Verifikation beim quadratischen und polynomialen Eigenwertproblem

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik der
Universität Karlsruhe (TH)
genehmigte

DISSERTATION

von

Dipl.-Math. Friederike Decker
aus Ribnitz

Tag der mündlichen Prüfung:	12. Februar 2009
Referent:	Prof. Dr. G. Mayer, Universität Rostock
Korreferent:	Prof. Dr. G. Alefeld, Universität Karlsruhe (TH)

Danksagung

An dieser Stelle möchte ich mich bei allen herzlich bedanken, die mich bei der Entstehung der vorliegenden Arbeit unterstützt haben.

Mein Dank gilt besonders Herrn Prof. Dr. G. Mayer, der das Thema dieser Dissertation anregte und meine Arbeit betreute. Herrn Prof. Dr. G. Alefeld danke ich dafür, dass er mir die Möglichkeit gab, im Rahmen meiner Tätigkeit an seinem Lehrstuhl zu promovieren, und dass er das Korreferat übernahm.

Bei meinen Kollegen am Institut für Angewandte und Numerische Mathematik bedanke ich mich für die angenehme und freundschaftliche Arbeitsatmosphäre.

Schließlich danke ich meinen Eltern, meinen Geschwistern und meinem Verlobten Wolfgang Voos dafür, dass sie immer für mich da waren und sind.

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Inhalt der Arbeit

Diese Dissertation befasst sich vor allem mit dem quadratischen Eigenwertproblem (QEP), das folgendermaßen definiert ist:

Für das vom Parameter $\lambda \in \mathbb{C}$ abhängige quadratische Matrixpolynom

$$P(\lambda) = A_2\lambda^2 + A_1\lambda + A_0 \tag{1.1}$$

mit $A_2, A_1, A_0 \in \mathbb{C}^{n\times n}$ und $\det A_2 \neq 0$ werden komplexe Zahlen λ^* (Eigenwerte) und Vektoren $x^* \in \mathbb{C}^n$ mit $x^* \neq 0$ (Eigenvektoren) gesucht, so dass

$$P(\lambda^*)x^* = 0$$

gilt. Ein Eigenwert λ^* heißt einfach, wenn λ^* einfache Nullstelle des skalaren Polynoms $\det(P(\lambda))$ ist. In diesem Fall nennen wir auch ein Eigenpaar (x^*, λ^*) einfach.

Ein Anwendungsgebiet des QEPs stellen Systeme linearer Differentialgleichungen 2. Ordnung

$$A_2\,\ddot{q}(t) + A_1\,\dot{q}(t) + A_0\,q(t) = f(t)$$

dar, wobei $f : U \subseteq \mathbb{R} \to \mathbb{C}^n$ gegeben und $q : V \subseteq \mathbb{R} \to \mathbb{C}^n$ gesucht ist. Für ein Eigenpaar (x^*, λ^*) von (1.1) ist dann $q(t) = e^{\lambda^* t}x^*$ eine Lösung des zugehörigen homogenen Differentialgleichungssystems (siehe Kapitel 2.3). Zwei wichtige Gebiete, in denen lineare Differentialgleichungen 2. Ordnung auftauchen, sind die mechanische und elektrische Schwingung. Als Beispiel wird in Kapitel 9 die Oszillation bei Brücken näher beleuchtet.

Weitere Anwendungen, deren QEPs nicht unbedingt aus Differentialgleichungssystemen resultieren, findet man in der Akkustik, in der Strömungslehre, in der Optimierung; siehe [27], Kapitel 2.

Häufig treten zwei Spezialfälle des QEP auf. Zum einen besitzt das quadratische Matrixpolynom (1.1) meist reelle Matrizenkoeffizienten A_2, A_1 und A_0, die zudem oft symmetrisch, tridiagonal oder positiv definit sind. Diese Problemstellung wird als reelles QEP bezeichnet. Zum anderen benötigt man oft nur die reellen Eigenpaare eines reellen QEP.

Es gibt numerische Verfahren zur näherungsweisen Lösung des QEP. Einige davon werden in Kapitel 8.1 beschrieben. Man kann z.B. das QEP auf ein verallgemeinertes Eigenwertproblem

doppelter Dimension $Ax = \lambda Bx$ (mit $A, B \in \mathbb{C}^{2n\times 2n}$) zurückführen und dieses mit bekannten Methoden lösen. Es gibt aber auch Verfahren, die das QEP direkt angehen.

In dieser Dissertation werden Verfahren zur Verifikation (d.h. Nachweis und/oder Angabe von Fehlerschranken) von einfachen Eigenwerten und zugehörigen Eigenvektoren des QEP entwickelt, die geeignete Eigenpaarnäherungen verwenden. Außerdem können verbesserte Schranken für die Eigenpaare (bei komplexen Eigenpaaren für Real– und Imaginärteil separat) iterativ berechnet werden. Mit ihrer Hilfe können dann in der Regel die meisten führenden Stellen in einem gegebenen Gleitpunktsystem garantiert werden. Es wird dazu die reelle Intervallrechnung genutzt, auf welche in Kapitel 3 näher eingegangen wird.

Als Ausgangspunkt zur Entwicklung dieser Verfahren wurde das im Übersichtsartikel von G. Mayer [18] beschriebene Verifikationsverfahren für reelle einfache Eigenpaare des einfachen Eigenwertproblems $Ax = \lambda x$ (mit $A \in \mathbb{R}^{n\times n}$) verwendet. Dieser Übersichtsartikel greift unter anderem auf Arbeiten von R. Krawczyk [14], S. Rump [22], [24] und G. Alefeld [2] zurück (siehe auch [19]). Außerdem wurde bereits in [3] ein Verifikationsverfahren für reelle einfache Eigenwerte des reellen verallgemeinerten Eigenwertproblems vorgestellt.

Zur Herleitung des Einschließungsverfahrens für reelle Eigenpaare beim reellen QEP wird in Kapitel 5 die Funktion

$$f(x,\lambda) = \begin{pmatrix} P(\lambda)x \\ e_s^T x - 1 \end{pmatrix}$$

($e_s \in \mathbb{R}^n$ ist der s–te Einheitsvektor, s geeignet) betrachtet. Jede Nullstelle (x^*, λ^*) von f ist Eigenpaar von $P(\lambda)$ mit durch $(x^*)_s = 1$ normiertem Eigenvektor. Für die Näherung $(\tilde{x}, \tilde{\lambda})$ eines solchen Eigenpaares (x^*, λ^*) kann man nun die Taylor–Entwicklung von f zur Entwicklungsstelle $(\tilde{x}, \tilde{\lambda})$ berechnen. Einige weitere Umformungen von f, u.a. die Multiplikation mit einer geeigneten Matrix $C \in \mathbb{R}^{(n+1)\times(n+1)}$, liefern dann ein Fixpunktproblem für eine Funktion $g : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$, das den Approximationsfehler $(x^* - \tilde{x}, \lambda^* - \tilde{\lambda})$ als Fixpunkt besitzt. Es wird gezeigt, dass für alle reellen $(n+1)$–komponentigen Intervallvektoren $[y]$

$$g([y]) \subseteq h([y]) \tag{1.2}$$

gilt, wobei die Funktion h ein kubisches System ist (siehe Kapitel 4). In Kapitel 4 wird bewiesen, dass unter einigen Voraussetzungen für gewisse $\beta > 0$ der Intervallvektor $[y]^{(0)} := [-\beta, \beta]e$ (mit $e = (1, \ldots, 1)^T \in \mathbb{R}^{n+1}$) die Beziehung

$$h([y]^{(0)}) \subseteq \operatorname{int}([y]^{(0)})$$

erfüllt. Zusammen mit (1.2) liefert dies die Existenz eines Fixpunkts von g in $[y]^{(0)}$ sowie Schranken, die dann mit Intervall–Fixpunktiteration bzgl. g und dem Startvektor $[y]^{(0)}$ verbessert werden können. Dies liefert uns eine neue Einschließung des Eigenpaares (x^*, λ^*).

In Kapitel 2 wird gezeigt, dass die Jacobimatrix $f'(x, \lambda)$ von f genau dann invertierbar ist, wenn $(x, \lambda) = (x^*, \lambda^*)$ ein einfaches Eigenpaar von $P(\lambda)$ ist. Aus Stetigkeitsgründen existiert dann für eine genügend gute Näherung $(\tilde{x}, \tilde{\lambda})$ eines einfachen Eigenpaares die Inverse von $f'(\tilde{x}, \tilde{\lambda})$. Somit kann man z.B. $C = f'(\tilde{x}, \tilde{\lambda})^{-1}$ wählen und damit die Voraussetzungen des eben beschriebenen Einschließungsverfahrens erfüllen.

Bei der Herleitung der Einschließungsverfahren für komplexe Eigenpaare des reellen QEP (Kapitel 6) bzw. des komplexen QEP (Kapitel 7) wird ähnlich vorgegangen. Jedoch liegt diesen Verfahren eine Funktion $f : \mathbb{R}^{2n+2} \to \mathbb{R}^{2n+2}$ zugrunde, deren Nullstellen die Gestalt $(\operatorname{Re}(x^*), \operatorname{Im}(x^*), \operatorname{Re}(\lambda^*), \operatorname{Im}(\lambda^*))$ besitzen, wobei (x^*, λ^*) ein Eigenpaar von $P(\lambda)$ mit

$(x^*)_s = 1$ ist.

Die verschiedenen Verfahren wurden in MATLAB mit Hilfe der Intervall-Toolbox INTLAB (Version 5.3) von S. Rump implementiert. In Kapitel 8.2 werden die Ergebnisse einiger Beispielrechnungen präsentiert.

Das QEP ist ein Spezialfall des polynomialen Eigenwertproblems (PEP) $Q(\lambda)x = \lambda x$, wobei

$$Q(\lambda) = A_l\lambda^l + A_{l-1}\lambda^{l-1} + \ldots + A_1\lambda + A_0$$

ein $n \times n$-Matrixpolynom der Ordnung l mit $\det A_l \neq 0$ ist. Auf die Herleitung von Verifikationsverfahren für reelle einfache Eigenpaare des reellen PEP wird in Kapitel 10 eingegangen. Für den kubischen Fall wird eine Beispielrechnung präsentiert.

Matrixpolynome und das PEP wurden umfassend in [10], [16] und [17] untersucht. Eine Zusammenfassung der Theorie zu den Matrixpolynomen, die in dieser Dissertation benötigt wird, ist in Kapitel 2 zu finden. Es gibt dort auch einige eigene Resultate der Verfasserin.

1.2 Notationen

Es gilt die Bezeichnung $e = (1, \ldots, 1)^T \in \mathbb{R}^n$. Des Weiteren ist e_s der s–te Einheitsvektor im $\mathbb{R}^n$. Die $n \times n$–Einheitsmatrix heißt I bzw. I_n. $(x)_s$ ist die s–te Komponente des Vektors $x \in \mathbb{C}^n$. Für die Matrix $A = (a_{ij}) \in \mathbb{C}^{n\times n}$ bezeichnet neben a_{ij} auch $(A)_{ij}$ den (i,j)-ten Eintrag. Außerdem ist $A_{*,j}$ die j–te Spalte von A.

$(d_{ij}) = \mathrm{diag}(d_{11}, \ldots, d_{nn}) \in \mathbb{C}^{n\times n}$ ist die Diagonalmatrix mit den Diagonaleinträgen d_{ii}, $i = 1, \ldots, n$, und $(d_{ij}) = \mathrm{tridiag}(d_{-1}, d_0, d_1) \in \mathbb{C}^{n\times n}$ die Tridiagonalmatrix mit den Einträgen $d_{i,i-1} = d_{-1},\ d_{ii} = d_0,\ d_{i,i+1} = d_1,\ i = 1, \ldots, n$ (wobei $d_{1,0}$ und $d_{n,n+1}$ nicht auftreten).

Die lineare Hülle der Vektoren $v_1, \ldots, v_n$ wird mit $\mathrm{span}\{v_1, \ldots, v_n\}$ bezeichnet.

Für $\lambda = \mathrm{Re}\,(\lambda) + i\,\mathrm{Im}\,(\lambda) \in \mathbb{C}$ ist $\overline{\lambda} = \mathrm{Re}\,(\lambda) - i\,\mathrm{Im}\,(\lambda)$ die zu λ konjugiert komplexe Zahl (analog für komplexe Vektoren).

Kapitel 2

Theorie zu Matrixpolynomen

Diesem Kapitel liegen die Bücher von Gohberg, Lancaster und Rodman [10], Lancaster [16] sowie Lancanster und Tismenetsky [17] zu Grunde. Da einige theoretische Ergebnisse zu den Matrixpolynomen später in dieser Arbeit benötigt werden, aber dem Leser vielleicht nicht so umfassend bekannt sind, werden sie hier ausführlich dargelegt. Kapitel 2.6 enthält ein eigenes Resultat.

2.1 Definitionen, Äquivalenzrelation, Linearisierung

Es sei

$$P(\lambda) = \sum_{i=0}^{l} A_i \lambda^i \,, \quad \lambda \in \mathbb{C},$$

ein $n \times n$–Matrixpolynom vom Grad l mit Koeffizientenmatrizen $A_l, \ldots, A_0 \in \mathbb{C}^{n \times n}$. Das Matrixpolynom $P(\lambda)$ wird manchmal auch λ–Matrix genannt.

Wir setzen weiter voraus, dass $\det A_l \neq 0$ gilt. $P(\lambda)$ heißt dann regulär. Außerdem sei $l \geq 1$.

Das polynomiale Eigenwertproblem lautet folgendermaßen:

Definition 1

Wenn für eine komplexe Zahl λ^* und einen komplexen n–komponentigen Vektor $x^* \neq 0$

$$P(\lambda^*)x^* = (A_l \lambda^{*l} + A_{l-1}\lambda^{*l-1} + \ldots + A_1\lambda^* + A_0)x^* = 0$$

gilt, dann heißt λ^* Eigenwert und x^* Eigenvektor von $P(\lambda)$; (x^*, λ^*) heißt Eigenpaar von $P(\lambda)$.

Lemma 1

λ^* ist genau dann ein Eigenwert von $P(\lambda)$, wenn $\det(P(\lambda^*)) = 0$ gilt.

Beweis:

λ^* ist ein Eigenwert von $P(\lambda)$ genau dann, wenn das homogene lineare Gleichungssystem

$P(\lambda^*)x = 0$ eine Lösung $x = x^* \neq 0$ besitzt. Dies ist genau dann der Fall, wenn die Determinante $\det(P(\lambda^*))$ gleich Null ist. □

Definition 2

Der Eigenwert λ^* von $P(\lambda)$ besitzt die algebraische Vielfachheit s, wenn λ^* s–fache Nullstelle des Polynoms $\det(P(\lambda))$ *ist.*
λ^* besitzt die geometrische Vielfachheit t, wenn $\dim(\mathrm{Ker}(P(\lambda^*))) = t$ *ist.*
λ^* heißt einfach, wenn λ^* die algebraische (und geometrische) Vielfachheit 1 *besitzt.*

Die Determinante $\det(P(\lambda))$ ist ein Polynom vom Grad ln mit führendem Koeffizienten $\det A_l$. (Dies kann man per vollständiger Induktion bzgl. n mit dem Laplaceschen Entwicklungssatz für Determinanten zeigen.) Deswegen besitzt ein reguläres $P(\lambda)$ höchstens ln verschiedene Eigenwerte. Eigenvektoren zu verschiedenen Eigenwerten können linear abhängig sein, wie man im folgenden Beispiel sieht.

Beispiel 1

Als Beispiel wird das 2×2–Matrixpolynom vom Grad 2

$$P(\lambda) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \lambda^2 + \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix} \lambda + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \lambda^2 & 0 \\ \lambda + 1 & \lambda(\lambda - 1) \end{pmatrix}$$

betrachtet. Hier ist $A_2 = I_2$ also $\det A_2 = 1$. Wegen $\det P(\lambda) = \lambda^3(\lambda - 1)$ besitzt $P(\lambda)$ nach Lemma 1 genau die Eigenwerte $\lambda_0 = 0$ und $\lambda_1 = 1$. $\lambda_0 = 0$ besitzt die algebraische Vielfachheit 3 *und $\lambda_1 = 1$ die algebraische Vielfachheit* 1*. Betrachtet man die Kerne (Nullräume) von*

$$P(0) = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \quad \textit{und} \quad P(1) = \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix},$$

so erkennt man, dass jeder Vektor der Form $\begin{pmatrix} 0 \\ \alpha \end{pmatrix}$, $\alpha \neq 0$, Eigenvektor von $P(\lambda)$ zu jedem der zwei Eigenwerte ist und dass keine weiteren Eigenvektoren existieren. Damit sind beide Eigenwerte geometrisch einfach.

Folgender Satz macht eine Aussage über die Eigenpaare reeller Matrixpolynome.

Satz 1

$P(\lambda) = \sum_{i=0}^{l} A_i \lambda^i$ sei ein reelles $n \times n$–Matrixpolynom, d.h. $A_0, \ldots, A_l \in \mathbb{R}^{n \times n}$.
Dann ist (x^*, λ^*) genau dann ein Eigenpaar von $P(\lambda)$, wenn $(\overline{x^*}, \overline{\lambda^*})$ ein Eigenpaar von $P(\lambda)$ ist.

Beweis:

Wenn die Koeffizientenmatrizen $A_0, \ldots, A_l$ reell sind, dann ist auch $\det(P(\lambda))$ ein Polynom mit reellen Koeffizienten, d.h. die Eigenwerte von $P(\lambda)$ sind dann entweder reell oder treten in konjugiert komplexen Paaren auf.

Für $M = (m_{ij}) \in \mathbb{C}^{n\times m}$ verwenden wir nun die Bezeichnung $\overline{M} := (\overline{m_{ij}})$ für die zugehörige konjugiert komplexe Matrix.

Es ist wegen $\overline{A_i} = A_i \in \mathbb{R}^{n\times n}$, $i = 0, \ldots, l$,

$$P(\overline{\lambda}) = \sum_{i=0}^{l} A_i \overline{\lambda}^i = \sum_{i=0}^{l} A_i \overline{\lambda^i} = \sum_{i=0}^{l} \overline{A_i}\ \overline{\lambda^i} = \sum_{i=0}^{l} \overline{A_i \lambda^i} = \overline{P(\lambda)}\,.$$

Die Behauptung folgt dann aus $P(\overline{\lambda})\overline{x} = \overline{P(\lambda)}\overline{x} = \overline{P(\lambda)x}$. □

Das $n \times n$–Matrixpolynom $E(\lambda)$ heißt invertierbar, wenn ein Matrixpolynom $F(\lambda)$ mit $E(\lambda)F(\lambda) = F(\lambda)E(\lambda) = I$ existiert.

Lemma 2

$E(\lambda)$ ist genau dann invertierbar, wenn $\det E(\lambda) = c$ *für eine von λ unabhängige Konstante $c \neq 0$.*

Beweis:

Wenn $\det E(\lambda) = c \neq 0$ gilt, sind die Einträge der zu $E(\lambda)$ inversen Matrix Minoren $(n-1)$–ter Ordnung (siehe Definition vor Satz 6) von $E(\lambda)$ multipliziert mit $\frac{1}{c} \neq 0$, also Polynome in λ. Damit ist $E(\lambda)^{-1}$ ein Matrixpolynom.

Wenn $E(\lambda)$ invertierbar ist, dann existiert ein Matrixpolynom $F(\lambda)$ mit $E(\lambda)F(\lambda) = I$, also gilt $\det E(\lambda) \det F(\lambda) = 1$. Dies ist nur möglich, wenn $\det E(\lambda)$ und $\det F(\lambda)$ jeweils eine Konstante ungleich Null ist. □

Definition 3

Die $m \times m$–Matrixpolynome $M_1(\lambda)$ und $M_2(\lambda)$ (die nicht regulär sein müssen) heißen äquivalent, wenn $m \times m$–Matrixpolynome $E(\lambda)$ und $F(\lambda)$ mit konstanten Determinanten ungleich Null existieren, so dass

$$M_1(\lambda) = E(\lambda)M_2(\lambda)F(\lambda)$$

gilt. Man schreibt dann $M_1(\lambda) \sim M_2(\lambda)$.

Eine $ln \times ln$–Matrix A heißt Linearisierung von $P(\lambda)$, falls

$$\lambda I_{ln} - A \sim \begin{pmatrix} P(\lambda) & O \\ O & I_{(l-1)n} \end{pmatrix}. \tag{2.1}$$

Die $ln \times ln$–Matrix

$$C_P := \begin{pmatrix} O & I_n & O & \ldots & O \\ O & O & I_n & \ldots & O \\ \vdots & \vdots & \vdots & & \vdots \\ O & O & O & \ldots & I_n \\ -A_l^{-1}A_0 & -A_l^{-1}A_1 & -A_l^{-1}A_2 & \ldots & -A_l^{-1}A_{l-1} \end{pmatrix}$$

heißt Begleitmatrix von $P(\lambda)$.

Die Äquivalenz von Matrixpolynomen ist eine Äquivalenzrelation (Reflexivität und Transitivität sind trivial, die Symmetrie erhält man mit Lemma 2).

Satz 2

C_P ist eine Linearisierung von $P(\lambda)$.

Beweis:

Die $ln \times ln$–Matrixpolynome $E(\lambda)$ und $F(\lambda)$ seien definiert durch

$$F(\lambda) := \begin{pmatrix} I_n & O & \dots & O & O \\ -\lambda I_n & I_n & \dots & O & O \\ \vdots & \vdots & & \vdots & \vdots \\ O & O & \dots & I_n & O \\ O & O & \dots & -\lambda I_n & I_n \end{pmatrix}, E(\lambda) := \begin{pmatrix} B_{l-1}(\lambda) & B_{l-2}(\lambda) & \dots & \dots & B_0(\lambda) \\ -I_n & O & \dots & \dots & O \\ O & -I_n & \dots & \dots & O \\ \vdots & & \ddots & & \vdots \\ O & \dots & & -I_n & O \end{pmatrix}$$

mit $B_0(\lambda) := A_l\,, \quad B_{r+1}(\lambda) := \lambda B_r(\lambda) + A_{l-r-1}$ für $r = 0, 1, \dots, l-2$. Es gilt dann $\lambda B_{l-1}(\lambda) + A_0 = P(\lambda)$. Es ist leicht zu sehen, dass $\det E(\lambda) \equiv \pm \det A_l \,(\neq 0)$ (Entwicklung von $E(\lambda)$ nach der letzten Block–Spalte) und $\det F(\lambda) \equiv 1$ gilt. $F(\lambda)$ ist damit nach Lemma 2 invertierbar. Es gilt $\det(F(\lambda)^{-1}) \equiv 1$.

Direkte Multiplikation zeigt, dass

$$E(\lambda)(\lambda I_{ln} - C_P) = \begin{pmatrix} P(\lambda) & O & \dots & O & O \\ -\lambda I_n & I_n & \dots & O & O \\ \vdots & \vdots & & \vdots & \vdots \\ O & O & \dots & I_n & O \\ O & O & \dots & -\lambda I_n & I_n \end{pmatrix} = \begin{pmatrix} P(\lambda) & O \\ O & I_{(l-1)n} \end{pmatrix} F(\lambda)$$

gilt, womit nach Multiplikation mit $F(\lambda)^{-1}$ von rechts der Satz bewiesen ist. □

Satz 3

λ^* ist Eigenwert von $P(\lambda)$ genau dann, wenn λ^* Eigenwert einer Linearisierung $A \in \mathbb{C}^{ln \times ln}$ von $P(\lambda)$ ist.

Beweis:

Für eine Linearisierung A von $P(\lambda)$ gilt nach Definition

$$\begin{pmatrix} P(\lambda) & O \\ O & I_{(l-1)n} \end{pmatrix} = E(\lambda)(\lambda I_{ln} - A)F(\lambda)$$

und damit

$$\det(P(\lambda)) = \det(E(\lambda)) \det(\lambda I_{ln} - A) \det(F(\lambda)) = c \cdot \det(\lambda I_{ln} - A)$$

für eine Konstante $c \neq 0$. Also haben $\det(P(\lambda))$ und $\det(\lambda I_{ln} - A)$ dieselben Nullstellen, womit der Satz bewiesen ist. □

Es ist anzumerken, dass die Eigenvektoren zum Eigenvektor λ^* von $P(\lambda)$ aus dem $\mathbb{C}^n$ stammen, die Eigenvektoren zu λ^* bzgl. einer Linearisierung $A \in \mathbb{C}^{ln \times ln}$ von $P(\lambda)$ dagegen Elemente des $\mathbb{C}^{ln}$ sind.

Beispiel 2

Für $P(\lambda)$ aus Beispiel 1 hat die Begleitmatrix die Gestalt

$$C_P = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 1 \end{pmatrix} .$$

Lemma 3

Es ist $\lambda I - A \sim \lambda I - B$ genau dann, wenn A und B zueinander ähnlich sind.

Beweis:

Um dieses Lemma zu beweisen, muss die Division für allgemeine (also nicht nur reguläre) Matrixpolynome $M(\lambda)$ eingeführt und bewiesen werden. Wir beschränken uns auf den Fall eines linearen Divisors $\lambda I + X$ ($X \in \mathbb{C}^{n\times n}$) :

Für $M(\lambda) = \sum_{i=0}^{l} A_i \lambda^i$ vom Grad l (d.h. $A_l \neq 0$) existieren genau ein $n \times n$–Matrixpolynom $Q_r(\lambda)$ vom Grad $l-1$ und eine Matrix $R_r \in \mathbb{C}^{n\times n}$, so dass

$$M(\lambda) = Q_r(\lambda)(\lambda I + X) + R_r \tag{2.2}$$

gilt. $Q_r(\lambda)$ heißt dann rechter Quotient und R_r rechter Rest bei Division von $M(\lambda)$ durch $\lambda I + X$. Analog existiert eine eindeutige Darstellung

$$M(\lambda) = (\lambda I + X)Q_l(\lambda) + R_l \tag{2.3}$$

($Q_l(\lambda)$ linker Quotient, R_l linker Rest).

Wir wollen nun (2.2) beweisen ((2.3) zeigt man analog):
Wenn $l = 0$ (d.h. $M(\lambda) \equiv$ const.), wählt man $Q_r(\lambda) \equiv O$ und $R_r = M(\lambda)$.
Sei nun also $l \geq 1$. Außerdem führen wir die Schreibweise $Q_r(\lambda) = \sum_{j=0}^{l-1} Q_j^{(r)} \lambda^j$ ein. Wenn man damit einen Koeffizientenvergleich für (2.2) durchführt, erhält man die zu (2.2) äquivalenten Gleichungen

$$A_l = Q_{l-1}^{(r)}\,, \quad A_{l-1} = Q_{l-2}^{(r)} + Q_{l-1}^{(r)} X\,,\ \ldots\ ,\ A_1 = Q_0^{(r)} + Q_1^{(r)} X\,, \quad A_0 = Q_0^{(r)} X + R_r\,.$$

Diese Gleichungen definieren der Reihe nach $Q_{l-1}^{(r)}, \ldots, Q_1^{(r)}, Q_0^{(r)}$ und R_r. Somit sind der rechte Quotient und der rechte Rest eindeutig bestimmt.

Nun zum eigentlichen Beweis des Lemmas:
Die Rückrichtung ist schnell gezeigt: wenn $A = TBT^{-1}$ (T invertierbar), dann liefert
$\lambda I - A = T(\lambda I - B)T^{-1}$ die Äquivalenz $\lambda I - A \sim \lambda I - B$.

Sei nun $\lambda I - A \sim \lambda I - B$. Dann gibt es Matrixpolynome $E(\lambda)$ und $F(\lambda)$ mit konstanten Determinanten ungleich Null, so dass

$$\lambda I - B = E(\lambda)(\lambda I - A)F(\lambda)\,. \tag{2.4}$$

$E(\lambda)$ ist nach Lemma 2 invertierbar. Linksseitige Division von $(E(\lambda))^{-1}$ durch $\lambda I - A$ und rechtsseitige Division von $F(\lambda)$ durch $\lambda I - B$ liefere

$$\begin{aligned} (E(\lambda))^{-1} &= (\lambda I - A)S(\lambda) + E_0\,, \\ F(\lambda) &= T(\lambda)(\lambda I - B) + F_0\,. \end{aligned} \tag{2.5}$$

Wenn man dies in die zu (2.4) äquivalente Gleichung

$$(E(\lambda))^{-1}(\lambda I - B) = (\lambda I - A)F(\lambda)$$

einsetzt, erhält man

$$\{(\lambda I - A)S(\lambda) + E_0\}(\lambda I - B) = (\lambda I - A)\{T(\lambda)(\lambda I - B) + F_0\}\,,$$

also

$$(\lambda I - A)(S(\lambda) - T(\lambda))(\lambda I - B) = (\lambda I - A)F_0 - E_0(\lambda I - B)\,.$$

Da der Grad des Matrixpolynoms auf der rechten Seite maximal 1 ist, folgt $S(\lambda) = T(\lambda)$ (für $S(\lambda) \neq T(\lambda)$ ist der Grad des Matrixpolynoms auf der linken Seite mindestens 2). Demnach gilt

$$(\lambda I - A)F_0 = E_0(\lambda I - B)\,,$$

also (nach Koeffizientenvergleich) $F_0 = E_0\,,\ AF_0 = E_0B$ und damit $AE_0 = E_0B\,.$

Es bleibt zu zeigen, dass E_0 nichtsingulär ist (denn dann ist $B = E_0^{-1}AE_0$). Zu diesem Zwecke dividiert man $E(\lambda)$ linksseitig durch $\lambda I - B$:

$$E(\lambda) = (\lambda I - B)U(\lambda) + R_0\,. \tag{2.6}$$

Mit (2.5) und (2.6) erhält man

$$\begin{aligned} I &= (E(\lambda))^{-1}E(\lambda) = \{(\lambda I - A)S(\lambda) + E_0\}\{(\lambda I - B)U(\lambda) + R_0\} \\ &= (\lambda I - A)S(\lambda)(\lambda I - B)U(\lambda) + (\lambda I - A)S(\lambda)R_0 + E_0(\lambda I - B)U(\lambda) + E_0R_0 \\ &= (\lambda I - A)S(\lambda)(\lambda I - B)U(\lambda) + (\lambda I - A)S(\lambda)R_0 + (\lambda I - A)F_0U(\lambda) + E_0R_0 \\ &= (\lambda I - A)[S(\lambda)(\lambda I - B)U(\lambda) + S(\lambda)R_0 + F_0U(\lambda)] + E_0R_0\,. \end{aligned}$$

Also muss das Matrixpolynom in den eckigen Klammern gleich Null sein und $E_0R_0 = I$, d.h. E_0 ist nichtsingulär. □

Satz 4

Die Matrix A ist genau dann eine Linearisierung von $P(\lambda)$, wenn A ähnlich zur Begleitmatrix C_P von $P(\lambda)$ ist.

