p)} | \\ &\leq 0.0273 \cdot \delta \qquad + \quad 0.012 \cdot \delta \\ &\leq 0.0393 \cdot \delta \end{aligned}$$

Dieselbe Abschätzung gilt für den Block A_{qq}. Alle anderen Diagonalblöcke bleiben bei einer Transformation für das Pivotpaar (p,q) unberührt. Damit kann der Eigenwert $\mu_i^{(p)}$ durch die beiden Transformationen nicht in einen Kreis zu einem anderen exakten Eigenwert λ als $\lambda_i^{(p)}$ wandern, denn es gilt

$$\begin{aligned} | \hat{\mu}_i^{(p)} - \lambda | \quad &\geq | \lambda_i^{(p)} - \lambda | - 2 \cdot | \hat{\mu}_i^{(p)} - \lambda_i^{(p)} | \\ &\geq \delta \qquad\qquad - 2 \cdot 0.0393 \cdot \delta = 0.9607 \cdot \delta, \end{aligned}$$

d.h. der Abstand zu jedem anderen exakten Eigenwert ist größer als der Kreis um den exakten Eigenwert, in dem die Näherung nach Satz (3.21) liegt.

●

3.14 Wachstum der Pivotzeilen bzw. gekoppelter Matrixblöcke

Wir benötigen abschließend noch eine Aussage, in welcher Weise sich die Zeilen bzw. Spalten verändern, die neben den Pivotblöcken bei einer Transformation zum Pivotpaar $(p,q) \in P_m$ ebenfalls betroffen sind.

Satz 3.24

Es gelte die Generalvoraussetzung (3.28). Die Parameter x_1 und x_2 seien durch das Newton-Verfahren (3.26) für das Pivotpaar $(p,q) \in P_m$ bestimmt. Dann folgen für die Pivotzeilen der Matrix A'' nach beiden Ähnlichkeitstransformationen

$$\| A''_{ip} \|_F^2 + \| A''_{iq} \|_F^2 \le (2 + 1.9\cdot 10^{-5})\cdot \left[\| A_{ip} \|_F^2 + \| A_{iq} \|_F^2 \right] \tag{3.66a}$$

$$\| A''_{pi} \|_F \le (1+7.1025\cdot 10^{-6})\cdot \| A_{pi} \|_F + (2+4.735\cdot 10^{-6})\cdot \| x^{(1)} \| \cdot \| A_{qi} \|_F \tag{3.66b}$$

für i = 1,...,m und mit

$$\| x^{(1)} \| \le \sum_{i=0}^{4} \left[2.05\cdot \| A \|_F \cdot \frac{1}{\delta} \right]^i \cdot \frac{1}{0.976\cdot \delta} \cdot \frac{\sqrt{2}}{2} \cdot S(A).$$

Beweis:

Wie wir bereits bei der Betrachtung der orthogonalen Transformation festgestellt haben, läßt die Transformation U die Norm gekoppelter Matrixblöcke nach Formel (3.11) unverändert, d.h.

$$\| A''_{ip} \|_F^2 + \| A''_{iq} \|_F^2 = \| A'_{ip} \|_F^2 + \| A'_{iq} \|_F^2 \qquad \text{für } i = 1,\dots,m$$

Weiterhin haben wir die Beziehung (3.57) bereits gezeigt. Es gilt:

$$\| A'_{ip} \|_F^2 \le 9.5\cdot 10^{-6}\cdot \| A_{iq} \|_F^2 + (2 + 9.5\cdot 10^{-6})\cdot \| A_{ip} \|_F^2$$

$$\| A'_{iq} \|_F^2 \le 9.5\cdot 10^{-6}\cdot \| A_{ip} \|_F^2 + (2 + 9.5\cdot 10^{-6})\cdot \| A_{iq} \|_F^2$$

Insgesamt folgt so

$$\| A''_{ip} \|_F^2 + \| A''_{iq} \|_F^2 \le \left[9.5\cdot 10^{-6} + (2 + 9.5\cdot 10^{-6}) \right] \cdot \left[\| A_{iq} \|_F^2 + \| A_{ip} \|_F^2 \right]$$

Für den Block A_{pi} gelten folgende Transformationsformeln:

$$A'_{pi} = T_{pp} \cdot A_{pi} + T_{pq} \cdot A_{qi}$$
$$A'_{qi} = T_{qp} \cdot A_{pi} + T_{qq} \cdot A_{qi}$$
$$A''_{pi} = U_{pp} \cdot A'_{pi} + U_{pq} \cdot A'_{qi}$$

Somit erhalten wir

$$\begin{aligned} \| A''_{pi} \|_F \leq & \left[\| U_{pp} \|_2 \cdot \| T_{pp} \|_2 + \| U_{pq} \|_2 \cdot \| T_{qp} \|_2 \right] \cdot \| A_{pi} \|_F \\ + & \left[\| U_{pp} \|_2 \cdot \| T_{pq} \|_2 + \| U_{pq} \|_2 \cdot \| T_{qq} \|_2 \right] \cdot \| A_{qi} \|_F \end{aligned}$$

Mit Hilfe der Abschätzungen der Blöcke T_{ij} bzw. U_{ij} nach (3.50) bzw. (3.59) ergibt sich mit $\| x^{(1)} \|^2 \leq 4.735 \cdot 10^{-6}$

$$\begin{aligned} \| U_{pp} \|_2 \cdot \| T_{pp} \|_2 + \| U_{pq} \|_2 \cdot \| T_{qp} \|_2 & \leq (1 + 2.3675 \cdot 10^{-6}) \cdot (1 + \| x^{(1)} \|^2) \\ & \leq (1 + 2.3675 \cdot 10^{-6}) \cdot (1 + 4.735 \cdot 10^{-6}) \end{aligned}$$

und

$$\| U_{pp} \|_2 \cdot \| T_{pq} \|_2 + \| U_{pq} \|_2 \cdot \| T_{qq} \|_2 \leq 2 \cdot (1 + 2.3675 \cdot 10^{-6}) \cdot \| x^{(1)} \|$$

Es folgt insgesamt

$$\begin{aligned} \| A''_{pi} \|_F \leq & \ (1 + 2.3675 \cdot 10^{-6}) \cdot (1 + 4.735 \cdot 10^{-6}) \cdot \| A_{pi} \|_F \\ & + 2 \cdot (1 + 2.3675 \cdot 10^{-6}) \cdot \| x^{(1)} \| \cdot \| A_{qi} \|_F \\ = & \ (1 + 7.1025 \cdot 10^{-6}) \cdot \| A_{pi} \|_F + (2 + 4.735 \cdot 10^{-6}) \cdot \| x^{(1)} \| \cdot \| A_{qi} \|_F \end{aligned}$$

●

3.15 Einhaltung der Generalvoraussetzung über den gesamten Zyklus

Die bisherigen Untersuchungen bezogen sich auf einzelne Transformationen an einem Pivotblock A_{pq} und deren Auswirkungen auf die Normen der Blöcke, der Außerdiagonale und der Norm der Gesamtmatrix A. Wir haben jetzt die Instrumente zur Verfügung gestellt, um die Wirkungen aller Transformationen eines vollen Zyklus aufzuaddieren. Wir wollen dabei alle bisher gezeigten Abschätzungen über den gesamten Zyklus verwenden. Unsere Abschätzungen zeigen, daß die Außerdiagonalnorm $S(A)$ in bestimmter Größenordnung wachsen kann genauso wie die Norm der Matrix A. Wir müssen daher unsere Generalvoraussetzung in dem Sinne nochmals verschärfen, daß sie in der ursprünglichen Form selbst auf dem letzten Pivotpaar eines Zyklus trotz vorangegangener Transformationen gültig bleibt. Die Schranke, die für die Außerdiagonale notwendig ist, wird in diesem Abschnitt konstruiert. In dem nachfolgenden Induktionsbeweis dürfen dann anschließend alle Abschätzungen unabhängig von dem gerade betrachteten Pivotpaar b-enutzt werden. Es gilt

Satz 3.25

Sei $\mu \in \mathbb{N}_0$. Wählen wir für die Startmatrix $A = A^{\mu}$ des $(\mu+1)$-ten Zyklus die Außerdiagonalnorm $S(A^{\mu})$ so klein, daß

$$S(A^{\mu}) \leq \epsilon \cdot \frac{1}{6.4 \cdot 10^4 \cdot \sqrt{m-1}} \cdot \frac{\delta^9}{\| A^0 \|_F^8} \tag{3.67a}$$

mit

$$\epsilon = 118.813^{-M} \cdot \left[1 + 2.1225 \cdot 10^{-5}\right]^{-R_M} \cdot \left[\frac{\| A^0 \|_F^5}{\delta^5}\right]^{-M} \text{ und } R_M := \frac{5 \cdot M \cdot (M-1)}{2}, \tag{3.67b}$$

so ist die Generalvoraussetzung (3.28) unabhängig vom betrachteten Pivotpaar $(p,q) \in P_m$ für alle Matrizen $A^{k,\mu}$ des $(\mu+1)$-ten Zyklus für $k = 1,\dots,M$ erfüllt, d.h. es gilt

$$S(A^{k,\mu}) \leq \frac{1}{6.4 \cdot 10^4 \cdot \sqrt{m-1}} \cdot \frac{\delta^9}{\| A^{k,\mu} \|_F^8} \qquad \text{für } k = 1,\dots,M.$$

Beweis:

Im Beweis ersetzen wir den oberen Index (k,μ) des Zyklus unter Vernachlässigung der Nummer μ des Zyklus durch die Nummer k der einzelnen Schritte des Zyklus. Unter der Annahme der Generalvoraussetzung (3.28) haben wir sowohl das Außerdiagonalnormwachstum als auch das Wachstum der Norm der Matrix A abgeschätzt. Es gilt nach (3.63) bzw. (3.58)

$$S(A^{k+1}) \leq 118.813 \cdot \frac{\| A^k \|_F^5}{\delta^5} \cdot S(A^k) \quad \text{für } k = 1,\ldots,M$$

$$\| A^{k+1} \|_F \leq (1 + 2.1225 \cdot 10^{-5}) \cdot \| A^k \|_F \quad \text{für } k = 1,\ldots,M$$

Letztere Abschätzung gilt, weil die Transformation U die Norm der Matrix A nicht verändert. Damit erhalten wir, falls für alle Schritte des Zyklus die Generalvoraussetzung gilt, folgenden maximalen Außerdiagonalzuwachs:

$$\begin{aligned}
S(A^M) &\leq 118.813 \cdot \frac{\| A^{M-1} \|_F^5}{\delta^5} \cdot S(A^{M-1}) \\
&\leq 118.813 \cdot \frac{\| A^{M-1} \|_F^5}{\delta^5} \cdot 118.813 \cdot \frac{\| A^{M-2} \|_F^5}{\delta^5} \cdot S(A^{M-2}) \\
&\leq \left[\frac{118.813}{\delta^5}\right]^M \cdot \left[\prod_{i=0}^{M-1} \| A^i \|_F^5\right] \cdot S(A^0) \\
&\leq \left[\frac{118.813}{\delta^5}\right]^M \cdot \left[\prod_{i=0}^{M-1} (1 + 2.1225 \cdot 10^{-5})^{5i} \cdot \| A^0 \|_F^5\right] \cdot S(A^0) \\
&\leq \left[\frac{118.813}{\delta^5}\right]^M \cdot \left[\prod_{i=0}^{M-1} (1 + 2.1225 \cdot 10^{-5})^{5i}\right] \cdot \| A^0 \|_F^{5M} \cdot S(A^0)
\end{aligned}$$

Mit $\sum_{i=0}^{M-1} 5i = 5 \cdot \sum_{i=0}^{M-1} i = \frac{5 \cdot M \cdot (M-1)}{2} = R_M$ folgt

$$S(A^M) \leq \left[\frac{118.813}{\delta^5} \cdot \| A^0 \|_F^5 \right]^M \cdot \left[1 + 2.1225 \cdot 10^{-5} \right]^{R_M} \cdot S(A^0)$$

Damit ist die maximale Außerdiagonalnorm nach Durchführung des gesamten Zyklus abgeschätzt durch die Norm bzw. die Außerdiagonalnorm der Startmatrix A^0 des Zyklus. Es gilt daher

$$S(A^j) \leq \frac{1}{6.4 \cdot 10^4 \cdot \sqrt{m-1}} \cdot \frac{\delta^9}{\| A_i^j \|_F^8} \leq \frac{1}{6.4 \cdot 10^4 \cdot \sqrt{m-1}} \cdot \frac{\delta^9}{\| A^0 \|_F^8}$$

für j = 1,...,M, falls

$$S(A^0) \leq \left[\frac{118.813}{\delta^5} \cdot \| A^0 \|_F^5 \right]^{-M} \cdot \left[1 + 2.1225 \cdot 10^{-5} \right]^{-R_M} \cdot \frac{1}{6.4 \cdot 10^4 \cdot \sqrt{m-1}} \cdot \frac{\delta^9}{\| A^0 \|_F^8}$$

•

3.16 Induktionsbeweis nach Ruhe für quadratische Abnahme der Außerdiagonale

Wir kommen nun zum eigentlichen Beweis der asymptotisch quadratischen Konvergenz. Wir setzen voraus, daß die Außerdiagonalnorm nach (3.63) folgende Größenordnung hat:

$$S(A^{\mu-1}) = \| A^{\mu-1} \| \cdot \tilde{\epsilon} \quad \text{mit}$$

$$\tilde{\epsilon} = \left[\frac{118.813}{\delta^5} \cdot \| A^0 \|_F^5\right]^{-M} \cdot \left[1 + 2.1225 \cdot 10^{-5}\right]^{-R_M} \cdot \epsilon \quad \text{und} \tag{3.68}$$

$$\epsilon = \frac{1}{6.4 \cdot 10^4 \cdot \sqrt{m-1}} \cdot \frac{\delta^9}{\| A^0 \|_F^8}$$

Die Behauptung ist, daß nach allen M Transformationen des μ-ten Zyklus die Außerdiagonalnorm von der Größenordnung $O(\tilde{\epsilon}^2)$ ist, d.h. daß gilt:

$$S(A^{M,\mu-1}) = S(A^{\mu}) = \text{Const} \cdot \tilde{\epsilon}^2 \quad \text{mit einer Konstanten Const} > 0.$$

Der Beweis der quadratischen Abnahme der Außerdiagonalnorm wird nach einer Idee von Ruhe [19] induktiv über die Anzahl der Blockzeilen der Matrix A geführt.

Der nachfolgende Satz besagt anschaulich, daß die Veränderungen der Normen der Blöcke A_{ij} durch alle Transformationen, die bis zum Ende einer jeden Zeile p für $p \in \{0,\dots, m-1\}$ durchgeführt wurden, sich so aufaddieren, daß alle Blöcke, die in bzw. oberhalb der p-ten Zeile in der oberen Dreieckshälfte stehen, quadratisch klein geworden sind.

Satz 3.26

Die Außerdiagonalnorm der Startmatrix $A^{\mu-1}$ genüge der Bedingung (3.68). Es sei weiter $k_p := \varphi(p,m)$ für $p \in \{0,\dots, m-1\}$ die Nummer der Transformation auf dem Pivotpaar (p,m). Dann existiert für alle $p \in \{0,\dots, m-1\}$ eine Konstante $C_p > 0$ mit

$$\sum_{i=1}^{p} \sum_{j=i+1}^{m} \| A_{ij}^{k_p,\mu-1} \|_F^2 \leq C_p \cdot \| A^0 \|_F^2 \cdot \tilde{\epsilon}^4 \tag{3.69}$$

Beweis:

Der Beweis erfolgt durch Induktion über p.

Induktionsanfang: (p = 0)

Für p=0 ist die Aussage leer, und wir wählen $C_0 = 1$.

Induktionsvoraussetzung:

Es existiere eine Konstante $C_{p-1} > 0$ mit

$$\sum_{i=1}^{p-1} \sum_{j=i+1}^{m} \| A_{ij}^{k_{p-1},\mu-1} \|_F^2 \leq C_{p-1} \cdot \| A^0 \|_F^2 \cdot \tilde{\epsilon}^4, \quad \text{wobei } k_{p-1} = k_p - (m-p) \text{ ist.}$$

Induktionsschritt:

Im Induktionsschritt haben wir die Auswirkungen der Transformationen der p-ten Zeile zu berücksichtigen, also die Folge

$$k_{p-1}+1 = \varphi(p,p+1), \varphi(p,p+2), \ldots, \varphi(p,p+m) = k_p$$

von Pivotpaaren mit den zugehörigen Transformationen. Wir betrachten zunächst die Veränderungen der ersten (p-1) Zeilen. Blöcke dieser Zeilen werden durch Transformationen der p-ten Pivotzeile entweder überhaupt nicht geändert oder gleich paarweise als gekoppelte Blöcke zweier Spalten. Die Veränderung gekoppelter Blöcke haben wir bereits abgeschätzt. Es gilt für das Pivotpaar $k_p = \varphi(p,m)$ nach (3.66a)

$$\| A_{ip}^{k_p,\mu-1} \|_F^2 + \| A_{im}^{k_p,\mu-1} \|_F^2 \leq (2 + 1.9 \cdot 10^{-5}) \cdot \left[\| A_{ip}^{k_p-1,\mu-1} \|_F^2 + \| A_{im}^{k_p-1,\mu-1} \|_F^2 \right]$$

für $i = 1,\ldots,p-1$.

$$\sum_{i=1}^{p-1} \sum_{j=i+1}^{m} \| A_{ij}^{k_p,\mu-1} \|_F^2$$

$$= \sum_{i=1}^{p-1} \sum_{\substack{j=i+1 \\ j \neq p}}^{m-1} \| A_{ij}^{k_p-1,\mu-1} \|_F^2 + \sum_{i=1}^{p-1} \left[\| A_{ip}^{k_p,\mu-1} \|_F^2 + \| A_{im}^{k_p,\mu-1} \|_F^2 \right]$$

$$\leq (2 + 1.9 \cdot 10^{-5}) \cdot \sum_{i=1}^{p-1} \sum_{j=i+1}^{m} \| A_{ij}^{k_p-1,\mu-1} \|_F^2$$

Betrachten wir anschließend das Pivotpaar $k_p - 1 = \varphi(p, m-1)$, so folgt analog

$$\sum_{i=1}^{p-1} \sum_{j=i+1}^{m} \| A_{ij}^{k_p-1,\mu-1} \|_F^2 \le (2 + 1.9 \cdot 10^{-5}) \cdot \sum_{i=1}^{p-1} \sum_{j=i+1}^{m} \| A_{ij}^{k_p-2,\mu-1} \|_F^2$$

und sukzessive für alle Pivotpaare der p-ten Zeile aufaddiert

$$\sum_{i=1}^{p-1} \sum_{j=i+1}^{m} \| A_{ij}^{k_p,\mu-1} \|_F^2 \le (2 + 1.9 \cdot 10^{-5})^{m-p} \cdot \sum_{i=1}^{p-1} \sum_{j=i+1}^{m} \| A_{ij}^{k_p-m+p,\mu-1} \|_F^2$$

Wegen $k_p - m + p = k_{p-1}$ kann man nun auf die rechte Doppelsumme die Induktionsvoraussetzung anwenden, und es folgt

$$\sum_{i=1}^{p-1} \sum_{j=i+1}^{m} \| A_{ij}^{k_p,\mu-1} \|_F^2 \le (2 + 1.9 \cdot 10^{-5})^{m-p} \cdot C_{p-1} \cdot \| \mathcal{A}^0 \|_F^2 \cdot \tilde{\epsilon}^4 .$$

Damit sind alle Blöcke der ersten (p-1) Zeilen der Matrix nach der Durchführung der Transformationen der p-ten Zeile quadratisch klein. Es bleibt das Verhalten der Blöcke der p-ten Zeile unter den Transformationen derselben Zeile abzuschätzen.