Beweis:

Ergibt sich direkt aus Lemma 3, Satz 2 und der Transitivität der Äquivalenzrelation $\sim$. □

2.2 Smith Form, Elementarteiler, partielle Vielfachheiten

In diesem Abschnitt muss das $n \times n$–Matrixpolynom $P(\lambda) = \sum_{i=0}^{l} A_i\lambda^i$ nicht regulär sein, d.h. $\det A_l = 0$ ist zugelassen.

Definition 4

Elementare Transformationen eines $n \times n$–Matrixpolynoms $P(\lambda)$ sind die folgenden:

- *die Vertauschung zweier Zeilen bzw. Spalten;*
- *die Addition einer mit einem skalaren Polynom multiplizierten Zeile bzw. Spalte zu einer anderen Zeile bzw. Spalte;*
- *die Multiplikation einer Zeile bzw. Spalte mit einer komplexen Zahl ungleich Null.*

Jede der elementaren Transformationen entspricht der Multiplikation von $P(\lambda)$ mit einer invertierbaren Matrix:

- die Vertauschung der Zeilen (Spalten) i und j von $P(\lambda)$ entspricht der Multiplikation von links (rechts) mit

$$
\begin{array}{c}
\\
\\
\\
i \rightarrow \\
\\
\\
\\
j \rightarrow \\
\\
\\
\end{array}
\begin{pmatrix}
1 & & & & & & & & & & \\
& \ddots & & & & & & & & & \\
& & 1 & & & & & & & & \\
\dots & & & 0 & \dots & \dots & \dots & 1 & & & \\
& & & \vdots & 1 & & & \vdots & & & \\
& & & \vdots & & \ddots & & \vdots & & & \\
& & & \vdots & & & 1 & \vdots & & & \\
\dots & & & 1 & \dots & \dots & \dots & 0 & & & \\
& & & & & & & & 1 & & \\
& & & & & & & & & \ddots & \\
& & & & & & & & & & 1
\end{pmatrix}
\tag{2.7}
$$

- die Addition der mit dem Polynom $f(\lambda)$ multiplizierten j–ten Zeile zu der i–ten Zeile entspricht der Multiplikation von links mit

$$
\begin{array}{c}
\\
\\
i \rightarrow \\
\\
\\
\\
\\
\\
\\
\end{array}
\begin{array}{c}
j \downarrow \\
\\
\\
\\
\\
\\
\\
\\
\\
\end{array}
\begin{pmatrix}
1 & & & & & & & & \\
& \ddots & & & & & & & \\
\dots & & 1 & & \dots & & f(\lambda) & & \\
& & & \ddots & & & & & \\
& & & & 1 & & \vdots & & \\
& & & & & \ddots & & & \\
& & & & & & 1 & & \\
& & & & & & & \ddots & \\
& & & & & & & & 1
\end{pmatrix}
\tag{2.8}
$$

- die Addition der mit dem Polynom $f(\lambda)$ multiplizierten j–ten Spalte zu der i–ten Spalte entspricht der Multiplikation von rechts mit

$$
\begin{array}{c}
\\
\\
\\
\\
\\
\\
j \rightarrow \\
\\
\\
\end{array}
\begin{array}{c}
\begin{array}{c} i \downarrow \end{array} \\
\left(\begin{array}{ccccccccc}
1 & & \vdots & & & & & & \\
& \ddots & & & & & & & \\
& & 1 & & & & & & \\
& & & \ddots & & & & & \\
& & \vdots & & 1 & & & & \\
& & & & & \ddots & & & \\
\ldots & & f(\lambda) & & \ldots & & 1 & & \\
& & & & & & & \ddots & \\
& & & & & & & & 1
\end{array}\right)
\end{array}
\tag{2.9}
$$

- die Multiplikation der i–ten Zeile (Spalte) mit einer komplexen Zahl $a \neq 0$ entspricht der Multiplikation von links (rechts) mit

$$
\begin{array}{c}
\\
\\
i \rightarrow \\
\\
\\
\\
\\
\end{array}
\begin{array}{c}
\left(\begin{array}{ccccccc}
1 & & & & & & \\
& \ddots & & & & & \\
& & 1 & & & & \\
\ldots & & & a & & & \\
& & & & 1 & & \\
& & & & & \ddots & \\
& & & & & & 1
\end{array}\right)
\end{array}
\tag{2.10}
$$

(Leere Stellen in den Matrizen (2.7)–(2.10) sind Nullen.)

Matrizen der Form (2.7)–(2.10) heißen elementar. Es ist offensichtlich, dass die Determinante einer elementaren Matrix eine Konstante ungleich Null ist.

Als Vorbereitung auf den folgenden Satz sei erwähnt, dass ein skalares Polynom $b_m x^m + b_{m-1} x^{m-1} + \ldots + b_0$ monisch heißt, wenn für den führenden Koeffizienten des Polynoms $b_m = 1$ gilt.

Satz 5

Jedes $n \times n$–Matrixpolynom $P(\lambda)$ ist äquivalent zu einem diagonalen Matrixpolynom

$$D(\lambda) = diag(d_1(\lambda), \ldots, d_r(\lambda), 0, \ldots, 0) \tag{2.11}$$

mit monischen skalaren Polynomen $d_i(\lambda)$, wobei $d_i(\lambda)$ teilbar durch $d_{i-1}(\lambda)$ ist für $i = 2, \ldots, r$.

(2.11) wird als Smith Form des Matrixpolynoms $P(\lambda)$ bezeichnet.

Es wird später gezeigt, dass die Smith Form eines Matrixpolynoms eindeutig bestimmt ist.

Beweis:

Es ist zu zeigen, dass es $n \times n$–Matrixpolynome $E(\lambda)$ und $F(\lambda)$ mit $D(\lambda) = E(\lambda)P(\lambda)F(\lambda)$ und $\det E(\lambda) = \text{const} \neq 0$, $\det F(\lambda) = \text{const} \neq 0$ gibt. Da die Determinante einer elementaren

Matrix eine Konstante ungleich Null ist, genügt es, Folgendes zu zeigen: durch Anwendung einer Folge von Elementartransformationen (d.h. Multiplikation mit elementaren Matrizen von links und rechts) kann das $n \times n$–Matrixpolynom $P(\lambda)$ in $D(\lambda)$ (siehe (2.11)) überführt werden. Dies wird per Induktion über n gezeigt.

Für $n = 1$ ist $P(\lambda) \equiv 0$ oder $P(\lambda) = c \cdot d_1(\lambda)$ mit $c \neq 0$ und $d_1(\lambda)$ monisch , d.h. die Behauptung stimmt.

Sei nun $n > 1$. Für $P(\lambda) = O$ ist die Behauptung trivial. Also sei $P(\lambda) \neq O$. Wir nehmen an, dass der Satz für $n - 1$ bewiesen ist.

Schritt 1:
Der (i, j)–te Eintrag von $P(\lambda)$ sei ungleich Null und kleinsten Grades in $P(\lambda)$. Durch Zeilen- und Spaltenvertauschungen in $P(\lambda)$ bringt man dieses Element an die Stelle $(1, 1)$ und bezeichnet die so entstandene Matrix als $(a_{ij}(\lambda))$. Für jeden Eintrag der ersten Zeile und ersten Spalte werden nun Quotient und Rest bei Division durch $a_{11}(\lambda)$ bestimmt:

$$\begin{aligned} a_{1j}(\lambda) &= a_{11}(\lambda)q_{1j}(\lambda) + r_{1j}(\lambda)\,, \quad j = 2, 3, \ldots, n\,, \\ a_{i1}(\lambda) &= a_{11}(\lambda)q_{i1}(\lambda) + r_{i1}(\lambda)\,, \quad i = 2, 3, \ldots, n\,. \end{aligned}$$

Für jedes $i \neq 1$ und jedes $j \neq 1$ subtrahiert man nun $q_{1j}(\lambda)$ mal die erste von der j–ten Spalte und $q_{i1}(\lambda)$ mal die erste von der i–ten Zeile (elementare Transformationen vom Typ 2). Dadurch werden die Einträge $a_{1j}(\lambda)$ bzw. $a_{i1}(\lambda)$ durch die Polynome $r_{1j}(\lambda)$ bzw. $r_{i1}(\lambda)$ ersetzt $(i, j = 2, \ldots, n)$, welche entweder das Nullpolynom oder kleineren Grades als $a_{11}(\lambda)$ sind. Wenn diese Polynome nicht alle gleich Null sind, nutzt man Zeilen– und Spaltenvertauschungen, um $a_{11}(\lambda)$ mit einem Eintrag $r_{1j}(\lambda)$ oder $r_{i1}(\lambda)$ kleinsten Grades auszutauschen.

Dieser Prozess, der den Grad der Nichtdiagonalelemente der ersten Zeile und Spalte so verkleinert, dass er kleiner als der Grad des Eintrags an der Stelle $(1, 1)$ ist, wird wiederholt. Da der Grad des Eintrags an der Stelle $(1, 1)$ in jedem Schritt verkleinert wird (und dies nur endlich oft passieren kann), erhalten wir nach endlich vielen Schritten ein Matrixpolynom der Form

$$\begin{pmatrix} a_{11}^{(1)}(\lambda) & 0 & \ldots & 0 \\ 0 & a_{22}^{(1)}(\lambda) & \ldots & a_{2n}^{(1)}(\lambda) \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(1)}(\lambda) & \ldots & a_{nn}^{(1)}(\lambda) \end{pmatrix}. \tag{2.12}$$

Schritt 2:
In der Matrix (2.12) kann es Einträge $a_{ij}^{(1)}(\lambda) \neq 0\,,\ 2 \leq i, j \leq n$, geben, deren Grad kleiner als der von $a_{11}^{(1)}(\lambda)$ ist. Wenn $a_{ij}^{(1)}(\lambda)$ solch ein Eintrag ist, wendet man Schritt 1 erneut an und erhält dann eine Matrix der Form (2.12), für die der Grad des Eintrags $(1, 1)$ wiederum verkleinert ist. Wenn man also Schritt 1 auf diese Weise genügend oft wiederholt, erhält man eine Matrix

$$\begin{pmatrix} a_{11}^{(2)}(\lambda) & 0 & \ldots & 0 \\ 0 & a_{22}^{(2)}(\lambda) & \ldots & a_{2n}^{(2)}(\lambda) \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)}(\lambda) & \ldots & a_{nn}^{(2)}(\lambda) \end{pmatrix} \tag{2.13}$$

für die $a_{11}^{(2)}(\lambda) \neq 0$ kleinsten Grades ist.

Schritt 3:
In der Matrix (2.13) kann es Einträge ungleich Null geben, die nicht durch $a_{11}^{(2)}(\lambda)$ teilbar sind. Wenn $a_{ij}^{(2)}(\lambda)$ ein solcher ist, addiert man die Spalte j zu der Spalte 1 und berechnet dann die Quotienten und Reste der Einträge der neuen Spalte 1 bei Division durch $a_{11}^{(2)}(\lambda)$. Zumindest der zu $a_{ij}^{(2)}(\lambda)$ gehörige Rest ist ungleich Null. Es werden dann die Schritte 1 und 2 durchgeführt. Man erhält wiederum eine Matrix der Form (2.13), für die der Grad des Eintrags $(1,1)$ verkleinert ist. Schritt 3 wird nun solange wie möglich durchgeführt. Dies kann nur endlich oft passieren. Man erhält dann eine Matrix

$$\begin{pmatrix} a_{11}^{(3)}(\lambda) & 0 & \dots & 0 \\ 0 & a_{22}^{(3)}(\lambda) & \dots & a_{2n}^{(3)}(\lambda) \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(3)}(\lambda) & \dots & a_{nn}^{(3)}(\lambda) \end{pmatrix} \tag{2.14}$$

für die jedes $a_{ij}^{(3)}(\lambda)$ durch $a_{11}^{(3)}(\lambda)$ teilbar ist.

Schritt 4:
Man multipliziert die erste Spalte von (2.14) so mit einer Konstante ungleich Null, dass der führende Koeffizient von $a_{11}^{(3)}(\lambda)$ zu 1 wird. $a_{11}^{(3)}(\lambda)$ ist dann also monisch. Das so entstandene Matrixpolynom wird mit $P_3(\lambda)$ bezeichnet. Nun kann man auf das $(n-1)\times(n-1)$–Matrixpolynom

$$P_4(\lambda) = \frac{1}{a_{11}^{(3)}(\lambda)} \begin{pmatrix} a_{22}^{(3)}(\lambda) & \dots & a_{2n}^{(3)}(\lambda) \\ \vdots & & \vdots \\ a_{n2}^{(3)}(\lambda) & \dots & a_{nn}^{(3)}(\lambda) \end{pmatrix}$$

die Induktionsannahme anwenden, d.h. es gilt $P_4(\lambda) \sim \operatorname{diag}(e_1(\lambda), \dots, e_s(\lambda), 0, \dots, 0) =: D_4(\lambda)$ mit monischen skalaren Polynomen $e_i(\lambda)$, wobei $e_i(\lambda)$ teilbar durch $e_{i-1}(\lambda)$ ist für $i = 2, \dots, s$. Damit gilt

$$\begin{aligned} P(\lambda) \;&\sim\; P_3(\lambda) \;=\; a_{11}^{(3)}(\lambda) \begin{pmatrix} 1 & 0 \\ 0 & P_4(\lambda) \end{pmatrix} \sim a_{11}^{(3)}(\lambda) \begin{pmatrix} 1 & 0 \\ 0 & D_4(\lambda) \end{pmatrix} \\ &= \operatorname{diag}(a_{11}^{(3)}(\lambda)\,,\, a_{11}^{(3)}(\lambda)e_1(\lambda)\,,\, \dots, a_{11}^{(3)}(\lambda)e_s(\lambda)\,,\, 0\,,\, \dots\,, 0), \end{aligned}$$

womit mit $d_1(\lambda) := a_{11}^{(3)}(\lambda)$ und $d_i(\lambda) := a_{11}^{(3)}(\lambda)e_{i-1}(\lambda)$ für $i = 2, \dots, s+1$ der Satz bewiesen ist. □

Es folgt nun die Definition der Minoren von $A \in \mathbb{C}^{n\times n}$. Für $1 \le k \le n$ und zwei Sequenzen $\underline{i} = (1 \le i_1 < i_2 < \dots < i_k \le n)$ und $\underline{j} = (1 \le j_1 < j_2 < \dots < j_k \le n)$ erhält man die $k \times k$–Matrix $A^{\underline{i},\underline{j}}$ aus A durch Streichen aller Zeilen außer denen mit Index $i_1, \dots, i_k$ und Streichen aller Spalten außer denen mit Index $j_1, \dots, j_k$. Die Determinante $\det(A^{\underline{i},\underline{j}})$ wird dann als Minor k–ter Ordnung bezeichnet. Es gibt $\binom{n}{k}^2$ Minoren k–ter Ordnung. Analog sind die Minoren eines $n \times n$–Matrixpolynoms $P(\lambda)$ vom Grad l definiert, welche dann skalare Polynome sind, die höchstens den Grad kl besitzen.

In Vorbereitung auf Satz 7 wird folgender Satz erwähnt:

Satz 6 *(Satz von Binet–Cauchy, [9])*

Es seien $A, B \in \mathbb{C}^{n\times n}$, $1 \le k \le n$, $\underline{i} = (1 \le i_1 < i_2 < \ldots < i_k \le n)$ *und* $\underline{j} = (1 \le j_1 < j_2 < \ldots < j_k \le n)$. *Dann gilt*

$$\det\left((AB)^{\underline{i},\underline{j}}\right) = \sum_{\underline{c}=(1\le c_1<c_2<\ldots<c_k\le n)} \det\left(A^{\underline{i},\underline{c}}\right)\cdot\det\left(B^{\underline{c},\underline{j}}\right).$$

Satz 7

$P(\lambda)$ sei ein $n\times n$–Matrixpolynom. $p_k(\lambda)$ sei der monische größte gemeinsame Teiler der Minoren k–ter Ordnung von $P(\lambda)$ wenn diese nicht alle gleich Null sind. Sonst sei $p_k(\lambda)\equiv 0$. Desweiteren sei $p_0(\lambda)\equiv 1$ und $D(\lambda) = diag(d_1(\lambda),\ldots,d_r(\lambda),0,\ldots,0)$ die Smith Form von $P(\lambda)$.
Dann ist r die größte natürliche Zahl, für die $p_r(\lambda)\not\equiv 0$ gilt. Außerdem gilt

$$d_i(\lambda) = \frac{p_i(\lambda)}{p_{i-1}(\lambda)}, \qquad i = 1,\ldots,r, \tag{2.15}$$

d.h. die $d_i(\lambda)$ sind eindeutig bestimmt.

Die Diagonalelemente $d_1(\lambda),\ldots,d_r(\lambda)$ der Smith Form werden als invariante Polynome von $P(\lambda)$ bezeichnet.

Eine Folgerung aus Satz 7 ist, dass die Smith Form von $P(\lambda)$ eindeutig bestimmt ist.

Beweis:

$P(\lambda)$ und $Q(\lambda)$ seien äquivalent, d.h. $P(\lambda) = E(\lambda)Q(\lambda)F(\lambda)$ mit $\det E(\lambda) = c_1 \neq 0$ und $\det F(\lambda) = c_2 \neq 0$. Dann stimmt der größte gemeinsame Teiler $p_{k,1}(\lambda)$ der Minoren k–ter Ordnung von $P(\lambda)$ mit dem größten gemeinsamen Teiler $p_{k,2}(\lambda)$ der Minoren k–ter Ordnung von $Q(\lambda)$ überein, wie gleich gezeigt wird.

Wenn man nämlich die Binet–Cauchy Formel zweimal auf $P(\lambda) = E(\lambda)Q(\lambda)F(\lambda)$ anwendet, kann man einen Minor $m_k(\lambda)$ k–ter Ordnung von $P(\lambda)$ mit Hilfe von Minoren $q_{k,s}(\lambda)$ k–ter Ordnung von $Q(\lambda)$ folgendermaßen ausdrücken (nach Umordnung):

$$m_k(\lambda) = \sum_s a_s(\lambda)q_{k,s}(\lambda)b_s(\lambda),$$

wobei $a_s(\lambda)$ bzw. $b_s(\lambda)$ geeignete Minoren k–ter Ordnung von $E(\lambda)$ bzw. $F(\lambda)$ sind. Jeder gemeinsame Teiler der Minoren $q_{k,s}(\lambda)$ k–ter Ordnung von $Q(\lambda)$ ist damit ein Teiler von $m_k(\lambda)$. Also ist $p_{k,2}(\lambda)$ ein Teiler von $p_{k,1}(\lambda)$.

$Q(\lambda) = E^{-1}(\lambda)P(\lambda)F^{-1}(\lambda)$ liefert mit derselben Argumentation, dass $p_{k,1}(\lambda)$ ein Teiler von $p_{k,2}(\lambda)$ ist. Weil $p_{k,1}(\lambda)$ und $p_{k,2}(\lambda)$ monisch sind, gilt dann also $p_{k,1}(\lambda) = p_{k,2}(\lambda)$.

Auf die gleiche Weise zeigt man, dass die größte natürliche Zahl r_1 mit $p_{r_1,1}(\lambda)\neq 0$ genau die größte natürliche Zahl r_2 mit $p_{r_2,2}(\lambda)\neq 0$ ist.

Nun verwendet man diese Aussagen für $P(\lambda)$ und $Q(\lambda) = D(\lambda)$. Es folgt, dass es ausreicht, die Aussagen des Satzes für $P(\lambda) = D(\lambda)$ zu zeigen. Es ist klar, dass für $1 \le s \le r$

$$d_1(\lambda)d_2(\lambda)\cdot\ldots\cdot d_s(\lambda)$$

der größte gemeinsame Teiler der Minoren s–ter Ordnung ungleich Null von $D(\lambda)$ ist. Also ist $p_s(\lambda) = d_1(\lambda) \cdot \ldots \cdot d_s(\lambda)$, $s = 1, \ldots, r$, und (2.15) folgt. Für $s > r$ sind alle Minoren s–ter Ordnung von $D(\lambda)$ gleich Null, d.h. $p_s(\lambda) \equiv 0$. □

Bemerkung 1

Für ein reguläres $n \times n$–Matrixpolynom $P(\lambda)$ l–ten Grades (d.h. $\det A_l \neq 0$*) ist der monische größte Teiler des Minors n–ter Ordnung* $\det(P(\lambda))$

$$p_n(\lambda) = \frac{1}{\det A_l} \det(P(\lambda)) \not\equiv 0\,,$$

also gilt nach Satz 7 in der Smith Form (2.11) *von $P(\lambda)$ die Beziehung $r = n$.*

Beispiel 3

Für das reguläre 2×2–Matrixpolynom vom Grad 2

$$P(\lambda) = \begin{pmatrix} \lambda^2 & 0 \\ \lambda(\lambda+1) & \lambda(\lambda-1) \end{pmatrix}$$

soll mit Hilfe von Satz 7 die Smith Form gefunden werden.

Es ist $p_1(\lambda) = \lambda$ und $p_2(\lambda) = \det P(\lambda) = \lambda^3(\lambda - 1)$. Damit lauten die invarianten Polynome $d_1(\lambda) = \lambda$ und $d_2(\lambda) = \lambda^2(\lambda - 1)$. Die Smith Form von $P(\lambda)$ ist folglich

$$D(\lambda) = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda^2(\lambda-1) \end{pmatrix}.$$

Lemma 4

Für die Anzahl r der invarianten Polynome von $P(\lambda)$ gilt

$$r = \max_{\lambda \in \mathbb{C}} \{\operatorname{rang} P(\lambda)\}\,.$$

Beweis:

Wenn man sich den Beweis von Satz 5 anschaut, sieht man, dass $P(\lambda)$ durch Anwendung einer Folge von elementaren Transformationen in $D(\lambda)$ (siehe (2.11)) überführt wird. Für ein festes $\lambda^* \in \mathbb{C}$ sind dies elementare Zeilen– und Spaltentransformationen, die den Rang einer Matrix nicht ändern. Also gilt $\operatorname{rang} P(\lambda^*) = \operatorname{rang} D(\lambda^*)$ für jedes $\lambda^* \in \mathbb{C}$.

Wenn λ^* keine Nullstelle eines der invarianten Polynome von $P(\lambda)$ ist, gilt trivialerweise $\operatorname{rang} D(\lambda^*) = r$. Sonst gilt $\operatorname{rang} D(\lambda^*) < r$. Damit ist die Behauptung bewiesen. □

Man kann jedes invariante Polynom als Produkt von Linearfaktoren darstellen:

$$d_i(\lambda) = (\lambda - \lambda_{i1})^{\alpha_{i1}} \cdot \ldots \cdot (\lambda - \lambda_{i,k_i})^{\alpha_{ik_i}}\,, \quad i = 1, \ldots, r$$

mit verschiedenen komplexen Zahlen $\lambda_{i1}, \ldots, \lambda_{i,k_i}$ und positiven natürlichen Zahlen $\alpha_{i1}, \ldots, \alpha_{i,k_i}$. Ein Faktor $(\lambda - \lambda_{ij})^{\alpha_{ij}}$, $j \in \{1, \ldots, k_i\}$, $i \in \{1, \ldots, r\}$, heißt Elementarteiler von $P(\lambda)$ zu λ_{ij}. Ein Elementarteiler heißt linear, wenn $\alpha_{ij} = 1$ ist. Für $\alpha_{ij} > 1$ heißt er nichtlinear.

Satz 8

$P(\lambda)$ sei ein $n \times n$–Matrixpolynom mit $\det P(\lambda) \not\equiv 0$.
Dann existiert für jedes $\lambda_0 \in \mathbb{C}$ die Darstellung

$$P(\lambda) = E_{\lambda_0}(\lambda)\, diag((\lambda-\lambda_0)^{\nu_1}, \ldots, (\lambda-\lambda_0)^{\nu_n})\, F_{\lambda_0}(\lambda) \tag{2.16}$$

mit Matrixpolynomen $E_{\lambda_0}(\lambda)$ und $F_{\lambda_0}(\lambda)$, die für $\lambda = \lambda_0$ invertierbar sind, und $\nu_1 \leq \ldots \leq \nu_n$ ($\nu_i \in \mathbb{N}_0$). Die $\nu_i\,, i = 1, \ldots, n$, sind (wenn man alle Nullen weglässt) genau die Grade der Elementarteiler von $P(\lambda)$ zu λ_0.

$\nu_1, \ldots, \nu_n$ werden partielle Vielfachheiten von $P(\lambda)$ zu λ_0 genannt. Die Darstellung (2.16) wird als lokale Smith Form von $P(\lambda)$ zu λ_0 bezeichnet.

Zur Erinnerung: Elementarteiler von $P(\lambda)$ zu λ_0 besitzen die Form $(\lambda-\lambda_0)^m$.

Beweis:

Die Existenz der Darstellung (2.16) folgt aus der Smith Form auf die folgende Weise: Die Diagonalmatrix $D(\lambda)$ sei die Smith Form von $P(\lambda)$ mit

$$P(\lambda) = E(\lambda)D(\lambda)F(\lambda) \tag{2.17}$$

und $\det E(\lambda) \equiv \text{const} \neq 0$, $\det F(\lambda) \equiv \text{const} \neq 0$.

Weil nach Voraussetzung $\det D(\lambda) = \det E(\lambda)^{-1} \det P(\lambda) \det F(\lambda)^{-1} \not\equiv 0$ gilt, kann $D(\lambda)$ keine Nullen auf der Diagonalen besitzen, d.h. es gilt $D(\lambda) = \text{diag}(d_1(\lambda), \ldots, d_n(\lambda))$. Für $\lambda_0 \in \mathbb{C}$ kann jedes $d_i(\lambda)$ folgendermaßen faktorisiert werden:

$$d_i(\lambda) = (\lambda-\lambda_0)^{\nu_i}\tilde{d}_i(\lambda)\,, \quad i = 1, \ldots, n,$$

mit $\tilde{d}_i(\lambda_0) \neq 0$ und $\nu_i \geq 0$. Weil $d_i(\lambda)$ Teiler von $d_{i+1}(\lambda)$ ist, gilt $\nu_i \leq \nu_{i+1}$. (2.16) folgt nun aus (2.17), wenn man

$$E_{\lambda_0}(\lambda) = E(\lambda)\text{diag}(\tilde{d}_1(\lambda), \ldots, \tilde{d}_n(\lambda))\,, \quad F_{\lambda_0}(\lambda) = F(\lambda)$$

setzt. Es gilt $E_{\lambda_0}(\lambda_0)^{-1} = \text{diag}(1/\tilde{d}_1(\lambda_0), \ldots 1/\tilde{d}_n(\lambda_0))E(\lambda_0)^{-1}$.

Es bleibt zu zeigen, dass die $\nu_i\,, i = 1, \ldots, n$, (wenn man alle Nullen weglässt) genau die Grade der Elementarteiler von $P(\lambda)$ zu λ_0 sind. Zu diesem Zwecke zeigt man, dass jede Faktorisierung von $P(\lambda)$ des Typs (2.16) mit $\nu_1 \leq \ldots \leq \nu_n$ impliziert, dass ν_j die Vielfachheit von λ_0 als Nullstelle von $d_j(\lambda)\,,\ j = 1, \ldots, n$, ist (wobei $D(\lambda) = \text{diag}(d_1(\lambda), \ldots, d_n(\lambda))$ die Smith Form von $P(\lambda)$ mit (2.17) ist).

Aus (2.16) und (2.17) erhält man

$$\begin{pmatrix} d_1(\lambda) & 0 & \ldots & 0 \\ 0 & d_2(\lambda) & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & d_n(\lambda) \end{pmatrix} = \tilde{E}_{\lambda_0}(\lambda) \begin{pmatrix} (\lambda-\lambda_0)^{\nu_1} & 0 & \ldots & 0 \\ 0 & (\lambda-\lambda_0)^{\nu_2} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & (\lambda-\lambda_0)^{\nu_n} \end{pmatrix} \tilde{F}_{\lambda_0}(\lambda), \tag{2.18}$$

wobei $\tilde{E}_{\lambda_0}(\lambda) = E(\lambda)^{-1}E_{\lambda_0}(\lambda)\,,\ \tilde{F}_{\lambda_0}(\lambda) = F_{\lambda_0}(\lambda)F(\lambda)^{-1}$ Matrixpolynome sind, die für $\lambda = \lambda_0$ invertierbar sind. Wenn man darauf die Binet–Cauchy Formel für Minoren i_0–ter Ordnung von

Matrixprodukten anwendet, erhält man

$$d_1(\lambda)d_2(\lambda)\cdot\ldots\cdot d_{i_0}(\lambda)=\sum_{i,j,k} m_{i,\tilde{E}}(\lambda)m_{j,D_{\lambda_0}}(\lambda)m_{k,\tilde{F}}(\lambda)\,,\quad i_0=1,\ldots,n, \tag{2.19}$$

wobei $m_{i,\tilde{E}}(\lambda)$ (bzw. $m_{j,D_{\lambda_0}}(\lambda)$, $m_{k,\tilde{F}}(\lambda)$) Minoren i_0–ter Ordnung von $\tilde{E}_{\lambda_0}(\lambda)$ (bzw. $\mathrm{diag}((\lambda-\lambda_0)^{\nu_1},\ldots,(\lambda-\lambda_0)^{\nu_n})$, $\tilde{F}_{\lambda_0}(\lambda)$) sind und die Summation über eine gewisse Menge von Tripeln (i,j,k) erfolgt. Aus (2.19) und der Bedingung $\nu_1\le\ldots\le\nu_n$ folgt, dass λ_0 eine mindestens $(\nu_1+\nu_2+\ldots+\nu_{i_0})$–fache Nullstelle des Produkts $d_1(\lambda)d_2(\lambda)\cdot\ldots\cdot d_{i_0}(\lambda)$ ist (da λ_0 mindestens $(\nu_1+\nu_2+\ldots+\nu_{i_0})$–fache Nullstelle eines beliebigen Minors i_0–ter Ordnung $m_{j,D_{\lambda_0}}(\lambda)$ von $\mathrm{diag}((\lambda-\lambda_0)^{\nu_1},\ldots,(\lambda-\lambda_0)^{\nu_n})$ ist).

Man kann (2.18) umschreiben in

$$\begin{aligned}(\tilde{E}_{\lambda_0}(\lambda))^{-1}&\begin{pmatrix} d_1(\lambda) & 0 & \ldots & 0\\ 0 & d_2(\lambda) & & 0\\ \vdots & & \ddots & \vdots\\ 0 & 0 & \ldots & d_n(\lambda)\end{pmatrix}(\tilde{F}_{\lambda_0}(\lambda))^{-1}\\ &=\begin{pmatrix}(\lambda-\lambda_0)^{\nu_1} & 0 & \ldots & 0\\ 0 & (\lambda-\lambda_0)^{\nu_2} & & 0\\ \vdots & & \ddots & \vdots\\ 0 & 0 & \ldots & (\lambda-\lambda_0)^{\nu_n}\end{pmatrix}.\end{aligned} \tag{2.20}$$

Für ein beliebiges $n\times n$–Matrixpolynom $A(\lambda)$ sind die Einträge der inversen Matrix Minoren $(n-1)$–ter Ordnung von $A(\lambda)$ geteilt durch $\det A(\lambda)$, also rationale Funktionen in λ. $(\tilde{E}_{\lambda_0}(\lambda))^{-1}$ und $(\tilde{F}_{\lambda_0}(\lambda))^{-1}$ sind also rationale Matrixfunktionen, die wegen $\det(\tilde{E}_{\lambda_0}(\lambda_0))$, $\det(\tilde{F}_{\lambda_0}(\lambda_0))\neq 0$ für $\lambda=\lambda_0$ definiert und invertierbar sind.

Wenn man die Binet–Cauchy Formel auf (2.20) anwendet und berücksichtigt, dass $d_i(\lambda)$ ein Teiler von $d_{i+1}(\lambda)$ $(i=1,\ldots,n-1)$ ist, erhält man

$$(\lambda-\lambda_0)^{\nu_1+\nu_2+\ldots+\nu_{i_0}}=d_1(\lambda)d_2(\lambda)\cdot\ldots\cdot d_{i_0}(\lambda)\Phi_{i_0}(\lambda) \tag{2.21}$$

wobei $\Phi_{i_0}(\lambda)$ eine rationale Funktion ist, die für $\lambda=\lambda_0$ definiert ist (d.h. λ_0 ist keine Polstelle von $\Phi_{i_0}(\lambda)$).

Angenommen λ_0 ist eine Nullstelle von $d_1(\lambda)d_2(\lambda)\cdot\ldots\cdot d_{i_0}(\lambda)$ mit Vielfachheit größer als $(\nu_1+\nu_2+\ldots+\nu_{i_0})$. Wegen (2.21) muss dann λ_0 Polstelle von $\Phi_{i_0}(\lambda)$ sein, Widerspruch.

Also ist λ_0 eine $(\nu_1+\nu_2+\ldots+\nu_{i_0})$–fache Nullstelle von $d_1(\lambda)d_2(\lambda)\cdot\ldots\cdot d_{i_0}(\lambda)\,, i_0=1,\ldots,n$. Deswegen ist ν_i genau die Vielfachheit von λ_0 als Nullstelle von $d_i(\lambda)\,, i=1,\ldots,n$. □

Wenn λ_0 kein Eigenwert von $P(\lambda)$ ist, d.h. keine Nullstelle eines invarianten Polynoms von $P(\lambda)$ ist, sind nach Definition alle partiellen Vielfachheiten von $P(\lambda)$ zu λ_0 gleich Null.

Bemerkung 2

Für ein reguläres $n\times n$–Matrixpolynom $P(\lambda)$ l–ten Grades (d.h. $\det A_l\neq 0$) gilt nach Satz 5 und Bemerkung 1

$$c\cdot\det(P(\lambda))=\det(D(\lambda))=\prod_{i=1}^{n} d_i(\lambda)\,,$$

wobei $D(\lambda)$ die Smith Form von $P(\lambda)$ und $c \neq 0$ ist.
Für die algebraische Vielfachheit σ eines Eigenwertes λ_0 von $P(\lambda)$ gilt dann also

$$\sigma = \sum_{i=1}^{n} \nu_i \,,$$

wobei ν_i, $i = 1, \ldots, n$, die partiellen Vielfachheiten von $P(\lambda)$ zu λ_0 sind (siehe Satz 8).

Beispiel 4

Für

$$P(\lambda) = \begin{pmatrix} \lambda^2 & 0 \\ \lambda(\lambda+1) & \lambda(\lambda-1) \end{pmatrix}$$

aus Beispiel 3 ist $\det P(\lambda) = \lambda^3(\lambda-1)$, *also sind $\lambda = 0$ und $\lambda = 1$ die Eigenwerte von $P(\lambda)$. Nach Beispiel 3 lauten die Elementarteiler von $P(\lambda)$ λ, λ^2 und $\lambda - 1$. Damit ist nach Satz 8*

$$\begin{pmatrix} \lambda & 0 \\ 0 & \lambda^2 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & \lambda - 1 \end{pmatrix} \text{ bzw. } \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

die lokale Smith Form von $P(\lambda)$ zu $\lambda_0 = 0$, $\lambda_0 = 1$ bzw. $\lambda_0 \neq 0, 1$.

2.3 Jordanketten und Lösungen von linearen Differentialgleichungen

In diesem Abschnitt ist $P(\lambda) = \sum_{i=0}^{l} A_i \lambda^i$ wieder ein reguläres $n \times n$–Matrixpolynom l–ten Grades, d.h. es gilt $\det A_l \neq 0$. Dann ist

$$P^{(k)}(\lambda) = \left(\frac{d^{\,k} P(\lambda)_{ij}}{d\,\lambda^k} \right)_{i,j=1,\ldots,n} \tag{2.22}$$

die k–te Ableitung von $P(\lambda)$ nach λ (wobei das skalare Polynom höchstens l–ten Grades $P(\lambda)_{ij}$ der (i,j)–te Eintrag von $P(\lambda)$ ist). Die erste Ableitung z.B. lautet

$$P'(\lambda) = \sum_{i=1}^{l} i A_i \lambda^{i-1} \,.$$

Definition 5

Eine Folge von n-komponentigen Vektoren $x_0, x_1, \ldots, x_k$ $(x_0 \neq 0)$, für die

$$\sum_{j=0}^{i} \frac{1}{j!} P^{(j)}(\lambda^*) x_{i-j} = 0 \,, \qquad i = 0, \ldots, k \,, \tag{2.23}$$

gilt, heißt Jordankette der Länge $k+1$ für $P(\lambda)$ und den Eigenwert λ^* von $P(\lambda)$. Die Vektoren $x_1, \ldots, x_k$ werden als verallgemeinerte Eigenvektoren bezeichnet.

Diese Definition einer Jordankette bei Matrixpolynomen ist eine Verallgemeinerung des Begriffs Jordankette für quadratische Matrizen. Für eine Jordankette $x_0, \ldots, x_k$ von $A \in \mathbb{C}^{n\times n}$ zum Eigenwert λ^* gilt

$$Ax_0 = \lambda^* x_0\,, \quad Ax_1 = \lambda^* x_1 + x_0\,, \ \ldots\ , \quad Ax_k = \lambda^* x_k + x_{k-1}\,.$$

Dies ist genau die Definition einer Jordankette des linearen Matrixpolynoms $I\lambda - A$.

Aus der Definition einer Jordankette von $P(\lambda)$ ergibt sich, dass der Vektor x_0 ein Eigenvektor von $P(\lambda)$ zum Eigenwert λ^* ist. Für ein Matrixpolynom vom Grad $l > 1$ müssen die Vektoren einer Jordankette nicht linear unabhängig sein. Sogar der Nullvektor ist als verallgemeinerter Eigenvektor zulässig. Dazu gibt es später ein Beispiel.

Wir betrachten nun die zu $P(\lambda)$ gehörige homogene lineare Differentialgleichung

$$P(\frac{d}{dt})u(t) = A_l u^{(l)}(t) + A_{l-1}u^{(l-1)}(t) + \ldots + A_1 u'(t) + A_0 u(t) = 0\,, \tag{2.24}$$

wobei $u : D \subseteq \mathbb{C}^n \to \mathbb{C}^n$ eine vektorwertige Funktion ist.

Satz 9

Die vektorwertige Funktion

$$u_k(t) = \left(\frac{t^k}{k!}x_0 + \frac{t^{k-1}}{(k-1)!}x_1 + \ldots + x_k\right)e^{\lambda^* t} \tag{2.25}$$

für ein $k \in \mathbb{N}_0$ ist genau dann eine Lösung der Differentialgleichung (2.24)*, wenn die Vektoren $x_0, x_1, \ldots, x_k$ eine Jordankette von $P(\lambda)$ zum Eigenwert λ^* bilden.*

Beweis:

$u(t) = u_k(t)$ sei gegeben durch (2.25). Dann gilt

$$\begin{aligned}\left(\frac{d}{dt} - \lambda^* I\right)u(t) &= u'(t) - \lambda^* u(t)\\ &= \left(\frac{t^{k-1}}{(k-1)!}x_0 + \ldots + x_{k-1}\right)e^{\lambda^* t} + \lambda^* u(t) - \lambda^* u(t)\\ &= \left(\frac{t^{k-1}}{(k-1)!}x_0 + \ldots + x_{k-1}\right)e^{\lambda^* t}\end{aligned}$$

und allgemeiner

$$\left(\frac{d}{dt} - \lambda^* I\right)^j u(t) = \left(\frac{t^{k-j}}{(k-j)!}x_0 + \frac{t^{k-j-1}}{(k-j-1)!}x_1 + \ldots + x_{k-j}\right)e^{\lambda^* t} \tag{2.26}$$

für $j = 0, \ldots, k$ und

$$\left(\frac{d}{dt} - \lambda^* I\right)^j u(t) = 0 \qquad \text{für } \ j = k+1,\, k+2, \ldots\,. \tag{2.27}$$

Die Taylorreihe von $P(\lambda)$ an der Stelle λ^* lautet

$$P(\lambda) = P(\lambda^*) + P'(\lambda^*)(\lambda - \lambda^*) + \frac{1}{2!}P''(\lambda^*)(\lambda - \lambda^*)^2 + \ldots + \frac{1}{l!}P^{(l)}(\lambda^*)(\lambda - \lambda^*)^l\,.$$

Wenn man nun λ durch $\dfrac{d}{dt}$ ersetzt, erhält man

$$P\left(\frac{d}{dt}\right)u(t) = P(\lambda^*)u(t) + P'(\lambda^*)\left(\frac{d}{dt} - \lambda^* I\right)u(t) + \ldots + \frac{1}{l!}P^{(l)}(\lambda^*)\left(\frac{d}{dt} - \lambda^* I\right)^l u(t)\,.$$

Das Einsetzen von (2.26) und (2.27) in diese Gleichung liefert

$$\begin{aligned} P\left(\frac{d}{dt}\right)u(t) &= \left\{\frac{t^k}{k!}P(\lambda^*)x_0 + \frac{t^{k-1}}{(k-1)!}(P(\lambda^*)x_1 + P'(\lambda^*)x_0) + \ldots \right. \\ &\quad \left. +(P(\lambda^*)x_k + P'(\lambda^*)x_{k-1} + \ldots + \frac{1}{k!}P^{(k)}(\lambda^*)x_0)\right\}e^{\lambda^* t}\,, \end{aligned}$$

was genau dann gleich Null ist, wenn $x_0, \ldots, x_k$ eine Jordankette von $P(\lambda)$ zu λ^* bilden. $\square$

Es ist leicht zu sehen, dass für eine Jordankette $x_0, x_1, \ldots, x_k$ von $P(\lambda)$ zum Eigenwert λ^* die k Lösungen der Differentialgleichung (2.24)

$$u_0(t) = x_0 e^{\lambda^* t}\,, \quad u_1(t) = (tx_0 + x_1)e^{\lambda^* t}\,, \quad \ldots\,, \quad u_k(t) = \left(\sum_{j=0}^{k}\frac{t^j}{j!}x_{k-j}\right)e^{\lambda^* t}$$

linear unabhängig sind.

Beispiel 5

Es wird das Matrixpolynom

$$P(\lambda) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\lambda^2 + \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix}\lambda + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \lambda^2 & 0 \\ \lambda+1 & \lambda(\lambda-1) \end{pmatrix}$$

aus Beispiel 1 mit den Eigenwerten $\lambda_0 = 0$ und $\lambda_1 = 1$ und jeweils zugehörigem Eigenvektor $x_0 = (\,0\;\;1\,)^T$ ($\alpha = 1$ in Beispiel 1) betrachtet. Es soll jeweils eine Jordankette zu den beiden Eigenwerten bestimmt werden.

Für $\lambda_0 = 0$ und $i = 1$ liefert (2.23) *das Gleichungssystem*

$$0 = P(0)x_1 + P'(0)x_0 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}x_1 + \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

für das jeder Vektor $x_1 = (\,1\;\;\beta\,)^T$, $\beta \in \mathbb{C}$, Lösung ist. Es wird hier $\beta = 0$ und damit $x_1 = (\,1\;\;0\,)^T$ gewählt. Für $i = 2$ liefert (2.23)

$$0 = P(0)x_2 + P'(0)x_1 + \frac{1}{2}P''(0)x_0 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}x_2 + \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

und dies wiederum den Vektor $x_2 = (\,-2\;\;\gamma\,)^T$, $\gamma \in \mathbb{C}$. Wir wählen $x_2 = (\,-2\;\;0\,)^T$. Für $i = 3$ erhält man wegen $P^{(3)}(\lambda) = O$ das Gleichungssystem

$$0 = P(0)x_3 + P'(0)x_2 + \frac{1}{2}P''(0)x_1 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}x_3 + \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} -2 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

welches nicht lösbar ist. Also bilden

$$x_0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix},\ x_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix},\ x_2 = \begin{pmatrix} -2 \\ 0 \end{pmatrix}$$

eine Jordankette der Länge 3 *für* $P(\lambda)$ *und* $\lambda_0 = 0$.

Für $\lambda_1 = 1$ *und* $i = 1$ *liefert* (2.23) *das Gleichungssystem*

$$0 = P(1)x_1 + P'(1)x_0 = \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix} x_1 + \begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

welches unlösbar ist. Damit ist $x_0 = (\,0\ \ 1\,)^T$ *eine Jordankette der Länge* 1 *für* $P(\lambda)$ *und* $\lambda_1 = 1$, *die nicht verlängert werden kann.*

Lösungen des zu $P(\lambda)$ *gehörigen homogenen Systems von Differentialgleichungen*

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} u''(t) + \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix} u'(t) + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} u(t) = 0$$

sind dann nach Satz 9

$$u_0(t) = \begin{pmatrix} 0 \\ 1 \end{pmatrix},\ u_1(t) = \begin{pmatrix} 1 \\ t \end{pmatrix},\ u_2(t) = \begin{pmatrix} t-2 \\ \frac{1}{2}t^2 \end{pmatrix} \qquad (\lambda^* = 0)$$

und

$$v_0(t) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} e^t \qquad (\lambda^* = 1)\,.$$

Diese vier Lösungen sind linear unabhängig.

2.4 Nullstellenpolynome

In diesem Abschnitt ist $P(\lambda)$ ein nicht notwendig reguläres $n \times n$–Matrixpolynom mit $\det P(\lambda) \not\equiv 0$, für das Eigenpaare und Jordanketten analog zum regulären Fall definiert werden können.

Außerdem sei ein n–komponentiges Vektorpolynom vom Grad l definiert durch

$$p(\lambda) = \sum_{i=0}^{l} v_i \lambda^i\,, \quad \lambda \in \mathbb{C},$$

mit Koeffizientenvektoren $v_0, \ldots, v_l \in \mathbb{C}^n$. Die komplexe Zahl λ^* heißt k–fache Nullstelle ($k \geq 1$) des Vektorpolynoms $p(\lambda)$, falls ein Vektorpolynom $\tilde{p}(\lambda)$ existiert mit $p(\lambda) = (\lambda - \lambda^*)^k \cdot \tilde{p}(\lambda)$ und $\tilde{p}(\lambda^*) \neq 0$.

Definition 6

$P(\lambda)$ sei ein (nicht notwendig reguläres) $n \times n$–Matrixpolynom. Ein n–komponentiges Vektorpolynom $\varphi(\lambda)$ mit $\varphi(\lambda^*) \neq 0$ und $P(\lambda^*)\varphi(\lambda^*) = 0$ heißt Nullstellenpolynom von $P(\lambda)$ zum Eigenwert λ^*. Wenn λ^* k–fache Nullstelle ($k \geq 1$) des Vektorpolynoms $P(\lambda)\varphi(\lambda)$ ist, heißt k die Ordnung von $\varphi(\lambda)$.

Man betrachtet nun die (endliche) Taylorentwicklung von $\varphi(\lambda)$ zur Entwicklungsstelle λ^*

$$\varphi(\lambda) = \sum_{j=0}^{q} (\lambda - \lambda^*)^j \varphi_j \,, \tag{2.28}$$

wobei $\varphi_0, \ldots, \varphi_q$ die entsprechenden Taylorkoeffizienten sind. Außerdem kann man auch $P(\lambda)$ nach λ^* entwickeln:

$$P(\lambda) = \sum_{j=0}^{l} \frac{1}{j!} P^{(j)}(\lambda^*)(\lambda - \lambda^*)^j \,. \tag{2.29}$$

Lemma 5

k sei die Ordnung des Nullstellenpolynoms $\varphi(\lambda)$ von $P(\lambda)$ zum Eigenwert λ^*. Dann bilden die Vektoren $\varphi_0\,,\ \varphi_1\,,\ \ldots\,,\ \varphi_{k-1}$ aus (2.28) eine Jordankette von $P(\lambda)$ zu λ^*.

Beweis:

$$P(\lambda)\varphi(\lambda) = \left[\sum_{j=0}^{l} \frac{1}{j!} P^{(j)}(\lambda^*)(\lambda - \lambda^*)^j\right] \left[\sum_{j=0}^{q} (\lambda - \lambda^*)^j \varphi_j\right] \,.$$

Man multipliziere nun die rechte Seite aus. Weil $P(\lambda)\varphi(\lambda) = (\lambda - \lambda^*)^k \tilde{\varphi}(\lambda)$ mit $\tilde{\varphi}(\lambda^*) \neq 0$ müssen die Koeffizienten bzgl. $(\lambda - \lambda^*)^j\,,\ j = 0, \ldots, k-1$, der rechten Seite gleich Null sein. Dies liefert nach Definition (2.23) die Behauptung. □

Bemerkung 3

Umgekehrt gilt: wenn $\varphi_0\,,\ \varphi_1\,,\ \ldots\,,\ \varphi_{k-1}$ eine Jordankette von $P(\lambda)$ zu λ^* bilden, ist ein Vektorpolynom vom Typ

$$\varphi(\lambda) = \sum_{j=0}^{k-1} (\lambda - \lambda^*)^j \varphi_j + (\lambda - \lambda^*)^k \psi(\lambda)$$

($\psi(\lambda)$ ein beliebiges Vektorpolynom) ein Nullstellenpolynom mindestens der Ordnung k von $P(\lambda)$ zu λ^*.

Dies gilt wegen $\varphi(\lambda^*) = \varphi_0$, (2.29) und der Definition (2.23) einer Jordankette.

Satz 10

$P(\lambda), A(\lambda)$ und $B(\lambda)$ seien $n \times n$–Matrixpolynome ($P(\lambda)$ nicht notwendig regulär). $A(\lambda^*)$ und $B(\lambda^*)$ seien invertierbar für $\lambda^* \in \mathbb{C}$.
Dann sind $y_0\,,\ \ldots\,,\ y_k$ eine Jordankette von $A(\lambda)P(\lambda)B(\lambda)$ zu λ^* genau dann, wenn die Vektoren

$$z_j = \sum_{i=0}^{j} \frac{1}{i!} B^{(i)}(\lambda^*) y_{j-i} \,, \quad j = 0, \ldots, k \tag{2.30}$$

eine Jordankette von $P(\lambda)$ zu λ^* bilden.

Beweis:

Vorbemerkung: Es sei $B(\lambda) = \sum_{i=0}^{p} \frac{1}{i!} B^{(i)}(\lambda^*)(\lambda - \lambda^*)^i$ die Taylor–Entwicklung von $B(\lambda)$ zu λ^* und $\psi(\lambda) := \sum_{j=0}^{k} (\lambda - \lambda^*)^j y_j$. Dann gilt wegen (2.30)

$$B(\lambda)\psi(\lambda) = \sum_{j=0}^{k} (\lambda - \lambda^*)^j z_j + (\lambda - \lambda^*)^{k+1} \theta_1(\lambda) =: \varphi(\lambda)\,.$$

Angenommen $z_0\,, \ldots\,, z_k$ aus (2.30) bilden eine Jordankette von $P(\lambda)$ zu λ^*. Dann ist $\varphi(\lambda)$ nach Bemerkung 3 ein Nullstellenpolynom von $P(\lambda)$ und $P(\lambda)\varphi(\lambda) = (\lambda - \lambda^*)^{k+1}\theta_2(\lambda)$ für ein Vektorpolynom $\theta_2(\lambda)$. Es gilt

$$[A(\lambda)P(\lambda)B(\lambda)]\psi(\lambda) = A(\lambda)[P(\lambda)\varphi(\lambda)] = (\lambda - \lambda^*)^{k+1} A(\lambda)\theta_2(\lambda)\,.$$

Wegen $\psi(\lambda^*) = y_0 = B(\lambda^*)^{-1} z_0 \neq 0$ und $[A(\lambda^*)P(\lambda^*)B(\lambda^*)]\psi(\lambda^*) = 0$ ist $\psi(\lambda)$ also ein Nullstellenpolynom von $A(\lambda)P(\lambda)B(\lambda)$ von mindestens Ordnung $k+1$. Also bilden $y_0\,, \ldots\,, y_k$ nach Lemma 5 eine Jordankette von $A(\lambda)P(\lambda)B(\lambda)$ zu λ^*.

Umgekehrt sei $y_0\,, \ldots\,, y_k$ eine Jordankette von $A(\lambda)P(\lambda)B(\lambda)$ zu λ^*. Dann ist nach Bemerkung 3 $\psi(\lambda)$ Nullstellenpolynom von $A(\lambda)P(\lambda)B(\lambda)$ mit

$$A(\lambda)P(\lambda)B(\lambda)\psi(\lambda) = (\lambda - \lambda^*)^{k+1}\theta_3(\lambda)$$

für ein Vektorpolynom $\theta_3(\lambda)$. Für $\varphi(\lambda) = B(\lambda)\psi(\lambda)$ gilt dann

$$P(\lambda)\varphi(\lambda) = (\lambda - \lambda^*)^{k+1} A^{-1}(\lambda)\theta_3(\lambda)\,,$$

wobei $P(\lambda)\varphi(\lambda)$ ein Vektorpolynom ist ($A^{-1}(\lambda)$ ist rationale Matrixfunktion). Weil $A(\lambda^*)$ nichtsingulär ist, ist die rechte Seite ein Vektorpolynom mit der Nullstelle $\lambda = \lambda^*$ von mindestens Vielfachheit $k+1$. Wegen $\varphi(\lambda^*) = B(\lambda^*)y_0 \neq 0$ und $P(\lambda^*)\varphi(\lambda^*) = 0$ ist $\varphi(\lambda)$ dann ein Nullstellenpolynom von $P(\lambda)$ von mindestens Ordnung $k+1$. Nach Lemma 5 sind dann $z_0\,, \ldots\,, z_k$ eine Jordankette von $P(\lambda)$ zu λ^*. □

2.5 Kanonische Menge von Jordanketten

Es wird nun eine kanonische Menge von Jordanketten eines regulären Matrixpolynoms $P(\lambda)$ konstruiert. Jedes Matrixpolynom besitzt also eine kanonische Menge von Jordanketten (die jedoch nicht eindeutig bestimmt ist, wie später gezeigt wird).

Es sei daran erinnert, dass ein reguläres Matrixpolynom $P(\lambda)$ wegen $\det A_l \neq 0$ höchstens nl verschiedene Eigenwerte besitzt.

Lemma 6

Die Ordnung k eines beliebigen Nullstellenpolynoms von $P(\lambda)$ zum Eigenwert λ^* ist höchstens so groß wie die algebraische Vielfachheit von λ^*.

Beweis:

Wegen Satz 10 in Verbindung mit Lemma 5 und Bemerkung 3 genügt es, dies für die lokale Smith Form $S(\lambda) = \mathrm{diag}((\lambda-\lambda^*)^{\nu_1}, \ldots, (\lambda-\lambda^*)^{\nu_n})$, $0 \le \nu_1 \le \ldots \le \nu_i \le \ldots \le \nu_n$, (siehe (2.16)) zu beweisen. Es ist $\nu_n \ge 1$ weil λ^* Eigenwert von $P(\lambda)$ ist, d.h. mindestens ein linearer Elementarteiler von $P(\lambda)$ zu λ^* existiert. Ein beliebiges Nullstellenpolynom $\varphi(\lambda)$ von $S(\lambda)$ besitzt mindestens die Ordnung 1. Der Index $i = i^* \in \{1, \ldots, n\}$ sei der kleinste, für den $\varphi(\lambda^*)_i \neq 0$ gilt (existiert wegen $\varphi(\lambda^*) \neq 0$). Dann gilt also

$$\begin{aligned} S(\lambda)\varphi(\lambda) &= ((\lambda-\lambda^*)^{\nu_1}\varphi(\lambda)_1, \ldots, (\lambda-\lambda^*)^{\nu_i}\varphi(\lambda)_i, \ldots, (\lambda-\lambda^*)^{\nu_n}\varphi(\lambda)_n)^T \\ &= (\lambda-\lambda^*)^{\nu_{i^*}}\tilde{\varphi}(\lambda) \end{aligned}$$

mit $\tilde{\varphi}(\lambda^*) \neq 0$, also $1 \le k = \nu_{i^*} \le \nu_1 + \ldots + \nu_n$ (wobei $\det S(\lambda) = (\lambda-\lambda^*)^{\nu_1+\ldots+\nu_n}$). □

Korollar 1

Die Länge einer Jordankette von $P(\lambda)$ zum Eigenwert λ^* ist höchstens so groß wie die algebraische Vielfachheit von λ^*.

Beweis:

σ sei die algebraische Vielfachheit von λ^*.

Angenommen es gibt eine Jordankette von $P(\lambda)$ zum Eigenwert λ^* der Länge $k > \sigma$. Dann existiert nach Bemerkung 3 ein Nullstellenpolynom mindestens der Ordnung $k > \sigma$, was Lemma 6 widerspricht. □

λ^* sei ein fester Eigenwert von $P(\lambda)$. In der anschließenden Konstruktion einer kanonischen Menge von Jordanketten gehören alle Nullstellenpolynome zu diesem λ^*.

Es sei $\varphi_1(\lambda) = \sum_{j=0}^{\kappa_1-1}(\lambda-\lambda^*)^j\varphi_{1j}$ ein Nullstellenpolynom größter Ordnung κ_1. Aus Lemma 6 folgt, dass die Ordnungen der Nullstellenpolynome nach oben beschränkt sind, also existiert solch ein $\varphi_1(\lambda)$. Lemma 5 liefert, dass φ_{10} Eigenvektor zu λ^* ist. $\varphi_2(\lambda) = \sum_{j=0}^{\kappa_2-1}(\lambda-\lambda^*)^j\varphi_{2j}$ sei ein Nullstellenpolynom größter Ordnung unter allen Nullstellenpolynomen, deren Eigenvektoren nicht ein skalares Vielfaches von φ_{10} sind. Insbesondere gilt $\kappa_2 \le \kappa_1$.

Wenn $\varphi_1(\lambda), \ldots, \varphi_{s-1}(\lambda)$ bereits gewählt wurden, wobei

$$\varphi_i(\lambda) = \sum_{j=0}^{\kappa_i-1}(\lambda-\lambda^*)^j\varphi_{ij}\,, \quad i = 1, \ldots, s-1,$$

dann sei $\varphi_s(\lambda) = \sum_{j=0}^{\kappa_s-1}(\lambda-\lambda^*)^j\varphi_{sj}$ ein Nullenstellenpolynom größter Ordnung κ_s unter allen Nullstellenpolynomen, deren Eigenvektor kein Element der linearen Hülle der Eigenvektoren $\varphi_{10}, \ldots, \varphi_{s-1,0}$ ist.

Dieser Prozess wird weitergeführt bis die Menge $\mathrm{Ker}\, P(\lambda^*)$ aller Eigenvektoren von $P(\lambda)$ zu λ^* erschöpft ist. Also wurden r Nullstellenpolynome

$$\varphi_i(\lambda) = \sum_{j=0}^{\kappa_i-1}(\lambda-\lambda^*)^j\varphi_{ij}\,, \quad i = 1, \ldots, r,$$

konstruiert, wobei $r = \dim(\operatorname{Ker} P(\lambda^*))$.

Man sagt dann, dass die Jordanketten (siehe Lemma 5)

$$\varphi_{10}, \ldots, \varphi_{1,\kappa_1-1}\,, \quad \varphi_{20}, \ldots, \varphi_{2,\kappa_2-1}\,, \quad \varphi_{r0}, \ldots, \varphi_{r,\kappa_r-1}$$

eine kanonische Menge von Jordanketten von $P(\lambda)$ zu λ^* bilden.

Solch eine kanonische Menge von Jordanketten ist nicht eindeutig. Man kann z.B. $\varphi_2(\lambda)$ durch $\varphi_2(\lambda) + \sum_{j=0}^{\kappa_2-1}(\lambda-\lambda^*)^j\varphi_{1j}$ in der obigen Konstruktion ersetzen und erhält eine andere kanonische Menge. Es wird nun gezeigt, dass die Zahlen $\kappa_1, \ldots, \kappa_r$ eindeutig bestimmt sind, also nicht von der Wahl einer kanonischen Menge von Jordanketten von $P(\lambda)$ zu λ^* abhängen.

Satz 11

Es sei $P(\lambda)$ ein reguläres $n \times n$–Matrixpolynom.
Dann sind die Längen $\kappa_1, \ldots, \kappa_r$ der Jordanketten in einer kanonischen Menge von Jordanketten von $P(\lambda)$ zum Eigenwert λ^* genau die partiellen Vielfachheiten ungleich Null von $P(\lambda)$ zu λ^*.