Wir schauen zunächst die Veränderungen an, die ein Block der p-ten Zeile nach dessen Verwendung als Pivotblock erfährt. Der Block A_{pi} wird in der $k_{p-1}+(i-p) = k_p - m + p + (i-p) = (k_p - m + i)$-ten Transformation des aktuellen Zyklus als Pivotblock verwendet. Anschließend wird er noch weitere (m-i)-mal mittransformiert, wobei jede einzelne Transformation nach Formel (3.66b) durch

$$\begin{aligned} \| A_{pi}^{k_p-m+i+j,\mu-1} \|_F \le & (1 + 7.1025 \cdot 10^{-6}) \cdot \| A_{pi}^{k_p-m+i+j-1,\mu-1} \|_F \\ & + (2 + 4.735 \cdot 10^{-6}) \cdot \| \mathcal{X}^{k_p-m+i+j-1,\mu-1} \| \cdot \| A_{qi}^{k_p-m+i+j-1,\mu-1} \|_F \end{aligned}$$

für j = 1,...,(m-i) und mit (3.39)

$$\begin{aligned} \| \mathcal{X}^{k_p-m+i+j-1,\mu-1} \| \le & \sum_{r=0}^{4} \left[2.05 \cdot \| \mathcal{A}^{k_p-m+i+j-1,\mu-1} \|_F \cdot \frac{1}{\delta} \right]^r \cdot \\ & \cdot \frac{1}{0.976 \cdot \delta} \cdot \frac{\sqrt{2}}{2} \cdot S(\mathcal{A}^{k_p-m+i+j-1,\mu-1}). \end{aligned}$$

abgeschätzt werden kann. Aufaddieren ergibt für die Veränderungen insgesamt.

$$\| A_{pi}^{k_p,\mu-1} \|_F \le (1+7.1025\cdot10^{-6})^{m-i} \cdot \| A_{pi}^{k_p-m+i,\mu-1} \|_F + \sum_{j=i+1}^{m} (1+7.1025\cdot10^{-6})^{m-j} \cdot (2+4.735\cdot10^{-6}) \cdot \| x^{k_p-m+i+j-1,\mu-1} \| \cdot \| A_{j\,i}^{k_p-m+i+j-1,\mu-1} \|_F$$

Die Norm der Parameter im $(k_p - m + i + j - 1)$-ten Schritt läßt sich unter Verwendung der Abschätzung (3.58) für den maximalen Normzuwachs wie folgt abschätzen:

$$\| x^{k_p-m+i+j-1,\mu-1} \| \le$$

$$\le \sum_{r=0}^{4} \left[2.05 \cdot \| A^{k_p-m+i+j-1,\mu-1} \|_F \cdot \frac{1}{\delta} \right]^r \cdot \frac{1}{0.976\cdot\delta} \cdot \frac{\sqrt{2}}{2} \cdot S(A^{k_p-m+i+j-1,\mu-1}).$$

$$\le \sum_{r=0}^{4} \left[2.05 \cdot (1 + 2.1225\cdot10^{-5})^{k_p-m+j-1} \cdot \| A^0 \|_F \cdot \frac{1}{\delta} \right]^r \cdot \frac{1}{0.976\cdot\delta} \cdot \frac{\sqrt{2}}{2} \cdot S(A^{k_p-m+i+j-1,\mu-1})$$

Es folgt

$$\| A_{pi}^{k_p,\mu-1} \|_F \le$$

$$\le (1+7.1025\cdot10^{-6})^{m-i} \cdot \| A_{pi}^{k_p-m+i,\mu-1} \|_F +$$

$$+ \sum_{j=i+1}^{m} \Bigg\{ (1+7.1025\cdot10^{-6})^{m-j} \cdot (2+4.735\cdot10^{-6}) \cdot$$

$$\cdot \sum_{r=0}^{4} \Bigg\{ \left[2.05 \cdot (1 + 2.1225\cdot10^{-5})^{k_p-m+j-1} \cdot \| A^0 \|_F \cdot \frac{1}{\delta} \right]^r \cdot$$

$$\cdot \frac{1}{0.976\cdot\delta} \cdot \frac{\sqrt{2}}{2} \cdot S(A^{k_p-m+i+j-1,\mu-1}) \Bigg\} \cdot \| A_{j\,i}^{k_p-m+i+j-1,\mu-1} \|_F \Bigg\}$$

Quadrieren der Ungleichung und Anwendung der Cauchy - Schwarzschen Ungleichung liefert

$$\| A_{pi}^{k_p,\mu-1} \|_F^2 \leq$$
$$\leq 2\cdot(1+7.1025\cdot 10^{-6})^{2(m-i)} \cdot \| A_{pi}^{k_p-m+i,\mu-1} \|_F^2 +$$
$$+ 2\cdot\Bigg\{ \sum_{j=i+1}^{m} \Bigg[(1+7.1025\cdot 10^{-6})^{m-j} \cdot (2+4.735\cdot 10^{-6}) \cdot$$
$$\cdot \sum_{r=0}^{4} \Bigg\{ \left[2.05\cdot(1 + 2.1225\cdot 10^{-5})^{k_p-m+j-1} \cdot \| A^0 \|_F \cdot \frac{1}{\delta} \right]^r \cdot$$
$$\cdot \frac{1}{0.976\cdot\delta} \cdot \frac{\sqrt{2}}{2} \cdot S(A^{k_p-m+j-1,\mu-1}) \Bigg\} \cdot \| A_{ji}^{k_p-m+j-1,\mu-1} \|_F \Bigg] \Bigg\}^2 .$$

Setzen wir den in der zweiten Summe auftretenden Faktor

$$\alpha_j := (2+4.735\cdot 10^{-6}) \cdot \sum_{r=0}^{4} \left[2.05\cdot(1+2.1225\cdot 10^{-5})^{k_p-m+j-1} \cdot \| A^0 \|_F \cdot \frac{1}{\delta} \right]^r \cdot \frac{1}{0.976\cdot\delta} \cdot \frac{\sqrt{2}}{2} ,$$

so folgt mit $\max_{j=i+1}^{m} \alpha_j = \alpha_m := \tilde{\alpha}$

$$\| A_{pi}^{k_p,\mu-1} \|_F^2 \leq (1+7.1025\cdot 10^{-6})^{2m} \cdot$$
$$\left\{ 2\cdot \| A_{pi}^{k_p-m+i,\mu-1} \|_F^2 + 4\cdot\tilde{\alpha}^2 \left[\cdot \sum_{j=i+1}^{m} \left[S(A^{k_p-m+j-1,\mu-1}) \cdot \| A_{ji}^{k_p-m+j-1,\mu-1} \|_F \right] \right]^2 \right\} .$$

Der erste Summand in der großen geschweiften Klammer stellt den Block A_{pi} nach Verwendung als Pivotblock dar, und der zweite Summand beinhaltet die Veränderungen des Blockes durch die nachfolgenden Transformationcn in der p - ten Zeile. Die Größenord -

nung des Blockes A_{pi} nach der näherungsweisen Annullierung durch das Newton-Verfahren haben wir bereits abgeschätzt. Nach (3.49) erhalten wir

$$\| F(x^{k_p-m+i,\mu-1}) \| = \| A_{pi}^{k_p-m+i,\mu-1} \|_F \leq$$
$$\leq 2.99138 \cdot 10^5 \cdot \frac{\| A^{k_p-m+i-1,\mu-1} \|_F^{13}}{\delta^{14}} \cdot S^2(A^{k_p-m+i-1,\mu-1})$$
$$\leq 2.99138 \cdot 10^5 \cdot (1 + 2.1225 \cdot 10^{-5})^{k_p-m+i-i} \cdot \frac{\| A^0 \|_F^{13}}{\delta^{14}} \cdot S^2(A^{k_p-m+i-1,\mu-1})$$

Nun ist

$$S(A^{k_p-m+i-1,\mu-1}) \leq$$

$$\leq \left[\frac{118.813}{\delta^5} \right]^{(k_p-m+i-1)} \cdot \left[\prod_{i=0}^{(k_p-m+i-1)} (1 + 2.1225 \cdot 10^{-5})^{5i} \right] \cdot \| A^0 \|_F^{5(k_p-m+i-1)} \cdot S(A^0)$$
$$:= \tilde{K}_{k_p-m+i-1} \cdot S(A^0)$$

Für alle $i \in \{1,\dots,m\}$ gilt $\tilde{K}_{k_p-m+i-1} \leq \tilde{K}_M$. Es folgt insgesamt

$$S(A^{k_p-m+i-1,\mu-1}) \leq \tilde{K}_{k_p-m+i-1} \cdot S(A^0) \leq \tilde{K}_M \cdot S(A^0) \leq \tilde{K}_M \cdot \| A^0 \| \cdot \tilde{\epsilon}$$

Für die Größenordnung des Pivotblockes nach allen Transformationen der p-ten Zeile ergibt sich

$$\| A_{pi}^{k_p-m+i,\mu-1} \|_F \leq 2.99138 \cdot 10^5 \cdot (1 + 2.1225 \cdot 10^{-5})^{k_p-m+i-1} \cdot \frac{\| A^0 \|_F^{15}}{\delta^{14}} \cdot \tilde{K}_M^{\,2} \cdot \tilde{\epsilon}^2$$

Damit ist der Pivotblock ohne weitere Transformationen der p-ten Zeile ebenfalls quadratisch klein bzgl. $\tilde{\epsilon}$. Es bleibt die Störung durch die restlichen Transformationen der p-ten Zeile abzuschätzen. Es gilt

$$\left[\sum_{j=i+1}^{m}\left[S(A^{k_p-m+j-1,\mu-1})\cdot\|A_{j\,i}^{k_p-m+j-1,\mu-1}\|_F\right]\right]^2\leq$$

$$\leq\max\left[S^2(A^{k,\mu-1})\mid k\in\{k_p-m+i,\ldots,k_p-1\}\right]\cdot\left[\sum_{j=i+1}^{m}\|A_{j\,i}^{k_p-m+j-1,\mu-1}\|_F\right]^2$$

$$\leq\max\left[S^2(A^{k,\mu-1})\mid k\in\{k_p-m+i,\ldots,k_p-1\}\right]\cdot(m-i)\cdot\sum_{j=i+1}^{m}\|A_{j\,i}^{k_p-m+j-1,\mu-1}\|_F^2$$

$$\leq\max\left[S^2(A^{k,\mu-1})\mid k\in\{k_p-m+i,\ldots,k_p-1\}\right]\cdot(m-i)\cdot\frac{1}{2}\cdot S^2(A^{k_p-m+j-1,\mu-1})$$

$$\leq\frac{(m-i)}{2}\cdot\max\left[S^4(A^{k,\mu-1})\mid k\in\{1,\ldots,M\}\right]$$

$$\leq\frac{(m-i)}{2}\cdot\leq\tilde{K}_M^{\,4}\cdot\|A^0\|_F^4\cdot\tilde{\epsilon}^4$$

Insgesamt können wir jetzt beide Summanden zusammenfassen und erhalten

$$\|A_{pi}^{k_p,\mu-1}\|_F^2\leq(1+7.1025\cdot10^{-6})^{2m}\cdot$$

$$\left\{2\cdot2.99138^2\cdot10^{10}\cdot(1+2.1225\cdot10^{-5})^{2M}\cdot\frac{\|A^0\|_F^{30}}{\delta^{28}}\cdot\tilde{K}_M^{\,4}\cdot\tilde{\epsilon}^4+\right.$$

$$\left.4\cdot\tilde{\alpha}^2\cdot\frac{(m-i)}{2}\cdot\tilde{K}_M^{\,4}\cdot\|A^0\|_F^4\cdot\tilde{\epsilon}^4\right\}\leq$$

$$\leq(1+7.1025\cdot10^{-6})^{2m}\cdot$$

$$\left\{2.99138^2\cdot10^{10}\cdot(1+2.1225\cdot10^{-5})^{2M}+2\cdot\tilde{\alpha}^2\cdot\frac{(m-i)}{2}\cdot\frac{\delta^{28}}{\|A^0\|_F^{26}}\right\}\cdot2\cdot\frac{\|A^0\|_F^{30}}{\delta^{28}}\cdot\tilde{K}_M^{\,4}\cdot\tilde{\epsilon}^4$$

$$\leq(1+7.1025\cdot10^{-6})^{2m}\cdot$$

$$\left\{2.99138^2\cdot10^{10}\cdot(1+1.132\cdot10^{-5})^{2M}+\tilde{\alpha}^2\cdot(m-i)\cdot2^{13}\cdot\delta^2\right\}\cdot2\cdot\frac{\|A^0\|_F^{30}}{\delta^{28}}\cdot\tilde{K}_M^{\,4}\cdot\tilde{\epsilon}^4$$

Damit ist jedes Element der p-ten Zeile nach sämtlichen Transformationen der p-ten

Zeile quadratisch klein bezüglich $\tilde{\epsilon}$. Summation der p-ten Zeile ergibt

$$\sum_{j=p+1}^{m} \| A_{pi}^{k_p,\mu-1} \|_F^2 \leq (1+7.1025\cdot 10^{-6})^{2m} \cdot \Big[17.897\cdot 10^{10} \cdot (1+2.1225\cdot 10^{-5})^{2M} \cdot (m-p) \cdot$$

$$+ \tilde{\alpha}^2 \cdot (m-p)\cdot(m-p-1)\cdot 2^{13}\cdot \delta^2 \Big] \cdot \frac{\| A^0 \|_F^{30}}{\delta^{28}} \cdot \tilde{K}_M^{\,4} \cdot \tilde{\epsilon}^4$$

$$:= Q_p \cdot \| A^0 \|^2 \cdot \tilde{\epsilon}^4.$$

Dies ist die fehlende Abschätzung für die p-te Zeile. Zusammen mit der Abschätzung für die ersten (p-1) Zeilen bekommen wir jetzt

$$\sum_{i=1}^{p} \sum_{j=i+1}^{m} \| A_{ij}^{k_p,\mu-1} \|_F^2 \leq (2 + 1.9\cdot 10^{-5})^{m-p} \cdot C_{p-1} \cdot \| A^0 \|_F^2 \cdot \tilde{\epsilon}^4 + Q_p \cdot \| A^0 \|^2 \cdot \tilde{\epsilon}^4$$

$$= \Big[(2 + 1.9\cdot 10^{-5})^{m-p} \cdot C_{p-1} + Q_p \Big] \cdot \| A^0 \|_F^2 \cdot \tilde{\epsilon}^4$$

$$:= C_p \cdot \| A^0 \|_F^2 \cdot \tilde{\epsilon}^4.$$

Damit ist der Induktionsschritt durchgeführt.

•

Mit Hilfe des soeben bewiesenen Satzes ist es nun leicht zu zeigen, daß nach Durchführung des gesamten Zyklus die Außerdiagonalnorm quadratisch klein geworden ist. Wir formulieren dieses abschließende Ergebnis.

Satz 3.27

Es gelten die Voraussetzungen (3.68). Dann folgt

$$S(A^{M,\mu-1}) = S(A^{\mu}) = C_{m-1} \cdot \tilde{\epsilon}^2 \quad \text{mit}$$

$$C_{m-1} = (2+1.9\cdot 10^{-5})^{M} \cdot (1+7.1025\cdot 10^{-6})^{2m} \cdot$$

$$\cdot \Big[17.897\cdot 10^{10} \cdot (1+2.1225\cdot 10^{-5})^{2M} \cdot M + \tilde{\alpha}^2 \cdot M^2 \cdot 2^{12} \cdot \delta^2 \Big] \cdot \frac{\| A^0 \|_F^{28}}{\delta^{28}} \cdot \tilde{K}_M^{\,4}$$

Beweis:

Die Aussage des letzten Satzes für den letzten Index ist äquivalent zur Aussage

$$S^2(A^{M,\mu-1})) = 2\cdot \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \| A_{ij}^{k_p} \|_F^2 \leq 2 \cdot C_{m-1} \cdot \| A^0 \|_F^2 \cdot \tilde{\epsilon}^4$$

Die Konstante C_{m-1} kann wie folgt abgeschätzt werden

$$\begin{aligned} C_{m-1} &= \left[(2 + 1.9\cdot 10^{-5}) \cdot C_{m-2} + Q_{m-1} \right] \\ &= \sum_{p=1}^{m-1} (2 + 1.9\cdot 10^{-5})^{\frac{m-p+1}{2}\cdot(m-p)} \cdot Q_p \\ &\leq (2 + 1.9\cdot 10^{-5})^{M} \cdot \sum_{p=1}^{m-1} Q_p \end{aligned}$$

Es gilt weiter

$$\sum_{p=1}^{m-1} Q_p = \sum_{p=1}^{m-1} \Bigg\{ (1+7.1025\cdot 10^{-6})^{2m} \cdot$$

$$\left[17.897\cdot 10^{10}\cdot(1+2.1225\cdot 10^{-5})^{2M}\cdot(m-p)+\tilde{\alpha}^2\cdot(m-p)\cdot(m-p-1)\cdot 2^{13}\cdot \delta^2 \right] \cdot \frac{\| A^0 \|_F^{28}}{\delta^{28}} \cdot \tilde{K}_M^{\,4} \Bigg\}$$

$$\leq (1+7.1025\cdot 10^{-6})^{2m} \cdot \left[17.897\cdot 10^{10}\cdot(1+2.1225\cdot 10^{-5})^{2M}\cdot M+\tilde{\alpha}^2\cdot M^2\cdot 2^{12}\cdot \delta^2 \right] \cdot \frac{\| A^0 \|_F^{28}}{\delta^{28}} \cdot \tilde{K}_M^{\,4}$$

Somit ergibt sich als Schranke für C_{m-1}

$$C_{m-1} = (2+1.9\cdot 10^{-5})^{M}\cdot(1+7.1025\cdot 10^{-6})^{2m}\cdot$$
$$\cdot\left[17.897\cdot 10^{10}\cdot(1+2.1225\cdot 10^{-5})^{2M}\cdot M+\tilde{\alpha}^{2}\cdot M^{2}\cdot 2^{12}\cdot \delta^{2}\right]\cdot\frac{\| A^{0} \|_F^{28}}{\delta^{28}}\cdot\tilde{K}_M^{4}$$

●

Mit dieser letzten Abschätzung ist gezeigt, daß ein Zyklus unter der Voraussetzung (3.68) die Außerdiagonalnorm quadratisch klein macht. Diese Voraussetzung ist normabhängig, d.h. bei wachsender Norm der Matrix wird die Forderung an die Außerdiagonale der Matrix stärker. Leider ist das Verfahren nicht normreduzierend. Deswegen bleibt selbst bei potentieller Verkleinerung der Außerdiagonalen die Voraussetzung (3.68) für die quadratische Abnahme der Außerdiagonalen nicht automatisch für alle Zyklen erfüllt.

Im nächsten Abschnitt wird deswegen eine Schranke für die Außerdiagonalnorm konstruiert, die selbst bei sukzessivem Normwachstum durch die vollständigen Zyklen immer noch die Erfüllung der Voraussetzung (3.68) für die quadratische Abnahme der Außerdiagonalnorm $S(A)$ für beliebig viele Zyklen garantiert.

3.17 Asymptotisch quadratische Konvergenz

Dieser Abschnitt verschärft die Forderung an die Größenordnung der Außerdiagonale S(A) in der Form, daß die Voraussetzung (3.68) für alle Zyklen erfüllt bleibt und damit das Verfahren asymptotisch quadratisch konvergiert.

Zunächst schätzen wir im nachfolgenden Satz das Wachstum der Diagonalblöcke sowie deren Henrici - Maße während eines vollständigen Zyklus ab.