Beweis:

Es wird die lokale Smith Form von $P(\lambda)$ verwendet:

$$P(\lambda) = E_{\lambda^*}(\lambda)D_{\lambda^*}(\lambda)F_{\lambda^*}(\lambda),$$

wobei $E_{\lambda^*}(\lambda)$ und $F_{\lambda^*}(\lambda)$ Matrixpolynome sind, die für λ^* invertierbar sind,

$$D_{\lambda^*}(\lambda) = \operatorname{diag}((\lambda-\lambda^*)^{\nu_i})_{i=1}^n$$

und $0 \le \nu_1 \le \ldots \le \nu_n$ die partiellen Vielfachheiten von $P(\lambda)$ zu λ^* sind. Eine kanonische Menge von Jordanketten von $D_{\lambda^*}(\lambda)$ (die genauso definiert ist wie für Matrixpolynome $\sum_{i=0}^{l} A_i\lambda^i$ mit $\det A_l \neq 0$) kann man leicht auf die zuvor beschriebene Art konstruieren:
wenn $0 = \nu_1 = \ldots = \nu_{i_0} < \nu_{i_0+1}$ gilt (wobei $\nu_n \ge 1$ ist), dann ist $\dim(\operatorname{Ker} D_{\lambda^*}(\lambda^*)) = n - i_0$.

Es gilt $D_{\lambda^*}(\lambda)e_j = (\lambda-\lambda^*)^{\nu_j}e_j$ für $j = i_0+1, \ldots, n$ (e_j bezeichnet den j–ten Einheitsvektor im $\mathbb{C}^n$). Also ist $\varphi_i(\lambda) = e_{n+1-i}$ ein Nullstellenpolynom mit Ordnung ν_{n+1-i} für $i = 1, \ldots, n-i_0$. Dann ist

$$e_n, 0, \ldots, 0, \qquad e_{n-1}, 0, \ldots, 0, \qquad \ldots, \qquad e_{i_0+1}, 0, \ldots, 0$$

eine kanonische Menge, wobei die Jordankette $e_j, 0, \ldots, 0$ die Länge ν_j besitzt ($j = i_0+1, \ldots, n$). Analog kann man sich überlegen, dass $\varphi(\lambda)$ genau dann ein Nullstellenpolynom von $D_{\lambda^*}(\lambda)$ ist, wenn $\varphi(\lambda^*) = (0, \ldots, 0, \alpha_i^{(i)}, \alpha_{i+1}^{(i)}, \ldots, \alpha_n^{(i)})^T =: v^{(i)}$ mit $\alpha_i^{(i)} \neq 0$ und $i \in \{i_0+1, \ldots, n\}$ (und Ordnung ν_i). Also hat jede kanonische Menge von Jordanketten von $D_{\lambda^*}(\lambda)$ die Gestalt

$$v^{(i)},\ v_{i,1},\ \ldots,\ v_{i,\nu_i-1}, \quad i = n, \ldots, i_0+1$$

mit geeigneten Vektoren $v_{i,1},\ \ldots,\ v_{i,\nu_i-1}$. Somit stimmt die Behauptung des Satzes für $D_{\lambda^*}(\lambda)$.

Desweiteren ist das System

$$\psi_{i0}, \ldots, \psi_{i,\kappa_i-1}\,, \quad i = 1, \ldots, r$$

genau dann eine kanonische Menge von Jordanketten von $P(\lambda)$ zu λ^*, wenn das System

$$\varphi_{i0}, \ldots, \varphi_{i,\kappa_i-1}\,, \quad i = 1, \ldots, r$$

eine kanonische Menge von Jordanketten von $D_{\lambda^*}(\lambda)$ zu λ^* ist, wobei

$$\varphi_{ij} = \sum_{m=0}^{j} \frac{1}{m!} F_{\lambda^*}^{(m)}(\lambda^*)\psi_{i,j-m}\,, \quad j = 0, \ldots, \kappa_i - 1\,, \quad i = 1, \ldots, r\,.$$

Dies folgt aus Satz 10, der Definition einer kanonischen Menge von Jordanketten und unter Beachtung von $\varphi_{i0} = F_{\lambda^*}(\lambda^*)\psi_{i0}\,,\ i = 1, \ldots, r$ (d.h. $\psi_{10}, \ldots, \psi_{r0}$ sind linear unabhängig genau dann, wenn $\varphi_{10}, \ldots, \varphi_{r0}$ linear unabhängig sind). Die Längen der Jordanketten einer kanonischen Menge von $P(\lambda)$ und $D_{\lambda^*}(\lambda)$ zu λ^* sind also identisch, und damit folgt die Behauptung des Satzes. □

Mit Bemerkung 2 erhält man dann das folgende Korollar.

Korollar 2

$P(\lambda)$ sei ein reguläres $n \times n$–Matrixpolynom. Dann stimmt die Summe $\sum_{i=1}^{r} \kappa_i$ der Längen der Jordanketten in einer kanonischen Menge von Jordanketten zum Eigenwert λ^* überein mit der algebraischen Vielfachheit von λ^*.

Der folgende Satz zeigt, dass ein kanonisches System die Rolle einer Basis für die Menge aller Jordanketten von $P(\lambda)$ zum Eigenwert λ^* spielt.

Es sei μ die Länge der längsten Jordankette von $P(\lambda)$ zu λ^* (siehe Korollar 1). Der Unterraum $\mathcal{N} \subset \mathbb{C}^{n\mu}$ bestehe aus allen Vektoren $(y_0^T, \ldots, y_{\mu-1}^T)^T$ (wobei $y_i \in \mathbb{C}^n\,,\ i = 0, \ldots, \mu - 1$), für die

$$\begin{pmatrix} P(\lambda^*) & 0 & \ldots & 0 \\ P'(\lambda^*) & P(\lambda^*) & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ \frac{1}{(\mu-1)!}P^{(\mu-1)}(\lambda^*) & \frac{1}{(\mu-2)!}P^{(\mu-2)}(\lambda^*) & \ldots & P(\lambda^*) \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{\mu-1} \end{pmatrix} = 0 \tag{2.31}$$

gilt. Wenn man sich die Definition (2.23) von Jordanketten von $P(\lambda)$ zu λ^* in Erinnerung ruft, ist es offensichtlich, dass $\mathcal{N}$ aus allen Jordanketten von $P(\lambda)$ zu λ^* besteht (wenn man, falls vorhanden, die ersten Nullvektoren $y_0 = \ldots = y_i = 0$ in $(y_0^T, \ldots, y_{\mu-1}^T)^T \in \mathcal{N}$ weglässt).

Satz 12

$$\varphi_{i0}, \ldots, \varphi_{i,\mu_i-1}\,, \quad i = 1, \ldots, s \tag{2.32}$$

sei eine Menge von Jordanketten des regulären $n \times n$–Matrixpolynoms $P(\lambda)$ zum Eigenwert λ^*. Dann sind äquivalent:

(i) die Menge (2.32) ist kanonisch;

(ii) die Eigenvektoren $\varphi_{10}, \ldots, \varphi_{s0}$ sind linear unabhängig und $\sum_{i=1}^{s} \mu_i = \sigma$, wobei σ die algebraische Vielfacheit von λ^* ist;

(iii) die Vektoren

$$\gamma_{ij} = (0^T, \ldots, 0^T, \varphi_{i0}^T, \ldots, \varphi_{ij}^T)^T\,, \quad j = 0, \ldots, \mu_i - 1\,, \quad i = 1, \ldots, s, \tag{2.33}$$

wobei $\mu - (j+1)$ die Anzahl der Nullvektoren vor φ_{i0} in γ_{ij} ist, bilden eine Basis für $\mathcal{N}$.

Beweis:

Der Teil $(i) \Rightarrow (ii)$ folgt aus der Definition einer kanonischen Menge von Jordanketten und aus Korollar 2.

Wie gleich gezeigt wird, genügt es, $(ii) \Rightarrow (iii)$ und $(iii) \Rightarrow (i)$ für ein spezielles diagonales Matrixpolynom zu beweisen.

Für die Matrix aus (2.31) wird als Abkürzung $\mathbf{P}$ geschrieben. Analog wird mit $\mathbf{A}$ bzw. $\mathbf{B}$ die entsprechende Matrix für das Matrixpolynom $A(\lambda)$ bzw. $B(\lambda)$ (ebenfalls an der Stelle λ^*) bezeichnet. $A(\lambda)$ und $B(\lambda)$ seien nichtsingulär für $\lambda = \lambda^*$. Dann sind auch $\mathbf{A}$ und $\mathbf{B}$ nichtsingulär (wegen $\det \mathbf{A} = (\det A(\lambda^*))^\mu \neq \mathbf{0}$, analog für $\mathbf{B}$). Desweiteren sei $\tilde{\mathcal{N}} \subset \mathbb{C}^{n\mu}$ der Unterraum, der durch (2.31) definiert ist, wenn dort $P(\lambda)$ durch $\tilde{P}(\lambda) = A(\lambda)P(\lambda)B(\lambda)$ ersetzt wird ($\tilde{\mathbf{P}}$ sei die entsprechende Matrix). Es gilt

$$\tilde{\mathbf{P}} = \mathbf{APB},$$

was man durch direktes Ausrechnen zeigen kann. Außerdem gilt

$$\mathcal{N} = \mathbf{B}\tilde{\mathcal{N}} \tag{2.34}$$

(nach Satz 10 ist μ auch die Länge der längsten Jordankette von $\tilde{P}(\lambda)$ zu λ^*):

$$\begin{aligned} z \in \mathcal{N} \quad &\Leftrightarrow \quad \mathbf{P}z = 0 \quad \Leftrightarrow \quad \mathbf{A^{-1}\tilde{P}B^{-1}}z = 0 \quad \Leftrightarrow \quad \mathbf{\tilde{P}B^{-1}}z = 0 \\ &\Leftrightarrow \quad \tilde{z} := \mathbf{B^{-1}}z \in \tilde{\mathcal{N}} \quad \Leftrightarrow \quad z \in \mathbf{B}\tilde{\mathcal{N}}. \end{aligned}$$

Aus der lokalen Smith Form (2.16) von $P(\lambda)$ zu λ^* und aus (2.34) folgt die Reduktion auf den Fall

$$D(\lambda) = \operatorname{diag}((\lambda - \lambda^*)^{\nu_i})_{i=1}^n\,, \qquad \nu_1 \geq \ldots \geq \nu_n \geq 0\,. \tag{2.35}$$

Also müssen die verbliebenen Aussagen von Satz 12 nur noch für den Fall (2.35) bewiesen werden.

Zuerst soll $(ii) \Rightarrow (iii)$ gezeigt werden. Für die Ketten (2.32) gelte also, dass $\varphi_{10}, \ldots, \varphi_{s0}$ linear unabhängig sind und $\sum_{i=1}^s \mu_i = \sigma$ erfüllt ist. Die Vektoren

$$\gamma_{ij} = (0^T, \ldots, 0^T, \varphi_{i0}^T, \ldots, \varphi_{ij}^T)^T\,, \quad j = 0, \ldots, \mu_i - 1\,, \quad i = 1, \ldots, s$$

sind dann Elemente von $\mathcal{N}$ und linear unabhängig (da die Vektoren $\varphi_{10}, \ldots, \varphi_{s0}$ linear unabhängig sind). Wegen der speziellen Gestalt (2.35) von $D(\lambda)$ ist leicht zu sehen (man betrachte die Anzahl der Nullzeilen in der Matrix (2.31); $(\lambda - \lambda^*)^{\nu_i}$ induziert Nullzeilen in den Positionen $i + s \cdot n$, $s = 0, 1, \ldots, \nu_i - 1$), dass

$$\dim \mathcal{N} = \sum_{j=1}^n \nu_j\,,$$

was mit der algebraischen Vielfachheit σ von λ^* ($\det D(\lambda) = (\lambda - \lambda^*)^{\sum_{j=1}^n \nu_j}$) übereinstimmt. Damit gilt nach Voraussetzung $\dim \mathcal{N} = \sum_{i=1}^s \mu_i$. Also bilden die γ_{ij}, $j = 0, \ldots, \mu_i - 1$, $i = 1, \ldots, s$, eine Basis von $\mathcal{N}$ und damit ist (iii) bewiesen.

Es verbleibt, $(iii) \Rightarrow (i)$ zu beweisen. O.B.d.A. kann man annehmen, dass die μ_i absteigend geordnet sind: $\mu_1 \geq \ldots \geq \mu_s$. Desweiteren nehmen wir an, dass für die partiellen Vielfachheiten ν_i aus (2.35) $\nu_r > \nu_{r+1} = \ldots = \nu_n = 0$ gilt. Wir wollen nun zeigen, dass $s = r$ und $\mu_i = \nu_i$ für $i = 1, \ldots, s$ gilt.

Die Dimension der Unterraums, der von allen Vektoren aus $\mathcal{N}$ der Form $(0^T, \ldots, 0^T, x^T)^T$ mit $x \in \mathbb{C}^n$ aufgepannt wird, ist r. Wegen der speziellen Gestalt (2.35) von $D(\lambda)$ sieht man dies leicht über die Definition von $\mathcal{N}$. Also ist $s = r$

Es wird nun die Dimension des Unterraumes U_j betrachtet, der von allen Vektoren aus $\mathcal{N}$ der Form $(0^T, \ldots, 0^T, x_0^T, \ldots, x_{j-1}^T)^T$ aufgespannt wird.

e_i sei der i–te Einheitsvektor im $\mathbb{C}^n$. Es ist wegen der einfachen Gestalt (2.35) von $D(\lambda)$ leicht zu sehen, dass $\mathrm{Ker}(D(\lambda^*)) = \mathrm{span}\{e_i \,|\, i = 1, \ldots, r\}$. Es ist

$$e_i, 0, \ldots, 0 \quad (k \text{ Nullvektoren, wobei } 0 \le k \le \nu_i - 1) \tag{2.36}$$

für $i \in \{1, \ldots, r\}$ eine Jordankette von $D(\lambda)$.

Angenommen, es gäbe eine längere Jordankette $e_i, v_1, \ldots, v_q$ mit $q \ge \nu_i$. Dann existiert nach Bemerkung 3 ein Nullstellenpolynom $\varphi(\lambda) = e_i + \sum_{j=1}^{q} (\lambda - \lambda^*)^j v_j$ mindestens der Ordnung $q + 1 > \nu_i$ von $D(\lambda)$, d.h.

$$(D(\lambda)\varphi(\lambda))^{(\nu_i)}(\lambda^*) = 0\,.$$

Dies ist aber ein Widerspruch zu $(D(\lambda)\varphi(\lambda))_i^{(\nu_i)}(\lambda^*) \neq 0$ (wegen $(D(\lambda)\varphi(\lambda))_i = (\lambda - \lambda^*)^{\nu_i} + \sum_{j=1}^{q} (\lambda - \lambda^*)^{j+\nu_i} (v_j)_i$).

Also sind für $i = 1, \ldots, r$ die Jordanketten (2.36) von $D(\lambda)$ höchstens von der Länge ν_i . Weil $\mathcal{N}$ genau aus den Jordanketten von $D(\lambda)$ zu λ^* besteht, gilt dann

$$\dim(U_j) = \sum_{p \ge j} j \cdot |\{i \,|\, \nu_i = p\}| + \sum_{p=1}^{j-1} p \cdot |\{i \,|\, \nu_i = p\}|\,, \quad j = 1, \ldots, n\,, \tag{2.37}$$

wobei $|M|$ die Mächtigkeit der endlichen Menge M bezeichnet.

Andererseits ist in (iii) vorausgesetzt, dass (2.33) eine Basis von $\mathcal{N}$ ist. Dies liefert

$$\dim(U_j) = \sum_{p \ge j} j \cdot |\{i \,|\, \mu_i = p\}| + \sum_{p=1}^{j-1} p \cdot |\{i \,|\, \mu_i = p\}|\,, \quad j = 1, \ldots, n\,. \tag{2.38}$$

Vergleicht man (2.37) und (2.38), erhält man $\mu_i = \nu_i$, $i = 1, \ldots, s = r$. (i) folgt nun aus der Definition einer kanonischen Menge von Jordanketten. □

Satz 13

$$\varphi_{i0}, \ldots, \varphi_{i,\mu_i - 1}\,, \quad i = 1, \ldots, s \tag{2.39}$$

sei eine Menge von Jordanketten des regulären $n \times n$–Matrixpolynoms $P(\lambda)$ zum Eigenwert λ^* mit linear unabhängigen Eigenvektoren $\varphi_{10}, \ldots, \varphi_{s0}$. Dann gilt

$$\sum_{i=1}^{s} \mu_i \le \sigma\,, \tag{2.40}$$

wobei σ die algebraische Vielfacheit von λ^* ist. In (2.40) *gilt Gleichheit genau dann, wenn die Menge* (2.39) *kanonisch ist.*

Beweis:

Der Beweis nutzt die gleichen Ideen wie der Beweis von Satz 12 . □

Beispiel 6

Es wird das reguläre quadratische 2×2*–Matrixpolynom*

$$P(\lambda) = \begin{pmatrix} \lambda^2 & -\lambda \\ 0 & \lambda^2 \end{pmatrix}$$

mit $\det P(\lambda) = \lambda^4$ *betrachtet. Die algebraische Vielfachheit des einzigen Eigenwertes* $\lambda^* = 0$ *ist also* $\sigma = 4$ *. Wegen*

$$P(0) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad P'(0) = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}, \quad \frac{1}{2} P''(0) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

ist leicht nachzurechnen (siehe (2.23)*), dass*

$$x_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \; x_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, x_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

eine Jordankette der Länge $\mu_1 = 3$ *zu* $\lambda^* = 0$ *und*

$$y_0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

eine Jordankette der Länge $\mu_2 = 1$ *zu* $\lambda^* = 0$ *ist. Satz* 12 *liefert nun, dass diese beiden Jordanketten eine kanonische Menge von Jordanketten zu* $\lambda^* = 0$ *bilden.*

Korollar 3

λ^* sei ein einfacher Eigenwert eines regulären $n \times n$–Matrixpolynoms $P(\lambda)$.
Dann gibt es bis auf skalare Vielfache nur einen Eigenvektor von $P(\lambda)$ zu λ^* und keine Jordankette der Länge $k > 1$ (d.h. es existieren keine verallgemeinerten Eigenvektoren von $P(\lambda)$ zu λ^*).

Beweis:

Weil λ^* geometrisch einfach ist, existiert bis auf skalare Vielfache genau ein Eigenvektor x^0 zu λ^* . Es gibt also eine Jordankette $x^0, x^1, \ldots, x^{k-1}$ der Länge $k \geq 1$. Wegen der Einfachheit von λ^* und Satz 13 existieren aber für $P(\lambda)$ und λ^* außer der Jordankette x^0 der Länge 1 keine weiteren Jordanketten. □

2.6 Äquivalenzaussage

Der folgende Satz ist ein eigenes Resultat, dessen Beweisführung sich an [26] anlehnt. Er ist später in dieser Dissertation von Bedeutung.

Es gelten die Bezeichnungen: für $A \in \mathbb{C}^{n \times n}$ sei $A_{*,j}$ die j–te Spalte von A und e_s sei der s–te Einheitsvektor im $\mathbb{C}^n$.

Satz 14

(x^*, λ^*) sei Eigenpaar eines regulären $n \times n$–Matrixpolynoms $P(\lambda)$ und es gelte $(x^*)_s \neq 0$. Dann sind äquivalent:

i) λ^* *ist ein einfacher Eigenwert von* $P(\lambda)$.

ii) $B^* := \begin{pmatrix} P(\lambda^*) & P'(\lambda^*)x^* \\ {e_s}^T & 0 \end{pmatrix} \in \mathbb{C}^{(n+1)\times(n+1)}$ *ist nichtsingulär.*

iii) $B^{**} := (P(\lambda^*)_{*,1}, \ldots, P(\lambda^*)_{*,s-1}, P(\lambda^*)_{*,s+1}, \ldots, P(\lambda^*)_{*,n}, P'(\lambda^*)x^*) \in \mathbb{C}^{n\times n}$ *ist nichtsingulär.*

Beweis:

i) $\Rightarrow$ ii):

Angenommen B^* ist singulär. Dann existiert ein $z = (z_1, \ldots, z_{n+1})^T \in \mathbb{C}^{n+1}, z \neq 0$, mit

$$B^* z = 0\,. \tag{2.41}$$

Wir definieren nun $\tilde{z} := (z_1, \ldots, z_n)^T \in \mathbb{C}^n$. (2.41) liefert dann

$$\begin{aligned} P(\lambda^*)\tilde{z} + z_{n+1}P'(\lambda^*)x^* &= 0 \qquad \text{und} \\ z_s &= 0\,. \end{aligned} \tag{2.42}$$

1. Fall: $z_{n+1} = 0$. Dann ist $\tilde{z} \neq 0$ wegen $z \neq 0$. (2.42) liefert somit $P(\lambda^*)\tilde{z} = 0$, d.h. $\tilde{z}$ ist ein Eigenvektor von $P(\lambda)$ zum Eigenwert λ^*. Wegen $(\tilde{z})_s = z_s = 0$ und $(x^*)_s \neq 0$ sind $\tilde{z}$ und x^* also linear unabhängige Eigenvektoren von $P(\lambda)$ zu λ^*. Dies widerspricht nach Korollar 3 der Einfachheit von λ^*.

2. Fall: $z_{n+1} \neq 0$. Dann ist (2.42) äquivalent zu $P(\lambda^*)\hat{z} + P'(\lambda^*)x^* = 0$, wobei $\hat{z} := \frac{1}{z_{n+1}}\tilde{z}$. Es existiert also eine Jordankette mindestens der Länge 2. Dies widerspricht nach Korollar 3 der Einfachheit von λ^*.

ii) $\Rightarrow$ iii):

B^* ist nichtsingulär, d.h. $\det(B^*) \neq 0$. Wenn man $\det(B^*)$ nach der $(n+1)$–ten Zeile entwickelt, erhält man

$$\det(B^*) = (-1)^{n+1+s}\det(B^{**}),$$

also ist auch $\det(B^{**}) \neq 0$ und somit B^{**} nichtsingulär.

iii) $\Rightarrow$ i):

Angenommen λ^* ist ein algebraisch t–facher Eigenwert von $P(\lambda)$ mit $t \geq 2$.

1. Fall: λ^* ist geometrisch einfach. Dann gibt es (bis auf skalare Vielfache) nur den Eigenvektor x^* zu λ^*. Nach Satz 13 besteht dann eine kanonische Menge von $P(\lambda)$ zu λ^* genau aus einer Jordankette $x^*, x_1, \ldots, x_{t-1}$ der Länge $t \geq 2$. Es gilt also

$$\begin{aligned} P(\lambda^*)x^* &= 0\,, \\ P(\lambda^*)x_1 + P'(\lambda^*)x^* &= 0\,. \end{aligned} \tag{2.43}$$

Wegen $\det(P(\lambda^*)) = 0$ gibt es einen Vektor $0 \neq y^* \in \mathbb{C}^n$ mit $(y^*)^H P(\lambda^*) = 0$. Aus (2.43) folgt

$$0 = (y^*)^H(P(\lambda^*)x_1 + P'(\lambda^*)x^*) = (y^*)^H P(\lambda^*)x_1 + (y^*)^H P'(\lambda^*)x^* = (y^*)^H P'(\lambda^*)x^*\,.$$

Also haben wir

$$(y^*)^H B^{**} = ((y^*)^H P(\lambda^*)_{*,1}, \ldots, (y^*)^H P(\lambda^*)_{*,n}, (y^*)^H P'(\lambda^*)x^*) = 0\,.$$

Wegen $y^* \neq 0$ ist damit B^{**} singulär, Widerspruch zu (iii).

2. Fall: λ^* ist ein geometrisch mehrfacher Eigenwert. Dann existieren mindestens zwei linear unabhängige Eigenvektoren x^* und z^* von $P(\lambda)$ zu λ^*. Es gilt also $P(\lambda^*)x^* = 0$ und $P(\lambda^*)z^* = 0$ für die linear unabhängigen Vektoren x^* und z^*, d.h.

$$\operatorname{rang}(P(\lambda^*)) = \dim(\operatorname{span}\{P(\lambda^*)_{*,1}, \ldots, P(\lambda^*)_{*,n}\}) \leq n-2\,.$$

Es folgt

$$\dim(\operatorname{span}\{P(\lambda^*)_{*,1}, \ldots, P(\lambda^*)_{*,s-1}, P(\lambda^*)_{*,s+1}, \ldots, P(\lambda^*)_{*,n}\}) \leq n-2\,,$$

also

$$\operatorname{rang}(B^{**}) = \dim(\operatorname{span}\{P(\lambda^*)_{*,1}, \ldots, P(\lambda^*)_{*,s-1}, P(\lambda^*)_{*,s+1}, \ldots, P(\lambda^*)_{*,n}, P'(\lambda^*)x^*\}) \leq n-1\,.$$

B^{**} ist damit singulär, was ein Widerspruch zur Voraussetzung (iii) ist. □

2.7 Das quadratische Eigenwertproblem

Es wird im Folgenden das reguläre quadratische Matrixpolynom

$$P(\lambda) = A_2\lambda^2 + A_1\lambda + A_0 \tag{2.44}$$

mit Koeffizientenmatrizen $A_2, A_1, A_0 \in \mathbb{C}^{n\times n}$ und $\det A_2 \neq 0$ betrachtet. Das quadratische Eigenwertproblem (QEP) liegt nun darin, Eigenwerte $\lambda^* \in \mathbb{C}$ und zugehörige Eigenvektoren $x^* \in \mathbb{C}^n$ von $P(\lambda)$ zu finden:

$$P(\lambda^*)x^* = (A_2\lambda^{*2} + A_1\lambda^* + A_0)x^* = 0\,.$$

Das QEP besitzt höchstens $2n$ verschiedene Eigenwerte. Wenn die Koeffizientenmatrizen A_0, A_1 und A_2 reell sind, gilt Satz 1. Es ist

$$C_P = \begin{pmatrix} O & I_n \\ -A_2^{-1}A_0 & -A_2^{-1}A_1 \end{pmatrix} \in \mathbb{C}^{2n\times 2n}$$

die Begleitmatrix (2.44) von $P(\lambda)$.

Satz 15

λ^* ist Eigenwert von $P(\lambda)$ mit Eigenvektor x^* genau dann, wenn λ^* Eigenwert der Linearisierung $A = TC_PT^{-1} \in \mathbb{C}^{2n\times 2n}$ von $P(\lambda)$ mit Eigenvektor $T\begin{pmatrix} x^* \\ \lambda^* x^* \end{pmatrix} \in \mathbb{C}^{2n}$ ist.

Beweis:

Nach Satz 3 wissen wir, dass λ^* genau dann Eigenwert von $P(\lambda)$ ist, wenn λ^* Eigenwert von A ist. Außerdem wissen wir, dass jede Linearisierung A von $P(\lambda)$ ähnlich zu C_P ist (Satz 4). A und C_P besitzen das gleiche Spektrum. Weiter gilt

$$\begin{aligned}(\lambda^* I_{2n} - A)Ty &= (\lambda^* TT^{-1} - TC_P T^{-1})Ty = T(\lambda^* I_{2n} - C_P)T^{-1}Ty \\ &= T(\lambda^* I_{2n} - C_P)y\,,\end{aligned}$$

d.h. Ty ist Eigenvektor zum Eigenwert λ^* von A genau dann, wenn y Eigenvektor zum Eigenwert λ^* von C_P ist. $y =: (y_1^T, y_2^T)^T$ wiederum ist genau dann Eigenvektor von C_P zu λ^*, wenn

$$0 = (\lambda^* I_{2n} - C_P)y = \lambda^* \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - C_P \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \lambda^* y_1 - y_2 \\ \lambda^* y_2 + A_2^{-1}A_0 y_1 + A_2^{-1}A_1 y_2 \end{pmatrix}$$

erfüllt ist. Dies ist äquivalent zu $y_2 = \lambda^* y_1$ und $0 = \lambda^* y_2 + A_2^{-1}A_0 y_1 + A_2^{-1}A_1 y_2 = A_2^{-1}(A_2\lambda^* y_2 + A_1 y_2 + A_0 y_1)$. Dies wiederum ist genau dann der Fall, wenn $P(\lambda^*)y_1 = 0$, also y_1 Eigenvektor von $P(\lambda)$ zu λ^* ist.

Mit $y_1 = x^*$ ist somit die Behauptung des Satzes gezeigt. □

Bemerkung 4

λ^* ist Eigenwert von $P(\lambda)$ mit Eigenvektor x^* genau dann, wenn λ^* Eigenwert der Begleitmatrix $C_P \in \mathbb{C}^{2n\times 2n}$ mit Eigenvektor $\begin{pmatrix} x^* \\ \lambda^* x^* \end{pmatrix} \in \mathbb{C}^{2n}$ ist.

Nun wird der Begriff Linearisierung (2.1) verallgemeinert.

Definition 7

Das lineare $2n \times 2n$–Matrixpolynom $A - \lambda B$ heißt Linearisierung des quadratischen $n \times n$–Matrixpolynoms $P(\lambda)$, falls

$$\begin{pmatrix} P(\lambda) & O \\ O & I_n \end{pmatrix} = E(\lambda)(A - \lambda B)F(\lambda) \tag{2.45}$$

gilt, wobei $E(\lambda)$ und $F(\lambda)$ $2n \times 2n$–Matrixpolynome mit konstanten Determinanten ungleich Null sind.

Wenn man $A = -\hat{A}$ und $B = -I_{2n}$ wählt, erhält man dann die Linearisierung $\hat{A} \in \mathbb{C}^{2n\times 2n}$ von $P(\lambda)$ gemäß Definition (2.1).

Für eine Linearisierung $A - \lambda B$ von $P(\lambda)$ gilt $\det(P(\lambda)) = k \cdot \det(A - \lambda B)$ für ein $0 \neq k \in \mathbb{C}$. $P(\lambda)$ und das verallgemeinerte Eigenwertproblem (GEP) $Ax = \lambda Bx$ besitzen also dieselben Eigenwerte. In der Praxis wird oft eine Linearisierung der Form

$$\begin{pmatrix} O & N \\ -A_0 & -A_1 \end{pmatrix} - \lambda \begin{pmatrix} N & O \\ O & A_2 \end{pmatrix}$$

für beliebiges invertierbares $N \in \mathbb{C}^{n\times n}$ verwendet. Dass dies eine Linearisierung von $P(\lambda)$ ist, kann man zeigen, indem man in (2.45)

$$E(\lambda) = \begin{pmatrix} -(A_1 + \lambda A_2)N^{-1} & -I_n \\ N^{-1} & 0 \end{pmatrix}, \quad F(\lambda) = \begin{pmatrix} I_n & O \\ \lambda I_n & I_n \end{pmatrix}$$

wählt.

Oft wird $N = c \cdot I_n$ mit $0 \neq c \in \mathbb{R}$ gewählt (z.B. $c = 1, ||A_0||$ oder $||A_2||$).
Für diesen Fall gilt dann:

Satz 16

Es sei $0 \neq c \in \mathbb{C}$. *Dann gilt:*

λ^* *ist Eigenwert von* $P(\lambda)$ *mit Eigenvektor* x^* *genau dann, wenn* λ^* *Eigenwert des GEP*

$\begin{pmatrix} O & cI_n \\ -A_0 & -A_1 \end{pmatrix} y = \lambda \begin{pmatrix} cI_n & O \\ O & A_2 \end{pmatrix} y$ *mit Eigenvektor* $y^* = \begin{pmatrix} x^* \\ \lambda^* x^* \end{pmatrix} \in \mathbb{C}^{2n}$ *ist.*

Beweis:

Ähnlich Satz 15. □

Kapitel 3

Reelle Intervallrechnung

Dieses Kapitel soll eine Einführung in die Intervallrechnung geben und später genutzte Hilfsmittel bereitstellen. Es lehnt sich dabei sehr an [1] an.

Es sei

$$\mathbf{IR} := \{[a] := [\underline{a}\,, \overline{a}] \mid \underline{a}\,, \overline{a} \in \mathbb{R}\,,\ \underline{a} \leq \overline{a}\}$$

die Menge aller reellen abgeschlossenen Intervalle. Dann heißt $\inf([a]) := \underline{a}$ Infimum und $\sup([a]) := \overline{a}$ Supremum von $[a]$. Zwei Intervalle heißen gleich, wenn sie die mengentheoretische Gleichheit besitzen. Punktintervalle (d.h. degenerierte Intervalle) werden immer mit dem reellen Element identifiziert, das sie enthalten. $[a] \in \mathbf{IR}$ heißt symmetrisch, falls $[a] = -[a]$, d.h. $[a] = [-r, r]$ für eine reelle Zahl $r \geq 0$.

Wir bezeichnen mit $[A] = ([a]_{ij}) = ([\underline{a}_{ij}\,, \overline{a}_{ij}])$ die Elemente der Menge $\mathbf{IR}^{n\times m}$ aller $n \times m$–Matrizen mit Einträgen aus $\mathbf{IR}$. $\mathbf{IR}^n$ sei analog die Menge aller reellen n–komponentigen Intervallvektoren. Offensichtlich ist $\mathbb{R}^{n\times m}$ isomorph zur Menge der Intervallmatrizen der Form $([a_{ij}\,, a_{ij}])$ (Punktmatrizen). Aufgrund dieser Isomorphie können sämtliche für $\mathbb{R}^{n\times m}$ gültigen Aussagen auf die Menge der Punktmatrizen übertragen werden. Analoges gilt für $\mathbb{R}^n$ und die Menge der Punktvektoren (Spezialfall $m = 1$).

Wir sagen, dass die Intervallmatrix $[A] \in \mathbf{IR}^{n\times m}$ die Punktmatrix $\dot{A} \in \mathbb{R}^{n\times m}$ enthält, falls $\dot{A} = (\dot{a}_{ij}) \in [A] = ([a]_{ij})$, d.h. $\dot{a}_{ij} \in [a]_{ij}$ für alle i, j. $[A]$ enthält $[B] = ([b]_{ij}) \in \mathbf{IR}^{n\times m}$, falls $[B] \subseteq [A]$, d.h. $[b]_{ij} \subseteq [a]_{ij}$ für alle i, j.

Für die zweistellige Operation $* \in \{+, -, \cdot, /\}$ auf $\mathbb{R}$ führen wir nun die Verknüpfung

$$[a] * [b] := \{a * b \mid a \in [a]\,, b \in [b]\} \tag{3.1}$$

für $[a], [b] \in \mathbf{IR}$ ein. Dabei muss im Fall '/' gewährleistet sein, dass $0 \notin [b]$ ist.

Da die Funktion $f(x\,, y) = x * y\,,\ * \in \{+, -, \cdot, /\}$, auf der kompakten Menge
$[a] \times [b] \subset \mathbb{R} \times \mathbb{R}$ ($0 \notin [b]$ für '/') jeweils eine stetige Funktion zweier Veränderlicher ist, nimmt sie nach dem Extremalsatz (Satz 36.3 in [12]) sowohl ihr Minimum als auch ihr Maximum an. Es wird nach dem Zwischenwertsatz (Satz 35.6 in [12]) aber auch jeder Wert zwischen den Extrema angenommen, weshalb $[a] * [b]$ jeweils wieder ein abgeschlossenes reelles Intervall ist. Die Menge $\mathbf{IR}$ ist also bzgl. jeder der oben aufgeführten Verknüpfungen abgeschlossen (d.h. $[a] * [b] \in \mathbf{IR}$ für alle zulässigen $[a], [b] \in \mathbf{IR}$).

$(\mathbf{IR}, +)$ bzw. $(\mathbf{IR} \setminus \{0\}, \cdot)$ ist eine kommutative Halbgruppe mit neutralem Element 0 bzw. 1: nicht für jedes $[a] \in \mathbf{IR}$ bzw. $[a] \in \mathbf{IR} \setminus \{0\}$ existiert ein inverses Element in $\mathbf{IR}$ bzw. $\mathbf{IR} \setminus \{0\}$. Für die Verbindung dieser beiden Operationen gilt die Subdistributivität:

$$[a] \cdot ([b] + [c]) \subseteq [a] \cdot [b] + [a] \cdot [c] \quad \forall\, [a], [b], [c] \in \mathbf{IR} \tag{3.2}$$

(siehe [1], Satz 4 auf S. 3 ff.).

Weiterhin ist anzumerken, dass $[a] * [b]$ leicht explizit mittels der Schranken $\underline{a}, \overline{a}, \underline{b}, \overline{b}$ berechnet werden kann ([1], S. 2):

$$\begin{aligned}
[a] + [b] &= [\underline{a} + \underline{b}\,, \overline{a} + \overline{b}]\,, \\
[a] - [b] &= [\underline{a} - \overline{b}\,, \overline{a} - \underline{b}] = [a] + (-1) \cdot [b]\,, \\
[a] \cdot [b] &= [\min\{\underline{a} \cdot \underline{b}\,, \underline{a} \cdot \overline{b}\,, \overline{a} \cdot \underline{b}\,, \overline{a} \cdot \overline{b}\}\,, \max\{\underline{a} \cdot \underline{b}\,, \underline{a} \cdot \overline{b}\,, \overline{a} \cdot \underline{b}\,, \overline{a} \cdot \overline{b}\}]\,, \\
[a]/[b] &= [a] \cdot [1/\overline{b}\,, 1/\underline{b}] \quad \text{für } 0 \notin [b]\,.
\end{aligned} \tag{3.3}$$

Die Intervallmatrixoperationen $+, -, \cdot$ werden analog zu den entsprechenden reellen Matrixoperationen definiert:

$$\begin{aligned}
[A] \pm [B] &:= ([a]_{ij} \pm [b]_{ij}) \quad \text{für } [A], [B] \in \mathbf{IR}^{n \times m}\,, \\
\alpha \cdot [A] &:= (\alpha \cdot [a]_{ij}) \quad \text{für } \alpha \in \mathbb{R}\,,\ [A] \in \mathbf{IR}^{n \times m}\,, \\
[A] \cdot [B] &:= (\sum_{k=1}^{m} [a]_{ik} \cdot [b]_{kj}) \quad \text{für } [A] \in \mathbf{IR}^{n \times m}\,,\ [B] \in \mathbf{IR}^{m \times l}\,, \\
[A] \cdot [X] &:= (\sum_{k=1}^{m} [a]_{ik} \cdot [x]_{k}) \quad \text{für } [A] \in \mathbf{IR}^{n \times m}\,,\ [X] \in \mathbf{IR}^{m}\,.
\end{aligned}$$

$(\mathbf{IR}^{n \times m}, +)$ ist eine kommutative Halbgruppe mit der Nullmatrix O als neutralem Element (da dies für $(\mathbf{IR}, +)$ der Fall ist).

$(\mathbf{IR}^{n \times n} \setminus \{O\}, \cdot)$ ist für $n \geq 2$ nicht kommutativ (da bereits $\mathbb{R}^{n \times n}\,(\subset \mathbf{IR}^{n \times n})$ bzgl. der Multiplikation nicht kommutativ ist), nicht assoziativ (siehe [1], Bsp. auf S. 151), besitzt aber die Einheitsmatrix I als neutrales Element.

Für die Verbindung von Intervallmatrixaddition und –multiplikation gilt die Subdistributivität (weil bereits $\mathbf{IR}$ subdistributiv ist):

$$\begin{aligned}
([A] + [B])[C] &\subseteq [A][C] + [B][C] \quad \forall\, [A], [B] \in \mathbf{IR}^{n \times m}\,,\ [C] \in \mathbf{IR}^{m \times l}\,, \\
[D]([A] + [B]) &\subseteq [D][A] + [D][B] \quad \forall\, [A], [B] \in \mathbf{IR}^{n \times m}\,,\ [D] \in \mathbf{IR}^{o \times n}\,.
\end{aligned} \tag{3.4}$$

Falls $r(x)$ eine stetige einstellige Operation auf $\mathbb{R}$ ist, dann sei durch

$$r([a]) := [\min_{x \in [a]} r(x)\,,\ \max_{x \in [a]} r(x)] \quad (= \{r(x) \mid x \in [a]\}) \tag{3.5}$$

die zugehörige einstellige Operation auf $\mathbf{IR}$ erklärt. Beispiele für solche einstelligen Operationen auf $\mathbf{IR}$ (bzw. auf einer geeigneten Teilmenge von $\mathbf{IR}$) sind $[a]^k$ $(k \in \mathbb{R})$, $e^{[a]}$, $\ln[a]$, $\sin[a]$, $\cos[a]$ usw.. Mit $r([A]) := (r([a]_{ij}))$ definieren wir weiterhin die zugehörige einstellige Operation auf $\mathbf{IR}^{n \times m}$.

Lemma 7

Es seien $[a],[b],[c],[d] \in \mathbf{IR}$ *und* r *eine einstellige Operation auf* $\mathbf{IR}$*. Dann gilt:*

a) Aus $[a] \subseteq [c]$, $[b] \subseteq [d]$ *folgt* $[a] * [b] \subseteq [c] * [d]$ *für* $* \in \{+,-,\cdot,/\}$.

b) Aus $[a] \subseteq [c]$ *folgt* $r([a]) \subseteq r([c])$.

Beweis:

a) Es gilt aufgrund der Definition der Verknüpfung $*$

$$[a] * [b] \;=\; \{x = a * b \mid a \in [a], b \in [b]\} \;\subseteq\; \{y = c * d \mid c \in [c], d \in [d]\} \;=\; [c] * [d].$$

b) Es gilt aufgrund der Definition der einstelligen Operation r auf $\mathbf{IR}$

$$r([a]) \;=\; \{x = r(a) \mid a \in [a]\} \;\subseteq\; \{y = r(c) \mid c \in [c]\} \;=\; r([c]).$$

□

Definition 8

Für $[a],[b] \in \mathbf{IR}$ *seien die nichtnegativen, reellen Zahlen*

$$|[a]| := \max\{|\underline{a}|, |\overline{a}|\} \quad (= \max_{x\in[a]} |x|) \qquad \textit{(Betrag)} \tag{3.6}$$

$$\mathrm{rad}([a]) := \tfrac{1}{2}(\overline{a} - \underline{a}) \quad (= \max_{x,y\in[a]} \tfrac{1}{2}|x-y|) \qquad \textit{(Radius)} \tag{3.7}$$

$$q([a],[b]) := \max\{|\underline{a} - \underline{b}|, |\overline{a} - \overline{b}|\} \qquad \textit{(Hausdorff–Abstand)}$$

definiert, wobei der Hausdorff–Abstand nach [1], S.14, *eine Metrik in* $\mathbf{IR}$ *ist.*

Für Intervallmatrizen $[A],[B] \in \mathbf{IR}^{n\times m}$ *seien dann die nichtnegativen, reellen* $n \times m$*–Matrizen*

$$|[A]| := (|[a]_{ij}|), \quad \mathrm{rad}([A]) := (\mathrm{rad}([a]_{ij})), \quad q([A],[B]) := (q([a]_{ij},[b]_{ij}))$$

definiert.

Die reelle Zahl

$$\mathrm{mid}([a]) := \frac{\underline{a} + \overline{a}}{2} \tag{3.8}$$

heißt Mittelpunkt des Intervalls $[a] \in \mathbf{IR}$ *und* $\mathrm{mid}([A]) := (\mathrm{mid}([a]_{ij})) \in \mathbb{R}^{n\times m}$ *Mittelpunkt der Intervallmatrix* $[A] \in \mathbf{IR}^{n\times m}$.

Für $[a] \in \mathbf{IR}$ *heißt das offene Intervall*

$$\mathrm{int}([a]) = (\underline{a}, \overline{a}) \tag{3.9}$$

das Innere von $[a]$*. Analog sei das Innere einer Intervallmatrix definiert.*

Nun folgen einige Lemmata, die in einem späteren Kapitel benötigt werden.

Lemma 8

Gegeben sei das symmetrische Intervall $[-r,r]$ *mit* $r \geq 0$ *und* $a,b \geq 0$ *sowie* $c \in \mathbb{R}$. *Dann gilt:*

a) $(a+b)\cdot[-r,r] = a\cdot[-r,r] + b\cdot[-r,r]$.

b) $c\cdot[-r,r] = |c|\cdot[-r,r]$.

Beweis:

a) Weil $a+b \geq 0$ gilt mit (3.3)

$$(a+b)\cdot[-r,r] = [-(a+b)r,(a+b)r] = [-ar-br,ar+br] = [-ar,ar]+[-br,br] = a[-r,r]+b[-r,r]\,.$$

b) ergibt sich direkt aus den Regeln für die Multiplikation in (3.3). □

Lemma 9

Es seien $[a],[b] \in \mathbf{IR}$. *Dann gilt*

a) $[a] \subseteq [b]$ *genau dann, wenn* $|\mathrm{mid}([a]) - \mathrm{mid}([b])| + \mathrm{rad}([a]) \leq \mathrm{rad}([b])$.

b) $[a] \subseteq \mathrm{int}([b])$ *genau dann, wenn* $|\mathrm{mid}([a]) - \mathrm{mid}([b])| + \mathrm{rad}([a]) < \mathrm{rad}([b])$.

Beweis:

a) Es sei $[a] = [\underline{a},\overline{a}]$, $[b] = [\underline{b},\overline{b}]$. Nach Definition von mid und rad ((3.8) und (3.7)) gilt

$$\underline{a} = \mathrm{mid}([a]) - \mathrm{rad}([a]) \quad \text{und} \quad \overline{a} = \mathrm{mid}([a]) + \mathrm{rad}([a])$$

und Analoges für $[b]$. Dies liefert

$$\begin{aligned}
[a] \subseteq [b] \;&\Leftrightarrow\; \underline{b} \leq \underline{a} \;\wedge\; \overline{a} \leq \overline{b} \\
&\Leftrightarrow\; \mathrm{mid}([b]) - \mathrm{rad}([b]) \leq \mathrm{mid}([a]) - \mathrm{rad}([a]) \;\wedge\; \mathrm{mid}([a]) + \mathrm{rad}([a]) \leq \mathrm{mid}([b]) + \mathrm{rad}([b]) \\
&\Leftrightarrow\; -\mathrm{mid}([b]) + \mathrm{rad}([b]) \geq -\mathrm{mid}([a]) + \mathrm{rad}([a]) \;\wedge\; \mathrm{mid}([a]) + \mathrm{rad}([a]) \leq \mathrm{mid}([b]) + \mathrm{rad}([b]) \\
&\Leftrightarrow\; \mathrm{mid}([a]) - \mathrm{mid}([b]) \geq -(\mathrm{rad}([b]) - \mathrm{rad}([a])) \;\wedge\; \mathrm{mid}([a]) - \mathrm{mid}([b]) \leq \mathrm{rad}([b]) - \mathrm{rad}([a]) \\
&\Leftrightarrow\; -(\mathrm{rad}([b]) - \mathrm{rad}([a])) \leq \mathrm{mid}([a]) - \mathrm{mid}([b]) \leq \mathrm{rad}([b]) - \mathrm{rad}([a]) \\
&\Leftrightarrow\; |\mathrm{mid}([a]) - \mathrm{mid}([b])| \leq \mathrm{rad}([b]) - \mathrm{rad}([a]) \\
&\Leftrightarrow\; |\mathrm{mid}([a]) - \mathrm{mid}([b])| + \mathrm{rad}([a]) \leq \mathrm{rad}([b])\,.
\end{aligned}$$

b) folgt aus a) und der Definition des Inneren int (3.9) eines Intervalls. □

Lemma 10

Es seien $[a],[b] \in \mathbf{IR}$ *und* $k \in \mathbb{R}$. *Dann gilt*

a) $\mathrm{rad}([a]+[b]) = \mathrm{rad}([a]) + \mathrm{rad}([b])$,

b) $\mathrm{rad}(k\cdot[a]) = |k|\mathrm{rad}([a])$,

c) $\operatorname{rad}([a]\cdot[b]) \le \operatorname{rad}([a])|[b]| + |[a]|\operatorname{rad}([b])\,,$

d) $[a]\subseteq[b] \quad\Longrightarrow\quad \operatorname{rad}([a]) \le \operatorname{rad}([b])\,,$

e) $|[a]\cdot[b]| = |[a]|\cdot|[b]|\,.$

Beweis:

a) $\operatorname{rad}([a]+[b]) \overset{(3.3)}{=} \operatorname{rad}([\underline{a}+\underline{b}\,,\overline{a}+\overline{b}]) \overset{(3.7)}{=} \frac{(\overline{a}+\overline{b})-(\underline{a}+\underline{b})}{2} = \frac{\overline{a}-\underline{a}}{2} + \frac{\overline{b}-\underline{b}}{2} \overset{(3.7)}{=} \operatorname{rad}([a]) + \operatorname{rad}([b])$

b) Fall $k \ge 0$:
$\operatorname{rad}(k\cdot[a]) \overset{(3.3)}{=} \operatorname{rad}([k\underline{a}\,,\,k\overline{a}]) \overset{(3.7)}{=} \frac{k\overline{a}-k\underline{a}}{2} = k\cdot\operatorname{rad}([a]) = |k|\cdot\operatorname{rad}([a])$
Fall $k < 0$:
$\operatorname{rad}(k\cdot[a]) \overset{(3.3)}{=} \operatorname{rad}([k\overline{a}\,,\,k\underline{a}]) \overset{(3.7)}{=} \frac{k\underline{a}-k\overline{a}}{2} = -k\cdot\operatorname{rad}([a]) = |k|\cdot\operatorname{rad}([a])$

c) $\operatorname{rad}([a][b]) \overset{(3.1)}{=} \operatorname{rad}\big(\{x\cdot y \mid x\in[a],\, y\in[b]\}\big)$
$\overset{(3.7)}{=} \max_{x,x'\in[a]\,,y,y'\in[b]} |xy-x'y'|/2$
$= \max_{x,x'\in[a]\,,y,y'\in[b]} |xy-xy'+xy'-x'y'|/2$
$\le \max_{x,x'\in[a]\,,y,y'\in[b]} \{|x(y-y')| + |(x-x')y'|\}/2$
$\le \max_{x\in[a]\,,y,y'\in[b]} |x||y-y'|/2 + \max_{x,x'\in[a]\,,y'\in[b]} |x-x'||y'|/2$
$= (\max_{x\in[a]}|x|)(\max_{y,y'\in[b]}|y-y'|/2) + (\max_{x,x'\in[a]}|x-x'|/2)(\max_{y'\in[b]}|y'|)$
$\overset{(3.6),(3.7)}{=} |[a]|\operatorname{rad}([b]) + \operatorname{rad}([a])|[b]|$

d) $[a]\subseteq[b] \quad\Rightarrow\quad \underline{b}\le\underline{a} \text{ und } \overline{a}\le\overline{b} \quad\Rightarrow\quad -\underline{a}\le-\underline{b} \text{ und } \overline{a}\le\overline{b} \quad\Rightarrow\quad \frac{1}{2}(\overline{a}-\underline{a}) \le \frac{1}{2}(\overline{b}-\underline{b})$
$\Rightarrow\quad \operatorname{rad}([a]) \le \operatorname{rad}([b])$

e) Wegen $[a][b] = \{xy \mid x\in[a],\, y\in[b]\}$ gilt
$|[a]\cdot[b]| \overset{(3.6)}{=} \max_{x\in[a],y\in[b]} |xy| = \max_{x\in[a],y\in[b]}(|x|\cdot|y|) = (\max_{x\in[a]}|x|)(\max_{y\in[b]}|y|)$
$\overset{(3.6)}{=} |[a]|\cdot|[b]|\,.$

□

Definition 9

Es sei $[a]^{(k)} = [\underline{a}^{(k)}, \overline{a}^{(k)}] \in \mathbf{IR}\,,\ k\ge 0,$ *und* $[a]^{\infty} = [\underline{a}\,,\overline{a}] \in \mathbf{IR}\,.$
Dann heißt die Intervallfolge $([a]^{(k)})_{k\ge0}$ *konvergent gegen den Grenzwert* $[a]^{\infty}$*, wenn*

$$\lim_{k\to\infty} \underline{a}^{(k)} = \underline{a} \quad \textit{und} \quad \lim_{k\to\infty} \overline{a}^{(k)} = \overline{a}\,.$$

Die Folge von Intervallmatrizen $([A]^{(k)})_{k\ge0}$ *(wobei* $[A]^{(k)} = ([a]_{ij}^{(k)}) \in \mathbf{IR}^{n\times m}$*) heißt konvergent, wenn für alle* $(i,j)\in\{1,\dots,n\}\times\{1,\dots,m\}$ *die Folge der* (i,j)*–ten Einträge* $([a]_{ij}^{(k)})_{k\ge0}$ *konvergent ist.*

Es ist klar, dass $([a]^{(k)})_{k\geq 0}$ genau dann konvergent gegen $[a]^{\infty}$ ist, wenn $\lim_{k\to\infty}[a]^{(k)} = [a]^{\infty}$ bzgl. des Hausdorffabstands q (d.h. $\lim_{k\to\infty} q([a]^{(k)}, [a]^{\infty}) = 0$). Analoges gilt für die Konvergenz bei Intervallmatrizen.

Lemma 11

Jede Intervallfolge $([a]^{(k)})_{k\geq 0}$ *mit*

$$[a]^{(0)} \supseteq [a]^{(1)} \supseteq [a]^{(2)} \supseteq \ldots$$

ist konvergent.

Beweis:
Es sei $[a]^{(k)} = [\underline{a}^{(k)}, \overline{a}^{(k)}]$ für $k \geq 0$. Es werden nun die Folgen der Schranken

$$\underline{a}^{(0)} \leq \underline{a}^{(1)} \leq \underline{a}^{(2)} \leq \underline{a}^{(3)} \leq \ldots \leq \overline{a}^{(3)} \leq \overline{a}^{(2)} \leq \overline{a}^{(1)} \leq \overline{a}^{(0)}$$

betrachtet. Die Folge der unteren Schranken $(\underline{a}^{(k)})_{k\geq 0}$ ist eine monoton wachsende und durch $\overline{a}^{(0)}$ nach oben beschränkte Folge reeller Zahlen. Sie konvergiert somit gegen eine reelle Zahl $\underline{a}$. Ebenso konvergiert die monoton fallende und nach unten beschränkte Folge $(\overline{a}^{(k)})_{k\geq 0}$ gegen eine reelle Zahl $\overline{a}$. Wegen $\underline{a}^{(k)} \leq \overline{a}^{(k)}$ für $k \geq 0$ gilt außerdem $\underline{a} \leq \overline{a}$. Somit ist $([a]^{(k)})_{k\geq 0}$ nach Definition konvergent. □

Es werden von nun an stetige Funktionen $f : D_1 \times \cdots \times D_m \to \mathbb{R}$ mit Definitionsbereich $D_1 \times \cdots \times D_m \subseteq \mathbb{R}^m$ betrachtet. Ein zu f gehöriger Funktionsausdruck $f(x_1, \ldots, x_m)$ sei eine Rechenvorschrift, mit welcher zu jedem Argument $(x_1, \ldots, x_m)$ der zugehörige Funktionswert von f bestimmt werden kann. Es wird vorausgesetzt, dass im Funktionsausdruck $f(x_1, \ldots, x_m)$ nur endlich viele arithmetische Grundoperationen sowie stetige einstellige Operationen auftreten. Außerdem darf f von endlich vielen reellen Konstanten $c_1 \in C_1, \ldots, c_r \in C_r$ ($C_i, i = 1, \ldots, r$ gegebene Parameterbereiche) abhängig sein: $f(x_1, \ldots, x_m; c_1, \ldots, c_r)$. Falls bei der Ersetzung aller Operanden durch Intervalle, welche im Definitionsbereich enthalten sind, und aller Operationen durch die entsprechenden Intervalloperationen gemäß (3.1) und (3.5) ein definierter Intervallausdruck

$$f([x]_1, \ldots, [x]_m; [c]_1, \ldots, [c]_r) \in \mathbf{IR} \tag{3.10}$$

entsteht, so wird dieser als intervallmäßige Auswertung von f bezeichnet. Der Funktionsausdruck von f heißt dann intervallmäßig auswertbar.

Man kann leicht zeigen, dass die Intervalloperationen (3.1) und (3.5) stetig bzgl. der Metrik Hausdorff–Abstand sind (siehe [1]). Damit ist dann auch die intervallmäßige Auswertung stetig in diesem Sinne.

Lemma 12

Der Funktionsausdruck $f(x_1, \ldots, x_m; c_1, \ldots, c_r)$ *von* $f : D_1 \times \cdots \times D_m \to \mathbb{R}$ *mit* $c_l \in C_l$ $(l = 1, \ldots, r)$ *sei intervallmäßig auswertbar.*

Dann gilt für alle $[y]_k \subseteq [x]_k \subseteq D_k$ *(*$k = 1, \ldots, m$*) und* $[d]_l \subseteq [c]_l \subseteq C_l$ *(*$l = 1, \ldots, r$*) die Teilmengeneigenschaft*

$$f([y]_1, \ldots, [y]_m;\, [d]_1, \ldots, [d]_r) \subseteq f([x]_1, \ldots, [x]_m;\, [c]_1, \ldots, [c]_r)\,.$$

Beweis:

Ergibt sich direkt aus Lemma 7 und der Tatsache, dass der Funktionausdruck intervallmäßig auswertbar ist. □

Der Funktionsausdruck von $F = (f_1, \ldots, f_m)^T : D \to \mathbb{R}^m$ mit $D \subseteq \mathbb{R}^m$ heißt intervallmäßig auswertbar, wenn für $i = 1, \ldots, m$ der entsprechende Funktionsausdruck von $f_i : D \to \mathbb{R}$ intervallmäßig auswertbar ist. Dann gilt auch für F die Teilmengeneigenschaft.

Satz 17

Der Funktionsausdruck der Abbildung

$$F = \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix} : \mathbb{R}^m \to \mathbb{R}^m$$

mit $f_i : \mathbb{R}^m \to \mathbb{R}$ *für* $i = 1, \ldots, m$ *sei wie oben beschrieben intervallmäßig auswertbar.*
Für die Iterationsvorschrift im $\mathbf{IR}^m$

$$[x]^{(k+1)} = F([x]^{(k)})\,, \quad k \geq 0,$$

sei

$$[x]^{(1)} \subseteq [x]^{(0)} \in \mathbf{IR}^m$$

erfüllt. Dann gilt:

a) Die Iteriertenfolge $([x]^{(k)})_{k \geq 0}$ *ist konvergent gegen den Grenzwert* $[x]^* \in \mathbf{IR}^m$ *, für den*

$$[x]^* = F([x]^*)$$

gilt.

b) Für jeden Fixpunkt $x^* \in [x]^{(0)}$ *von* F *gilt*

$$x^* \in [x]^{(k)} \quad \textit{für } k \in \mathbb{N}_0$$

und damit

$$x^* \in [x]^*\,.$$

Beweis:

a)

Nach Voraussetzung gilt $[x]^{(1)} \subseteq [x]^{(0)}$. Weil die intervallmäßige Auswertung die Teilmengeneigenschaft besitzt, folgt

$$[x]^{(2)} = F([x]^{(1)}) \subseteq F([x]^{(0)}) = [x]^{(1)} \subseteq [x]^{(0)}\,.$$

Durch vollständige Induktion lässt sich dann

$$\ldots \subseteq [x]^{(3)} \subseteq [x]^{(2)} \subseteq [x]^{(1)} \subseteq [x]^{(0)} \tag{3.11}$$

beweisen. Aus Lemma 11 (angewandt auf jede Komponente in (3.11)) folgt die Konvergenz der Iteriertenfolge gegen ein Element $[x]^* \in \mathbf{IR}^m$. Wegen der Stetigkeit der intervallmäßigen Auswertung von F folgt

$$[x]^* = \lim_{k\to\infty} [x]^{(k)} = \lim_{k\to\infty} [x]^{(k+1)} = \lim_{k\to\infty} F([x]^{(k)}) = F([x]^*)\,.$$

b)

Es sei $x^* \in [x]^{(0)}$ mit $F(x^*) = x^*$. Dann folgt wiederum aufgrund der Teilmengeneigenschaft der intervallmäßigen Auswertung

$$x^* = F(x^*) \in F([x]^{(0)}) = [x]^{(1)}\,.$$

Dies liefert mit vollständiger Induktion $x^* \in [x]^{(k)}$ für $k \geq 0$ und somit auch $x^* \in [x]^*$. □

Kapitel 4

Kubische Systeme

Im Folgenden sei $h : \mathbb{R}^m \to \mathbb{R}^m$ eine Funktion der Gestalt

$$h(x) = r + Sx + Tx^2 + Ux^3\,, \tag{4.1}$$

wobei $r \in \mathbb{R}^m$, $S \in \mathbb{R}^{m\times m}$,

$$T : \begin{cases} \mathbb{R}^m \times \mathbb{R}^m & \to & \mathbb{R}^m \\ (x,y) & \mapsto & \left(\sum_{j=1}^{m}\sum_{k=1}^{m} t_{ijk}x_k y_j\right)_{i=1,\ldots,m} \end{cases},$$

$$U : \begin{cases} \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^m & \to & \mathbb{R}^m \\ (x,y,z) & \mapsto & \left(\sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m} u_{ijkl}x_l y_k z_j\right)_{i=1,\ldots,m} \end{cases}$$

und $Tx^2 := T(x,x)$, $Ux^3 := U(x,x,x)$.

Die Matrix S entspricht einer linearen Abbildung. Die bilineare Abbildung T wird mit $(t_{ijk}) \in \mathbb{R}^{m\times m\times m}$ und die trilineare Abbildung U mit $(u_{ijkl}) \in \mathbb{R}^{m\times m\times m\times m}$ identifiziert.