Satz 3.28

Es gelte für A^{l-1} die Generalvoraussetzung (3.28). A^{l-1} sei einem vollständigen Zyklus der hyperbolischen Transformation T sowie der orthogonalen Transformation U unterworfen. Nach $M = \frac{m\cdot(m-1)}{2}$ Transformationen erhalten wir A^l. Dann gilt für jeden Diagonalblock A_{jj}^l

$$\left.\begin{aligned} \| A_{jj}^l \|_F &\le \| A_{jj}^{l-1} \|_F + D_j^{l-1} \cdot S^2(A^{l-1}) \\ \Delta_F(A_{jj}^l) &\le \Delta_F(A_{jj}^{l-1}) + 2 \cdot D_j^{l-1} \cdot S^2(A^{l-1}) \end{aligned}\right] \quad \text{für } j = 1,..,m \qquad (3.71a)$$

mit

$$D_j^{l-1} :=$$

$$\sum_{k=1}^{m-1} \left\{ (1+2.1225\cdot 10^{-5})^{9(\Psi_j(k)-1)+\Sigma_k} \cdot \| A^{l-1} \|_F^{9+\Sigma_k} \cdot \left[\frac{118.813}{\delta^5}\right]^{2(\Psi_j(k)-1)} \right\} \cdot \frac{5952.93}{\delta^{10}} \qquad (3.71b)$$

und

$$\Sigma_k := 10 \cdot \sum_{l=1}^{k-1} (\psi_j(l) - 1) \text{ sowie } \psi_j \text{ gemäß (3.72).}$$

Beweis:

Die Diagonalblöcke A_{jj}^{l-1} werden für $j = 1,...,m$ jeweils $(m-1)$-mal im l-ten Zyklus transformiert. Die Nummern der Transformationen sind mit (3.15) gegeben durch

$$\psi_j(k) := \begin{cases} \varphi(k,j) & \text{für } 1 \le k < j \\ \varphi(j,k+1) & \text{für } j \le k \le m-1 \end{cases} \qquad (3.72)$$

Die Veränderungen des Diagonalblockes A_{jj}^{l-1} im $\psi_j(k)$-ten Schritt des Zyklus sind nach (3.61) und (3.53) gegeben durch die Matrizen $W_{jj}^{\Psi_j(k)-1,l-1}$ bzw. $W'^{\Psi_j(k)-1,l-1}_{jj}$. Es gilt

$$A_{jj}^{\Psi_j(k),l-1} = A_{jj}^{\Psi_j(k)-1,l-1} + W_{jj}^{\Psi_j(k)-1,l-1} + W'^{\Psi_j(k)-1,l-1}_{jj} \quad \text{für } k = 1,\ldots,m-1$$

Mit $A_{jj}^{\Psi_j(k)-1,l-1} = A_{jj}^{\Psi_j(k-1),l-1}$ folgt

$$A_{jj}^{l} = A_{jj}^{\Psi_j(m-1),l-1} = A_{jj}^{l-1} + \underbrace{\sum_{k=1}^{m-1}\left[W_{jj}^{\Psi_j(k)-1,l-1} + W'^{\Psi_j(k)-1,l-1}_{jj}\right]}_{:=\tilde{W}_{jj}^{l-1}}$$

Nach (3.61) bzw. (3.53) erhalten wir

$$\| W_{jj}^{\Psi_j(k)-1,l-1} \|_F \leq 1992.2 \cdot \frac{\| A^{\Psi_j(k)-1,l-1}\|_F^9}{\delta^{10}} \cdot S^2(A^{\Psi_j(k)-1,l-1})$$

$$\| W'^{\Psi_j(k)-1,l-1}_{jj} \|_F \leq 3960.73 \cdot \frac{\| A^{\Psi_j(k)-1,l-1}\|_F^9}{\delta^{10}} \cdot S^2(A^{\Psi_j(k)-1,l-1})$$

$$\| A_{jj}^{l} \|_F - \| A_{jj}^{l-1} \|_F \leq \sum_{k=1}^{m-1} 5952.93 \cdot \frac{\| A^{\Psi_j(k)-1,l-1}\|_F^9}{\delta^{10}} \cdot S^2(A^{\Psi_j(k)-1,l-1})$$

Für die maximale Normerhöhung pro Transformation gilt

$$\| A_{jj}^{\Psi_j(k)-1,l-1} \|_F \leq (1 + 2.1225\cdot 10^{-5})^{\Psi_j(k)-1} \cdot \| A_{jj}^{l-1} \|_F$$

und weiter

$$S(A^{\Psi_j(k)-1,l-1}) \leq$$

$$\left[\frac{118.813}{\delta^5}\right]^{\Psi_j(k)-1} \cdot \left[\prod_{l=1}^{k-1}(1+2.1225\cdot 10^{-5})^{5\Psi_j(l)-1} \cdot \| A^{l-1} \|_F^{5\Psi_j(l)-1}\right] \cdot S(A^{l-1})$$

Einsetzen liefert

$$\| \tilde{W}_{jj}^{l-1} \|_F \leq$$

$$\sum_{k=1}^{m-1} \left\{ (1+2.1225\cdot 10^{-5})^{9(\Psi_j(k)-1)+\Sigma_k} \cdot \| A^{l-1} \|_F^{9+\Sigma_k} \cdot \left[\frac{118.813}{\delta^5}\right]^{2(\Psi_j(k)-1)} \right\} \cdot \frac{5952.93}{\delta^{10}} \cdot S^2(A^{l-1})$$

$$:= D_j \cdot S^2(A^{l-1}) \qquad \text{mit } \Sigma_k := 10 \cdot \sum_{l=1}^{k-1} (\psi_j(l) - 1)$$

Für die Veränderung der Norm des Diagonalblockes A_{jj}^{l-1} bzw. des Henrici-Maßes durch den gesamten Zyklus folgen somit ähnliche Formeln:

$$\| A_{jj}^{l} \|_F \leq \| A_{jj}^{l-1} \|_F + \| \tilde{W}_{jj}^{l-1} \|_F \quad \text{bzw.}$$

$$\Delta_F(A_{jj}^{l}) \leq \Delta_F(A_{jj}^{l-1}) + 2 \cdot \| \tilde{W}_{jj}^{l-1} \|_F$$

also

$$\left.\begin{aligned} \| A_{jj}^{l} \|_F &\leq \| A_{jj}^{l-1} \|_F + D_j^{l-1} \cdot S^2(A^{l-1}) \\ \Delta_F(A_{jj}^{l}) &\leq \Delta_F(A_{jj}^{l-1}) + 2 \cdot D_j^{l-1} \cdot S^2(A^{l-1}) \end{aligned}\right\} \quad \text{für } j = 1,..,m$$

•

Nach der Abschätzung des Wachstums der Frobenius-Norm bzw. der Henrici-Maße der Diagonalblöcke nach einem vollen Zyklus kommen wir nun zur Berechnung von oberen Schranken für die Gesamtnorm der Matrix abhängig von der Größenordnung der Außerdiagonalen. Wir erhalten

Satz 3.29

$A \in \mathbb{R}^{2m,2m}$ sei eine 2x2-Blockmatrix, die die Generalvoraussetzung (3.28) erfüllt. Mit λ_φ für $\varphi = 1,\dots,n = 2\cdot m$ bezeichnen wir die Eigenwerte von A. $A_{jj} \in \mathbb{R}^{2,2}$ für $j = 1,\dots,m$ seien die Diagonalblöcke von A und $D = \mathrm{diag}\,(A)$ die Blockdiagonale von A. Dann gilt

$$\| A \|_F \le \left[\sum_{\varphi=1}^{n} |\lambda_\varphi|^2 \right]^{\frac{1}{2}} + \left[\sum_{j=1}^{m} \Delta_F^2(A_{jj}) \right]^{\frac{1}{2}} + R_1(S(A)) + R_2(S(A),\Delta_F(D)) \tag{3.73a}$$

wobei

$$R_1(\, S(A)\,) := \Big[1 + m\cdot(m-1) \Big]^{\frac{1}{2}} \cdot S(A) + \sqrt{\sqrt{2\cdot(m-1)}} \cdot \sqrt{S(A)} \cdot \left[\sum_{\varphi=1}^{n} |\lambda_\varphi| \right]^{\frac{1}{2}} \tag{3.73b}$$

$$\begin{aligned} R_2(\, S(A),\Delta_F(D)\,) := {} & \sqrt{2\cdot(m-1)} \cdot S(A)^{\frac{3}{4}} \cdot \sum_{j=1}^{m} \Delta_F(A_{jj})^{\frac{1}{4}} \\ & + \sqrt{(m-1)}\cdot\sqrt{S(A)} \cdot \sum_{j=1}^{m} \sqrt{\Delta_F(A_{jj})} \\ & + \sqrt{\sqrt{2\cdot(m-1)}} \cdot S(A)^{\frac{1}{4}} \cdot \sum_{j=1}^{m} \Big[|\lambda_{k(j,1)}| + |\lambda_{k(j,2)}| \Big]^{\frac{1}{2}} \cdot \Delta_F(A_{jj})^{\frac{1}{4}} \end{aligned} \tag{3.73c}$$

Beweis:

Wir teilen die Matrix $A \in \mathbb{R}^{2m,2m}$ in die Blockdiagonale $D = \mathrm{diag}\,(A)$ und die Außerdiagonale $AD := A - D$ auf und erhalten so für die Frobenius-Norm von A

$$\| A \|_F^2 = \| D \|_F^2 + \| AD \|_F^2 = \| D \|_F^2 + S^2(A)$$

und

$$\| D \|_F^2 = \sum_{j=1}^{m} \| A_{jj} \|_F^2$$

Nun gilt speziell für die Frobenius-Norm

$$\| A_{jj} \|_F^2 = \sum_{i=1}^{2} | \mu_i^{(j)} |^2 + \Delta_F^2(A_{jj}),$$

wobei $\mu_i^{(j)}$ für i = 1,2 die beiden Eigenwerte von A_{jj} sind. Diese beiden Eigenwerte können wir in Abhängigkeit von der Außerdiagonalen $S(A)$ als Näherungen für die beiden zugehörigen exakten Eigenwerte $\lambda_{k(j,1)}$ bzw. $\lambda_{k(j,2)}$ der Matrix A auffassen. Es gilt nach Theorem (3.1)

$$| \mu_i^{(j)} - \lambda_{k(j,i)} | \le \left[1 + \left(\sqrt{y_j + \tfrac{1}{4}} - \tfrac{1}{2}\right) \right] \cdot \sum_{\substack{k=1 \\ k \neq j}}^{m} \| A_{jk} \|_F .$$

mit $y_j = \dfrac{\Delta_F(A_{jj})}{\sum\limits_{\substack{k=1 \\ k \neq j}}^{m} \| A_{jk} \|_F}$.

Wegen $\sum\limits_{\substack{k=1 \\ k \neq j}}^{m} \| A_{jk} \|_F \le \sqrt{\dfrac{m-1}{2}} \cdot S(A)$ folgt für j = 1,...,m

$$| \mu_i^{(j)} - \lambda_{k(j,i)} | \le \frac{1}{2} \cdot \sqrt{\frac{m-1}{2}} \cdot S(A) + \sqrt{\Delta_F(A_{jj}) \cdot \sqrt{\frac{m-1}{2}} \cdot S(A) + \frac{m-1}{8} \cdot S^2(A)}$$

$$\le \sqrt{\frac{m-1}{2}} \cdot \sqrt{S(A)} \cdot \left[\sqrt{S(A)} + \sqrt{\Delta_F(A_{jj})} \right]$$

Damit erhalten wir

$$\| A_{jj} \|_F^2 \le \sum_{i=1}^{2} \left[|\mu_i^{(j)} - \lambda_{k(j,i)}|^2 + 2 \cdot |\mu_i^{(j)} - \lambda_{k(j,i)}| \cdot | \lambda_{k(j,i)} | + |\lambda_{k(j,i)}|^2 \right]$$

$$+ \Delta_F^2(A_{jj})$$

$$\le \sum_{i=1}^{2} \frac{m-1}{2} \cdot S(A) \cdot \left[\sqrt{S(A)} + \sqrt{\Delta_F(A_{jj})} \right]^2$$

$$+ \sum_{i=1}^{2} \left[|\lambda_{k(j,i)}| \cdot \sqrt{2 \cdot (m-1)} \cdot \sqrt{S(A)} \cdot \left[\sqrt{S(A)} + \sqrt{\Delta_F(A_{jj})} \right] + |\lambda_{k(j,i)}|^2 \right]$$

$$+ \Delta_F^2(A_{jj})$$

$$\le \sum_{i=1}^{2} |\lambda_{k(j,i)}|^2 + \Delta_F^2(A_{jj}) +$$

$$+ (m-1)\cdot S(A)\cdot\left[S(A) + 2\cdot\sqrt{\Delta_F(A_{jj})\cdot S(A)} + \Delta_F(A_{jj})\right]$$

$$+ \sqrt{2\cdot(m-1)}\cdot\sqrt{S(A)}\cdot\sum_{i=1}^{2}\left[|\lambda_{k(j,i)}|\cdot\left[\sqrt{S(A)} + \sqrt{\Delta_F(A_{jj})}\right]\right]$$

$$\| D\|_F^2 = \sum_{j=1}^{m} \| A_{jj} \|_F^2$$

$$\le \sum_{j=1}^{m}\sum_{i=1}^{2}|\lambda_{k(j,i)}|^2 + \sum_{j=1}^{m}\Delta_F^2(A_{jj}) + m\cdot(m-1)\cdot S^2(A)$$

$$+ 2\cdot(m-1)\cdot S(A)\cdot\sqrt{S(A)}\cdot\sum_{j=1}^{m}\sqrt{\Delta_F(A_{jj})} + (m-1)\cdot S(A)\cdot\sum_{j=1}^{m}\Delta_F(A_{jj})$$

$$+ S(A)\cdot\sqrt{2\cdot(m-1)}\cdot\sum_{i=1}^{2}\sum_{j=1}^{m}|\lambda_{k(j,i)}|$$

$$+ \sqrt{2\cdot(m-1)\cdot S(A)}\cdot\sum_{j=1}^{m}\left[|\lambda_{k(j,1)}| + |\lambda_{k(j,2)}|\right]\cdot\sqrt{\Delta_F(A_{jj})}$$

Mit der Abschätzung $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ ergibt die Addition von $S^2(A)$ zu $\| D\|_F^2$ die Behauptung.

●

Wir werden die Abschätzung (3.72) jeweils auf die Endmatrizen A^l der Zyklen mit $l \in \mathbb{N}$ anwenden. Die Formel hat als Grundlage das absolute Minimum der Frobenius-Norm einer Matrix, nämlich die Wurzel aus der Quadratsumme der Eigenwerte der Matrix. Als Korrekturterme treten drei unterschiedlich wirkende Terme auf. Der erste Term ist die Wurzel aus der Quadratsumme der Henrici-Maße der Blockdiagonale. Dieser Term trägt zur Normerhöhung bei und wächst von Zyklus zu Zyklus an. Der zweite Term $R_1(S(A))$ ist der Korrekturterm der Außerdiagonale. $R_1(S(A))$ für sich allein betrachtet liefert bei fallender Außerdiagonalnorm monoton fallende obere Schranken für die Frobenius-Norm der iterierten Matrizen. Der Term $R_2(S(A), \Delta_F(D))$ beinhaltet die Mischterme zwischen Außerdiagonalnorm $S(A)$ und Henrici-Maßen $\Delta_F(A_{jj})$ der Blockdiagonale. Zunächst ist also nicht klar, ob man die Gültigkeit der Voraussetzung (3.68) für lie quadratische Abnahme der Außerdiagonale über beliebig viele Zyklen garantieren kann, da das Wechselspiel zwischen dem Wachstum der $\Delta_F(A_{jj})$ und der Abnahme von $S(A)$ nicht bekannt ist. Dies liegt insbesondere daran, daß die Abschätzungen, die das Wachstum beider Größen beschreiben, Terme beinhalten, die von der aktuellen Norm der Matrix abhängen. Nach (3.71) gilt im l-ten Zyklus bei Einhaltung der Voraussetzung (3.68), die wir auch in der Form

$$S(A^{l-1}) \le \epsilon^{l-1} \quad \text{mit}$$

$$\epsilon^{l-1} = 118.813^{-M} \cdot \left[1+2.1225\cdot 10^{-5}\right]^{-K_M} \cdot \frac{\delta^2}{6.4\cdot 10^4 \cdot \sqrt{m-1}} \cdot \left[\frac{\delta}{\| A^{l-1}\|_F}\right]^{5m+7} \tag{3.74}$$

$$:= \tilde{E} \cdot \| A^{l-1}\|_F^{-5m-7}$$

schreiben können, für das Wachstum der Henrici-Maße $\Delta_F(A_{jj}^{l-1})$

$$\Delta_F(A_{jj}^{l}) \le \Delta_F(A_{jj}^{l-1}) + 2\cdot D_j^{l-1} \cdot (\epsilon^{l-1})^2$$

mit

$$D_j^{l-1} := \sum_{k=1}^{m-1} \left\{ (1+2.1225\cdot 10^{-5})^{9(\Psi_j(k)-1)+\Sigma_k} \cdot \| A^{l-1} \|_F^{9+\Sigma_k} \cdot \left[\frac{118.813}{\delta^5}\right]^{2(\Psi_j(k)-1)} \right\} \cdot \frac{5952.93}{\delta^{10}}$$

und $\Sigma_k := 10 \cdot \sum_{r=1}^{k-1} (\psi_j(r) - 1)$ sowie ψ_j gemäß (3.72).

Nach einigen Umformungen erhält man für die Terme D_j^{l-1} die Form

$$D_j^{l-1} = \tilde{D}_j \cdot \| A^{l-1} \|_F^{9+\Sigma_{m-1}} \quad \text{mit } \tilde{D}_j \in \mathbb{R} \text{ und für alle } j = 1,\dots,m. \tag{3.75}$$

Ebenso erhalten wir mit (3.70) für die Außerdiagonalnormabnahme im l-ten Zyklus

$$S(A^{l}) \le C^{l-1} \cdot (\epsilon^{l-1})^2 := \epsilon^{l} \quad \text{mit}$$

$$C^{l-1} := (2+1.9\cdot 10^{-5})^{M}\cdot(1+7.1025\cdot 10^{-6})^{2m}\cdot$$

$$\cdot\left[17.897\cdot 10^{10}\cdot(1+2.1225\cdot 10^{-5})^{2M}\cdot M+\tilde{\alpha}^{2}\cdot M^{2}\cdot 2^{12}\cdot \delta^{2}\right]\cdot\frac{\| A^{l-1}\|_F^{26}}{\delta^{28}}\cdot \tilde{K}_M^{\,4}$$

, wobei

$$\tilde{\alpha} := (2+4.735\cdot 10^{-6})\cdot \sum_{r=0}^{4}\left[2.05\cdot(1+2.1225\cdot 10^{-5})^{M}\cdot\| A^{l-1}\|_F\cdot\frac{1}{\delta}\right]^{r}\cdot\frac{1}{0.976\cdot\delta}\cdot\frac{\sqrt{2}}{2}$$

und

$$\tilde{K}_M \leq \left[\frac{118.813}{\delta^{5}}\right]^{\tilde{M}}\cdot\left[\prod_{i=0}^{\tilde{M}}(1+2.1225\cdot 10^{-5})^{5i}\right]\cdot\| A^{l-1}\|_F^{5M}$$

Nach einigen Umformungen erkennt man, daß auch hier der Term C^{l-1} von der Form

$$C^{l-1} = \tilde{C}\cdot\| A^{l-1}\|_F^{20M+34} \quad \text{mit } \tilde{C}\in\mathbb{R} \tag{3.76}$$

ist.

Damit sind alle drei Terme C^i, D^i_j sowie E^i als Funktionen der aktuellen Frobenius-Norm $\| A^i\|_F$ von der Form

$$C^i,\ D^i_j,\ E^i = \text{Const.}\cdot\| A^i\|_F^{s} \quad \text{mit geeigneten } s\in\mathbb{Z}. \tag{3.77}$$

Damit sehen wir, daß einerseits die Schranken für die Henrici-Maße der Blockdiagonalen (3.71) bzw. für die Verkleinerung der Außerdiagonalen (3.70) bei wachsender Matrix-Norm immer gröber werden, andererseits die Voraussetzung (3.74) für quadratische Außerdiagonalnormabnahme bei wachsender Norm eine immer kleiner werdende Außerdiagonalnorm zu Beginn eines Zyklus fordert. Für E^i ist die Potenz s der Matrix-Norm negativ, ansonsten positiv.

An der Abschätzung (3.73a) für die Frobenius-Norm erkennt man, daß das eigentliche Problem im Wachstum der Henrici-Maße der Blockdiagonale liegt. Wählt man die Außerdiagonale ϵ^0 so klein, daß quadratische Abnahme von $S(A)$ einsetzt, also

$$C^0\cdot\epsilon^0 = \tilde{C}\cdot\| A^0\|_F^{20M+34}\cdot\epsilon^0 \leq \frac{1}{2} < 1$$

und blieben die Henrici-Maße der Blockdiagonale beschränkt, so konvergierte das Verfahren asymptotisch quadratisch. Damit die Henrici-Maße beschränkt bleiben, ist es sinnvoll, eine Forderung zu stellen, die ihr Wachstum einschränken. Die Größenordnung des Wachstums ist nach Satz (3.28) gegeben durch

$$\Delta_F(A^1_{jj}) \le \Delta_F(A^0_{jj}) + 2 \cdot D^0_j \cdot (\epsilon^0)^2$$

Wir fordern zusätzlich

$$2 \cdot D^0_j \cdot \epsilon^0 = 2 \cdot \tilde{D}_j \cdot \| A^0 \|^6_F \cdot \epsilon^0 = \frac{1}{2} < 1$$

Beide Forderungen sind erfüllbar, und der zweite Zyklus ist durchführbar und führt zur quadratischen Abnahme der Außerdiagonale. Wir erhalten folglich

$$S(A^1) \le \epsilon^1 \le C^0 \cdot (\epsilon^0)^2 \le \frac{1}{2} \cdot \epsilon^0$$

Gleichzeitig bekommen wir eine obere Schranke für die neuen Henrici-Maße. Es gilt nach (3.27) und (3.71a)

$$\Delta_F(A^1_{jj}) \le \Delta_F(A^0_{jj}) + 2 \cdot D^0_j \cdot (\epsilon^0)^2 \le \Delta_F(A^0_{jj}) + \frac{1}{2} \cdot \epsilon^0$$

Mit der neuen Schranke für $\Delta_F(A^1_{jj})$ ergibt sich aus Satz (3.29) eine obere Schranke für $\| A^1 \|_F$, und wir können erneut fordern:

Wähle $\epsilon^0 \in \mathbb{R}$ so klein, daß die Forderungen

$$\begin{array}{llll}
\epsilon^1 & < \epsilon^0 & \le \tilde{E} & \cdot \left[\max\limits_{k=0}^{1} \| A^k \|_F \right]^{-5m-7} \\
2 \cdot D^1_j \cdot \epsilon^1 & < 2 \cdot D^1_j \cdot \epsilon^0 & \le 2 \cdot \tilde{D}_j & \cdot \left[\max\limits_{k=0}^{1} \| A^k \|_F \right]^{9+\Sigma_{m-1}} \cdot \epsilon^0 \le \frac{1}{2} \text{ für alle } j = 1,\ldots,m \\
C^1 \cdot \epsilon^1 & < C^1 \cdot \epsilon^0 & \le \tilde{C} & \cdot \left[\max\limits_{k=0}^{1} \| A^k \|_F \right]^{20M+34} \cdot \epsilon^0 \le \frac{1}{2}
\end{array}$$

erfüllt sind

Damit wären die beiden ersten Zyklen mit quadratischer Außerdiagonalnormabnahme verbunden. Diese Vorgehensweise ist beliebig oft wiederholbar, sodaß für jedes $l \in \mathbb{N}$ ein $\epsilon^0 \in \mathbb{R}$ berechenbar ist, daß die ersten l Zyklen quadratische Außerdiagonalnormabnahme erzeugen. Dies ist allerdings nur unter Berechnung ständig wachsender Schranken für die Normen der Matrizen A^l möglich. Der nachfolgende Satz beschreibt die Situation nach l Zyklen des Verfahrens.

Satz 3.30

Sei $l \in \mathbb{N}$ beliebig und $A^0 \in \mathbb{R}^{2m,2m}$ eine 2x2-Blockmatrix. Die Eigenwerte von A^0 bezeichnen wir mit λ_φ für $\varphi = 1,\dots,n = 2\cdot m$. Ferner seien $A^0_{jj} \in \mathbb{R}^{2,2}$ für $j = 1,\dots,m$ die Diagonalblöcke von A^0. Es gelte $S(A^0) \le \epsilon^0$, wobei ϵ^0 so gewählt ist, daß folgende drei Bedingungen erfüllt sind:

$$\epsilon^0 \le \tilde{E} \cdot \left[\max_{k=0}^{l-1} \| A^k \|_F\right]^{-5m-7} \tag{3.78a}$$

$$D_j^{l-1}\cdot\epsilon^0 \le \tilde{D}_j \cdot \left[\max_{k=0}^{l-1} \| A^k \|_F\right]^{9+\Sigma_{m-1}} \cdot\epsilon^0 = \frac{1}{2} < 1 \text{ für alle } j = 1,\dots,m \tag{3.78b}$$

$$C^{l-1}\cdot\epsilon^0 \le \tilde{C} \cdot \left[\max_{k=0}^{l-1} \| A^k \|_F\right]^{20M+34} \cdot\epsilon^0 = \frac{1}{2} < 1. \tag{3.78c}$$

Dann sind die ersten l Zyklen durchführbar mit quadratischer Außerdiagonalnormabnahme, d.h.

$$S(A^i) \le \epsilon^i \le \prod_{k=1}^{i-1} (C^{k-1})^{2^{i-k}} \cdot (\epsilon^0)^{2^i} \le \left[\frac{1}{2}\right]^{2^i - 1} \cdot \epsilon^0 \quad \text{für } i = 1,\dots,l \tag{3.79}$$

Weiterhin gilt für die Henrici-Maße der Blockdiagonale

$$\Delta_F(A^i_{jj}) \le \Delta_F(A^0_{jj}) + \epsilon^0\cdot\sum_{k=1}^{i}\left[\frac{1}{2}\right]^{2^k-1} \quad \text{für } i = 1,\dots,l. \tag{3.80}$$

Für die Frobenius-Normen der Matrizen A^i erhalten wir mit $D^0 = \mathrm{diag}(A^0)$ für $i = 1,\dots,l$

$$\| A^i \|_F \le \left[\sum_{\varphi=1}^{n} |\lambda_\varphi|^2\right]^{\frac{1}{2}} + \left[\sum_{j=1}^{m} \Delta_F^2(A^0_{jj})\right]^{\frac{1}{2}} + R_1^i(\epsilon^0) + R_2^i(\epsilon^0, \Delta_F(D^0)), \tag{3.81a}$$

wobei

$$R_1^i(\epsilon^0) := \left[1 + m\cdot(m-1)\right]^{\frac{1}{2}} \cdot \left[\frac{1}{2}\right]^{2^{i-1}-1} \cdot \epsilon^0$$
$$+ \sqrt{\sqrt{2\cdot(m-1)}} \cdot \left\{\left[\frac{1}{2}\right]^{2^{i-1}-1} \cdot \epsilon^0\right\}^{\frac{1}{2}} \cdot \left[\sum_{\varphi=1}^{n} |\lambda_\varphi|\right]^{\frac{1}{2}} \tag{3.81b}$$
$$+ \sqrt{m} \cdot \epsilon^0 \cdot \sum_{k=1}^{i} \left[\frac{1}{2}\right]^{2^k-1}$$

$$R_2^i(\epsilon^0, \Delta_F(D^0)) =$$
$$= \sqrt{2\cdot(m-1)} \cdot \left\{\left[\frac{1}{2}\right]^{2^{i-1}-1} \cdot \epsilon^0\right\}^{\frac{3}{4}} \cdot \sum_{j=1}^{m}\left[\Delta_F(A_{jj}^0) + \epsilon^0 \cdot \sum_{k=1}^{i}\left[\frac{1}{2}\right]^{2^k-1}\right]^{\frac{1}{4}} \tag{3.81c}$$
$$+ \sqrt{(m-1)} \cdot \left\{\left[\frac{1}{2}\right]^{2^{i-1}-1} \cdot \epsilon^0\right\}^{\frac{1}{2}} \cdot \sum_{j=1}^{m}\left[\Delta_F(A_{jj}^0) + \epsilon^0 \cdot \sum_{k=1}^{i}\left[\frac{1}{2}\right]^{2^k-1}\right]^{\frac{1}{2}}$$
$$+ \sqrt{\sqrt{2\cdot(m-1)}} \cdot \left\{\left[\frac{1}{2}\right]^{2^{i-1}-1} \cdot \epsilon^0\right\}^{\frac{1}{4}} \cdot$$
$$\cdot \sum_{j=1}^{m}\left[|\lambda_{k(j,1)}| + |\lambda_{k(j,2)}|\right]^{\frac{1}{2}} \cdot \left\{\Delta_F(A_{jj}^0) + \epsilon^0 \cdot \sum_{k=1}^{i}\left[\frac{1}{2}\right]^{2^k-1}\right\}^{\frac{1}{4}}$$
$$+ \left\{\epsilon^0 \cdot \sum_{k=1}^{i}\left[\frac{1}{2}\right]^{2^k-2}\right]^{\frac{1}{2}} \cdot \left[\sum_{j=1}^{m} \Delta_F(A_{jj}^0)\right]^{\frac{1}{2}}$$

Beweis:

Es gelten wegen Voraussetzung (3.78) für alle $i \in \{1,\dots,l-1\}$ folgende Beziehungen

$$\epsilon^i \leq \left[\frac{1}{2}\right]^i \cdot \epsilon^0 \leq \epsilon^0 \leq \tilde{E} \cdot \left[\max_{k=0}^{l-1} \| A^k \|_F\right]^{-5m-7} \leq \tilde{E} \cdot \| A^i \|_F^{-5m-7}$$

$$2\cdot D_j^i \cdot \epsilon^i \leq \left[\frac{1}{2}\right]^{i-1} \cdot D_j^i \cdot \epsilon^0 \leq D_j^{l-1} \cdot \epsilon^0 \leq \tilde{D}_j \cdot \left[\max_{k=0}^{l-1} \| A^k \|_F\right]^{9+\Sigma_{m-1}} \cdot \epsilon^0 = \frac{1}{2} < 1$$

für alle $j = 1,\dots,m$

$$C^i \cdot \epsilon^i \leq \left[\frac{1}{2}\right]^i \cdot C^i \cdot \epsilon^0 \leq C^{l-1} \cdot \epsilon^0 \leq \tilde{C} \cdot \left[\max_{k=0}^{l-1} \| A^k \|_F\right]^{20M+34} \cdot \epsilon^0 = \frac{1}{2} < 1$$

Die erste Ungleichung besagt gerade, daß die Voraussetzung für die quadratische Abnahme der Außerdiagonale bis zum l-ten Zyklus gültig ist, während die beiden übrigen Abschätzungen für alle l Zyklen das Wachstum von $S(A^{i})$ bzw. der $\Delta_F(A^{i}_{jj})$ bestimmen. Wir erhalten mit Satz (3.28) für die Henrici-Maße der Diagonalblöcke

$$\Delta_F(A^{i}_{jj}) \leq \Delta_F(A^{0}_{jj}) + \sum_{k=1}^{i} 2 \cdot D_j^{k-1} \prod_{l=1}^{k-1} (C^{l-1})^{2^{k-1}} \cdot (\epsilon^0)^{2^k} \qquad \text{für } i = 1,\ldots,l$$

Wegen (3.78b) folgt damit für alle $j \in \{1,\ldots,m\}$

$$\Delta_F(A^{i}_{jj}) \leq \Delta_F(A^{0}_{jj}) + \epsilon^0 \cdot \sum_{k=1}^{i} \left[\frac{1}{2}\right]^{2^k-1} \qquad \text{für } i = 1,\ldots,l$$

Schließlich bekommen wir wegen (3.78c)

$$S(A^i) \leq \epsilon^i \leq \prod_{k=1}^{i-1} (C^{k-1})^{2^{i-k}} \cdot (\epsilon^0)^{2^i} \leq \left[\frac{1}{2}\right]^{2^i-1} \cdot \epsilon^0 \qquad \text{für } i = 1,\ldots,l$$

Die Abschätzung (3.81) für die Frobenius-Norm der iterierten Matrizen A^i erhalten wir nun durch Einsetzen der beiden obigen Abschätzungen in (3.73). Dabei muß wiederum von der Formel $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ Gebrauch gemacht werden.

●

Mit der Abschätzung (3.81) der Frobenius-Normen der Matrix A können wir nun für beliebiges $l \in \mathbb{N}$ eine obere Schranke für $\| A^{l} \|_F$ konstruieren, die unabhängig von der Nummer des aktuellen Zyklus ist. Mit dieser oberen Schranke ist es möglich, die Forderungen an $S(A^{0}) \leq \epsilon^0$ so zu verschärfen, daß asymptotisch quadratische Konvergenz garantiert ist.

Satz 3.31

Sei $A^0 \in \mathbb{R}^{2m,2m}$ eine 2x2-Blockmatrix. Die Eigenwerte von A^0 bezeichnen wir mit λ_φ für $\varphi = 1,\dots,n = 2\cdot m$. Ferner seien $A^0_{jj} \in \mathbb{R}^{2,2}$ für $j = 1,\dots,m$ die Diagonalblöcke von A^0. Es gelte $S(A^0) \leq \epsilon^0$, wobei ϵ^0 so gewählt ist, daß folgende drei Bedingungen erfüllt sind

$$\epsilon^0 \leq \tilde{E} \cdot \left[\| A \|_{F,\max} \right]^{-5m-7} \tag{3.82a}$$

$$\tilde{D}_j \cdot \left[\| A \|_{F,\max} \right]^{9+\Sigma_{m-1}} \cdot \epsilon^0 \leq \frac{1}{2} \quad \text{für alle } j = 1,\dots,m \tag{3.82b}$$

$$\tilde{C} \cdot \left[\| A \|_{F,\max} \right]^{20M+34} \cdot \epsilon^0 \leq \frac{1}{2}. \tag{3.82c}$$

Hierbei ist

$$\| A \|_{F,\max} = \left[\sum_{\varphi=1}^{n} |\lambda_\varphi|^2 \right]^{\frac{1}{2}} + \left[\sum_{j=1}^{m} \Delta_F^2(A^0_{jj}) \right]^{\frac{1}{2}} + R_1^{\max}(\epsilon^0) + R_2^{\max}(\epsilon^0, \Delta_F(D^0)) \tag{3.83a}$$

mit $D^0 = \operatorname{diag}(A^0)$

$$R_1^{\max}(\epsilon^0) := \epsilon^0 \cdot \left\{ \left[1 + m\cdot(m-1) \right]^{\frac{1}{2}} + \sqrt{m} \cdot 0.633 \right\} + (\epsilon^0)^{\frac{1}{2}} \cdot \sqrt{\sqrt{2\cdot(m-1)}} \cdot \left[\sum_{\varphi=1}^{n} |\lambda_\varphi| \right]^{\frac{1}{2}} \tag{3.83b}$$

$$\begin{aligned} R_2^{\max}(\epsilon^0, \Delta_F(D^0)) = {} & \sqrt{2\cdot(m-1)\cdot(\epsilon^0)^{\frac{3}{4}}} \cdot \sum_{j=1}^{m} \left[\Delta_F(A^0_{jj}) + 0.633 \cdot \epsilon^0 \right]^{\frac{1}{4}} \\ & + \sqrt{(m-1)\cdot(\epsilon^0)^{\frac{1}{2}}} \cdot \left\{ \left[\sum_{j=1}^{m} \left[\Delta_F(A^0_{jj}) + 1.266 \cdot \epsilon^0 \right] \right]^{\frac{1}{2}} + \left[\sum_{j=1}^{m} \Delta_F(A^0_{jj}) \right]^{\frac{1}{2}} \right\} \\ & + (\epsilon^0)^{\frac{1}{4}} \cdot \sqrt{\sqrt{2\cdot(m-1)}} \cdot \sum_{j=1}^{m} \left\{ \left[|\lambda_{k(j,1)}| + |\lambda_{k(j,2)}| \right]^{\frac{1}{2}} \cdot \left[\Delta_F(A^0_{jj}) + 0.633 \cdot \epsilon^0 \right]^{\frac{1}{4}} \right\} \end{aligned} \tag{3.83c}$$

Dann ist das Verfahren asymptotisch quadratisch konvergent. Die Abnahme der Außerdiagonalnorm $S(A)$ ist gegeben durch

$$S(A^i) \le \epsilon^i \le \prod_{k=1}^{i-1} (C^{k-1})^{2^{i-k}} \cdot (\epsilon^0)^{2^i} \le \left[\frac{1}{2}\right]^{2^i-1} \cdot \epsilon^0 \quad \text{für } i \ge 1 \tag{3.84}$$

Weiterhin gilt für die Henrici-Maße $\Delta_F(A_{jj})$ für $j \in \{1,\ldots,m\}$

$$\Delta_F(A_{jj}^i) \le \Delta_F(A_{jj}^0) + 0.633 \cdot \epsilon^0 \qquad \text{für } i \ge 1 \tag{3.85}$$

<u>Beweis:</u>

Der Beweis folgt unmittelbar aus der Abschätzung (3.81) des vorangehenden Satzes. Die beiden Terme $R_1^i(\epsilon^0)$ und $R_2^i(\epsilon^0, \Delta_F(D^0))$ aus (3.81) sind gegeben durch

$$\begin{aligned} R_1^i(\epsilon^0) :=\ & \Big[1 + m\cdot(m-1)\Big]^{\frac{1}{2}} \cdot \left[\frac{1}{2}\right]^{2^{i-1}-1} \cdot \epsilon^0 \\ & + \sqrt{\sqrt{2\cdot(m-1)}} \cdot \left\{\left[\frac{1}{2}\right]^{2^{i-1}-1} \cdot \epsilon^0\right\}^{\frac{1}{2}} \cdot \left[\sum_{\varphi=1}^{n} |\lambda_\varphi|\right]^{\frac{1}{2}} \\ & + \sqrt{m} \cdot \epsilon^0 \cdot \sum_{k=1}^{i} \left[\frac{1}{2}\right]^{2^k-1} \end{aligned}$$

$$\begin{aligned} & R_2^i(\epsilon^0, \Delta_F(D^0)) = \\ & = \sqrt{2\cdot(m-1)} \cdot \left\{\left[\frac{1}{2}\right]^{2^{i-1}-1} \cdot \epsilon^0\right\}^{\frac{3}{4}} \cdot \sum_{j=1}^{m} \left\{ \Delta_F(A_{jj}^0) + \epsilon^0 \cdot \sum_{k=1}^{i} \left[\frac{1}{2}\right]^{2^k-1} \right\}^{\frac{1}{4}} \\ & + \sqrt{(m-1)} \cdot \left\{\left[\frac{1}{2}\right]^{2^{i-1}-1} \cdot \epsilon^0\right\}^{\frac{1}{2}} \cdot \sum_{j=1}^{m} \left\{ \Delta_F(A_{jj}^0) + \epsilon^0 \cdot \sum_{k=1}^{i} \left[\frac{1}{2}\right]^{2^k-1} \right\}^{\frac{1}{2}} \\ & + \sqrt{\sqrt{2\cdot(m-1)}} \cdot \left\{\left[\frac{1}{2}\right]^{2^{i-1}-1} \cdot \epsilon^0\right\}^{\frac{1}{4}} \cdot \\ & \cdot \sum_{j=1}^{m} \Big[|\lambda_{k(j,1)}| + |\lambda_{k(j,2)}|\Big]^{\frac{1}{2}} \cdot \left\{ \Delta_F(A_{jj}^0) + \epsilon^0 \cdot \sum_{k=1}^{i} \left[\frac{1}{2}\right]^{2^k-1} \right\}^{\frac{1}{4}} \\ & + \left\{ \epsilon^0 \cdot \sum_{k=1}^{i} \left[\frac{1}{2}\right]^{2^k-2} \right\}^{\frac{1}{2}} \cdot \left[\sum_{j=1}^{m} \Delta_F(A_{jj}^0) \right]^{\frac{1}{2}} \end{aligned}$$

Es ist $\left[\frac{1}{2}\right]^{2^{i-1}-1} \leq 1$ für alle $i \geq 1$ und weiter gilt

$$\sum_{k=1}^{i}\left[\frac{1}{2}\right]^{2^k-1} \leq 2\cdot\sum_{k=0}^{\infty}\left[\frac{1}{2}\right]^{2^{k+1}} \approx 0.632843 \leq 0.633.$$

Setzt man diese Abschätzungen in die beiden obigen Terme ein, so erhält man $R_1^{max}(\epsilon^0)$ und $R_2^{max}(\epsilon^0,\Delta_F(D^0))$, und als Normabschätzung für alle iterierten Matrizen A^i mit $i \geq 1$ ergibt sich

$$\| A^i \|_F \leq \| A \|_{F,max} = \left[\sum_{\varphi=1}^{n} |\lambda_\varphi|^2\right]^{\frac{1}{2}} + \left[\sum_{j=1}^{m} \Delta_F^2(A_{jj}^0)\right]^{\frac{1}{2}} + R_1^{max}(\epsilon^0) + R_2^{max}(\epsilon^0,\Delta_F(D^0))$$

Bei Verwendung von $\| A \|_{F,max}$ in den Forderungen an ϵ^0 haben wir dann ein $\epsilon^0 > 0$ konstruiert, das für alle Zyklen die Voraussetzung (3.68) garantiert und damit asymptotisch quadratische Konvergenz.

•

Bemerkung:

Die Abschätzung der Norm der iterierten Matrizen A^i zeigt, daß das bisher beschriebene Verfahren das absolute Minimum der Norm der Matrix A, nämlich die Wurzel aus der Quadratsumme der Eigenwerte von A, nicht erreichen kann. Unsere Abschätzung besagt selbst bei verschwundener Außerdiagonale $S(A) = 0$

$$\| A \|_F \leq \left[\sum_{\varphi=1}^{n} |\lambda_\varphi|^2\right]^{\frac{1}{2}} + \left[\sum_{j=1}^{m} \Delta_F^2(A_{jj}^0)\right]^{\frac{1}{2}}.$$

Dies liegt daran, daß die Blockdiagonale der Matrix überhaupt nicht bearbeitet wird.

Das Minimum der Norm kann aber nur erreicht werden, wenn sämtliche Diagonalblöcke A_{jj} ebenfalls diagonalisiert sind bzw. einem komplex konjugierten Eigenwertpaar entsprechen. Solange keine Maßnahmen getroffen werden, die die Norm der Blockdiagonale minimieren, wird immer der Störterm der Henrici - Maße der Blockdiagonale vorhanden bleiben.

Der letzte Abschnitt behandelt eine Transformation, die die Blockdiagonale auf Murnaghan - Form bringt, falls dies möglich ist.

3.18 Normreduzierung der Blockdiagonalen

Im letzten Abschnitt beschreiben wir den Einbau einer weiteren hyperbolischen Ähnlichkeitstransformation. Durch diese Transformation kann im Fall diagonalisierbarer Diagonalblöcke die Matrix auf Murnaghan-Form gebracht werden, und im Fall defektiver Diagonalblöcke wird zumindest eine Normreduzierung des jeweiligen defektiven Blockes erreicht und eine Normalisierung angestrebt.