Wir wollen nun für eine gegebene Funktion h der Gestalt (4.1) einen m–dimensionalen symmetrischen Intervallvektor konstruieren, der unter gewissen Bedingungen einen Fixpunkt von h enthält. Es sei erwähnt, dass der Funktionsausdruck (4.1) von h intervallmäßig auswertbar ist. Da später symmetrische Intervalle verwendet werden, spielt bei der Auswertung eine spezielle Klammerung – etwa ähnlich dem Horner–Schema – keine Rolle.

Es bleibt zu bemerken, dass das System $x = h(x)$ äquivalent zum kubischen System $x - h(x) = 0$ ist. Diese Tatsache gibt dem Kapitel seinen Namen. Im Folgenden wird auch die Funktion h aus (4.1) selbst als kubisches System bezeichnet.

In Vorbereitung auf Satz 20 soll an dieser Stelle an die Vorzeichenregel von Descartes, die eine Aussage über die maximale Anzahl der positiven Nullstellen eines reellen Polynoms macht, und an die Lösungsformel für Nullstellen kubischer Polynome (Cardanische Formeln) erinnert werden. Außerdem wird der Brouwersche Fixpunktsatz benötigt.

Satz 18 *(Vorzeichenregel von Descartes, [5])*

Die Anzahl aller positiven Nullstellen eines reellen Polynoms (gezählt entsprechend ihrer Vielfachheit) ist gleich der Zahl der Vorzeichenwechsel seiner Koeffizientenfolge oder um eine gerade natürliche Zahl kleiner als diese.

Zur Veranschaulichung drei kleine Beispiele:

- $x^3-x^2+x-1 = (x-1)(x^2+1)$ besitzt 3 VZ–Wechsel und eine einfache positive Nullstelle.
- $x^3-x^2-x+1 = (x-1)^2(x+1)$ besitzt 2 VZ–Wechsel und eine doppelte positive Nullstelle.
- $x^3 - 2x^2 - x + 2 = (x-1)(x-2)(x+1)$ besitzt 2 VZ–Wechsel und zwei einfache positive Nullstellen.

Bemerkung 5 *(Cardanische Formeln, [7])*

Die kubische Gleichung

$$ay^3 + by^2 + cy + d = 0 \tag{4.2}$$

mit $a, b, c, d \in \mathbb{R}$ und $a \neq 0$ kann mittels Division durch a und Substitution von $y = z - \frac{b}{3a}$ in die Form

$$z^3 + pz + q = 0 \tag{4.3}$$

gebracht werden, wobei

$$p = \frac{c}{a} - \frac{b^2}{3a^2} \quad \textit{und} \quad q = \frac{2b^3}{27a^3} - \frac{bc}{3a^2} + \frac{d}{a}\,.$$

$D = \left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3$ heißt dann die Diskriminante von (4.2).

Das Lösungsverhalten von (4.2) hängt vom Vorzeichen der Diskriminante ab:

- *$D > 0$: Es gibt genau eine reelle Lösung und zwei (konjugiert) komplexe Lösungen.*
- *$D = 0$: Es gibt entweder eine doppelte und eine einfache reelle Lösung oder $x = 0$ als dreifache reelle Lösung.*
- *$D < 0$: Es gibt drei verschiedene reelle Lösungen.*

Im Fall $D < 0$ sind

$$z_k = 2\sqrt{-\frac{p}{3}}\cos\left(\frac{\alpha + 2k\pi}{3}\right), \quad k = 0, 1, 2,$$

mit

$$\alpha = \arccos\left(-\frac{\sqrt{27}q}{2\sqrt{-p^3}}\right) \in (0, \pi)$$

die drei reellen Lösungen von (4.3). Die drei reellen Lösungen von (4.2) lauten dann

$$y_k = z_k - \frac{b}{3a}\,,\ k = 0, 1, 2.$$

Wegen $\alpha \in (0, \pi)$ gilt $z_1 < z_2 < z_0$ und damit auch

$$y_1 < y_2 < y_0\,.$$

Satz 19 *(Brouwerscher Fixpunktsatz, [13])*

Jede stetige Selbstabbildung f einer konvexen, kompakten und nichtleeren Teilmenge des $\mathbb{R}^n$ (versehen mit irgendeiner Norm) besitzt mindestens einen Fixpunkt.

Satz 20

Die Funktion h sei definiert wie in (4.1); $g : \mathbb{R}^m \to \mathbb{R}^m$ sei eine stetige Funktion, deren Funktionsausdruck intervallmäßig auswertbar ist, und für die

$$g([x]) \subseteq h([x]) \quad \textit{für alle } [x] \in \mathbf{IR}^m \tag{4.4}$$

gilt.

Desweiteren sei $\varphi := ||r||_\infty\,, \quad \sigma := ||S||_\infty\,,$

$$t_\infty := \max_{1\le i\le m}\left(\sum_{j=1}^{m}\sum_{k=1}^{m}|t_{ijk}|\right) \le \tau\,, \quad u_\infty := \max_{1\le i\le m}\left(\sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m}|u_{ijkl}|\right) \le \gamma\,,$$

$$p := \frac{\sigma-1}{\gamma} - \frac{\tau^2}{3\,\gamma^2} \;\textit{ und }\; q := \frac{2\,\tau^3}{27\,\gamma^3} - \frac{(\sigma-1)\tau}{3\,\gamma^2} + \frac{\varphi}{\gamma}\,.$$

Außerdem seien die Bedingungen

$$\varphi > 0,\ \gamma > 0\,,\ \sigma < 1 \tag{4.5}$$

und

$$D := \frac{q^2}{4} + \frac{p^3}{27} < 0 \tag{4.6}$$

erfüllt.

Mit $\alpha := \arccos\left(-\dfrac{\sqrt{27}q}{2\sqrt{-p^3}}\right)$ *gilt dann für*

$$\beta^- := 2\sqrt{\frac{-p}{3}}\cos\left(\frac{\alpha+4\pi}{3}\right) - \frac{\tau}{3\,\gamma} \qquad \textit{und} \qquad \beta^+ := 2\sqrt{\frac{-p}{3}}\cos\left(\frac{\alpha}{3}\right) - \frac{\tau}{3\,\gamma}$$

die Eigenschaft

$$0 < \beta^- < \beta^+\,.$$

Außerdem folgt:

a) Für beliebiges $\beta \in (\beta^-, \beta^+)$ besitzt die Funktion g mindestens einen Fixpunkt x^* in $[x]^{(0)} := [-\beta, \beta]e \in \mathbf{IR}^m$ und die Iteriertenfolge

$$[x]^{(k+1)} := g([x]^{(k)})\,, \qquad k = 0, 1, \ldots, \tag{4.7}$$

konvergiert gegen einen Intervallvektor $[x]^* \in \mathbf{IR}^m$ mit

$$x^* \in [x]^* \subseteq [x]^{(k)} \subseteq [x]^{(k-1)} \subseteq \ldots \subseteq [x]^{(0)}\,, k \in \mathbb{N}. \tag{4.8}$$

b) Für beliebiges $\beta \in (\beta^-, \sqrt{\frac{-p}{3}} - \frac{\tau}{3\,\gamma})$ besitzt die Funktion g einen eindeutigen Fixpunkt x^* in $[x]^{(0)} := [-\beta, \beta]e \in \mathbf{IR}^m$. (4.8) gilt mit $[x]^* = x^*$, d.h. (4.7) konvergiert gegen x^*.

Beweis:

a)

Es sei

$$T([x],[y]) := \sum_{j=1}^{m}\sum_{k=1}^{m} t_{ijk}[x]_k[y]_j \quad \text{mit} \quad T[x]^2 := T([x],[x]) \quad \text{und}$$

$$U([x],[y],[z]) := \sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m} u_{ijkl}[x]_l[y]_k[z]_j \quad \text{mit} \quad U[x]^3 := U([x],[x],[x])\,.$$

Wir zeigen, dass $[x] := [x]^{(0)} = [-\beta,\beta]e$ für $\beta > 0$ die Inklusion

$$h([x]) = r + S[x] + T[x]^2 + U[x]^3 \subseteq \operatorname{int}([x]) \tag{4.9}$$

erfüllt, wenn $\beta \in (\beta^-, \beta^+)$. In diesem Fall gilt nach Voraussetzung (4.4) auch $g([x]) \subseteq \operatorname{int}([x])$ also $[x]^{(1)} \subseteq [x]^{(0)}$. Die Existenz eines Fixpunktes $x^* \in [x]$ von g ist dann garantiert durch den Brouwerschen Fixpunktsatz (Satz 19), da trivialerweise $g(x) \in g([x])\ \forall x \in [x] \in \mathbf{IR}^m$ gilt. Mit Satz 17 folgt außerdem der Rest der Behauptung, da der Funktionsausdruck von $g : \mathbb{R}^m \to \mathbb{R}^m$ intervallmäßig auswertbar ist.

Zum Beweis von (4.9): die Inklusion (4.9) ist äquivalent zu

$$\begin{aligned}
\operatorname{int}([-\beta,\beta]e) \quad &\supseteq \quad r + \left(\sum_{j=1}^{m} s_{ij}[-\beta,\beta]\right) + \left(\sum_{j=1}^{m}\sum_{k=1}^{m} t_{ijk}[-\beta,\beta][-\beta,\beta]\right) \\
&\qquad + \left(\sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m} u_{ijkl}[-\beta,\beta][-\beta,\beta][-\beta,\beta]\right) \\
&\overset{\text{L. 8b)}}{=} \quad r + \left(\sum_{j=1}^{m} |s_{ij}|[-\beta,\beta]\right) + \left(\sum_{j=1}^{m}\sum_{k=1}^{m} |t_{ijk}|[-\beta^2,\beta^2]\right) \\
&\qquad + \left(\sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m} |u_{ijkl}|[-\beta^3,\beta^3]\right) \\
&\overset{\text{L. 8a) \& Ind.}}{=} \quad r + \left([-\beta,\beta]\sum_{j=1}^{m} |s_{ij}|\right) + \left([-\beta^2,\beta^2]\sum_{j=1}^{m}\sum_{k=1}^{m} |t_{ijk}|\right) \\
&\qquad + \left([-\beta^3,\beta^3]\sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m} |u_{ijkl}|\right) \\
&= \quad r + [-\beta,\beta]|S|e + [-\beta^2,\beta^2]|T|e^2 + [-\beta^3,\beta^3]|U|e^3
\end{aligned} \tag{4.10}$$

mit $|T| := (|t_{ijk}|)$ und $|U| := (|u_{ijkl}|)$.

Lemma 9 b), angewandt auf jede Komponente von (4.10), liefert die Äquivalenz von (4.10) und

$$|r| + \beta|S|e + \beta^2|T|e^2 + \beta^3|U|e^3 < \beta e\,. \tag{4.11}$$

Die Ungleichung (4.11) ist sicherlich erfüllt, wenn

$$||r||_\infty + \beta||S||_\infty + \beta^2 t_\infty + \beta^3 u_\infty < \beta\,,$$

d.h. wenn

$$||r||_\infty + \beta(||S||_\infty - 1) + \beta^2 t_\infty + \beta^3 u_\infty < 0 \tag{4.12}$$

gilt. Wegen $||r||_\infty = \varphi,\ ||S||_\infty = \sigma,\ t_\infty \leq \tau$ und $u_\infty \leq \gamma$ ist

$$p_1(\beta) := \varphi + \beta(\sigma - 1) + \beta^2\tau + \beta^3\gamma < 0 \tag{4.13}$$

mit $\beta > 0$ eine für (4.12) und damit für (4.9) hinreichende Bedingung.

τ ist nichtnegativ. Es gilt per Voraussetzung (4.5) $\gamma > 0$, $\varphi > 0$ und $\sigma - 1 < 0$. $p_1(\beta)$ ist ein reelles kubisches Polynom, dessen Nullstellen mit den Cardanischen Formeln (Bemerkung 5) berechnet werden können. Die Diskriminante $D = \frac{q^2}{4} + \frac{p^3}{27}$ von p_1 ist nach Voraussetzung (4.6) negativ. Deswegen besitzt p_1 genau drei verschiedene reelle Nullstellen. Diese sind

$$\beta_k = 2\sqrt{-\frac{p}{3}}\cos\left(\frac{\alpha + 2k\pi}{3}\right) - \frac{\tau}{3\gamma}\,, \quad k = 0, 1, 2,$$

mit $\beta_1 < \beta_2 < \beta_0$.

Es gilt $\lim_{\beta\to\pm\infty} p_1(\beta) = \pm\infty$. Daraus folgt, dass $p_1(\beta) < 0$ für $\beta \in (\beta_2, \beta_0)$ gilt. $p_1(0) = \varphi > 0$, deswegen hat p_1 mindestens eine negative Nullstelle, d.h. $\beta_1 < 0$. Außerdem ist $\beta = 0$ keine Nullstelle von p_1. Die Vorzeichenregel von Descartes (Satz 18) garantiert, dass p_1 entweder zwei oder keine positive Nullstellen besitzt (da es bei den Koeffizienten von p_1 genau zwei Vorzeichenwechsel gibt: $\varphi > 0, \sigma - 1 < 0, \tau \geq 0, \gamma > 0$). Angenommen alle Nullstellen von p_1 sind negativ. Dann hat das Polynom $p_1(\beta) = \gamma(\beta - \beta_0)(\beta - \beta_1)(\beta - \beta_2)$ ausschließlich positive Koeffizienten im Widerspruch zu $\sigma - 1 < 0$. Also besitzt p_1 genau zwei positive Nullstellen, welche β_2 und β_0 sein müssen. Wir können also $\beta^- := \beta_2$ und $\beta^+ := \beta_0$ setzen und haben damit $p_1(\beta) < 0$ für $\beta \in (\beta^-, \beta^+)$ mit $0 < \beta^- < \beta^+$, also (4.13) und damit (4.9).

b)

Es ist $(\beta^-, \sqrt{\frac{-p}{3}} - \frac{\tau}{3\gamma}) \subseteq (\beta^-, \beta^+)$, denn $\beta^- < \sqrt{\frac{-p}{3}} - \frac{\tau}{3\gamma} < \beta^+$ (wegen $\alpha \in (0, \pi)$). Es gilt also a) für $\beta \in (\beta^-, \sqrt{\frac{-p}{3}} - \frac{\tau}{3\gamma})$.

$[x]^*$ sei der Grenzwert der Iterierten $[x]^{(k)}$ aus (4.7). Mit (4.7) und (4.4) gilt für $k \in \mathbb{N}_0$

$$[x]^{(k+1)} = g([x]^{(k)}) \subseteq h([x]^{(k)}) = r + S[x]^{(k)} + T([x]^{(k)})^2 + U([x]^{(k)})^3.$$

Für $k \to \infty$ erhält man

$$[x]^* = g([x]^*) \subseteq h([x]^*) = r + S[x]^* + T([x]^*)^2 + U([x]^*)^3.$$

Für die Radien liefert dies mit Lemma 10 a), b) und d)

$$\begin{aligned}
\mathrm{rad}([x]^*) &\leq \mathrm{rad}(r + S[x]^* + T([x]^*)^2 + U([x]^*)^3) \\
&= \left(\sum_{j=1}^{m} |s_{ij}|\mathrm{rad}([x]_j^*)\right) + \left(\sum_{j=1}^{m}\sum_{k=1}^{m} |t_{ijk}|\mathrm{rad}([x]_k^*[x]_j^*)\right) \\
&\quad + \left(\sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m} |u_{ijkl}|\mathrm{rad}([x]_l^*[x]_k^*[x]_j^*)\right).
\end{aligned} \tag{4.14}$$

Es wird $r_\infty := ||\mathrm{rad}([x]^*)||_\infty = \max_{i=1}^m \mathrm{rad}([x]_i^*)$ definiert. Weil $[x]^* \subseteq [x]^{(0)} = [-\beta, \beta]e$, gilt $|[x]_j^*| \leq \beta$ für $j = 1, \ldots, m$.

Wegen Lemma 10 c) gilt für beliebiges $j, k, l \in \{1, \ldots, m\}$

$$\begin{aligned} \mathrm{rad}([x]_k^*[x]_j^*) &\leq \mathrm{rad}([x]_k^*)|[x]_j^*| + |[x]_k^*|\mathrm{rad}([x]_j^*) \leq \beta(\mathrm{rad}([x]_k^*) + \mathrm{rad}([x]_j^*)) \\ &\leq \beta\, 2\, r_\infty \end{aligned}$$

und damit

$$\begin{aligned} \mathrm{rad}([x]_l^*[x]_k^*[x]_j^*) &\leq \mathrm{rad}([x]_l^*)\,|[x]_k^*[x]_j^*| + |[x]_l^*|\,\mathrm{rad}([x]_k^*[x]_j^*) \\ &\leq \mathrm{rad}([x]_l^*)\,|[x]_k^*|\,|[x]_j^*| + |[x]_l^*|\,\beta\, 2\, r_\infty \\ &\leq \beta^2(\mathrm{rad}([x]_l^*) + 2\, r_\infty) \\ &\leq \beta^2\, 3\, r_\infty. \end{aligned}$$

Aus (4.14) erhält man dann

$$\mathrm{rad}([x]^*) \leq r_\infty \left(\sum_{j=1}^m |s_{ij}| \right) + 2\,\beta\, r_\infty \left(\sum_{j=1}^m \sum_{k=1}^m |t_{ijk}| \right) + 3\,\beta^2\, r_\infty \left(\sum_{j=1}^m \sum_{k=1}^m \sum_{l=1}^m |u_{ijkl}| \right). \quad (4.15)$$

Sei nun $r_\infty = \mathrm{rad}([x]_{i'}^*)$. Es gilt dann mit (4.15)

$$\begin{aligned} r_\infty &= \mathrm{rad}([x]_{i'}^*) \leq r_\infty \sum_{j=1}^m |s_{i'j}| + 2\,\beta\, r_\infty \sum_{j=1}^m \sum_{k=1}^m |t_{i'jk}| + 3\,\beta^2\, r_\infty \sum_{j=1}^m \sum_{k=1}^m \sum_{l=1}^m |u_{i'jkl}| \\ &\leq r_\infty ||S||_\infty + 2\,\beta\, r_\infty t_\infty + 3\,\beta^2\, r_\infty u_\infty \\ &\leq r_\infty\, \sigma + 2\,\beta\, r_\infty\, \tau + 3\,\beta^2\, r_\infty\, \gamma \end{aligned} \quad (4.16)$$

für $\beta \in (\beta^-, \sqrt{\frac{-p}{3}} - \frac{\tau}{3\gamma})$.

Angenommen es gilt $r_\infty > 0$. Dann ist die Aussage (4.16) äquivalent zu

$$1 \leq \sigma + 2\,\tau\,\beta + 3\,\gamma\,\beta^2 \quad \overset{\gamma>0}{\Longleftrightarrow} \quad 0 \leq \frac{\sigma - 1}{3\gamma} + \frac{2\tau}{3\gamma}\beta + \beta^2 =: p_2(\beta) \quad (4.17)$$

für $\beta \in (\beta^-, \sqrt{\frac{-p}{3}} - \frac{\tau}{3\gamma})$. Die Nullstellen $\tilde{\beta}_1$ und $\tilde{\beta}_2$ des quadratischen Polynoms $p_2(\beta)$ sind

$$\tilde{\beta}_{2/1} = -\frac{\tau}{3\gamma} \pm \sqrt{\frac{\tau^2}{9\gamma^2} - \frac{\sigma - 1}{3\gamma}} = -\frac{\tau}{3\gamma} \pm \sqrt{-\frac{p}{3}}$$

und es gilt

$$p_2(\beta) < 0 \qquad \text{für} \quad \beta \in (\tilde{\beta}_1, \tilde{\beta}_2).$$

Wegen $\alpha \in (0, \pi)$ gilt $\tilde{\beta}_1 < \beta^- < \tilde{\beta}_2 = \sqrt{\frac{-p}{3}} - \frac{\tau}{3\gamma} < \beta^+$ und damit

$$p_2(\beta) < 0 \qquad \text{für} \quad \beta \in (\beta^-, \sqrt{\frac{-p}{3}} - \frac{\tau}{3\,\gamma}),$$

was ein Widerspruch zu (4.17) ist. Also ist $r_\infty = 0$.

Da jeder Fixpunkt von g, der in $[x]^{(0)}$ enthalten ist, in $[x]^{(k)}$ bleibt (für $k \in \mathbb{N}$), liefert uns $r_\infty = 0$ die Eindeutigkeit des Fixpunkts x^* von g in $[x]^{(0)}$.

Wegen $x^* \in [x]^*$ (nach a)) und $r_\infty = 0$ gilt dann $[x]^* = x^* \in \mathbb{R}^m$. □

Bemerkung 6

Der in Satz 20 definierte Ausdruck t_∞ bzw. u_∞ ist eine obere Schranke für die von der Maximumsnorm im $\mathbb{R}^n$ induzierte Norm des bilinearen bzw. trilinearen Operators T bzw. U .

Bemerkung 7

Satz 20 gilt natürlich auch für $g = h$.

Bemerkung 8

Wenn die Voraussetzungen von Satz 20 bis auf (4.6) *gelten und $\varphi > 0$ hinreichend klein ist, dann ist $D < 0$ erfüllt.*

Um dies nachzuvollziehen, wird das Polynom (4.13)

$$p_1(\beta) = \varphi + \beta(\sigma - 1) + \beta^2\tau + \beta^3\gamma = \varphi + \beta\{(\sigma - 1) + \beta\tau + \beta^2\gamma\}$$

aus dem Beweis von Satz 20 betrachtet. Für $\varphi > 0$ hinreichend klein und $\beta' > 0$ sehr klein ist dann wegen $\sigma - 1 < 0$

$$p_1(\beta') = \varphi + \beta'\{(\sigma - 1) + \beta'\tau + \beta'^2\gamma\} < 0\,.$$

Weil aber wegen $\gamma > 0$ außerdem $\lim_{\beta\to\infty} p_1(\beta) = \infty$ gilt, besitzt dann p_1 eine positive Nullstelle $\beta^* > \beta'$. Da die Vorzeichenregel von Descartes (Satz 18) besagt, dass p_1 entweder zwei oder keine positive reelle Nullstelle besitzt ($p_1(\beta)$ hat genau zwei Vorzeichenwechsel bei seinen Koeffizienten: $\varphi > 0, \sigma - 1 < 0, \tau \geq 0, \gamma > 0$), muss p_1 also genau zwei positive reelle Nullstellen haben. Es kann sich dabei nicht um eine doppelte Nullstelle handeln, da $p_1(0) = \varphi > 0$, $p_1(\beta') < 0$ und $p_1(\beta) > 0$ für $\beta >> \beta'$. Deswegen kann man für die Diskriminante D von p_1 die Fälle $D > 0$ und $D = 0$ ausschließen (siehe Bemerkung 5), es muss also $D < 0$ gelten.

Bemerkung 9

Je größer die Konstanten τ und γ für eine Funktion g mit $\sigma \approx 0$ sind, welche die Voraussetzungen von Satz 20 erfüllt, desto kleiner ist φ.

Es gelte $\sigma - 1 \approx -1$. Wenn die Voraussetzungen von Satz 20 erfüllt sind, gilt

$$p_1(\beta') = \varphi + \beta'(\sigma - 1) + (\beta')^2\tau + (\beta')^3\gamma = \varphi + \underbrace{(\sigma - 1 + \tau\beta' + \gamma(\beta')^2)}_{=:q(\beta')}\beta' < 0 \qquad (4.18)$$

für $\beta' \in (\beta^-, \beta^+)$ (siehe (4.13) im Beweis von Satz 20). Wegen $\varphi, \beta' > 0$ muss dann $q(\beta') < 0$ sein. Je größer $\tau \geq 0$ und $\gamma > 0$ sind, desto kleiner muss β' sein, damit $q(\beta') < 0$ erfüllt ist. Dann ist aber auch $\beta' q(\beta') < 0$ betragsmäßig sehr klein. Wegen (4.18) muss dann wiederum φ sehr klein sein.

Abschließend noch ein Lemma, das ebenfalls in den folgenden Kapiteln benötigt wird.

Lemma 13

Für $C, B \in \mathbb{R}^{m\times m}$ gelte $||I_m - CB||_\infty < 1$.
Dann sind die Matrizen C und B invertierbar.

Beweis:

Nach Voraussetzung gilt

$$|1 - \lambda_{CB}| \leq \rho(I_m - CB) \leq ||I_m - CB||_\infty < 1\,,$$

wobei ρ den Spektralradius bezeichnet und λ_{CB} ein beliebiger Eigenwert der Matrix CB ist. Damit ist $\lambda_{CB} \neq 0$ und somit CB regulär. Wegen $0 \neq \det(CB) = \det(C)\cdot\det(B)$ gilt dies auch für C und B. □

Kapitel 5

Einschließung von reellen Eigenpaaren des reellen QEP

Betrachtet wird im Folgenden ein Spezialfall des reellen quadratischen Eigenwertproblems (QEP): Für das reelle reguläre quadratische $n \times n$–Matrixpolynom

$$P(\lambda) = A_2\lambda^2 + A_1\lambda + A_0 \tag{5.1}$$

mit Koeffizientenmatrizen $A_2, A_1, A_0 \in \mathbb{R}^{n\times n}$ und $\det A_2 \neq 0$ werden reelle Eigenwerte λ^* und zugehörige reelle Eigenvektoren x^* gesucht, d.h. es soll

$$P(\lambda^*)x^* = 0$$

gelten.

Es werden konkret die folgenden Probleme angegangen:

- Man finde ein Intervall $[\lambda]$, das mindestens (bzw. genau) einen reellen Eigenwert λ^* des QEPs enthält. Dabei soll $\mathrm{rad}([\lambda])$ hinreichend klein sein.
- Man finde einen Intervallvektor $[x]$, der mindestens (bzw. genau) einen zum Eigenwert λ^* zugehörigen Eigenvektor x^* enthält. Insbesondere soll $\mathrm{rad}([x])$ hinreichend klein sein.

Sei nun die nichtlineare Funktion $f : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ definiert durch

$$f(x,\lambda) := \begin{pmatrix} P(\lambda)x \\ {e_s}^T x - 1 \end{pmatrix}, \qquad x \in \mathbb{R}^n,\ \lambda \in \mathbb{R}, \tag{5.2}$$

wobei e_s der s–te Einheitsvektor im $\mathbb{R}^n$ ist. Folgendes Lemma ergibt sich dann sofort.

Lemma 14

(x^*, λ^*) ist genau dann ein Eigenpaar von $P(\lambda)$, das die Normalisierung $(x^*)_s = 1$ erfüllt, wenn (x^*, λ^*) eine Nullstelle von f ist.

$(\tilde{x}, \tilde{\lambda})$ sei nun eine Näherung eines Eigenpaares (x^*, λ^*) von $P(\lambda)$ mit $(x^*)_s = 1$ (d.h. Näherung einer Nullstelle von f). Es wird

$$(\Delta x, \Delta\lambda) := (x - \tilde{x}, \lambda - \tilde{\lambda})$$

definiert und anschließend $f(x,\lambda) = f(\tilde{x}+\Delta x, \tilde{\lambda}+\Delta\lambda)$ berechnet:

$$\begin{aligned} f(x,\lambda) &= f(\tilde{x}+\Delta x, \tilde{\lambda}+\Delta\lambda) \\ &= \begin{pmatrix} \{A_2(\tilde{\lambda}+\Delta\lambda)^2 + A_1(\tilde{\lambda}+\Delta\lambda) + A_0\}(\tilde{x}+\Delta x) \\ {e_s}^T(\tilde{x}+\Delta x) - 1 \end{pmatrix} \\ &= \begin{pmatrix} \{P(\tilde{\lambda}) + \Delta\lambda\, P'(\tilde{\lambda}) + (\Delta\lambda)^2 A_2\}(\tilde{x}+\Delta x) \\ {e_s}^T\tilde{x} + {e_s}^T\Delta x - 1 \end{pmatrix} \\ &= f(\tilde{x},\tilde{\lambda}) + \begin{pmatrix} P(\tilde{\lambda})\Delta x + \{\Delta\lambda\, P'(\tilde{\lambda}) + (\Delta\lambda)^2 A_2\}(\tilde{x}+\Delta x) \\ {e_s}^T\Delta x \end{pmatrix} \\ &= f(\tilde{x},\tilde{\lambda}) + \begin{pmatrix} P(\tilde{\lambda}) & P'(\tilde{\lambda})\tilde{x} + P'(\tilde{\lambda})\Delta x + \Delta\lambda\, A_2\tilde{x} + \Delta\lambda\, A_2\Delta x \\ {e_s}^T & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} \end{aligned} \tag{5.3}$$

mit $P'(\tilde{\lambda}) = 2\tilde{\lambda}A_2 + A_1$ und $\frac{1}{2}P''(\tilde{\lambda}) = A_2$, siehe (2.22).

Dies entspricht genau der Taylor-Entwicklung von f an der Stelle $(\tilde{x},\tilde{\lambda})$

$$\begin{aligned} f(x,\lambda) = f(\tilde{x},\tilde{\lambda}) &+ \left(\frac{\partial f_i}{\partial x_j}(\tilde{x},\tilde{\lambda})\right)_{i,j=1,\ldots,n+1} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} + \frac{1}{2}\left(\frac{\partial^2 f_i}{\partial x_k \partial x_j}(\tilde{x},\tilde{\lambda})\right)_{i,j,k=1,\ldots,n+1} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix}^2 \\ &+ \frac{1}{6}\left(\frac{\partial^3 f_i}{\partial x_l \partial x_k \partial x_j}(\tilde{x},\tilde{\lambda})\right)_{i,j,k,l=1,\ldots,n+1} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix}^3 \end{aligned} \tag{5.4}$$

(mit $x = (x_i)_{i=1,\ldots,n}$ und $x_{n+1} = \lambda$), welche hier ein kubisches System (4.1) in $(\Delta x^T, \Delta\lambda)^T$ mit $m = n+1$ ist. Aus dem Vergleich von (5.3) und (5.4) erhält man, dass die Jacobi–Matrix von f an der Stelle $(\tilde{x},\tilde{\lambda})$ die Gestalt

$$f'(\tilde{x},\tilde{\lambda}) := \left(\frac{\partial f_i}{\partial x_j}(\tilde{x},\tilde{\lambda})\right)_{i,j=1,\ldots,n+1} = \begin{pmatrix} P(\tilde{\lambda}) & P'(\tilde{\lambda})\tilde{x} \\ {e_s}^T & 0 \end{pmatrix} \tag{5.5}$$

besitzt.