Für jeden Diagonalblock A_{pp} mit $p = 1,...,m$ sind die (i,j) -Restriktionen der Transformationsmatrix $S^p(z)$ gegeben durch

$$S^p_{pp} := \begin{bmatrix} \cosh z & \sinh z \\ \sinh z & \cosh z \end{bmatrix} \quad \text{mit } z \in \mathbb{R} \tag{3.86}$$

$$S^p_{ij} := I \cdot \delta_{ij} \qquad \text{für } (i,j) \neq (p,p)$$

$S^p(z)$ ist J-orthogonal. Falls es nicht von Interesse ist, auf welchem Pivotblock die Transformation S wirkt, lassen wir den oberen Index p weg. Die Matrix S ist stets regulär, und sie hat als Funktion des Parameters $z \in \mathbb{R}$ folgende Eigenschaften

$$\begin{aligned} S(0,0) &= I & S(\tilde{z}+z) &= S(\tilde{z}) \cdot S(z) \\ S^{-1}(z) &= S(-z) & S(z)^T &= S(z) \end{aligned} \tag{3.87}$$

Für feste Parameter $z \in \mathbb{R}$ und ein festen Index p sei die Ähnlichkeitstransformation auf A beschrieben durch

$$A''' := (S^p)^{-1}(z) \cdot A'' \cdot (S^p)(z) \tag{3.88}$$

Dann ist A''' wieder J-symmetrisch, und es gelten folgende Transformationsvorschriften für die einzelnen (i,j)-Restriktionen der Matrix A''

$$\begin{aligned} A'''_{pi} &= S_{pp}(-z) \cdot A''_{pi} \quad \text{für } i = 1, ...m \text{ und } i \neq p \\ A'''_{pp} &= S_{pp}(-z) \cdot A''_{pp} \cdot S_{pp}(z) \end{aligned} \tag{3.89}$$

Für spätere Betrachtungen benötigen wir ebenfalls die Transformationsformeln einzelner Elemente der Matrix $A'' = \left[a''_{ij} \right]_{1 \le i,j \le n}$. Es gilt

$$\left.\begin{aligned} a'''_{pi} &= a''_{pi} \cdot ch - a''_{p+m,i} \cdot sh \\ a'''_{p+m,i} &= a''_{p+m,i} \cdot ch - a''_{pp} \cdot ch \end{aligned}\right] \quad \text{für } i = 1,\ldots,m \text{ und } i \neq p \tag{3.90a}$$

Für die Elemente des Diagonalblockes A_{pp} bekommen wir

$$\begin{aligned} a'''_{pp} &= a''_{pp} \cdot ch^2 + 2 \cdot a''_{p,p+m} \cdot ch \cdot sh - a''_{p+m,p+m} \cdot sh^2 \\ a'''_{p,p+m} &= a''_{p,p+m} \cdot ch^2 + (a''_{pp} - a''_{p+m,p+m}) \cdot ch \cdot sh + a''_{p,p+m} \cdot sh^2 \\ a'''_{p+m,p+m} &= a''_{p+m,p+m} \cdot ch^2 - 2 \cdot a''_{p,p+m} \cdot ch \cdot sh - a''_{p,p} \cdot sh^2 \end{aligned} \tag{3.90b}$$

Die übrigen Matrixelemente bleiben unverändert. Hierbei sei $ch := \cosh(z)$ und $sh := \sinh(z)$.

Im weiteren lassen wir zur Vereinfachung der Notation die 3 Striche zur Erkennung der durch S transformierten Elemente weg. Eine kurze Rechnung ergibt für die Frobenius Norm des Diagonalblockes A'''_{pp} nach der Transformation mit S^p

$$\begin{aligned} \| A_{pp} \|_F^2 = \cosh^2(2 \cdot z) &\cdot (a_{pp}^2 + 2 \cdot a_{p,p+m}^2 + a_{p+m,p+m}^2) \cdot \\ &+ \sinh^2(2 \cdot z) \cdot (2 \cdot a_{p,p+m}^2 - a_{pp} \cdot a_{p+m,p+m}) \\ &+ \sinh(4 \cdot z) \cdot 2 \cdot a_{p,p+m} \cdot (a_{pp} - a_{p+m,p+m}) \end{aligned} \tag{3.91}$$

Für $2 \cdot |a_{p,p+m}| \neq | a_{pp} - a_{p+m,p+m} |$, also für nicht defektive Diagonalblöcke, ist die Norm des Diagonalblockes A_{pp} als Funktion des Paramters z streng konvex, ansonsten konvex.

Die Aufgabe der hyperbolischen Diagonaltransformation ist es, die Norm der Blockdiagonale zu reduzieren. Für nicht defektive Diagonalblöcke ist dieses Problem einfach zu lösen, denn diese sind auf Murnaghan - Form transformierbar, und das Minimum der Norm wird gerade für dieses Form angenommen. Die Norm des transformierten Diagonalblockes A'_{pp} als Funktion des Winkels z der hyperbolischen Transformation hat folgende Gestalt:

$$\begin{aligned} \| A'_{pp} \|_F^2 = \cosh^2(2\cdot z) & \cdot (a_{pp}^2 + 2\cdot a_{p,p+m}^2 + a_{p+m,p+m}^2)\cdot \\ + \sinh^2(2\cdot z) & \cdot 2\cdot(a_{p,p+m}^2 - a_{pp}\cdot a_{p+m,p+m}) \\ + \sinh(4\cdot z) & \cdot 2\cdot a_{p,p+m}\cdot(a_{pp} - a_{p+m,p+m}) \quad := f(z) \end{aligned}$$

Zur Abkürzung setzen wir

$$\begin{aligned} A &:= (a_{pp}^2 + 2\cdot a_{p,p+m}^2 + a_{p+m,p+m}^2) \\ B &:= 2\cdot(a_{p,p+m}^2 - a_{pp}\cdot a_{p+m,p+m}) \\ C &:= 2\cdot a_{p,p+m}\cdot(a_{pp} - a_{p+m,p+m}) \end{aligned} \tag{3.92a}$$

und erhalten

$$f(z) = A\cdot\cosh^2(2\cdot z) + B\cdot\sinh^2(2\cdot z) + C\cdot\sinh(4\cdot z). \tag{3.92b}$$

Das Minumum der Funktion ergibt sich durch Differenzieren

$$f'(z) = 2\cdot(A + B)\cdot\sinh(4\cdot z) + 4\cdot C\cdot\cosh(4\cdot z) = 0 \iff$$

$$\tanh(4\cdot z) = \frac{2\cdot C}{A + B} = \frac{4\cdot a_{p,p+m}\cdot(a_{pp} - a_{p+m,p+m})}{4\cdot a_{p,p+m}^2 + (a_{pp} - a_{p+m,p+m})^2} \tag{3.93}$$

Der Winkel ist für alle nichtdefektiven Diagonalblöcke stets wohldefiniert, denn der Quotient ist betragsmäßig stets kleiner gleich 1. Für solche Blöcke, die bereits diagonalisiert bzw. in Murnaghan - Form vorliegen, ist z = 0, und die minimale Norm ist bereits erreicht. Für alle diagonalisierbaren Blöcke liegt nach Durchführung der Transformation mit dem oben berechneten Winkel der zugehörige Diagonalblock in Murnaghan - Form vor.

Nur für defektive Blöcke mit $2\cdot|a_{p,p+m}| = | a_{pp} - a_{p+m,p+m} |$ gilt $| \tanh(4\cdot z) | = 1$, und der Winkel z divergiert. Das Minimum von $\| A'_{pp} \|_F$ wird dann im Unendlichen angenommen. In diesem Fall ist die Funktion f nicht mehr streng konvex.

Für solche defektiven Diagonalblöcke A_{pp} muß der Winkel z künstlich beschränkt werden, da sonst die Außerdiagonale in ihre Größenordnung zu stark verändert wird. Aber selbst für fast normale Diagonalblöcke kann der Winkel sehr groß werden. Dies ist genau dann der Fall, wenn die beiden zugehörigen Eigenwerte sehr nahe beieinanderliegen. Die eigentliche Problematik von großen Winkeln bei der Normreduzierung der Blockdiagonalen tritt aber erst dann auf, wenn neben den Eigenwerten zusätzlich die Eigenvektoren mitberechnet werden sollen. Diese ergeben sich aus den Spalten des Produkts aller Transformationsmatrizen. Die Genauigkeit der Eigenvektoren wird deswegen bei schlecht konditionierten Transformationsmatrizen in Mitleidenschaft gezogen. Im Falle von J-symmetrischen 2x2-Blöcken ist die Kondition der Transformationsmatrix leicht explizit berechenbar. Es gilt

Satz 3.32

Sei $A = \begin{bmatrix} a & b \\ -b & d \end{bmatrix}$ eine diagonalisierbare, J-symmetrische Matrix, also gelte $|a-d| \neq 2\cdot|b|$. A sei nicht normal, also $a \neq d$ und $b \neq 0$. Die Eigenwerte von A sind gegeben durch

$$\lambda_{1,2} = \frac{a+d}{2} \pm \sqrt{\left\{\frac{a-d}{2}\right\}^2 - b^2}.$$

Die Transformationsmatrix S, die A diagonalisiert, lautet

$$S = \begin{bmatrix} (\lambda_1 - a) & (\lambda_2 - a) \\ b & b \end{bmatrix} \text{und } S^{-1} = \frac{1}{b^2\cdot(\lambda_1 - \lambda_2)} \cdot \begin{bmatrix} b & (a-\lambda_2) \\ -b & (\lambda_1 - a) \end{bmatrix}$$

Für die Kondition der Transformation erhalten wir

$$\mathrm{Kond}_F(S) = \| S \|_F \cdot \| S^{-1} \|_F = \frac{|d-a|}{|b|} \cdot \left\{(a-d)^2 - 4\cdot b^2\right\}^{-\frac{1}{2}} \tag{3.94}$$

$$= \frac{|d-a|}{|b|} \cdot \frac{1}{|\lambda_1 - \lambda_2|}$$

Bemerkung:

Man sieht an der Konditionsformel, daß sich bei sehr nahe beieinanderliegenden Eigenwerten die Kondition der Transformation und damit die Genauigkeit der im Produkt der Transformationsmatrizen stehenden Eigenvektoren drastisch verschlechtern kann. Dieses Phänomen äußert sich in betragsmäßig großen Winkeln der hyperbolischen Diagonaltransformation S.

Zur Verhinderung der beiden oben beschriebenen Wirkungen beschränken wir uns auf Winkel z, die eine bestimmte Größenordnung nicht überschreiten. Wir verzichten also bei sehr nahe beieinanderliegenden Eigenvektoren bzw. bei Matrizen, die defektiv oder fast defektiv sind, auf eine Normalisierung der Blockdiagonale gemäß (3.93) und begnügen uns mit einer reinen Normreduzierung. Der Winkel z soll deshalb wie folgt bestimmt werden:

Für $\left| \dfrac{4 \cdot a_{p,p+m} \cdot (a_{pp} - a_{p+m,p+m})}{4 \cdot a_{p,p+m}^2 + (a_{pp} - a_{p+m,p+m})^2} \right| \leq \tanh(4) = 0.9993293$ sei z die Lösung von

$$\tanh(4 \cdot z) = \frac{4 \cdot a_{p,p+m} \cdot (a_{pp} - a_{p+m,p+m})}{4 \cdot a_{p,p+m}^2 + (a_{pp} - a_{p+m,p+m})^2}, \tag{3.95}$$

ansonsten sei $z = \operatorname{sign} \left\{ a_{p,p+m} \cdot (a_{pp} - a_{p+m,p+m}) \right\}$.

Mit diesem Ansatz folgt

$$|z| \leq 1 \text{ und damit } \cosh(z) \leq 1.5431 \text{ bzw. } |\sinh(z)| \leq 1.1753 \tag{3.96}$$

Diese Form der Winkelbestimmung garantiert gut konditionierte Transformationsmatrizen, und gleichzeitig ist die Wirkung der Transformation S auf die Außerdiagonalnorm gering. Die m Transformationen der Blockdiagonale kann die Außerdiagonalnorm nur minimal vergrößern, so daß die quadratische Konvergenz des Verfahrens erhalten bleibt. Es gilt

Satz 3.33

Sei mit $A \in \mathbb{R}^{2m,2m}$ eine 2x2-Blockmatrix gegeben und es gelte $S(A) \leq \epsilon$. Die Matrix A werde den m Transformationen $S^P(z)$ für p = 1,...,m unterworfen, wobei die Winkel gemäß (3.95) bestimmt werden, d.h.

$$A' = (S^P)^{-1} \cdot \ldots \cdot (S^1)^{-1} \cdot A \cdot S^1 \cdot \ldots \cdot S^P$$

Dann gilt

$$S(A') \leq 7.3891^m \cdot S(A). \tag{3.97}$$

Beweis:

Wir teilen A in die Blockdiagonale D und in $AD = A - D$ auf. Dann ergibt eine Transformation mit S^i für beliebiges $i \in \{1,\ldots,m\}$

$$\tilde{A} = (S^P)^{-1} \cdot A \cdot S^i = (S^i)^{-1} \cdot D \cdot S^i + (S^i)^{-1} \cdot AD \cdot S^i$$

Die Transformationsmatrix S^i ist selbst eine Blockdiagonalmatrix. Daher gilt

$$S(\tilde{A}) = \| (S^i)^{-1} \cdot AD \cdot S^i \|_F \leq \mathrm{Kond}_2(S^i) \cdot \| AD \|_F = \mathrm{Kond}_2(S^i) \cdot S(A)$$

Dabei ist von Bedeutung, daß man die Kondition auch in der Spektralnorm messen kann. Die Eigenwerte der Transformation S^i sind aber gerade gegeben durch (n-2) Einsen, $\lambda_+ = \cosh(z) + |\sinh(z)|$ und $\lambda_- = \cosh(z) - |\sinh(z)|$. Wegen $|z| \leq 1$ erhalten wir so

$$\mathrm{Kond}_2(S^i) \leq (\cosh(1) + |\sinh(1)|)^2 \leq e^2 \approx 7.3890561 \leq 7.3891$$

Diese Abschätzung gilt für alle m Transformationen, und es folgt die Behauptung.

•

Damit bleibt die durch die einzelnen Zyklen erzeugte Größenordnung der Außerdiagonale $S(A)$ auch nach Durchführung der m Normreduzierungen der Blockdiagonale erhalten. Dabei ist gleichzeitig gewährleistet, daß die Diagonalblöcke überall dort, wo dies ohne Verschlechterung der Genauigkeit der Eigenvektoren möglich ist, diagonalisiert werden. Insbesondere wenn die Eigenwerte gut getrennt liegen, ist mit einer asymptotisch quadratischen Konvergenz des Verfahrens gegen eine Diagonalmatrix zu rechnen.

Hier sind die theoretischen Überlegungen abgeschlossen. Im nächsten Kapitel beschäftigen wir uns mit der numerischen Praxis des vorgestellten Verfahrens.

4. Die Verfahren in der Praxis

Im 4. Kapitel wollen wir zunächst den Ursprung J-symmetrischer Matrizen erläutern und die Möglichkeiten darstellen, wie man aus dem quadratischen Eigenwertproblem durch entsprechende Transformationen das J-symmetrische Problem gewinnt. Anschließend erklären wir, wie das asymptotisch quadratisch konvergente primitive Verfahren des 3. Kapitels zur Beschleunigung eines J-symmetrischen Blockverfahrens von Hoppe [11] eingesetzt werden kann. Im 3. Abschnitt stellen wir die numerischen Ergebnisse dar.

4.1 Erzeugung J-symmetrischer Matrizen aus dem quadratischen Eigenwertproblem

J-symmetrische Eigenwertprobleme entstehen insbesondere bei der Betrachtung von schwingungsfähigen Systemen. Die Grundgesetze der Physik besagen, daß ein solches System durch den Erhalt der Gesamtenergie des Systems charakterisiert ist. Die Gesamtenergie ist gegeben durch die Summe aus kinetischer Energie T und der potentiellen Energie U.

$$T + U = \text{const.} \tag{4.1}$$

Für ein System mit n Freiheitsgraden kann man beide Energien über die Systemkoordinaten y_i, $i = 1,\dots,n$, als quadratische Formen mit reellen, symmetrischen, positiv semidefiniten Matrizen beschreiben. Es ist

$$\begin{aligned} U &= 0.5 \cdot y^T \cdot K \cdot y, \quad K \in \mathbb{R}^{n,n} \text{ symmetrisch, positiv semidefinit} \\ T &= 0.5 \cdot y^T \cdot M \cdot y, \quad M \in \mathbb{R}^{n,n} \text{ symmetrisch, positiv definit} \end{aligned} \tag{4.2}$$

Die Matrix M wird als Massenmatrix und die Matrix K als Steifigkeitsmatrix bezeichnet. Durch Differenzieren von (4.1) nach der Zeit erhält man die Differentialgleichung der ungedämpften Schwingung

$$M \cdot y'' + K \cdot y = 0 \tag{4.3}$$

Bei gedämpften Systemen tritt ein weiterer geschwindigkeitsabhängiger Term auf, der durch eine positiv definite, symmetrische Dämpfungsmatrix D gegeben ist, und wir erhalten

$$M \cdot y'' + D \cdot y' + K \cdot y = 0 \tag{4.4}$$

Der zeitliche Verlauf der Bewegung wird durch den Vektor y(t) beschrieben, der die n Systemkoordinaten $y_i(t)$ beinhaltet. Unter Verwendung des Ansatzes

$$y(t) := x \cdot e^{\lambda \cdot t} \tag{4.5}$$

erhalten wir das quadratische Eigenwertproblem

$$(M \cdot \lambda^2 + D \cdot \lambda + K) \cdot x = 0. \tag{4.6}$$

Untersuchungen des Wertebereiches der Matrizen M, D und K zeigen, daß im Fall positiv definiter Matrizen D und K die Lösungen der Differentialgleichung (4.4) stabil sind. Die Berechnung der $2 \cdot n$ Eigenwerte λ_i und der zugehörigen Eigenvektoren x_i ergeben gemäß (4.5) die Fundamentallösungen der Differentialgleichung der gedämpften Schwingung, und jede Lösung ist durch Vorgabe von geeigneten Anfangs- bzw. Randbedingungen linear aus diesen Fundamentallösungen kombinierbar. Das Auffinden der $2 \cdot n$ Eigenwerte des quadratischen Eigenwertproblems entspricht dabei der Berechnung der Grundschwingungen des schwingungsfähigen Systems.

Wir wollen nun beschreiben, wie das quadratische Eigenwertproblem (4.6) in ein gewöhnliches J-symmetrisches Eigenwertproblem umgewandelt werden kann und gehen dabei von positiv definiten, symmetrischen Matrizen M, D und K aus. Die einfache Transformation

$$z := \begin{bmatrix} \lambda \cdot x \\ x \end{bmatrix} \text{ und } \mu := \frac{1}{\lambda} \tag{4.7}$$

führt auf das verallgemeinerte Eigenwertproblem der doppelten Dimension $2 \cdot n$

$$S \cdot z = \mu \cdot T \cdot z \text{ mit } S := \begin{bmatrix} 0 & M \\ M & D \end{bmatrix} \text{ und } T := \begin{bmatrix} M & 0 \\ 0 & -K \end{bmatrix} \tag{4.8}$$

S ist dabei eine symmetrische, reelle Matrix der Dimension $2 \cdot n$. Die Matrix $T \in \mathbb{R}^{2n,2n}$ hingegen ist symmetrisch und besitzt wegen der positiven Definitheit der Matrizen M und K n negative und n positive, reelle Eigenwerte. Hierin tritt die Möglichkeit zutage, das verallgemeinerte Eigenwertproblem (4.8) in ein J-symmetrisches einfaches Eigenwertproblem der gleichen Dimension zu transformieren.

Als erste Variante dieser Transformation verwenden wir die positive Definitheit der Matrizen M und K, um sie nach Cholesky zu zerlegen. Wir setzen

$$K = L_1 \cdot L_1^T \qquad M = L_2 \cdot L_2^T \text{ mit } L_1, L_2 \text{ als linke, untere Dreiecksmatrizen} \tag{4.9}$$

und erhalten so

$$T = \begin{bmatrix} M & 0 \\ 0 & -K \end{bmatrix} = \begin{bmatrix} L_2 & 0 \\ 0 & L_1 \end{bmatrix} \cdot J \cdot \begin{bmatrix} L_2^T & 0 \\ 0 & L_1^T \end{bmatrix}$$

mit $J := \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}$, I Einheitsmatrix im $\mathbb{R}^{n,n}$.

Somit transformiert sich (4.8) in das Problem

$$\begin{bmatrix} 0 & L_2^{-1} \cdot M \cdot L_1^{-T} \\ L_1^{-1} \cdot M \cdot L_2^{-T} & L_1^{-1} \cdot D \cdot L_1^{-T} \end{bmatrix} \cdot z = \mu \cdot J \cdot z$$

$\Leftrightarrow$

$$\begin{bmatrix} 0 & L_2^{T} \cdot L_1^{-T} \\ -L_1^{-1} \cdot L_2 & -L_1^{-1} \cdot D \cdot L_1^{-T} \end{bmatrix} \cdot z = \mu \cdot z \tag{4.10}$$

Wir erhalten so eine spezielle Klasse J-symmetrischen Eigenwertproblemen (4.10), die aus dem ursprünglich gegebenen gedämpften Schwingungsproblem ableitbar sind. Die Eigenwertverteilung ist für diese Klasse leicht in Abhängigkeit von der Dämpfungsmatrix D charakterisierbar. Falls $D \equiv 0$ ist (keine Dämpfung), so erhalten wir ungedämpfte Schwingungen, und alle Eigenwerte sind rein imaginär. Sie liegen alle als komplex konjugierte Paare auf der imaginären Achse der Gaußschen Zahlenebene. Bei wachsender positiv definiter Matrix D (Energieverlust durch Dämpfung) wandern die Eigenwerte als komplex konjugierte Paare in die negative Halbebene, bis sie schließlich bei genügend starker Dämpfung zu doppelt reellen, negativen Eigenwerten entarten. Bei weiterer Zunahme der Dämpfung treten durch Trennung der doppelten Eigenwerte und Laufen auf der negativen, reellen Halbachse in entgegengesetzen Richtungen auch einfache, reelle Eigenwerte auf.

<u>Bemerkung</u>

Für die Transformation von (4.8) auf das J-symmetrische Problem ist nicht von Bedeutung, daß die Zerlegungen (4.9) auf Dreiecksmatrizen beruhen. Man kann vielmehr statt der normalen Cholesky-Zerlegung eine Zerlegung mit Spaltenpivotisierung durchführen. Dies führt zu einer Cholesky-Zerlegung mit einer Matrix L mit permutierten Zeilen, so daß die Dreiecksgestalt zerstört ist. Diese Vorgehensweise soll die Diagonalelemente in ihrer Größenordnung sortieren. Dies hat den Vorteil, daß naheliegende Eigenwerte in gleiche 2x2-Blöcke des J-symmetrischen Problems gelangen und so die Trennung der Eigenräume der J-symmetrischen Matrix verbessert wird.

Bemerkung

Die benutzte Transformation führt auf ein Eigenwertproblem zur Berechnung der Kehrwerte der gesuchten Eigenwerte der gedämpften Schwingung. Vom physikalischen Standpunkt werden in der Regel bei komplizierten schwingungsfähigen Systemen nur eine kleine Zahl von Grundschwingunen benötigt, die die größten Schwingungsamplituden besitzen. Der Übergang zu den Kehrwerten führt zu einer besseren Trennung dieser Grundschwingungen. Der hochfrequente Anteil der Schwingungen wird in die Nähe von Null transformiert und kann damit schlechter berechnet werden. Dies ist aber nicht so wichtig.

Man kann natürlich auch die Eigenwerte der inversen Matrix zum J-symmetrischen Problem (4.10) betrachten. Das inverse Problem ist gegeben durch

$$\begin{bmatrix} L_2^{-1} \cdot D \cdot L_2^{-T} & -L_2^{-1} \cdot L_1 \\ L^T \cdot L^{-T} & 0 \end{bmatrix} \cdot z = \lambda \cdot z \tag{4.11}$$

Eine etwas aufwendigere Technik, aus dem quadratischen Eigenwertproblem (4.6) ein J-symmetrisches Eigenwertproblem zu konstruieren, besteht darin, die symmetrischen, positiv definiten Matrizen K und M durch die symmetrische QR-Methode [29] simultan zu diagonalisieren, indem man zunächst das verallgemeinerte Eigenwertproblem

$$K \cdot x = \omega^2 \cdot M \cdot x \tag{4.12}$$

löst. Dieses Problem besitzt nur die halbe Dimension des ursprünglichen J-symmetrischen Problems. Man erhält die Modalmatrix Φ mit

$$\begin{aligned} &\Phi^{T} \cdot K \cdot \Phi = k = \operatorname{diag}(\omega_1^2,\ldots,\omega_n^2) && (\omega_i^2 \text{ Eigenwerte von (4.12))} \\ &\Phi^{T} \cdot M \cdot \Phi = I && I \ \text{ Einheitsmatrix im } \mathbb{R}^{n,n}. \end{aligned} \tag{4.13}$$

Das quadratisch Eigenwertproblem (4.6) ergibt sich damit zu

$$\begin{aligned} &\left\{ \Phi^{T} \cdot M \cdot \Phi \cdot \lambda^2 + \Phi^{T} \cdot D \cdot \Phi \cdot \lambda + \Phi^{T} \cdot K \cdot \Phi \right\} \cdot x = 0 \\ &\Leftrightarrow \\ &\left\{ \lambda^2 + \Phi^{T} \cdot D \cdot \Phi \cdot \lambda + k \right\} \cdot x = 0. \end{aligned} \tag{4.14}$$

Das mit diesem vortransformierten Problem gebildete J-symmetrische Eigenwertproblem lautet

$$\begin{bmatrix} 0 & k^{-\frac{1}{2}} \\ -k^{-\frac{1}{2}} & -k^{-\frac{1}{2}} \cdot \Phi^{T} \cdot D \cdot \Phi \cdot k^{-\frac{1}{2}T} \end{bmatrix} \cdot z = \mu \cdot z \tag{4.15}$$

Diese Vorgehensweise verspricht eine bessere Konvergenz der iterativen Verfahren, da die J-symmetrische Matrix bereits diagonalisierte Nebenblöcke $k^{-\frac{1}{2}}$ besitzt und der Matrixblock $-k^{-\frac{1}{2}} \cdot \Phi^{T} \cdot D \cdot \Phi \cdot k^{-\frac{1}{2}T}$ in der Regel bei nicht zu starken Kopplungen auch bereits eine diagonaldominante Matrix ist, so daß die Blockdiagonalisierung wesentlich schneller abläuft.

Bemerkung

Auch die Verwendung der symmetrischen QR-Methode erlaubt den Übergang zur inversen J-symmetrischen Matrix, die sich aufgrund der geleisteten Vorarbeit noch schneller berechnen läßt. Das inverse Problem lautet

$$\begin{bmatrix} -\Phi^{\mathrm{T}} \cdot D \cdot \Phi & k^{\frac{1}{2}} \\ -k^{\frac{1}{2}} & 0 \end{bmatrix} \cdot \mathrm{z} = \lambda \cdot \mathrm{z} \tag{4.16}$$

Unsere numerischen Tests in den nachfolgenden Abschnitten werden sich auf die in diesem Abschnitt vorgestellten J-symmetrischen Matrizen beziehen.

4.2 Kombination des primitiven Verfahrens und eines normreduzierenden Verfahrens

Das 3. Kapitel hat bewiesen, daß unser primitives Blockverfahren für J-symmetrische Matrizen asymptotisch quadratisches Konvergenzverhalten zeigt. Die Taktik, einzelne Matrixelemente des Pivotblockes zu annullieren, ist rechentechnisch einfach und deshalb vom Zeitaufwand recht gering. Allerdings ist das Verfahren nicht global konvergent. Ein primitives Verfahren wird in der Regel nur dann konvergieren, wenn die Außerdiagonalnorm der Matrix bereits genügend kleine Größenordnung hat.

Daher ist ein primitives Verfahren lediglich als Beschleuniger für global konvergente Verfahren einzusetzen, die allerdings kompliziertere und damit zeitintensive Parameterbestimmungen für die einzelnen Transformationen benutzen. Wir werden unser primitives Verfahren zusammen mit einem normreduzierenden Verfahren von Hoppe [11] einsetzen.

Für das in [11] behandelte Verfahren ist zwar ebenfalls lediglich asymptotisch quadratische Konvergenz nachgewiesen. Dennoch ist kein Fall aus der Praxis bekannt, wo das Verfahren nicht konvergierte. Das Verfahren benutzt dieselben Ähnlichkeitstransformationen wie unser primitives Verfahren und ebenfalls dieselbe Pivotstrategie. Allerdings wird die hyperbolische Transformation zur Normreduzierung der Gesamtmatrix eingesetzt. Die Parameter werden dabei so bestimmt, daß die Abnahme der Frobenius-Norm der Matrix durch die Transformation maximal wird. Dies bedingt, daß die Veränderungen von 2 Zeilen und 2 Spalten der Matrix als Funktion der hyperbolischen Winkel berechnet werden müssen und ein Extremum dieser Funktion in zwei Parametern aufgesucht werden muß. Zusätzlich wird eine dritte Transformation, die wir in Kapitel 3.18 beschrieben haben, ebenfalls zur Normreduktion eingesetzt. Sie dient der Normalisierung der Blockdiagonalen der Matrix. Hierbei tritt der gleiche Rechenaufwand wie bei der Normreduktion der Außerdiagonalen auf, wenngleich die Normreduktionsfunktion in diesem Fall nur von einem Parameter abhängt. Hoppe hat in seinen Algorithmen *Alg1,* *Alg2 bzw. Alg3* (vgl. [11] ,S. 124 ff.) diese Hauptdiagonalnormreduktion jeweils einmal pro Zyklus auf die Blockdiagonale angewendet. Dieselbe Strategie werden wir in unserer Kombination beider Algorithmen ebenfalls verfolgen. Unser Konvergenzbeweis des primitiven Verfahrens zeigt, daß die Transformation lediglich zur Erzeugung der Murnaghan - Form der Matrix von Bedeutung ist. Außerdem zeigen die Ausführungen in 3.18, daß statt der aufwendigen Normreduzierung ein einfacher Annullierungsschritt im Pivotblock zur Normalisierung genügt.

Das von uns getestete Verfahren hat damit folgenden Aufbau:

1) Wir benutzen gemäß Kapitel 3.2 zeilenzyklische Pivotstrategie
2) Auf jeden Pivotblock wird je eine hyperbolische und eine orthogonale Transformation gemäß Kapitel 3.1 angewendet
3) Die Parameterbestimmungen erfolgt stets zunächst mit der primitiven Methode. Treten hierbei große hyperbolische Parameter auf mit $|\tanh| >$THMAX, so werden die berechneten Parameter verworfen und sie werden nach der normreduzierenden Methode in [11] neu bestimmt.

Die Schranke THMAX ist somit die Einschaltschwelle für das primitive Verfahren. Eine Matrix, die von Beginn an kleine hyperbolische Winkel zuläßt, wird folglich rein annullierend mit dem primitiven Algorithmus auf Blockdiagonalgestalt gebracht.

Für das so kombinierte Verfahren bieten sich weitere Modifikationen an:

1) Zu Beginn des Verfahrens wird für eine bestimmte Zahl von Zyklen nur das normreduzierende Verfahren eingesetzt, um eine diagonaldominante Matrix als Ausgangsmatrix für das eigentliche Schwellenverfahren zu haben.
2) Nach Durchführung aller Transformationen einer Pivotzeile wird abschließend der zugehörige Hauptdiagonalblock gemäß 3.18 einmal transformiert.
3) Dem gesamten Verfahren wird eine einmalige Normreduzierung der Blockdiagonale vorgeschaltet mit dem Ziel, eine schon normalisierte Blockdiagonale als Start für die Iteration vorzufinden.

Zusammen mit den Modifikationen erhalten wir für das Kombinationsverfahren zwei Parameter, die das Verfahren wesentlich beeinflussen.

THMAX Einschaltschwelle der primitiven Transformationen

und

NZYKL Zahl der Startzyklen mit rein normreduzierender Parameterwahl

Unsere Tabellen mit den numerischen Resultaten im letzten Abschnitt werden sich auch auf diese beiden Parameter beziehen.

Schließlich ist noch eine Bemerkung zur Berechnung der Eigenvektoren zu machen. Zur Bestimmung der Eigenvektoren müssen die einzelnen Transformationsmatrizen einseitig aufmultipliziert werden. Man kann hierbei einigen Rechenaufwand sparen, wenn man neben der Matrix der Eigenvektoren einen zusätzlichen Vektor mitführt und die Elementarmatrizen nach der Fast-Givens-Methode (siehe Rath [18]) durchführt. Dabei werden die Elementartransformationen aufgespalten in der Form

$$\begin{bmatrix} \cos(x) & \sin(x) \\ -\sin(x) & \cos(x) \end{bmatrix} = \begin{bmatrix} \cos(x) & 0 \\ 0 & \cos(x) \end{bmatrix} \cdot \begin{bmatrix} 1 & \tan(x) \\ -\tan(x) & 1 \end{bmatrix} \text{für } |\cos(x)| \geq |\sin(x)|$$

bzw.

$$\begin{bmatrix} \cos(x) & \sin(x) \\ -\sin(x) & \cos(x) \end{bmatrix} = \begin{bmatrix} \sin(x) & 0 \\ 0 & \sin(x) \end{bmatrix} \cdot \begin{bmatrix} \cot(x) & 1 \\ 1 & \cot(x) \end{bmatrix} \text{für } |\cos(x)| < |\sin(x)|$$

(analog im hyperbolischen Fall, wobei wegen $\cosh(x) > |\sinh(x)|$ stets die erste Beziehung benutzt wird), wodurch die entstandenen Diagonalmatrizen in einem getrennten Vektor mit leichten Modifikationen aufmultipliziert werden können und der Rechenaufwand an Multiplikationen etwa halbiert wird. Diese Verringerung der arithmetischen Operationen wird insbesondere bei großen Dimensionen einen deutlichen Zeiteinsparungseffekt liefern.

Damit kommen wir zu den numerischen Testergebnissen.

4.3 Numerische Ergebnisse

Alle numerischen Ergebnisse wurden auf einem TANDON - AT Personal - Computer mit INTEL 80286 Prozessor sowie dem numerischen Coprozessor 80287 bei einem Systemtakt von 8 MHz erzielt. Alle Programme wurden in TURBO - PASCAL 5.0 geschrieben, wobei das EXTENDED - Format der reellen Zahlen benutzt wurde, so daß die Maschinengenauigkeit bei 10^{-20} liegt.

Die Tabellen dieses Abschnittes beinhalten folgende Werte:

ZYKL	Zahl der Zyklen
TRANS	Gesamtzahl aller Transformationen
NRED	Zahl der normreduzierenden Transformationen
ZEIT	Die CPU - Zeit in Sekunden
S(A)	Die Außerdiagonalnorm
JORTH	Die Abweichung der Modalmatrix von der J - Orthogonalität
RESID	Die Norm des Residuums des Eigenvektorsystem $\max_i \| A \cdot x_i - \lambda_i x_i \|_2$, wobei λ_i der zum Eigenvektor x_i gehörende Eigenwert sei.

Die Jacobi - ähnlichen primitiven Algorithmen sind zusätzlich durch die im letzten Abschnitt bereits erwähnten reellen Parameter

THMAX	Einschaltschwelle des primitiven Algorithmus
NZYKL	Zahl der anfänglichen Zyklen mit normreduzierenden Rotationen

gekennzeichnet. Die Namensgebung lautet stets

THMAX Schwelle NZYKL bzw. Primitiv NZYKL

Weiterhin sind in jeder Tabelle als Vergleich die Ergebnisse des QR - Algorithmus aus [29] sowie des rein normreduzierenden Hoppe - Verfahrens aus [11] mit angeführt.

Als erste Beispiele verwende ich diagonaldominante Matrizen mit gut getrennten Eigenwerten. Deshalb habe ich die Murnaghan-Form der Matrix vorgegeben und anschließend durch Ähnlichkeitstransformationen mit kleinen Parametern die Murnaghan-Form zerstört. Diese Vorgehensweise liefert Testmatrizen mit exakt vorgebbaren Eigenräumen und genau beschreibbaren Trennungseigenschaften der Eigenräume der gestörten Matrizen. Der Konvergenzbeweis hat bereits gezeigt, daß die Getrenntheit der Eigenräume von besonderer Bedeutung ist.

1. Matrix mit rein imaginären Eigenwerten.

a) 10x10-Matrix mit Eigenwerten $\pm 0, \pm i, \pm 2i, \ldots, \pm 4i$

b) 20x20-Matrix mit Eigenwerten $\pm 0, \pm i, \pm 2i, \ldots, \pm 9i$

c) 30x30-Matrix mit Eigenwerten $\pm 0, \pm i, \pm 2i, \ldots, \pm 14i$

Alle drei Matrizen wurden mithilfe der hyperbolischen bzw. orthogonalen Transformationen von ihrer Murnaghan-Form wegtransformiert, wobei ein vollständiger Zyklus mit zeilenzyklischer Pivotwahl und konstanten Parametern für beide Ähnlichkeitstransformationen durchgeführt wurde. Durch die Größenordnung der Parameter kann dann die Abweichung von der Murnaghan-Form kontrolliert werden.
Es ergeben sich folgende Tabellen

a) 10x10 Matrix mit imaginären Eigenwerten

<u>Tabelle 1a</u> Störung mit $\tanh(x) = \tan(x) = 10^{-4}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				2.09		3.416E-03	1.792E-18
Hoppe	3	40	40	1.92	1.059E-20	4.968E-20	8.673E-19
Primitiv 0	3	42	5	1.70	9.451E-21	4.629E-20	6.505E-19

Tabelle 1b — Störung mit $\tanh(x) = \tan(x) = 10^{-2}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				2.26		2.434E-01	5.901E-18
Hoppe	4	48	48	2.25	1.278E-25	3.866E-19	1.084E-18
Primitiv 0	4	51	5	2.04	2.542E-22	6.632E-19	1.192E-18

Tabelle 1c — Störung mit $\tanh(x) = \tan(x) = 10^{-1}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				2.25		4.734E-01	4.289E-18
Hoppe	4	57	57	2.63	5.996E-21	2.575E-19	1.627E-18
Primitiv 0	4	58	5	2.30	2.044E-20	3.432E-19	1.737E-18

b) 20x20 - Matrix mit imaginären Eigenwerten

Tabelle 1d — Störung mit $\tanh(x) = \tan(x) = 10^{-4}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				16.48		9.104E-04	8.116E-18
Hoppe	5	176	176	17.24	5.865E-20	9.298E-20	8.890E-18
Primitiv 0	5	174	10	14.39	8.289E-20	1.225E-18	6.938E-18

Tabelle 1e — Störung mit $\tanh(x) = \tan(x) = 10^{-2}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				16.48		4.567E-03	1.010E-17
Hoppe	4	193	193	18.84	5.672E-21	5.958E-19	5.638E-18
Primitiv 0	4	201	10	16.31	1.770E-21	4.313E-18	1.366E-17

Tabelle 1f — Störung mit $\tanh(x) = \tan(x) = 10^{-1}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				16.48		3.687E-02	5.789E-18
Hoppe	6	273	273	26.47	1.494E-20	3.002E-19	8.722E-18
Primitiv 0	7	272	10	22.41	7.502E-20	4.171E-20	1.051E-17

c) 30x30 - Matrix mit imaginären Eigenwerten

Tabelle 1g — Störung mit $\tanh(x) = \tan(x) = 10^{-4}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				52.95		1.447E-04	2.214E-17
Hoppe	4	383	383	54.49	1.076E-19	2.396E-19	1.517E-17
Primitiv 0	4	392	15	45.54	7.371E-20	4.612E-20	9.974E-18

Tabelle 1h — Störung mit $\tanh(x) = \tan(x) = 10^{-2}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				53.06		2.670E-03	2.068E-17
Hoppe	4	444	444	1:03.22	5.749E-20	1.892E-18	2.081E-17
Primitiv 0	4	451	15	52.89	5.523E-21	8.043E-19	2.645E-17

Tabelle 1i — Störung mit $\tanh(x) = \tan(x) = 10^{-1}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				53.33		1.334E-02	2.166E-17
Hoppe	6	590	590	1:23.98	5.122E-20	4.997E-19	2.224E-17
Primitiv 0	6	596	15	1:09.97	4.847E-20	1.712E-19	1.613E-17

Die Tabellen 1a - 1i machen deutlich, daß das primitive Verfahren dem normreduzierenden Verfahren bei gut getrennten imaginären Eigenwerten überlegen ist. Einerseits ist zu erkennen, daß die Rechenzeit bei zunehmender Störung der Murnaghan - Form der Matrix zunimmt. Andererseits bleibt der Zeitgewinn durch die primitive Winkelbestimmung stets erhalten und nimmt mit der Dimension der Matrix zu. Die Eigenwerte bzw. die Eigenvektoren werden stets bis auf Maschinengenauigkeit berechnet. Hierbei kann man einen nützlichen Nebeneffekt der Jacobi - ähnlichen Verfahren erkennen, der beim ansonst überlegenen QR - Verfahren nicht erzeugt wird. Die Eigenvektoren bleiben als Produkt der J - orthogonalen Matrizen der Ähnlichkeitstransformationen J - orthogonal, während der QR - Algorithmus dies nur unter zusätzlichem Rechenaufwand erreichen kann.