Die Multiplikation der rechten Seite von (5.3) mit einer Matrix[1] $-C = -(c_{ij}) \in \mathbb{R}^{(n+1)\times(n+1)}$ und anschließende Addition von $(\Delta x^T, \Delta\lambda)^T$ liefert die Funktion
$g:\ \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$

$$\begin{aligned} g(\Delta x,\Delta\lambda) &:= -Cf(\tilde{x},\tilde{\lambda}) + \\ &\left\{ I_{n+1} - C \begin{pmatrix} P(\tilde{\lambda}) & P'(\tilde{\lambda})\tilde{x} + P'(\tilde{\lambda})\Delta x + \Delta\lambda\, A_2\tilde{x} + \Delta\lambda\, A_2\Delta x \\ {e_s}^T & 0 \end{pmatrix} \right\} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix}. \end{aligned} \tag{5.6}$$

Wegen der Definition von g und Lemma 14 gilt dann das folgende Lemma.

Lemma 15

C sei regulär. Dann gilt:

(x^*,λ^*) ist genau dann ein Eigenpaar von $P(\lambda)$ mit $(x^*)_s = 1$, wenn $(\Delta x^*, \Delta\lambda^*)$ ein Fixpunkt von g ist.

[1] Die Matrix C wird später so gewählt, dass in (5.6) der Ausdruck in der geschweiften Klammer in der Nähe der Nullmatrix liegt.

Die Funktion g wird nun in eine für Intervall-Fixpunktiteration günstigere Gestalt gebracht.
Es seien definiert

$$C^* := (c_{ij})_{\substack{i=1,\dots,n+1\\ j=1,\dots,n}}, \quad D_1 := C^*A_1 \in \mathbb{R}^{(n+1)\times n} \quad \text{und} \quad D_2 := C^*A_2 \in \mathbb{R}^{(n+1)\times n}.$$

Dann gilt wegen (5.5)

$$\begin{aligned}
g(\Delta x, \Delta\lambda) &= -Cf(\tilde{x},\tilde{\lambda}) + \Big\{ I_{n+1} - Cf'(\tilde{x},\tilde{\lambda}) \\
&\quad -C \begin{pmatrix} O & P'(\tilde{\lambda})\Delta x + \Delta\lambda\, A_2\tilde{x} + \Delta\lambda\, A_2\Delta x \\ 0^T & 0 \end{pmatrix} \Big\} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} \\
&= -Cf(\tilde{x},\tilde{\lambda}) + \Big\{ I_{n+1} - Cf'(\tilde{x},\tilde{\lambda}) \\
&\quad -C \begin{pmatrix} O & (2\tilde{\lambda}A_2 + A_1)\Delta x + \Delta\lambda\, A_2\tilde{x} + \Delta\lambda\, A_2\Delta x \\ 0^T & 0 \end{pmatrix} \Big\} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} \\
&= -Cf(\tilde{x},\tilde{\lambda}) + \Big\{ I_{n+1} - Cf'(\tilde{x},\tilde{\lambda}) \\
&\quad - \begin{pmatrix} O & (2\tilde{\lambda}C^*A_2 + C^*A_1)\Delta x + \Delta\lambda\, C^*A_2\tilde{x} + \Delta\lambda\, C^*A_2\Delta x \end{pmatrix} \Big\} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} \\
&= -Cf(\tilde{x},\tilde{\lambda}) + \Big\{ I_{n+1} - Cf'(\tilde{x},\tilde{\lambda}) \\
&\quad - \begin{pmatrix} O & (2\tilde{\lambda}D_2 + D_1)\Delta x + (D_2\tilde{x})\Delta\lambda + (D_2\Delta x)\Delta\lambda \end{pmatrix} \Big\} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix}.
\end{aligned}$$

Der zuletzt erhaltene Funktionsausdruck von $g(\Delta x, \Delta\lambda)$ ist intervallmäßig auswertbar gemäß (3.10).

Es wird im Beweis des folgenden Satzes gezeigt, dass man unter gewissen Voraussetzungen Satz 20 auf die Funktion g anwenden kann. Ein bestimmter $(n+1)$–komponentiger Intervallvektor enthält dann also einen Fixpunkt von g. Dieser kann mit Hilfe von Intervall–Fixpunktiteration bzgl. g eng eingeschlossen werden. Mit Lemma 15 erhält man so eine Einschließung eines Eigenpaares (x^*, λ^*) mit $(x^*)_s = 1$ von $P(\lambda)$.

Satz 21

Es sei $P(\lambda)$ wie in (5.1), $\tilde{\lambda} \in \mathbb{R}$, $\tilde{x} \in \mathbb{R}^n$ und $C \in \mathbb{R}^{(n+1)\times(n+1)}$.
Für die wie folgt definierten Ausdrücke

$$\varphi := ||Cf(\tilde{x},\tilde{\lambda})||_\infty, \quad \sigma := ||I_{n+1} - Cf'(\tilde{x},\tilde{\lambda})||_\infty,$$

$$\tau := ||C||_\infty\{(2|\tilde{\lambda}| + ||\tilde{x}||_\infty)||A_2||_\infty + ||A_1||_\infty\}, \quad \gamma := ||C||_\infty||A_2||_\infty,$$

$$p := \frac{\sigma - 1}{\gamma} - \frac{\tau^2}{3\,\gamma^2} \;\; \textit{und} \;\; q := \frac{2\,\tau^3}{27\,\gamma^3} - \frac{(\sigma-1)\tau}{3\,\gamma^2} + \frac{\varphi}{\gamma}$$

seien die Bedingungen

$$\varphi > 0\,,\ \sigma < 1 \tag{5.7}$$

und

$$D := \frac{q^2}{4} + \frac{p^3}{27} < 0 \tag{5.8}$$

erfüllt.

Mit $\alpha := \arccos\left(-\dfrac{\sqrt{27}q}{2\sqrt{-p^3}}\right)$ *gilt dann für*

$$\beta^- := 2\sqrt{\frac{-p}{3}}\cos\left(\frac{\alpha+4\pi}{3}\right) - \frac{\tau}{3\,\gamma} \qquad \textit{und} \qquad \beta^+ := 2\sqrt{\frac{-p}{3}}\cos\left(\frac{\alpha}{3}\right) - \frac{\tau}{3\,\gamma} \tag{5.9}$$

die Eigenschaft

$$0 < \beta^- < \beta^+ .$$

Weiter sei

$$\begin{aligned} g(\Delta x, \Delta\lambda) \;\; := \;\; & -Cf(\tilde{x},\tilde{\lambda}) + \Big\{ I_{n+1} - Cf'(\tilde{x},\tilde{\lambda}) \\ & - \begin{pmatrix} O & (2\tilde{\lambda}D_2 + D_1)\Delta x + (D_2\tilde{x})\Delta\lambda + (D_2\Delta x)\Delta\lambda \end{pmatrix} \Big\} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} \end{aligned}$$

mit $D_1 := C^*A_1$, $D_2 := C^*A_2$, $C^* := (c_{ij})_{\substack{i=1,\ldots,n+1 \\ j=1,\ldots,n}}$ *und*

$$\begin{pmatrix} [\Delta x]^{(0)} \\ [\Delta\lambda]^{(0)} \end{pmatrix} := [-\beta,\beta]e \in \mathbf{IR}^{n+1}$$

für beliebiges $\beta \in (\beta^-, \beta^+)$. *Dann konvergiert die Iteriertenfolge*

$$\begin{pmatrix} [\Delta x]^{(k+1)} \\ [\Delta\lambda]^{(k+1)} \end{pmatrix} := g([\Delta x]^{(k)}, [\Delta\lambda]^{(k)}), \quad k \in \mathbb{N}_0,$$

gegen einen Intervallvektor $\begin{pmatrix} [\Delta x]^* \\ [\Delta\lambda]^* \end{pmatrix}$ *mit*

$$\begin{pmatrix} [\Delta x]^* \\ [\Delta\lambda]^* \end{pmatrix} \subseteq \begin{pmatrix} [\Delta x]^{(k+1)} \\ [\Delta\lambda]^{(k+1)} \end{pmatrix} \subseteq \begin{pmatrix} [\Delta x]^{(k)} \\ [\Delta\lambda]^{(k)} \end{pmatrix} \subseteq \ldots \subseteq \begin{pmatrix} [\Delta x]^{(0)} \\ [\Delta\lambda]^{(0)} \end{pmatrix}, \quad k \in \mathbb{N},$$

und es existiert mindestens ein reelles Eigenpaar (x^*, λ^*) *von* $P(\lambda)$ *mit* $(x^*)_s = 1$ *und*

$$\begin{pmatrix} x^* \\ \lambda^* \end{pmatrix} \in \begin{pmatrix} \tilde{x} \\ \tilde{\lambda} \end{pmatrix} + \begin{pmatrix} [\Delta x]^{(k)} \\ [\Delta\lambda]^{(k)} \end{pmatrix}, \qquad k \in \mathbb{N}_0 . \tag{5.10}$$

Unter der Einschränkung $\beta \in (\beta^-, \sqrt{\frac{-p}{3}} - \frac{\tau}{3\gamma})$ *ist dieses Eigenpaar* (x^*, λ^*) *eindeutig, und die Iterierten* $\begin{pmatrix} [\Delta x]^{(k)} \\ [\Delta\lambda]^{(k)} \end{pmatrix}$ *konvergieren gegen dessen Approximationsfehler* $\begin{pmatrix} \Delta x^* \\ \Delta\lambda^* \end{pmatrix} \in \mathbb{R}^{n+1}$.

Beweis:
Es sei $([\Delta x]^T, [\Delta\lambda])^T \in \mathbf{IR}^{n+1}$. Es gilt

$$\begin{aligned}
g([\Delta x],[\Delta\lambda]) &= -Cf(\tilde{x},\tilde{\lambda}) + \Big\{ I_{n+1} - Cf'(\tilde{x},\tilde{\lambda}) \\
&\quad - \begin{pmatrix} O & (2\tilde{\lambda}D_2 + D_1)[\Delta x] + (D_2\tilde{x})[\Delta\lambda] + (D_2[\Delta x])[\Delta\lambda] \end{pmatrix} \Big\} \begin{pmatrix} [\Delta x] \\ [\Delta\lambda] \end{pmatrix} \\
&\overset{(3.4)}{\subseteq} -Cf(\tilde{x},\tilde{\lambda}) + \Big(I_{n+1} - Cf'(\tilde{x},\tilde{\lambda}) \Big) \begin{pmatrix} [\Delta x] \\ [\Delta\lambda] \end{pmatrix} \\
&\quad - \begin{pmatrix} O & (2\tilde{\lambda}D_2 + D_1)[\Delta x] + (D_2\tilde{x})[\Delta\lambda] + (D_2[\Delta x])[\Delta\lambda] \end{pmatrix} \begin{pmatrix} [\Delta x] \\ [\Delta\lambda] \end{pmatrix} \\
&\overset{(3.2)}{\subseteq} -Cf(\tilde{x},\tilde{\lambda}) + \Big(I_{n+1} - Cf'(\tilde{x},\tilde{\lambda}) \Big) \begin{pmatrix} [\Delta x] \\ [\Delta\lambda] \end{pmatrix} \\
&\quad - \Big((2\tilde{\lambda}D_2 + D_1)[\Delta x] \Big)[\Delta\lambda] - (D_2\tilde{x})[\Delta\lambda][\Delta\lambda] - \Big((D_2[\Delta x])[\Delta\lambda] \Big)[\Delta\lambda] \\
&\overset{(3.2)}{\subseteq} -Cf(\tilde{x},\tilde{\lambda}) + \Big(I_{n+1} - Cf'(\tilde{x},\tilde{\lambda}) \Big) \begin{pmatrix} [\Delta x] \\ [\Delta\lambda] \end{pmatrix} \\
&\quad + \Big(\sum_{k=1}^{n} -(2\tilde{\lambda}D_2 + D_1)_{ik}[\Delta x]_k[\Delta\lambda] - (D_2\tilde{x})_i[\Delta\lambda][\Delta\lambda] \\
&\quad + \sum_{l=1}^{n} -(D_2)_{il}[\Delta x]_l[\Delta\lambda][\Delta\lambda] \Big)_{i=1,\ldots,n+1} \\
&=: r + S \begin{pmatrix} [\Delta x] \\ [\Delta\lambda] \end{pmatrix} + T \begin{pmatrix} [\Delta x] \\ [\Delta\lambda] \end{pmatrix}^2 + U \begin{pmatrix} [\Delta x] \\ [\Delta\lambda] \end{pmatrix}^3 =: h([\Delta x],[\Delta\lambda]).
\end{aligned}$$

$h(x) = r + Sx + Tx^2 + Ux^3\,, x \in \mathbb{R}^{n+1}$, ist ein kubisches System (4.1) mit $m = n+1$,
$r = -Cf(\tilde{x},\tilde{\lambda})\,,\ S := I_{n+1} - Cf'(\tilde{x},\tilde{\lambda})\,,\ T = (t_{ijk}) \in \mathbb{R}^{(n+1)\times(n+1)\times(n+1)}$ und
$U = (u_{ijkl}) \in \mathbb{R}^{(n+1)\times(n+1)\times(n+1)\times(n+1)}$, wobei

$$t_{ijk} = \begin{cases} -(2\tilde{\lambda}D_2 + D_1)_{ik}\,, & i = 1,\ldots,n+1;\ j = n+1;\ k = 1,\ldots,n, \\ -(D_2\tilde{x})_i\,, & i = 1,\ldots,n+1;\ j = k = n+1, \\ 0\,, & \text{sonst}, \end{cases}$$

$$u_{ijkl} = \begin{cases} -(D_2)_{il}\,, & i = 1,\ldots,n+1;\ j = k = n+1;\ l = 1,\ldots,n, \\ 0\,, & \text{sonst}. \end{cases}$$

Dann ist $||r||_\infty = ||Cf(\tilde{x},\tilde{\lambda})||_\infty = \varphi$ und $||S||_\infty = ||I_{n+1} - Cf'(\tilde{x},\tilde{\lambda})||_\infty = \sigma$. Weiter gilt

$$\begin{aligned}
t_\infty &:= \max_{1\le i\le n+1}\Big(\sum_{j=1}^{n+1}\sum_{k=1}^{n+1}|t_{ijk}|\Big)\\
&= \max_{1\le i\le n+1}\Big(\sum_{k=1}^{n}|(2\tilde{\lambda}D_2+D_1)_{ik}| + |(D_2\tilde{x})_i|\Big)\\
&\overset{\substack{\max\text{ für}\\ i\equiv i^*}}{} \sum_{k=1}^{n}|(2\tilde{\lambda}D_2+D_1)_{i^*k}| + |(D_2\tilde{x})_{i^*}| = \sum_{k=1}^{n}|(2\tilde{\lambda}C^*A_2+C^*A_1)_{i^*k}| + |(C^*A_2\tilde{x})_{i^*}|\\
&= \sum_{k=1}^{n}|2\tilde{\lambda}\sum_{l=1}^{n}c_{i^*l}(A_2)_{lk} + \sum_{l=1}^{n}c_{i^*l}(A_1)_{lk}| + |\sum_{l=1}^{n}c_{i^*l}(A_2\tilde{x})_l|\\
&\le 2|\tilde{\lambda}|\sum_{k=1}^{n}\sum_{l=1}^{n}|c_{i^*l}|\,|(A_2)_{lk}| + \sum_{k=1}^{n}\sum_{l=1}^{n}|c_{i^*l}|\,|(A_1)_{lk}| + \sum_{l=1}^{n}|c_{i^*l}|\,|(A_2\tilde{x})_l|\\
&= 2|\tilde{\lambda}|\sum_{l=1}^{n}|c_{i^*l}|\big(\sum_{k=1}^{n}|(A_2)_{lk}|\big) + \sum_{l=1}^{n}|c_{i^*l}|\big(\sum_{k=1}^{n}|(A_1)_{lk}|\big) + \sum_{l=1}^{n}|c_{i^*l}|\,|(A_2\tilde{x})_l|\\
&\le 2|\tilde{\lambda}|\,||A_2||_\infty\sum_{l=1}^{n}|c_{i^*l}| + ||A_1||_\infty\sum_{l=1}^{n}|c_{i^*l}| + ||A_2\tilde{x}||_\infty\sum_{l=1}^{n}|c_{i^*l}|\\
&\le 2|\tilde{\lambda}|\,||A_2||_\infty||C||_\infty + ||A_1||_\infty||C||_\infty + ||A_2\tilde{x}||_\infty||C||_\infty\\
&\le 2|\tilde{\lambda}|\,||A_2||_\infty||C||_\infty + ||A_1||_\infty||C||_\infty + ||A_2||_\infty||\tilde{x}||_\infty||C||_\infty\\
&= ||C||_\infty\big((2|\tilde{\lambda}| + ||\tilde{x}||_\infty)||A_2||_\infty + ||A_1||_\infty\big) = \tau\,.
\end{aligned}$$

Mit ähnlichen Schlussfolgerungen erhält man

$$\begin{aligned}
u_\infty &:= \max_{1\le i\le n+1}\Big(\sum_{j=1}^{n+1}\sum_{k=1}^{n+1}\sum_{l=1}^{n+1}|u_{ijkl}|\Big)\\
&= \max_{1\le i\le n+1}\Big(\sum_{l=1}^{n}|(D_2)_{il}|\Big)\\
&\le ||C||_\infty||A_2||_\infty = \gamma\,.
\end{aligned}$$

C ist wegen $\sigma = ||I_{n+1} - Cf'(\tilde{x},\tilde{\lambda})||_\infty < 1$ nach Lemma 13 regulär. Weil auch A_2 nach Voraussetzung regulär ist, sind beide Matrizen ungleich der Nullmatrix, d.h. $\gamma \neq 0$. Außerdem ist der Funktionsausdruck von $g : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ intervallmäßig auswertbar.

Nun kann man Satz 20 anwenden. Satz 20 a) liefert für $\beta \in (\beta^-, \beta^+)$ die Existenz mindestens eines Fixpunkts

$$\begin{pmatrix}\Delta x^*\\ \Delta\lambda^*\end{pmatrix} \in \begin{pmatrix}[\Delta x]^{(k)}\\ [\Delta\lambda]^{(k)}\end{pmatrix}, \quad k \in \mathbb{N}_0,$$

von g. Mit Lemma 15 folgt dann die Existenz mindestens eines Eigenpaares (x^*, λ^*) von $P(\lambda)$ mit (5.10) und $(x^*)_s = 1$. Für $\beta \in (\beta^-, \sqrt{\frac{-p}{3}} - \frac{\tau}{3\gamma})$ folgt aus Satz 20 b) und Lemma 15 analog der Rest des Satzes. □

Satz 22

Es sei (x^*, λ^*) ein Eigenpaar von $P(\lambda)$ mit $(x^*)_s \neq 0$. Dann gilt:

λ^* ist genau dann ein einfacher Eigenwert von $P(\lambda)$, wenn $f'(x^*, \lambda^*)$ aus (5.5) nichtsingulär ist.

Beweis:

Dies ist genau die Aussage von Satz 14, wobei $B^* = f'(x^*, \lambda^*)$. □

Bemerkung 10

Die Voraussetzungen des Satzes 21 sind erfüllt, wenn alle folgenden Bedingungen gelten:

a) $(\tilde{x}, \tilde{\lambda})$ ist eine hinreichend gute Näherung eines reellen Eigenpaares (x^*, λ^*) von $P(\lambda)$ mit $(x^*)_s = 1$; o.B.d.A. sei $(\tilde{x}, \tilde{\lambda})$ kein Eigenpaar von $P(\lambda)$.

b) $f'(\tilde{x}, \tilde{\lambda})$ ist regulär.

c) $C := f'(\tilde{x}, \tilde{\lambda})^{-1}$ oder C ist eine hinreichend gute Näherung von $f'(\tilde{x}, \tilde{\lambda})^{-1}$.

Wenn a) gilt und der Eigenwert λ^* zusätzlich einfach ist, impliziert dies b).
Wenn a) gilt und λ^* nicht einfach ist, dann ist $f'(\tilde{x}, \tilde{\lambda})$ fast singulär.

Wegen b) kann $C := f'(\tilde{x}, \tilde{\lambda})^{-1}$ (oder zumindest eine hinreichend gute Näherung von $f'(\tilde{x}, \tilde{\lambda})^{-1}$) gewählt werden (vgl. c)). Damit ist $\sigma = ||I_{n+1} - Cf'(\tilde{x}, \tilde{\lambda})||_\infty = 0$ (oder zumindest $\sigma \approx 0$) und die Bedingung $\sigma < 1$ aus (5.7) erfüllt.

Da nach Annahme $(x^*, \lambda^*) \neq (\tilde{x}, \tilde{\lambda})$ gilt, ist $\varphi \neq 0$, d.h. die zweite Bedingung aus (5.7) ist erfüllt.

Für die hinreichend gute Eigenpaarnäherung $(\tilde{x}, \tilde{\lambda})$ aus a) ist $\varphi = ||Cf(\tilde{x}, \tilde{\lambda})||_\infty$ hinreichend klein. Unter Verwendung des Beweises von Satz 21 kann man mit der Heuristik aus Bemerkung 8 folgern, dass $D < 0$ und damit die Bedingung (5.8) gilt. Damit sind alle Voraussetzungen von Satz 21 erfüllt.

Wenn a) gilt und λ^* einfach (bzw. nicht einfach) ist, ist die Matrix $f'(x^*, \lambda^*)$ nach Satz 22 regulär (bzw. singulär). Für die hinreichend gute Näherung $(\tilde{x}, \tilde{\lambda})$ von (x^*, λ^*) ist dann aus Stetigkeitsgründen $f'(\tilde{x}, \tilde{\lambda})$ regulär (bzw. fast singulär).

Bemerkung 11

Die Bedingungen aus Bemerkung 10 und somit die Voraussetzungen von Satz 21 seien erfüllt. Dann gilt:

Je größer die Konstanten τ und γ sind, desto besser muss die Näherung $(\tilde{x}, \tilde{\lambda})$ sein.

Wenn die Bedingungen aus Bemerkung 10 erfüllt sind, gilt $\sigma = ||I_{n+1} - Cf'(\tilde{x}, \tilde{\lambda})||_\infty \approx 0$ (oder sogar $\sigma = 0$). Unter Verwendung des Beweises von Satz 21 und Bemerkung 9 erhält man, dass je größer τ und γ werden, desto kleiner muss $\varphi = ||Cf(\tilde{x}, \tilde{\lambda})||_\infty$ sein. Um dies zu erreichen, muss die Näherung $(\tilde{x}, \tilde{\lambda})$ umso besser sein.

Lemma 16

Wenn $f'(\tilde{x}, \tilde{\lambda})$ *aus* (5.5) *regulär ist, gilt* $(f'(\tilde{x}, \tilde{\lambda})^{-1})_{s,*} = e_{n+1}^T$ *, d.h.*

$$||f'(\tilde{x}, \tilde{\lambda})^{-1}||_\infty \geq 1\,.$$

Beweis:

Es sei $J := f'(\tilde{x}, \tilde{\lambda}) \in \mathbb{R}^{(n+1)\times(n+1)}$ und $J^{-1} =: (v_1, \ldots, v_{n+1})$ mit $v_i \in \mathbb{R}^{n+1}$, $i = 1, \ldots, n+1$. Für die letzte Zeile von J gilt nach Definition $J_{n+1,*} = e_s^T$. Wegen

$$e_{n+1}^T = (J \cdot J^{-1})_{n+1,*} = J_{n+1,*} \cdot J^{-1} = e_s^T(v_1, \ldots, v_{n+1})$$

gilt $0 = e_s^T v_i = (v_i)_s$, $i = 1, \ldots, n$, und $1 = e_s^T v_{n+1} = (v_{n+1})_s$ und damit die Behauptung. $\square$

Kapitel 6

Einschließung von komplexen Eigenpaaren des reellen QEP

In diesem Kapitel wird das reelle quadratische Eigenwertproblem (QEP) betrachtet: Für das reguläre reelle quadratische $n \times n$–Matrixpolynom

$$P(\lambda) = A_2\lambda^2 + A_1\lambda + A_0 \tag{6.1}$$

mit reellen Koeffizientenmatrizen $A_2, A_1, A_0 \in \mathbb{R}^{n\times n}$ und $\det A_2 \neq 0$ werden komplexe Eigenwerte $\lambda^* = \lambda_1^* + i\lambda_2^*$ und zugehörige komplexe Eigenvektoren $x^* = x_1^* + ix_2^* \in \mathbb{C}^n$ gesucht (d.h. $P(\lambda^*)x^* = 0$ für $x^* \neq 0$).

Es werden jetzt in Analogie zum vorherigen Kapitel reelle Intervalle $[\lambda_1]^*, [\lambda_2]^*$ bzw. Intervallvektoren $[x_1]^*, [x_2]^*$ mit hinreichend kleinem Radius gesucht, die Realteil und Imaginärteil der Eigenwerte bzw. zugehöriger Eigenvektoren von $P(\lambda)$ einschließen:

$$\mathrm{Re}\,(\lambda^*) = \lambda_1^* \in [\lambda_1]^*, \quad \mathrm{Im}\,(\lambda^*) = \lambda_2^* \in [\lambda_2]^*, \quad \mathrm{Re}\,(x^*) = x_1^* \in [x_1]^*, \quad \mathrm{Im}\,(x^*) = x_2^* \in [x_2]^*\,.$$

Es ist klar, dass

$$P(\lambda^*)x^* = 0 \quad \text{genau dann gilt, wenn} \quad \mathrm{Re}\,(P(\lambda^*)x^*) = 0 \ \text{ und } \ \mathrm{Im}\,(P(\lambda^*)x^*) = 0\,. \tag{6.2}$$

Weiter gilt für die i–te Komponente des Realteils von $P(\lambda)x$

$$\begin{aligned}
\mathrm{Re}\,(P(\lambda)x)_i &= \mathrm{Re}\,((P(\lambda)x)_i) = \mathrm{Re}\,\Big(\sum_{j=1}^{n} P(\lambda)_{ij}x_j\Big) = \sum_{j=1}^{n} \mathrm{Re}\,(P(\lambda)_{ij}x_j) \\
&= \sum_{j=1}^{n} \{\mathrm{Re}\,(P(\lambda)_{ij})\mathrm{Re}\,(x_j) - \mathrm{Im}\,(P(\lambda)_{ij})\mathrm{Im}\,(x_j)\} \\
&= \sum_{j=1}^{n} \mathrm{Re}\,(P(\lambda))_{ij}\mathrm{Re}\,(x)_j - \sum_{j=1}^{n} \mathrm{Im}\,(P(\lambda))_{ij}\mathrm{Im}\,(x)_j \\
&= (\mathrm{Re}\,(P(\lambda))\mathrm{Re}\,(x))_i - (\mathrm{Im}\,(P(\lambda))\mathrm{Im}\,(x))_i
\end{aligned}$$

und analog $\mathrm{Im}\,(P(\lambda)x)_i = (\mathrm{Im}\,(P(\lambda))\mathrm{Re}\,(x))_i + (\mathrm{Re}\,(P(\lambda))\mathrm{Im}\,(x))_i$ für $i = 1, \ldots, n$.

Mit $\lambda = \lambda_1 + i\lambda_2$ und $x = x_1 + ix_2$ gilt dann also

$$\begin{aligned}
\mathrm{Re}\,(P(\lambda)x) &= \mathrm{Re}\,(P(\lambda))\mathrm{Re}\,(x) - \mathrm{Im}\,(P(\lambda))\mathrm{Im}\,(x) = \mathrm{Re}\,(P(\lambda))x_1 - \mathrm{Im}\,(P(\lambda))x_2\,, \\
\mathrm{Im}\,(P(\lambda)x) &= \mathrm{Im}\,(P(\lambda))\mathrm{Re}\,(x) + \mathrm{Re}\,(P(\lambda))\mathrm{Im}\,(x) = \mathrm{Im}\,(P(\lambda))x_1 + \mathrm{Re}\,(P(\lambda))x_2\,,
\end{aligned} \tag{6.3}$$

wobei $\mathrm{Re}\,(P(\lambda)) = \mathrm{Re}\,(P(\lambda_1 + i\lambda_2)) = (\lambda_1^2 - \lambda_2^2)A_2 + \lambda_1 A_1 + A_0$ und
$\mathrm{Im}\,(P(\lambda)) = \mathrm{Im}\,(P(\lambda_1 + i\lambda_2)) = 2\lambda_1\lambda_2 A_2 + \lambda_2 A_1\,.$
Für

$$R(P(\lambda)) := \begin{pmatrix} \mathrm{Re}\,(P(\lambda)) & -\mathrm{Im}\,(P(\lambda)) \\ \mathrm{Im}\,(P(\lambda)) & \mathrm{Re}\,(P(\lambda)) \end{pmatrix} \in \mathbb{R}^{2n\times 2n} \tag{6.4}$$

ist dann wegen (6.2) und (6.3) das folgende Lemma gültig.

Lemma 17

$P(\lambda^*)x^* = 0$ *genau dann, wenn* $R(P(\lambda^*)) \begin{pmatrix} x_1^* \\ x_2^* \end{pmatrix} = 0\,.$

Es wird nun eine nichtlineare Funktion $f:\ \mathbb{R}^{2n+2} \to \mathbb{R}^{2n+2}$ definiert durch

$$f(x_1, x_2, \lambda_1, \lambda_2) := \begin{pmatrix} R(P(\lambda))\binom{x_1}{x_2} \\ {e_s}^T x_1 - 1 \\ {e_s}^T x_2 \end{pmatrix}, \quad x_1, x_2 \in \mathbb{R}^n,\ \lambda_1, \lambda_2 \in \mathbb{R}\ (x = x_1 + ix_2,\ \lambda = \lambda_1 + i\lambda_2), \tag{6.5}$$

wobei e_s der s–te Einheitsvektor im $\mathbb{R}^n$ ist. Folgendes Lemma ergibt sich dann sofort aus Lemma 17 .