2. Matrizen mit rein reellen Eigenwerten

a) 10x10 - Matrix mit Eigenwerten ± 0, ± 1, ± 2, ... ,± 4

b) 20x20 - Matrix mit Eigenwerten ± 0, ± 1, ± 2, ..., ± 9

c) 30x30 - Matrix mit Eigenwerten ± 0, ± 1, ± 2, ..., ± 14

a) 10x10 - Matrix mit rellen Eigenwerten

Tabelle 2a Störung mit $\tanh(x) = \tan(x) = 10^{-4}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				2.09		7.217E-04	1.772E-18
Hoppe	3	40	40	1.81	1.368E-20	4.539E-21	8.673E-19
Primitiv 0	3	43	5	1.76	1.338E-20	2.719E-20	1.105E-18

Tabelle 2b Störung mit $\tanh(x) = \tan(x) = 10^{-2}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				2.15		5.961E-02	2.338E-18
Hoppe	4	48	48	2.19	6.756E-20	1.638E-19	6.512E-19
Primitiv 0	4	51	5	1.98	1.138E-20	8.182E-19	2.176E-18

Tabelle 2c Störung mit $\tanh(x) = \tan(x) = 10^{-1}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				2.37		5.095E-02	2.898E-18
Hoppe	4	57	57	2.58	6.825E-20	2.371E-20	1.088E-18
Primitiv 0	4	58	5	2.25	1.635E-20	1.200E-19	2.191E-18

b) 20x20 Matrix mit rein rellen Eigenwerten

Tabelle 2d Störung mit $\tanh(x) = \tan(x) = 10^{-4}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				15.98		1.214E-04	2.120E-17
Hoppe	5	176	176	17.13	7.936E-20	1.019E-20	2.168E-18
Primitiv 0	5	175	10	14.45	7.755E-20	3.854E-18	5.777E-18
0.90 Schwelle 0	3	171	10	13.34	3.336E-19	3.854E-18	5.777E-18

Tabelle 2e — Störung mit tanh(x) = tan(x) = 10^{-2}

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				16.65		2.397E-01	1.759E-17
Hoppe	5	240	240	23.18	3.184E-21	5.787E-20	3.519E-18
Primitiv 0	6	245	10	20.04	8.116E-20	5.551E-20	5.247E-18
0.90 Schwelle 0	5	245	10	19.12	8.116E-20	5.551E-20	5.247E-18
0.80 Schwelle 0	5	245	10	19.11	8.116E-20	5.551E-20	5.247E-18

Tabelle 2f — Störung mit tanh(x) = tan(x) = 10^{-1}

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				17.85		8.531E-01	2.458E-17
Hoppe	7	359	359	34.38	1.687E-20	1.477E-18	1.072E-17
Primitiv 0	13	562	10	46.69	6.226E-20	1.063E-17	1.131E-15
Primitiv 1	8	397	65	32.18	1.369E-20	3.333E-18	2.117E-17
Primitiv 2	7	345	120	28.73	4.149E-20	2.290E-18	1.244E-17
Primitiv 3	7	350	175	29.88	2.963E-20	1.185E-18	1.213E-17
0.90 Schwelle 0	8	369	11	28.95	3.708E-20	7.182E-19	1.618E-17

c) 30x30 - Matrix mit reellen Eigenwerten

Tabelle 2g — Störung mit tanh(x) = tan(x) = 10^{-4}

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				51.52		1.961E-04	1.544E-17
Hoppe	4	383	383	54.00	1.093E-19	2.406E-19	7.806E-18
Primitiv 0	4	393	15	45.42	9.470E-20	4.304E-20	1.822E-17

Tabelle 2h — Störung mit tanh(x) = tan(x) = 10^{-2}

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				51.85		3.681E-03	2.624E-17
Hoppe	4	444	444	1:02.78	4.391E-20	5.984E-19	8.674E-18
Primitiv 0	4	453	15	52.90	5.633E-21	6.374E-18	1.303E-17

Tabelle 2i — Störung mit tanh(x) = tan(x) = 10^{-1}

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				52.46		1.060E-02	2.526E-17
Hoppe	6	590	590	1:23.43	8.489E-20	2.761E-19	8.926E-18
Primitiv 0	6	596	15	1:09.54	6.876E-20	9.883E-20	1.222E-17

Für den Fall gut getrennter reeller Eigenwerte ergeben sich ebenfalls Zeitgewinne durch das primitive Verfahren. Der Zeitgewinn nimmt mit wachsender Dimension der Matrix zu. Weiterhin ist zu erkennen, daß bei zunehmender Störung der Matrix das Konvergenzverhalten der Jacobi-ähnlichen Verfahren zunehmend schlechter wird, während der QR-Algorithmus von den Veränderungen nur wenig berührt wird.

3. Matrizen mit gemischten Eigenwerten

a) 10x10-Matrix mit Eigenwerten $\pm 0,\ -1\pm 1i,\ -2\pm 2i, \ldots, -4\pm 4i$

Tabelle 3a — Störung mit $\tanh(x) = \tan(x) = 10^{-4}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				2.52		5.233E-01	5.512E-18
Hoppe	4	60	60	2.80	8.073E-21	1.943E-20	1.791E-18
Primitiv 0	30	455	5	18.84	2.125E-01	Abbruch nach 30 Zyklen	
Primitiv 1	30	455	20	18.78	2.325E-02	Abbruch nach 30 Zyklen	
Primitiv 2	21	200	35	07.85	1.315E-24	4.039E-20	3.903E-04
Primitiv 3	5	64	49	02.85	2.612E-22	3.385E-20	1.754E-12

Tabelle 3b — Störung mit $\tanh(x) = \tan(x) = 10^{-2}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				2.08		5.340E-01	2.240E-18
Hoppe	4	50	50	2.30	1.460E-24	1.802E-20	1.578E-18
Primitiv 0	30	455	5	18.84	2.500E-02	Abbruch nach 30 Zyklen	
Primitiv 1	30	276	20	10.71	1.221E-20	1.025E-19	5.268E-04
Primitiv 2	7	65	35	2.64	5.591E-21	2.010E-19	4.479E-08
Primitiv 3	4	51	48	2.31	1.990E-24	1.238E-18	3.749E-17

Tabelle 3c — Störung mit $\tanh(x) = \tan(x) = 10^{-1}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				2.25		2.657E-01	2.584E-18
Hoppe	4	59	59	2.80	2.186E-20	1.508E-20	8.267E-18
Primitiv 0	30	455	5	18.89	2.125E-01	Abbruch nach 30 Zyklen	
Primitiv 1	30	455	20	18.79	2.889E-02	Abbruch nach 30 Zyklen	
Primitiv 2	21	200	35	7.86	1.315E-24	3.980E-20	3.903E-04
Primitiv 3	4	63	49	2.80	2.373E-22	3.403E-20	9.000E-11

b) 20x20 - Matrix mit Eigenwerten ± 0, - 1± 1i, - 2± 2i, ..., - 9± 9i

Tabelle 3d — Störung mit tanh(x) = tan(x) = 10^{-4}

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				15.16		2.645E-04	1.255E-17
Hoppe	3	168	168	15.54	5.021E-19	2.916E-19	4.465E-18
Primitiv 0	30	1660	10	2:14.02	2.382E-02	Abbruch nach 30 Zyklen	
Primitiv 1	30	1235	65	1:37.50	1.215E-07	Abbruch nach 30 Zyklen	
Primitiv 2	4	182	120	15.65	5.786E-20	2.340E-19	1.627E-09

Tabelle 3e — Störung mit tanh(x) = tan(x) = 10^{-2}

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				15.87		4.745E-04	1.026E-17
Hoppe	4	195	195	18.34	4.881E-22	1.336E-18	1.084E-17
Primitiv 0	30	1660	10	2:14.35	6.377E-02	Abbruch nach 30 Zyklen	
Primitiv 1	30	1594	65	2:08.03	3.785E-03	Abbruch nach 30 Zyklen	
Primitiv 2	19	463	120	36.47	9.224E-20	1.196E-18	2.213E-06
Primitiv 3	4	201	173	18.13	7.035E-22	1.669E-18	6.244E-13

Tabelle 3f — Störung mit tanh(x) = tan(x) = 10^{-1}

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				16.37		5.410E-04	1.538E-17
Hoppe	5	246	246	23.01	7.002E-20	1.409E-19	2.785E-17
Primitiv 0	30	1660	10	2:14.57	6.962E-01	Abbruch nach 30 Zyklen	
Primitiv 1	30	1660	65	2:14.57	1.858E-01	Abbruch nach 30 Zyklen	
Primitiv 2	30	1655	120	2:14.13	2.465E-03	Abbruch nach 30 Zyklen	
Primitiv 3	23	534	175	42.52	4.700E-20	1.053E-19	4.407E-06
Primitiv 4	5	251	228	22.85	6.055E-20	2.689E-19	5.045E-12

c) 30x30 - Matrix mit Eigenwerten ± 0, - 1± 1i, - 2± 2i, ..., - 14± 14i

Tabelle 3g Störung mit $\tanh(x) = \tan(x) = 10^{-4}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				55.58		4.136E-01	3.540E-17
Hoppe	6	616	616	1:26.89	6.050E-20	5.839E-19	2.135E-17
Primitiv 0	30	3615	15	7:13.48	1.451E+00	Abbruch nach 30 Zyklen	
Primitiv 1	30	3615	135	7:14.74	5.395E-01	Abbruch nach 30 Zyklen	
Primitiv 2	30	3615	255	7:15.89	3.852E-02	Abbruch nach 30 Zyklen	
Primitiv 3	30	2508	375	5:00.49	8.304E-07	Abbruch nach 30 Zyklen	
Primitiv 4	13	689	495	1:31.07	5.958E-20	6.821E-19	8.259E-08
0.50 Schwelle 3	30	2508	375	5:00.50	8.304E-07	Abbruch nach 30 Zyklen	
0.50 Schwelle 4	13	689	495	1:31.07	5.958E-20	6.821E-19	8.259E-08
0.50 Schwelle 5	6	620	597	1:26.56	4.978E-20	2.183E-19	3.524E-15

Tabelle 3h Störung mit $\tanh(x) = \tan(x) = 10^{-2}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				51.46		4.675E-03	2.522E-17
Hoppe	4	446	446	1:03.11	4.642E-20	4.122E-20	1.736E-17
Primitiv 0	30	3615	15	7:13.20	1.211E-01	Abbruch nach 30 Zyklen	
Primitiv 1	30	3615	135	7:12.32	1.601E-02	Abbruch nach 30 Zyklen	
Primitiv 2	30	1733	255	3:23.71	2.664E-08	Abbruch nach 30 Zyklen	
Primitiv 3	4	454	374	1:01.63	4.813E-20	1.919E-18	2.599E-10
0.30 Schwelle 3	4	453	374	1:01.57	4.497E-20	7.609E-19	1.749E-10

Die Tabellen dieses Abschnittes machen deutlich, daß das primitive Verfahren bei gemischten Eigenwerten nur bedingt einsatzfähig ist. Setzt man die primitive Winkelbestimmung von Anfang an ein, so konvergiert das Verfahren überhaupt nicht. Konvergenz erhält man nur dann, wenn die Matrix durch das normreduzierende Verfahren bereits genügend vorbereitet wurde. Man kann das Verfahren nur für die letzten beiden Zyklen verwenden. Dabei stellt sich die Konvergenz nur unter dem Nachteil der schlechten Kondition der Modalmatrix ein, was sich im Fehler des Eigenvektorsystems RESID ausdrückt. Zeitgewinn stellt sich, wie in der letzten Tabelle deutlich wird, erst bei größeren Dimensionen der Matrizen ein.

4. Matrix mit defektiven Eigenwerten

a) 10x10 - Matrix mit doppelten, defektiven Eigenwerten ± 0, ± 1, ...,± 4

Tabelle 4a Störung mit $\tanh(x) = \tan(x) = 10^{-4}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				3.68		1.152E+00	1.197E-10
Hoppe	12	156	156	6.92	1.057E-19	5.000E-01	2.044E-01
Primitiv 1	4	61	20	2.74	1.290E-23	5.000E-01	1.962E-01
0.90 Schwelle 0	3	45	5	1.93	9.036E-20	5.000E-01	6.553E-01

Tabelle 4b Störung mit $\tanh(x) = \tan(x) = 10^{-2}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				4.01		1.448E-05	2.821E-11
Hoppe	12	153	153	6.92	8.193E-20	5.000E-01	2.044E-01
Primitiv 1	4	63	20	2.75	2.063E-21	5.000E-01	1.962E-01
0.90 Schwelle 0	3	50	5	2.20	1.899E-19	4.999E-01	2.534E-01

Tabelle 4c Störung mit $\tanh(x) = \tan(x) = 10^{-1}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				5.00		8.209E-03	4.743E-10
Hoppe	11	150	150	6.70	5.900E-20	5.000E-01	2.647E-01
Primitiv 1	4	62	20	2.69	1.842E-20	5.000E-01	1.969E-01
0.90 Schwelle 0	5	68	5	2.96	1.857E-20	4.999E-01	9.832E-02

a) 20x20 - Matrix mit doppelten, defektiven Eigenwerten ± 0, ± 1, ...,± 9

Tabelle 4d Störung mit $\tanh(x) = \tan(x) = 10^{-4}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				29.23		4.340E-03	3.119E-10
Hoppe	13	596	596	54.48	1.841E-19	5.000E-01	3.562E-01
Primitiv 0	12	588	10	46.80	2.019E-19	1.975E-01	1.248E-03
Primitiv 1	11	575	65	46.53	8.776E-20	4.787E-01	4.758E-03
Primitiv 2	10	496	120	40.81	4.376E-19	4.912E-01	4.714E-03
Primitiv 3	10	502	175	42.24	5.748E-19	4.999E-01	5.318E-03
0.90 Schwelle 0	10	513	11	40.76	2.250E-19	4.495E-01	1.780E-03

<u>Tabelle 4e</u> Störung mit $\tanh(x) = \tan(x) = 10^{-2}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				29.01		3.187E-03	3.954E-10
Hoppe	13	597	597	54.65	2.608E-19	5.000E-01	3.167E-01
Primitiv 2	10	496	120	40.87	3.462E-20	4.994E-01	4.718E-03
Primitiv 3	10	515	175	43.01	4.033E-19	4.988E-01	5.082E-03
0.90 Schwelle 2	10	496	120	40.91	3.462E-20	4.994E-01	4.718E-03

<u>Tabelle 4f</u> Störung mit $\tanh(x) = \tan(x) = 10^{-1}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				31.97		1.236E-03	3.340E-10
Hoppe	13	608	608	55.64	2.114E-19	5.000E-01	3.189E-01
Primitiv 2	10	495	120	40.97	3.511E-20	4.922E-01	3.370E-03
Primitiv 3	10	521	175	43.67	5.331E-19	4.977E-01	4.580E-03
0.90 Schwelle 2	10	495	120	40.97	3.511E-20	4.922E-01	3.370E-03

a) 30x30 - Matrix mit doppelten, defektiven Eigenwerten $\pm 0, \pm 1, \ldots, \pm 14$

<u>Tabelle 4g</u> Störung mit $\tanh(x) = \tan(x) = 10^{-4}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				1:27.05		1.258E-03	6.467E-10
Hoppe	14	1331	1331	3:05.10	3.296E-19	5.000E-01	3.372E-01
Primitiv 3	10	1113	375	2:30.06	7.979E-19	4.998E-01	8.286E-03
Primitiv 4	12	1191	495	2:30.39	7.813E-19	4.999E-01	4.106E-03

<u>Tabelle 4h</u> Störung mit $\tanh(x) = \tan(x) = 10^{-2}$

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				1:25.02		1.020E-03	6.950E-10
Hoppe	14	1337	1337	3:06.08	3.656E-19	5.000E-01	3.679E-01
Primitiv 3	10	1131	375	2:21.98	8.613E-19.	4.885E-01	7.062E-03

Die vorangehenden Tabellen zeigen, daß das primitive Verfahren auch bei einfach defektiven Eigenwerten, die in einen Diagonalblock liegen, bei diagonaldominanten Matrizen Zeitvorteile liefert. Da keine Basis aus Eigenvektoren existiert, geht die J - Orthogonalität der Eigenvektoren natürlich überall verloren. Einzelne Beispiele zeigen, daß zu überlegen ist, das primitive Verfahren erst nach einigen Zyklen des normreduzierenden Ver fahrens zu verwenden. Dies ergibt in manchen Fällen die größten Zeitgewinne. Dies gilt insbesondere für größere Matrixdimensionen (Tabellen 4 g - i).

5. Diagonaldominante Zufallszahlenmatrizen

Wir haben eine Reihe von Zufallszahlentests durchgeführt. Beispielhaft sollen hier 3 Ergebnisse aufgeführt werden. Die Tests haben durchweg Vorteile für die primitive Methode bei diagonaldominanten Matrizen ergeben, obwohl die Eigenwerte statistisch gestreut sind. Die Testmatrizen wurden mithilfe eines Minimalstandard - Zufallszahlengenerators erzeugt und anschließend wurde die Außerdiagonaldominanz durch Division mit großen Zahlen erzeugt. Es ergaben sich folgende repräsentativen Ergebnisse

a) 10x10 - Zufallszahlenmatrizen

Tabelle 5a — Division der Nebendiagonale mit 10^2

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				2.85		4.527E-17	8.331E-19
Hoppe	4	62	62	2.75	3.218E-20	2.752E-20	4.160E-17
Primitiv 0	4	62	20	2.53	1.234E-20	4.960E-20	4.362E-19

Tabelle 5b — Division der Nebendiagonalemit 10^2

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				2.75		2.6601E-18	4.291E-19
Hoppe	6	69	69	3.18	9.557E-20	2.4394E-19	3.282E-19
Primitiv 0	5	69	20	2.80	4.328E-20	8.1315E-20	9.468E-19

a) 20x20 - Zufallszahlenmatrizen

Tabelle 5c	Division der Nebendiagonale mit 10^2						
Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				20.49		4.258E-17	1.255E-18
Hoppe	6	238	238	22.35	1.811E-18	1.484E-19	1.838E-11
Primitiv 0	5	236	65	19.12	1.066E-20	4.305E-19	8.695E-19
0.90 Schwelle 0	5	236	65	19.06	1.066E-20	4.305E-19	8.695E-19

6. Gedämpfte Schwingungsprobleme

a) Als erstes Beispiel für eine gedämpfte Schwingung wollen wir folgendes quadratisches Eigenwertproblem betrachten.

$$(M\cdot\lambda^2 + D\cdot\lambda + K)\cdot x = 0 \tag{4.17}$$

mit

$M = I$ Einheitsmatrix im $\mathbb{R}^{m,m}$

$K = K\cdot\mathrm{Diag}(0,...0,1,...,1,0,...,0) \in \mathbb{R}^{m,m}$ und

$$D = (n+1)^4\cdot\begin{bmatrix} 6 & -4 & 1 & & & & & 1 & -4 \\ -4 & 6 & -4 & 1 & & 0 & & & 1 \\ 1 & 4 & 6 & -4 & 1 & & & & \\ & & & & \cdots & 1 & -4 & 6 & -4 & 1 \\ 1 & & & 0 & & & 1 & -4 & 6 & -4 \\ -4 & 1 & & & & & & 1 & -4 & 6 \end{bmatrix} + C\cdot I \in \mathbb{R}^{m,m}$$

Wir betrachten dieses Problem für m = 10, K = 1 und C = 5, so daß alle Matrizen positiv semidefinit sind. Die Diagonale von K sei dabei zur Hälfte mit Einsen besetzt. Wir haben dieses Problem mithilfe von Cholesky in das J-symmetrische Eigenwertproblem der Dimension $n = 2\cdot m = 20$ gemäß den Ausführungen des letzten Abschnittes transformiert und erhalten folgende Resultate

Tabelle 6a

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				15.44		6.155E-17	2.620E-17
Hoppe	12	553	553	51.57	2.416E-18	1.948E-18	4.738E-17
Primitiv 5	11	552	329	48.94	3.059E-21	3.435E-18	4.159E-17
0.90 Schwelle 5	11	552	329	48.93	3.059E-21	3.435E-18	4.159E-17
0.80 Schwelle 4	11	553	275	48.50	5.779E-20	3.096E-18	2.248E-17
0.40 Schwelle 4	11	547	277	47.95	3.769E-20	3.391E-18	1.891E-17
0.10 Schwelle 4	11	545	281	47.90	4.314E-20	3.344E-18	2.465E-17

Es zeigt sich in diesem Fall, daß es am besten ist, in den ersten Zyklen ganz auf die primitive Winkelbestimmung zu verzichten und erst nach ca. 4-5 Zyklen die primitive Winkelbestimmung einzusetzen. Dabei sollte eine Schwelle von 0.8 für die Benutzung der primitiven Parameter verwendet werden.

b) Als weiteres Beispiel betrachten wir das übergedämpfte quadratische Eigenwertproblem (4.17) mit

$M = I$; $K = 4 \cdot I$ mit I als Einheitsmatrix im $\mathbb{R}^{n,n}$ und

$$D = \begin{bmatrix} 2 & -1 & & & & & 0 \\ -1 & 2 & -1 & & 0 & & \\ & -1 & 2 & -1 & & & \\ & & & \cdots & -1 & 2 & -1 \\ & & 0 & & -1 & 2 & -1 \\ 0 & & & & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{n,n}$$

Da M und D diagonal sind, entspricht der Erzeugung des J-symmetrischen Eigenwertproblems mit Hilfe des symmetrischen QR-Algorithmus der vollständigen Lösung des Problems. Deshalb haben wir hier nur die Erzeugung mithilfe der beiden Cholesky-Varianten untersucht für 4 verschiedene Dimensionen n =10,20,30 und 40. Wir erhalten

Tabelle 7a (n = 10 durch Cholesky erzeugt in 0.27 sec)

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				2.42		4.372E-17	4.868E-18
Hoppe	8	124	124	5.55	1.407E-19	4.912E-19	4.280E-18
Primitiv 0	8	112	20	4.56	1.283E-20	6.200E-19	7.724E-18

Tabelle 7b (n = 10 durch pivotisierendem Cholesky erzeugt in 0.99 sec)

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				2.42		1.912E-17	6.159E-18
Hoppe	9	132	132	5.88	6.710E-21	9.181E-19	6.745E-18
Primitiv 0	8	118	17	4.72	1.085E-20	8.538E-19	6.724E-18

Tabelle 8a (n = 20 durch Cholesky erzeugt in 2.27 sec)

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				15.43		1.396E-16	1.980E-17
Hoppe	11	589	589	55.09	1.594E-18	1.790E-18	2.084E-17
Primitiv 0	10	524	65	42.35	5.453E-20	9.918E-19	2.923E-17

Tabelle 8b (n = 20 durch pivotisierendem Cholesky erzeugt in 2.47 sec)

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				15.49		6.336E-17	1.325E-17
Hoppe	12	553	553	51.63	2.416E-18	2.652E-18	3.782E-17
Primitiv 0	10	497	54	39.98	5.648E-20	2.141E-18	1.422E-17

Tabelle 9a (n = 30 durch Cholesky erzeugt in 3.84 sec)

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				47.41		1.138E-15	1.399E-16
Hoppe	14	1452	1452	3:27.12	6.962E-14	1.678E-18	7.170E-14
Primitiv 0	12	1328	135	2:39.89	7.042E-20	2.843E-18	5.537E-17

Tabelle 9b (n = 20 durch pivotisierendem Cholesky erzeugt in 3.92 sec)

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				47.34		2.343E-16	6.432E-17
Hoppe	16	1598	1598	3:47.07	1.702E-14	5.783E-18	3.772E-11
Primitiv 0	11	1139	100	2:15.94	5.347E-20	2.398E-18	3.869E-17

Tabelle 10a (n = 40 durch Cholesky erzeugt in 8.24 sec)

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				1:42.33		2.170E-15	1.202E-16
Hoppe	15	2676	2676	8:33.39	6.023E-15	6.560E-18	4.412E-15
Primitiv 0	12	2303	230	6:08.45	1.195E-19	3.879E-18	1.415E-16

Tabelle 10b (n = 40 durch pivotisierendem Cholesky erzeugt in 9.58 sec)

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				1:42.22		8.431E-16	3.973E-17
Hoppe	14	2555	2555	8:10.59	4.389E-16	3.826E-18	3.979E-16
Primitiv 0	13	2465	174	6:31.40	9.003E-20	5.951E-17	5.857E-16

Es zeigt sich, daß die Erzeugung durch pivotisierenden Cholesky leichte Zeitvorteile gegenüber der normalen Cholesky - Zerlegung hat. Diese werden bei größeren Dimensionen um so deutlicher. Dabei spielt der leicht höhere Aufwand bei der Erzeugung des J - symmetrischen Problems keine Rolle. Die benötigte Zeit für den Aufbau des J - symmetrischen Problems ist in Klammern jeweils angegeben. Es zeigt sich, daß der primitive Algorithmus dem rein normreduzierenden Algorithmus von Hoppe deutlich überlegen ist.

c) Als letztes Beispiel will ich ein quadratisches Eigenwertproblem, das aus der numerischen Praxis stammt, verwenden. Dabei haben wir ein quadratisches Problem der Dimension m = 156 betrachtet, das der Diskretisierung eines Rohrleitungssystems mit Stoßdämpfern entspricht. Die Massenmatrix M ist eine Diagonalmatrix, die allerdings nicht mehr konstant mit Einsen auf der Diagonale besetzt ist. Die Dämpfungsmatrix ist eine Diagonalmatrix, die nur an 12 Stellen eine von Null verschiedene Zahl $C > 0$ hat. Hierin spiegelt sich die geringe Zahl von Stoßdämpfern des Systems wieder. Die Steifigkeitsmatrix K beschreibt das eigentliche Rohrleitungssystem und ist im wesentlichen eine symmetrische Bandmatrix. Da das volle Problem nach Umformung in ein J - symmetrisches Problem die Dimension $n = 2 \cdot m = 312$ besitzt, haben wir uns auf einen kleinen Ausschnitt um eine Position von Stoßdämpfern (Positionen der Konstante C in der Dämpfungsmatrix D) zurückgezogen. Diesen Ausschnitt der Dimension 20 haben wir für verschiedene Zahlen C > 0 (unterschiedliche Dämpfung) auf das J-symmetrische Problem mit allen 3 Verfahren des letzten Abschnittes transformiert. Wir erhalten folgende Tabellen

Tabelle 11a (C = 10 durch Cholesky erzeugt in 1.37 sec)

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				15.10		1.465E-12	8.581E-21
Hoppe	6	227	227	22.46	7.0655E-17	3.069E-16	2.126E-12
Primitiv 3	7	299	166	26.03	3.8844E-20	7.862E-20	1.580E-14
Primitiv 4	7	256	212	23.45	3.0838E-20	8.293E-20	2.943E-15
Primitiv 5	7	252	227	23.90	3.0838E-20	8.572E-20	1.257E-18
0.50 Schwelle 3	7	299	166	25.98	3.8844E-20	7.862E-20	1.580E-14
0.50 Schwelle 4	7	256	212	23.45	3.0838E-20	8.293E-20	2.943E-15
0.50 Schwelle 5	7	252	227	23.89	3.0838E-20	8.572E-20	1.257E-18

Tabelle 11b (C = 10 durch pivotisierenden Cholesky erzeugt in 1.59 sec)

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				15.11		2.575E-14	8.612E-21
Hoppe	7	317	317	30.27	3.259E-17	2.064E-14	6.469E-18
Primitiv 3	9	418	173	35.48	3.600E-20	2.150E-19	1.593E-17
Primitiv 4	9	386	228	33.61	4.464E-20	2.246E-19	3.690E-16
Primitiv 5	9	356	283	31.69	6.085E-20	2.056E-19	1.599E-15
0.50 Schwelle 3	9	418	173	35.49	3.600E-20	2.150E-19	1.593E-17
0.50 Schwelle 4	9	386	228	33.61	4.464E-20	2.246E-19	3.690E-16
0.50 Schwelle 5	9	356	283	31.69	6.085E-20	2.056E-19	1.599E-15

Tabelle 11c (C = 10 durch symmetrischen QR erzeugt in 2.14 sec)

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				14.94		2.879E-16	1.064E-20
Hoppe	5	146	146	14.88	1.201E-15	9.463E-15	4.243E-16
Primitiv 3	6	175	139	15.92	4.982E-20	4.076E-20	5.457E-16
Primitiv 4	6	172	146	16.37	4.982E-20	4.076E-20	1.906E-17
Primitiv 5	5	146	146	14.89	1.201E-15	9.463E-15	4.243E-16
0.50 Schwelle 3	6	175	139	15.93	4.982E-20	4.076E-20	5.457E-16
0.50 Schwelle 4	6	172	146	16.31	4.982E-20	4.076E-20	1.906E-17
0.50 Schwelle 5	5	146	146	14.88	1.201E-15	9.463E-15	4.243E-16

Tabelle 12a (C = 100 durch Cholesky erzeugt in 2.23 sec)

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR			15.98		2.250E-18	2.331E-20	
Hoppe	8	318	318	31.20	1.755E-16	2.636E-15	5.655E-17
Primitiv 3	9	391	166	33.23	3.848E-20	2.176E-19	2.904E-15
Primitiv 4	9	383	221	33.45	3.258E-20	3.056E-19	2.558E-14
Primitiv 5	9	359	275	32.24	1.122E-19	2.054E-19	1.766E-15
Primitiv 6	8	345	313	32.18	1.034E-19	3.402E-19	2.766E-18
Primitiv 7	9	346	318	32.79	1.034E-19	3.131E-19	2.766E-18
0.50 Schwelle 3	9	391	166	33.23	3.848E-20	2.176E-19	2.904E-15
0.50 Schwelle 4	9	383	221	33.45	3.258E-20	3.056E-19	2.558E-14
0.50 Schwelle 5	9	359	275	32.30	1.122E-19	2.054E-19	1.766E-15
0.50 Schwelle 8	8	318	318	31.25	1.755E-16	2.636E-15	5.655E-17
0.50 Schwelle 7	9	346	318	32.74	1.034E-19	3.131E-19	2.766E-18

Tabelle 12b (C = 100 durch pivotisierenden Cholesky erzeugt in 2.53 sec)

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				16.31		2.335E-18	1.890E-20
Hoppe	11	484	484	46.58	1.950E-16	1.507E-15	6.873E-17
Primitiv 3	30	1292	173	1:44.19	1.841E+08	Abbruch nach 30 Zyklen	
Primitiv 4	30	1338	228	1:48.91	9.073E+05	Abbruch nach 30 Zyklen	
Primitiv 5	30	1548	283	2:07.81	2.114E+00	Abbruch nach 30 Zyklen	
Primitiv 6	12	565	338	49.66	5.789E-20	4.516E-19	2.370E-18
Primitiv 7	12	552	393	49.32	5.373E-20	4.398E-19	3.714E-13
Primitiv 8	12	530	447	48.17	1.040E-20	4.391E-19	8.149E-15
0.50 Schwelle 3	30	1338	507	1:57.05	6.052E-02	Abbruch nach 30 Zyklen	
0.50 Schwelle 4	30	1487	448	2:07.59	5.256E-03	Abbruch nach 30 Zyklen	
0.50 Schwelle 5	30	823	284	1:07.12	1.134E-04	Abbruch nach 30 Zyklen	
0.50 Schwelle 6	12	565	338	49.66	5.789E-20	4.516E-19	2.370E-18
0.50 Schwelle 7	12	552	393	49.33	5.373E-20	4.398E-19	3.714E-13
0.50 Schwelle 8	12	530	447	48.17	1.040E-20	4.391E-19	8.149E-15

Tabelle 12c (C = 100 durch symmetrischen QR erzeugt in 2.41 sec)

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				17.14		7.323E-19	2.301E-20
Hoppe	6	247	247	24.05	1.444E-17	1.566E-15	1.405E-17
Primitiv 3	9	325	169	27.68	8.000E-20	6.271E-19	1.241E-14
Primitiv 4	8	284	216	25.49	9.862E-20	9.023E-19	5.833E-14
Primitiv 5	7	272	247	25.43	7.998E-20	1.078E-18	9.128E-18
0.50 Schwelle 3	9	325	169	27.69	8.000E-20	6.271E-19	1.241E-14
0.50 Schwelle 4	8	284	216	25.43	9.862E-20	9.023E-19	5.833E-14
0.50 Schwelle 5	7	272	247	25.43	7.998E-20	1.078E-18	9.128E-18

Tabelle 13a (C = 1000 durch Cholesky erzeugt in 1.49 sec)

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				14.22		1.316E-16	4.829E-19
Hoppe	7	257	257	25.71	3.017E-17	1.143E-14	2.275E-17
Primitiv 3	8	319	166	27.52	4.953E-20	4.786E-19	2.633E-18
Primitiv 4	6	295	220	26.37	2.659E-20	4.277E-19	2.593E-14
Primitiv 5	7	286	254	26.75	5.518E-20	4.092E-19	4.943E-16
Primitiv 6	8	288	257	27.46	7.206E-20	4.092E-19	4.943E-16
0.50 Schwelle 3	8	319	166	27.57	4.953E-20	4.786E-19	2.633E-18
0.50 Schwelle 4	6	295	220	26.42	2.659E-20	4.277E-19	2.593E-14
0.50 Schwelle 5	7	286	254	26.80	5.518E-20	4.092E-19	4.943E-16

Tabelle 13b (C = 1000 durch pivotisierendem Cholesky erzeugt in 1.58 sec)

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				14.23		4.239E-17	5.057E-19
Hoppe	10	457	457	43.88	8.679E-17	1.479E-14	2.945E-17
Primitiv 5	30	1575	283	2:09.68	1.513E-01	Abbruch nach 30 Zyklen	
Primitiv 6	13	610	338	52.72	2.484E-20	3.898E-19	8.581E-13
Primitiv 7	13	556	392	49.21	7.830E-20	4.848E-19	6.808E-14
Primitiv 8	12	498	437	45.59	3.997E-20	4.959E-19	1.639E-16
0.50 Schwelle 4	30	1507	413	2:07.76	5.884E-03	Abbruch nach 30 Zyklen	
0.50 Schwelle 5	24	1106	300	1:32.00	1.111E-19	9.495E-19	7.796E-16
0.50 Schwelle 6	13	610	338	52.73	2.484E-20	3.898E-19	8.581E-13
0.50 Schwelle 7	13	556	392	49.21	7.830E-20	4.848E-19	6.808E-14
0.50 Schwelle 8	12	498	437	45.59	3.997E-20	4.959E-19	1.639E-16

Tabelle 13c (C = 1000 durch symmetrischen QR erzeugt in 2.14 sec)

Algorithmus	ZYKL	TRANS	NRED	ZEIT	S(A)	JORTH	RESID
QR				14.50		1.319E-15	8.852E-19
Hoppe	9	342	342	33.89	7.209E-16	3.800E-14	3.009E-16
Primitiv 4	30	1400	229	1:52.43	1.187E-02	Abbruch nach 30 Zyklen	
Primitiv 5	11	444	281	38.72	4.386E-20	1.778E-19	4.227E-15
0.50 Schwelle 4	30	1148	254	1:32.99	1.372E-05	Abbruch nach 30 Zyklen	
0.50 Schwelle 5	11	444	281	38.72	4.386E-20	1.778E-19	4.227E-15

Die ersten drei angeführten Tabellen befassen sich mit dem nur gering gedämpften System. Während der QR-Algorithmus unabhängig von der Art der Erzeugung stets ähnliche Rechenzeiten benötigt, sind die Jacobi-ähnlichen Routinen stark von dem gewählten Erzeugungsverfahren abhängig. Es zeigt sich, daß die Erzeugung mit Hilfe der pivotisierenden Cholesky-Zerlegung die schlechtesten Ergebnisse liefert. Die aufwendigste Methode der Diagonalisierung mittels des symmetrischen QR-Algorithmus liefert das schnellste Verfahren. Leider ist in diesem konkreten Rechenbeispiel durch den Einsatz der primitiven Winkelbestimmungen kein Zeitgewinn zu erzielen, selbst wenn man die Schwellentaktik anwendet und erst in den letzten Zyklen auf die primitive Winkelwahl umschaltet. Weitere in den Tabellen nicht aufgeführte Tests haben gezeigt, daß sich selbst bei stärkster Herabsetzung der Einschaltschwelle THMAX auf 10^{-2} keine Verbesserung ergibt. Dieses Problem bleibt bestehen, wenn man zu stärkerer Dämpfung mit C = 100 bzw C = 1000 übergeht. In diesen Fällen besitzt die Matrix bereits 4 bzw. 8 abgedämpfte, reelle Eigenwerte. Die Erzeugung der J-symmetrischen Matrix mit pivotisierendem Cholesky führt hier sogar zum Abbruch der Iteration nach 30 erfolglosen Zyklen. Als Resultat kann man erkennen, daß der geringfügig höhere Rechenaufwand des symmetrischen QR-Verfahrens bei der Berechnung der J-symmetrischen Matrix in Kauf genommen werden sollte, falls die Eigenwerte mit Jacobi-ähnlichen Algorithmen berechnet werden. Die Zeitunterschiede sind für den Gesamtrechenaufwand ohne Bedeutung, da der symmetrische QR-Algorithmus nur auf Matrizen der halben Dimension ausgeführt werden braucht.

5.Literaturverzeichnis

[1] *Durand, E.* Solutions numériques des équations algébriques
Marson, Paris 1972

[2] *Erdmann, Andreas* On The Convergence of the Eberlein Method for Diagonalizing arbitrary complex Matrices
Sindelfingen, Juni 1987

[3] *Erdmann, Andreas* Über Jacobi - Ähnliche Verfahren zur Lösung des Eigenwertproblems nicht - normaler Matrizen
Dissertation, Fernuniversität Hagen, 1984

[4] *Forsythe, G.E.*, *Henrici, P.* The Cyclic Jacobi Method for Computing Principal Values of a Complex Matrix
Trans. Amer. Math. Soc. 94, 1 - 23 (1960)

[5] *Francis, J.G.F.* The QR Transformation, Parts I and II
Comp. J. 4., 265 - 271, 332 - 345 (1961/62)

[6] *Friedman, Bernard* Eigenvalues of Composite Matrices
Math. Proc. Camb. Phil. Soc. 57 (1961)

[7] *Golub, G.H.*, *Van Loan, C.F.* Matrix Compuations
North Oxford Academic Publishing,Oxford 1983

[8] *Henrici, P.* Bounds for iterates, inverses, spectral variation and fields of values of non - normal matrices
Numer. Math. 4, 24 - 41 (1962)

[9] *Henrici, P.*, *Zimmermann, K.* An Estimate for the Norms of Certain Cyclic Jacobi Operators
Linear Algebra and Its Appl. 1, 489 - 501 (1968)

[10] *Hoppe, P.*, *Veselić, K.* A Jacobi - like method for the symmetric indefinite problem $Sx = \lambda Tx$ with complex eigenvalues
Glasnik Mat. 19 (39), 383 - 390 (1983)

[11] *Hoppe, P.* Jacobi - Ähnliche Blockverfahren zur numerischen Behandlung des J - symmetrischen Algebraischen Eigenwertproblems
Dissertation Fernuniversität Hagen 1984

[12] *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)

[13] *Meyer, Dirk*; *Veselić, Kresimir* On some new inclusion Theorems for the eigenvalues of partitioned matrices
Numer. Math. 34, 431 - 437 (1980)

[14] *Meyer, Dirk* Über das Henrici - Maß und seine Anwendungen auf Gerschgorin - Sätze
Diplomarbeit Universität Dortmund 1979

[15] *Meyer, Dirk* Abschätzung des Konditionsmaßes einer Matrix, die eine beliebige n x n - Matrix auf Blockdiagonalgestalt überführt
Unveröffentliches Manuskript

[16] *Mirsky, L.* On the minimization of matrix norms
Amer. Math. Monthly 65, 106 - 107 (1958)

[17] *Ortega, J.M.*; *Rheinbold, W.C.* Iterative Solutions of nonlinear equations in several variables
Academic Press,N. Y. - San Francisco - London 1970

[18] *Rath, Wolfgang* Fast Givens Rotations for Orthogonal Similarity Transformations
Numerische Mathematik 40, 47 - 56 (1982)

[19] *Ruhe, A.* On the quadratic convergence of a generalization of the Jacobi method to arbitrary matrices
BIT 8, 210 - 231 (1968)

[20]	*Veselić, Kresimir*	Some Convergent Jacobi - like Procedures for Diagonalizing J - Symmetric Matrices Numerische Mathematik 27, 67 - 75 (1975)
[21]	*Veselić, Kresimir*	On a Class of Jacobi - Like Procedures for Diagonalizing Arbitrary Real Matrices Numerische Mathematik 33, 157 - 172 (1979)
[22]	*Veselić, Kresimir* *Wenzel, H.J.*	A Quadratically Convergent Jacobi - Like Method for Real Matrices With Complex Eigenvalues Numerische Mathematik 33, 425 - 435 (1979)
[23]	*Veselić, Kresimir*	A Global Jacobi method for a symmetric indefinite problem $Sx = \lambda Tx$ Comp - Meths.Appl.Mech.Engrg. 38, 273 - 290 (1983)
[24]	*Veselić, Kresimir*	An Eigenreduction Algorithm for Definite Matrix Pairs and its Applications to overdamped Linear Systems November 1987
[25]	*Veselić, Kresimir* *Zakrajsek, E.*	Unveröffentlichtes Manuskript
[26]	*Veselić, Kresimir*	On a new class of elementary matrices Numer. Math. 33, 173 - 180, (1979)
[27]	*Wenzel, H.J.*	Über quadratisch konvergente Jacobi - ähnliche Blockverfahren für beliebige reelle Matrizen mit komplexen Eigenwerten Dissertation, Fernuniversität Hagen, 1983
[28]	*Wilkinson, J.H.*	The Algebraic Eigenvalue Problem Clarendon Press Oxford 1965

[29] *Wilkinson, J.H.* *Reinsch, C.* Handbook of automatic Computation II
Linear Algebra
Springer Verlag, Berlin - Heidelberg - New York 1971

[30] *Zimmermann, K.* Zur Konvergenz eines Jacobiverfahrens für gewöhnliche und verallgemeinerte Eigenwertprobleme
Dissertation ETH Zürich 1969

[31] *Zurmühl, R.* Matrizen und ihre technische Anwendungen
Springer Verlag, Berlin - Göttingen - Heidelberg, 1964

Lebenslauf

29.01.1960	Geboren als Sohn des kaufmännischen Angestellten Wolfgang Zacharias und seiner Ehefrau, der Schneiderin Edith Zacharias, geb. Wolff
01.04.1966 - 10.07.1969	Besuch der Städtischen Volksschule Hansaschule in Dortmund
01.09.1969 - 08.06.1977	Besuch des Bert - Brecht - Gymnasiums in Dortmund
08.06.1977	Erlangung der allgemeinen Hochschulreife
13.09.1977 - 06.10.1983	Studium der Mathematik mit Nebenfach Physik an der Universität Dortmund
06.10.1983	Abschluß des Diplomstudienganges mit der Diplomhauptprüfung
10.10.1983 - 01.04.1988	Wissenschaftlicher Angestellter der Abteilung Mathematik der Universität Dortmund
seit 01.04.1988	Systemanalytiker in der zentralen Datenverarbeitung der Hoesch AG Dortmund