Lemma 18

$(x^*, \lambda^*) = (x_1^* + ix_2^*\,,\ \lambda_1^* + i\lambda_2^*)$ *ist genau dann ein komplexes Eigenpaar von* $P(\lambda)$*, das die Normalisierung* $(x^*)_s = 1$ *erfüllt, wenn* $(x_1^*, x_2^*, \lambda_1^*, \lambda_2^*)$ *eine Nullstelle von* f *ist.*

$(\tilde{x}, \tilde{\lambda}) = (\tilde{x}_1 + i\tilde{x}_2, \tilde{\lambda}_1 + i\tilde{\lambda}_2)$ sei nun Näherung eines Eigenpaares (x^*, λ^*) mit $(x^*)_s = 1$ (d.h. Näherung einer Nullstelle von f).
Es wird

$$\Delta x_1 := x_1 - \tilde{x}_1,\ \Delta x_2 := x_2 - \tilde{x}_2,\ \Delta\lambda_1 := \lambda_1 - \tilde{\lambda}_1,\ \Delta\lambda_2 := \lambda_2 - \tilde{\lambda}_2$$

definiert und anschließend $f(x_1, x_2, \lambda_1, \lambda_2) = f(\tilde{x}_1 + \Delta x_1, \tilde{x}_2 + \Delta x_2, \tilde{\lambda}_1 + \Delta\lambda_1, \tilde{\lambda}_2 + \Delta\lambda_2)$ berechnet:

$$\begin{aligned} f(x_1, x_2, \lambda_1, \lambda_2) \;=\; & f(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2) + \\ & \begin{pmatrix} R(P(\tilde{\lambda})) & R(P'(\tilde{\lambda})\tilde{x}) + R(P'(\tilde{\lambda})\Delta x) + M(\tilde{x}, \Delta x, \Delta\lambda) \\ {e_s}^T & 0^T \\ {e_{n+s}}^T & 0^T \end{pmatrix} \begin{pmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta\lambda_1 \\ \Delta\lambda_2 \end{pmatrix} \end{aligned} \tag{6.6}$$

mit

$$
\begin{aligned}
R(P'(\lambda)x) &:= \begin{pmatrix} \mathrm{Re}\,(P'(\lambda)x) & -\mathrm{Im}\,(P'(\lambda)x) \\ \mathrm{Im}\,(P'(\lambda)x) & \mathrm{Re}\,(P'(\lambda)x) \end{pmatrix} \in \mathbb{R}^{2n\times 2}, \qquad (6.7)\\
\mathrm{Re}\,(P'(\lambda)x) &= (2A_2\lambda_1 + A_1)x_1 - 2A_2\lambda_2 x_2 \in \mathbb{R}^n,\\
\mathrm{Im}\,(P'(\lambda)x) &= 2A_2\lambda_2 x_1 + (2A_2\lambda_1 + A_1)x_2 \in \mathbb{R}^n,\\
M(\tilde{x},\Delta x,\Delta\lambda) &:= \begin{pmatrix} M_{1,\tilde{x},\Delta x,\Delta\lambda} & -M_{2,\tilde{x},\Delta x,\Delta\lambda} \\ M_{2,\tilde{x},\Delta x,\Delta\lambda} & M_{1,\tilde{x},\Delta x,\Delta\lambda} \end{pmatrix} \in \mathbb{R}^{2n\times 2},\\
M_{1,\tilde{x},\Delta x,\Delta\lambda} &:= (A_2\tilde{x}_1)\Delta\lambda_1 + A_2\Delta x_1\Delta\lambda_1 - (A_2\tilde{x}_2)\Delta\lambda_2 - A_2\Delta x_2\Delta\lambda_2,\\
M_{2,\tilde{x},\Delta x,\Delta\lambda} &:= (A_2\tilde{x}_2)\Delta\lambda_1 + A_2\Delta x_2\Delta\lambda_1 + (A_2\tilde{x}_1)\Delta\lambda_2 + A_2\Delta x_1\Delta\lambda_2,
\end{aligned}
$$

wobei $P'(\lambda) = 2A_2\lambda + A_1 = (2A_2\lambda_1 + A_1) + i(2A_2\lambda_2)$, siehe (2.22).

(6.6) ist genau die Taylor-Entwicklung von f an der Stelle $(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2)$. Diese Tatsache liefert, dass die Jacobi–Matrix von f an der Stelle $(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2)$ die Gestalt

$$
J(\tilde{x},\tilde{\lambda}) := \left(\frac{\partial f_i}{\partial y_j}(\tilde{x}_1,\tilde{x}_2,\tilde{\lambda}_1,\tilde{\lambda}_2)\right)_{i,j=1,\ldots,2n+2} = \begin{pmatrix} R(P(\tilde{\lambda})) & R(P'(\tilde{\lambda})\tilde{x}) \\ {e_s}^T & 0^T \\ {e_{n+s}}^T & 0^T \end{pmatrix} \qquad (6.8)
$$

(mit $x_1 = (y_i)_{i=1,\ldots,n}$, $x_2 = (y_i)_{i=n+1,\ldots,2n}$, $y_{2n+1} = \lambda_1$ und $y_{2n+2} = \lambda_2$) besitzt. Wegen der speziellen Gestalt von $f(x_1, x_2, \lambda_1, \lambda_2)$ ist die Taylor-Entwicklung endlich.

Die Multiplikation der rechten Seite von (6.6) mit einer geeigneten Matrix
$-C = -(c_{ij}) \in \mathbb{R}^{(2n+2)\times(2n+2)}$ und eine anschließende Addition von $({\Delta x_1}^T, {\Delta x_2}^T, \Delta\lambda_1, \Delta\lambda_2)^T$ liefert die Funktion $g : \mathbb{R}^{2n+2} \to \mathbb{R}^{2n+2}$

$$
\begin{aligned}
g(\Delta x_1,\Delta x_2,\Delta\lambda_1,\Delta\lambda_2) := -Cf(\tilde{x}_1,\tilde{x}_2,\tilde{\lambda}_1,\tilde{\lambda}_2) + \Bigg\{ I_{2n+2} - \qquad (6.9)\\
C\begin{pmatrix} R(P(\tilde{\lambda})) & R(P'(\tilde{\lambda})\tilde{x}) + R(P'(\tilde{\lambda})\Delta x) + M(\tilde{x},\Delta x,\Delta\lambda) \\ {e_s}^T & 0^T \\ {e_{n+s}}^T & 0^T \end{pmatrix}\Bigg\} \begin{pmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta\lambda_1 \\ \Delta\lambda_2 \end{pmatrix}.
\end{aligned}
$$

Wegen der Definition von g und Lemma 18 gilt dann folgende Aussage.

Lemma 19

C sei regulär. Dann gilt:

$(x^*, \lambda^*) = (x_1^* + ix_2^*, \lambda_1^* + i\lambda_2^*)$ ist genau dann ein komplexes Eigenpaar von $P(\lambda)$ mit $(x^*)_s = 1$, wenn $(\Delta x_1^*, \Delta x_2^*, \Delta\lambda_1^*, \Delta\lambda_2^*)$ ein Fixpunkt von g ist.

Nun wird der Funktionsausdruck von g wieder in eine für Intervall-Fixpunktiteration günstigere Gestalt gebracht.

$$
C^* := (c_{ij})_{\substack{i=1,\ldots,2n+2 \\ j=1,\ldots,n}} \quad \text{bzw.} \quad C^{**} := (c_{ij})_{\substack{i=1,\ldots,2n+2 \\ j=n+1,\ldots,2n}}
$$

sei die Matrix mit den ersten n bzw. zweiten n Spalten von C. Weiterhin sei

$$\begin{aligned} D_1 &:= C^* A_1 \in \mathbb{R}^{(2n+2)\times n}, \qquad D_2 := C^* A_2 \in \mathbb{R}^{(2n+2)\times n}, \\ E_1 &:= C^{**} A_1 \in \mathbb{R}^{(2n+2)\times n}, \qquad E_2 := C^{**} A_2 \in \mathbb{R}^{(2n+2)\times n}. \end{aligned}$$

Für die Matrix $(C^* \ C^{**}) \in \mathbb{R}^{(2n+2)\times 2n}$ der ersten $2n$ Spalten von C und $v_1, v_2 \in \mathbb{R}^n$ gilt dann

$$(C^* \ C^{**}) \begin{pmatrix} v_1 & -v_2 \\ v_2 & v_1 \end{pmatrix} = \begin{pmatrix} C^* v_1 + C^{**} v_2 & -C^* v_2 + C^{**} v_1 \end{pmatrix} \in \mathbb{R}^{(2n+2)\times 2}. \tag{6.10}$$

Aus (6.9) erhält man damit und mit (6.8)

$$\begin{aligned} g(\Delta x_1, \Delta x_2, \Delta\lambda_1, \Delta\lambda_2) = -Cf(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2) + \Bigg\{ & I_{2n+2} - CJ(\tilde{x}, \tilde{\lambda}) \\ & -C \begin{pmatrix} O & R(P'(\tilde{\lambda})\Delta x) + M(\tilde{x}, \Delta x, \Delta\lambda) \\ 0^T & 0^T \\ 0^T & 0^T \end{pmatrix} \Bigg\} \begin{pmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta\lambda_1 \\ \Delta\lambda_2 \end{pmatrix} \\ = -Cf(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2) + \Big\{ & I_{2n+2} - CJ(\tilde{x}, \tilde{\lambda}) \\ & - \begin{pmatrix} O & N_1(\tilde{x}, \tilde{\lambda}, \Delta x, \Delta\lambda) & N_2(\tilde{x}, \tilde{\lambda}, \Delta x, \Delta\lambda) \end{pmatrix} \Big\} \begin{pmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta\lambda_1 \\ \Delta\lambda_2 \end{pmatrix} \end{aligned} \tag{6.11}$$

mit

$$\begin{aligned} N_1(\tilde{x}, \tilde{\lambda}, \Delta x, \Delta\lambda) &= C^* \big(\mathrm{Re}\,(P'(\tilde{\lambda})\Delta x) + M_{1,\tilde{x},\Delta x,\Delta\lambda} \big) + C^{**} \big(\mathrm{Im}\,(P'(\tilde{\lambda})\Delta x) + M_{2,\tilde{x},\Delta x,\Delta\lambda} \big) \\ &= (2D_2\tilde{\lambda}_1 + D_1)\Delta x_1 - 2D_2\tilde{\lambda}_2\Delta x_2 + (D_2\tilde{x}_1)\Delta\lambda_1 - (D_2\tilde{x}_2)\Delta\lambda_2 \\ &\quad + D_2\Delta x_1\Delta\lambda_1 - D_2\Delta x_2\Delta\lambda_2 \\ &\quad + 2E_2\tilde{\lambda}_2\Delta x_1 + (2E_2\tilde{\lambda}_1 + E_1)\Delta x_2 + (E_2\tilde{x}_2)\Delta\lambda_1 + (E_2\tilde{x}_1)\Delta\lambda_2 \\ &\quad + E_2\Delta x_2\Delta\lambda_1 + E_2\Delta x_1\Delta\lambda_2 \\ &= (D_2\tilde{x}_1 + E_2\tilde{x}_2)\Delta\lambda_1 + (E_2\tilde{x}_1 - D_2\tilde{x}_2)\Delta\lambda_2 + (2D_2\tilde{\lambda}_1 + 2E_2\tilde{\lambda}_2 + D_1)\Delta x_1 \\ &\quad + (2E_2\tilde{\lambda}_1 - 2D_2\tilde{\lambda}_2 + E_1)\Delta x_2 + (D_2\Delta x_1)\Delta\lambda_1 - (D_2\Delta x_2)\Delta\lambda_2 \\ &\quad + (E_2\Delta x_2)\Delta\lambda_1 + (E_2\Delta x_1)\Delta\lambda_2, \end{aligned}$$

$$\begin{aligned} N_2(\tilde{x}, \tilde{\lambda}, \Delta x, \Delta\lambda) &= -C^* \big(\mathrm{Im}\,(P'(\tilde{\lambda})\Delta x) + M_{2,\tilde{x},\Delta x,\Delta\lambda} \big) + C^{**} \big(\mathrm{Re}\,(P'(\tilde{\lambda})\Delta x) + M_{1,\tilde{x},\Delta x,\Delta\lambda} \big) \\ &= -2D_2\tilde{\lambda}_2\Delta x_1 - (2D_2\tilde{\lambda}_1 + D_1)\Delta x_2 - (D_2\tilde{x}_2)\Delta\lambda_1 - (D_2\tilde{x}_1)\Delta\lambda_2 \\ &\quad - D_2\Delta x_2\Delta\lambda_1 - D_2\Delta x_1\Delta\lambda_2 \\ &\quad + (2E_2\tilde{\lambda}_1 + E_1)\Delta x_1 - 2E_2\tilde{\lambda}_2\Delta x_2 + (E_2\tilde{x}_1)\Delta\lambda_1 - (E_2\tilde{x}_2)\Delta\lambda_2 \\ &\quad + E_2\Delta x_1\Delta\lambda_1 - E_2\Delta x_2\Delta\lambda_2 \\ &= (E_2\tilde{x}_1 - D_2\tilde{x}_2)\Delta\lambda_1 - (D_2\tilde{x}_1 + E_2\tilde{x}_2)\Delta\lambda_2 + (2E_2\tilde{\lambda}_1 - 2D_2\tilde{\lambda}_2 + E_1)\Delta x_1 \\ &\quad - (2D_2\tilde{\lambda}_1 + 2E_2\tilde{\lambda}_2 + D_1)\Delta x_2 + (E_2\Delta x_1)\Delta\lambda_1 - (E_2\Delta x_2)\Delta\lambda_2 \\ &\quad - (D_2\Delta x_2)\Delta\lambda_1 - (D_2\Delta x_1)\Delta\lambda_2. \end{aligned}$$

Der in (6.11) zuletzt erhaltene Funktionsausdruck von g ist intervallmäßig auswertbar. Es wird im Beweis des folgenden Satz gezeigt, dass man unter gewissen Voraussetzungen Satz 20 auf die Funktion g anwenden kann. Ein bestimmter $(2n+2)$–komponentiger Intervallvektor enthält dann also einen Fixpunkt von g. Dieser kann dann mit Hilfe von Intervall–Fixpunktiteration bzgl. g eng eingeschlossen werden. Mit Lemma 19 erhält man so Einschließungen von Real– und Imaginärteil eines komplexen Eigenwertes λ^* und des zugehörigen Eigenvektors x^* (mit $(x^*)_s = 1$) von $P(\lambda)$.

Satz 23

Es sei $P(\lambda)$ wie in (6.1), $\tilde{\lambda} = \tilde{\lambda}_1 + i\tilde{\lambda}_2 \in \mathbb{C}$, $\tilde{x} = \tilde{x}_1 + i\tilde{x}_2 \in \mathbb{C}^n$ und $C \in \mathbb{R}^{(2n+2)\times(2n+2)}$.
Für die wie folgt definierten Ausdrücke

$$\varphi := ||Cf(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2)||_\infty\,, \quad \sigma := ||I_{2n+2} - CJ(\tilde{x}, \tilde{\lambda})||_\infty\,,$$

$$\tau := 2||C||_\infty\{(2|\tilde{\lambda}_1| + 2|\tilde{\lambda}_2| + ||\tilde{x}_1||_\infty + ||\tilde{x}_2||_\infty)||A_2||_\infty + ||A_1||_\infty\}\,, \quad \gamma := 4||C||_\infty||A_2||_\infty\,,$$

$$p := \frac{\sigma - 1}{\gamma} - \frac{\tau^2}{3\,\gamma^2} \;\; \textit{und} \;\; q := \frac{2\,\tau^3}{27\,\gamma^3} - \frac{(\sigma-1)\tau}{3\,\gamma^2} + \frac{\varphi}{\gamma}$$

seien die Bedingungen

$$\varphi > 0\,,\ \sigma < 1 \tag{6.12}$$

und

$$D := \frac{q^2}{4} + \frac{p^3}{27} < 0 \tag{6.13}$$

erfüllt.

Mit $\alpha := \arccos\left(-\dfrac{\sqrt{27}q}{2\sqrt{-p^3}}\right)$ *gilt dann für*

$$\beta^- := 2\sqrt{\frac{-p}{3}}\cos\left(\frac{\alpha + 4\pi}{3}\right) - \frac{\tau}{3\,\gamma} \qquad \textit{und} \qquad \beta^+ := 2\sqrt{\frac{-p}{3}}\cos\left(\frac{\alpha}{3}\right) - \frac{\tau}{3\,\gamma}$$

die Eigenschaft

$$0 < \beta^- < \beta^+\,.$$

Weiter sei

$$g(\Delta x_1, \Delta x_2, \Delta\lambda_1, \Delta\lambda_2) = -Cf(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2) + \Big\{I_{2n+2} - CJ(\tilde{x}, \tilde{\lambda}) - \begin{pmatrix} O & N_1(\tilde{x}, \tilde{\lambda}, \Delta x, \Delta\lambda) & N_2(\tilde{x}, \tilde{\lambda}, \Delta x, \Delta\lambda) \end{pmatrix}\Big\} \begin{pmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta\lambda_1 \\ \Delta\lambda_2 \end{pmatrix} \tag{6.14}$$

mit

$$
\begin{aligned}
N_1(\tilde{x},\tilde{\lambda},\Delta x,\Delta\lambda) &= (D_2\tilde{x}_1+E_2\tilde{x}_2)\Delta\lambda_1+(E_2\tilde{x}_1-D_2\tilde{x}_2)\Delta\lambda_2+(2D_2\tilde{\lambda}_1+2E_2\tilde{\lambda}_2+D_1)\Delta x_1\\
&\quad +(2E_2\tilde{\lambda}_1-2D_2\tilde{\lambda}_2+E_1)\Delta x_2+(D_2\Delta x_1)\Delta\lambda_1-(D_2\Delta x_2)\Delta\lambda_2\\
&\quad +(E_2\Delta x_2)\Delta\lambda_1+(E_2\Delta x_1)\Delta\lambda_2\,,\\
N_2(\tilde{x},\tilde{\lambda},\Delta x,\Delta\lambda) &= (E_2\tilde{x}_1-D_2\tilde{x}_2)\Delta\lambda_1-(D_2\tilde{x}_1+E_2\tilde{x}_2)\Delta\lambda_2+(2E_2\tilde{\lambda}_1-2D_2\tilde{\lambda}_2+E_1)\Delta x_1\\
&\quad -(2D_2\tilde{\lambda}_1+2E_2\tilde{\lambda}_2+D_1)\Delta x_2+(E_2\Delta x_1)\Delta\lambda_1-(E_2\Delta x_2)\Delta\lambda_2\\
&\quad -(D_2\Delta x_2)\Delta\lambda_1-(D_2\Delta x_1)\Delta\lambda_2
\end{aligned}
$$

($D_1 := C^*A_1$, $D_2 := C^*A_2$, $E_1 := C^{**}A_1$, $E_2 := C^{**}A_2$) und

$$
\begin{pmatrix} [\Delta x_1]^{(0)}\\ [\Delta x_2]^{(0)}\\ [\Delta\lambda_1]^{(0)}\\ [\Delta\lambda_2]^{(0)} \end{pmatrix} := [-\beta,\beta]e \in \mathbf{IR}^{2n+2}
$$

für beliebiges $\beta \in (\beta^-,\beta^+)$. Dann konvergiert die Iteriertenfolge

$$
\begin{pmatrix} [\Delta x_1]^{(k+1)}\\ [\Delta x_2]^{(k+1)}\\ [\Delta\lambda_1]^{(k+1)}\\ [\Delta\lambda_2]^{(k+1)} \end{pmatrix} := g([\Delta x_1]^{(k)},[\Delta x_2]^{(k)},[\Delta\lambda_1]^{(k)},[\Delta\lambda_2]^{(k)})\,,\quad k\in\mathbb{N}_0,
$$

gegen einen Intervallvektor

$$
\begin{pmatrix} [\Delta x_1]^{*}\\ [\Delta x_2]^{*}\\ [\Delta\lambda_1]^{*}\\ [\Delta\lambda_2]^{*} \end{pmatrix} \subseteq \begin{pmatrix} [\Delta x_1]^{(k+1)}\\ [\Delta x_2]^{(k+1)}\\ [\Delta\lambda_1]^{(k+1)}\\ [\Delta\lambda_2]^{(k+1)} \end{pmatrix} \subseteq \begin{pmatrix} [\Delta x_1]^{(k)}\\ [\Delta x_2]^{(k)}\\ [\Delta\lambda_1]^{(k)}\\ [\Delta\lambda_2]^{(k)} \end{pmatrix} \subseteq \cdots \subseteq \begin{pmatrix} [\Delta x_1]^{(0)}\\ [\Delta x_2]^{(0)}\\ [\Delta\lambda_1]^{(0)}\\ [\Delta\lambda_2]^{(0)} \end{pmatrix},\quad k\in\mathbb{N},
$$

und es existiert mindestens ein komplexes Eigenpaar $(x_1^*+ix_2^*,\lambda_1^*+i\lambda_2^*)$ von $P(\lambda)$ mit $(x_1^*+ix_2^*)_s = 1$ und

$$
\begin{pmatrix} x_1^*\\ x_2^*\\ \lambda_1^*\\ \lambda_2^* \end{pmatrix} \in \begin{pmatrix} \tilde{x}_1\\ \tilde{x}_2\\ \tilde{\lambda}_1\\ \tilde{\lambda}_2 \end{pmatrix} + \begin{pmatrix} [\Delta x_1]^{(k)}\\ [\Delta x_2]^{(k)}\\ [\Delta\lambda_1]^{(k)}\\ [\Delta\lambda_2]^{(k)} \end{pmatrix},\quad k\in\mathbb{N}_0.
$$

Unter der Einschränkung $\beta \in (\beta^-,\sqrt{\frac{-p}{3}}-\frac{\tau}{3\gamma})$ ist dieses Eigenpaar eindeutig, und die Iterierten $\begin{pmatrix} [\Delta x_1]^{(k)}\\ [\Delta x_2]^{(k)}\\ [\Delta\lambda_1]^{(k)}\\ [\Delta\lambda_2]^{(k)} \end{pmatrix}$ *konvergieren gegen dessen Approximationsfehler* $\begin{pmatrix} \Delta x_1^*\\ \Delta x_2^*\\ \Delta\lambda_1^*\\ \Delta\lambda_2^* \end{pmatrix}$.

Beweis:

Es sei $([\Delta x_1]^T,[\Delta x_2]^T,[\Delta\lambda_1],[\Delta\lambda_2])^T \in \mathbf{IR}^{2n+2}$, $[\Delta x] := [\Delta x_1]+i[\Delta x_2]$ und $[\Delta\lambda] := [\Delta\lambda_1]+i[\Delta\lambda_2]$. Es wird die Abkürzung $[\Delta\lambda_k]^2 := [\Delta\lambda_k]\cdot[\Delta\lambda_k]$, $k=1,2$, verwendet.

Dann gilt

$$
\begin{aligned}
&g([\Delta x_1],[\Delta x_2],[\Delta\lambda_1],[\Delta\lambda_2]) \;=\\
&\qquad -Cf(\tilde{x}_1,\tilde{x}_2,\tilde{\lambda}_1,\tilde{\lambda}_2)+\Big\{I_{2n+2}-CJ(\tilde{x},\tilde{\lambda})\\
&\qquad -\begin{pmatrix} O & N_1(\tilde{x},\tilde{\lambda},[\Delta x],[\Delta\lambda]) & N_2(\tilde{x},\tilde{\lambda},[\Delta x],[\Delta\lambda])\end{pmatrix}\Big\}\begin{pmatrix}[\Delta x_1]\\[\Delta x_2]\\[\Delta\lambda_1]\\[\Delta\lambda_2]\end{pmatrix}\\
&\overset{(3.4)}{\subseteq} -Cf(\tilde{x}_1,\tilde{x}_2,\tilde{\lambda}_1,\tilde{\lambda}_2)+\Big(I_{2n+2}-CJ(\tilde{x},\tilde{\lambda})\Big)\begin{pmatrix}[\Delta x_1]\\[\Delta x_2]\\[\Delta\lambda_1]\\[\Delta\lambda_2]\end{pmatrix}\\
&\qquad -\begin{pmatrix} O & N_1(\tilde{x},\tilde{\lambda},[\Delta x],[\Delta\lambda]) & N_2(\tilde{x},\tilde{\lambda},[\Delta x],[\Delta\lambda])\end{pmatrix}\begin{pmatrix}[\Delta x_1]\\[\Delta x_2]\\[\Delta\lambda_1]\\[\Delta\lambda_2]\end{pmatrix}\\
&= -Cf(\tilde{x}_1,\tilde{x}_2,\tilde{\lambda}_1,\tilde{\lambda}_2)+\Big(I_{2n+2}-CJ(\tilde{x},\tilde{\lambda})\Big)([\Delta x_1]^T,[\Delta x_2]^T,[\Delta\lambda_1],[\Delta\lambda_2])^T\\
&\qquad -N_1(\tilde{x},\tilde{\lambda},[\Delta x],[\Delta\lambda])\cdot[\Delta\lambda_1]-N_2(\tilde{x},\tilde{\lambda},[\Delta x],[\Delta\lambda])\cdot[\Delta\lambda_2]\\
&\overset{(3.2)}{\subseteq} -Cf(\tilde{x}_1,\tilde{x}_2,\tilde{\lambda}_1,\tilde{\lambda}_2)+\Big(I_{2n+2}-CJ(\tilde{x},\tilde{\lambda})\Big)([\Delta x_1]^T,[\Delta x_2]^T,[\Delta\lambda_1],[\Delta\lambda_2])^T\\
&\qquad +\Big(-(D_2\tilde{x}_1+E_2\tilde{x}_2)_i[\Delta\lambda_1]^2-(E_2\tilde{x}_1-D_2\tilde{x}_2)_i[\Delta\lambda_2][\Delta\lambda_1]\\
&\qquad -\sum_{k=1}^{n}(2D_2\tilde{\lambda}_1+2E_2\tilde{\lambda}_2+D_1)_{ik}[\Delta x_1]_k[\Delta\lambda_1]-\sum_{k=1}^{n}(2E_2\tilde{\lambda}_1-2D_2\tilde{\lambda}_2+E_1)_{ik}[\Delta x_2]_k[\Delta\lambda_1]\\
&\qquad -\sum_{l=1}^{n}(D_2)_{il}[\Delta x_1]_l[\Delta\lambda_1]^2+\sum_{l=1}^{n}(D_2)_{il}[\Delta x_2]_l[\Delta\lambda_2][\Delta\lambda_1]-\sum_{l=1}^{n}(E_2)_{il}[\Delta x_2]_l[\Delta\lambda_1]^2\\
&\qquad -\sum_{l=1}^{n}(E_2)_{il}[\Delta x_1]_l[\Delta\lambda_2][\Delta\lambda_1]-(E_2\tilde{x}_1-D_2\tilde{x}_2)_i[\Delta\lambda_1][\Delta\lambda_2]+(D_2\tilde{x}_1+E_2\tilde{x}_2)_i[\Delta\lambda_2]^2\\
&\qquad -\sum_{k=1}^{n}(2E_2\tilde{\lambda}_1-2D_2\tilde{\lambda}_2+E_1)_{ik}[\Delta x_1]_k[\Delta\lambda_2]+\sum_{k=1}^{n}(2D_2\tilde{\lambda}_1+2E_2\tilde{\lambda}_2+D_1)_{ik}[\Delta x_2]_k[\Delta\lambda_2]\\
&\qquad -\sum_{l=1}^{n}(E_2)_{il}[\Delta x_1]_l[\Delta\lambda_1][\Delta\lambda_2]+\sum_{l=1}^{n}(E_2)_{il}[\Delta x_2]_l[\Delta\lambda_2]^2\\
&\qquad +\sum_{l=1}^{n}(D_2)_{il}[\Delta x_2]_l[\Delta\lambda_1][\Delta\lambda_2]+\sum_{l=1}^{n}(D_2)_{il}[\Delta x_1]_l[\Delta\lambda_2]^2\Big)_{i=1,\dots,2n+2}\\
&= \quad r+S\begin{pmatrix}[\Delta x_1]\\[\Delta x_2]\\[\Delta\lambda_1]\\[\Delta\lambda_2]\end{pmatrix}+T\begin{pmatrix}[\Delta x_1]\\[\Delta x_2]\\[\Delta\lambda_1]\\[\Delta\lambda_2]\end{pmatrix}^2+U\begin{pmatrix}[\Delta x_1]\\[\Delta x_2]\\[\Delta\lambda_1]\\[\Delta\lambda_2]\end{pmatrix}^3=:h([\Delta x_1],[\Delta x_2],[\Delta\lambda_1],[\Delta\lambda_2]).
\end{aligned}
\tag{6.15}
$$

$$h(x) = r + Sx + Tx^2 + Ux^3\,, \quad x \in \mathbb{R}^{2n+2},$$

ist ein kubisches System (4.1) mit $m = 2n+2,\ r = -Cf(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2)\,,$
$S := I_{2n+2} - CJ(\tilde{x}, \tilde{\lambda})\,,\ T = (t_{ijk}) \in \mathbb{R}^{(2n+2)\times(2n+2)\times(2n+2)}$ und
$U = (u_{ijkl}) \in \mathbb{R}^{(2n+2)\times(2n+2)\times(2n+2)\times(2n+2)}\,.$
Dabei ist für $i = 1, \ldots, 2n+2$

$$t_{ijk} = \begin{cases} -(D_2\tilde{x}_1 + E_2\tilde{x}_2)_i\,, & j = k = 2n+1, \\ (-E_2\tilde{x}_1 + D_2\tilde{x}_2)_i\,, & j = 2n+1;\ k = 2n+2, \\ (-E_2\tilde{x}_1 + D_2\tilde{x}_2)_i\,, & j = 2n+2;\ k = 2n+1, \\ (D_2\tilde{x}_1 + E_2\tilde{x}_2)_i\,, & j = k = 2n+2, \\ -(2D_2\tilde{\lambda}_1 + 2E_2\tilde{\lambda}_2 + D_1)_{ik}\,, & j = 2n+1;\ k = 1, \ldots, n, \\ (-2E_2\tilde{\lambda}_1 + 2D_2\tilde{\lambda}_2 - E_1)_{i,k-n}\,, & j = 2n+1;\ k = n+1, \ldots, 2n, \\ (-2E_2\tilde{\lambda}_1 + 2D_2\tilde{\lambda}_2 - E_1)_{ik}\,, & j = 2n+2;\ k = 1, \ldots, n, \\ (2D_2\tilde{\lambda}_1 + 2E_2\tilde{\lambda}_2 + D_1)_{i,k-n}\,, & j = 2n+2;\ k = n+1, \ldots, 2n, \\ 0\,, & \text{sonst}, \end{cases}$$

$$u_{ijkl} = \begin{cases} -(D_2)_{il}\,, & j = k = 2n+1;\ l = 1, \ldots, n, \\ (D_2)_{i,l-n}\,, & j = 2n+1;\ k = 2n+2;\ l = n+1, \ldots, 2n, \\ -(E_2)_{i,l-n}\,, & j = k = 2n+1;\ l = n+1, \ldots, 2n, \\ -(E_2)_{il}\,, & j = 2n+1;\ k = 2n+2;\ l = 1, \ldots, n, \\ (E_2)_{i,l-n}\,, & j = k = 2n+2;\ l = n+1, \ldots, 2n, \\ -(E_2)_{il}\,, & j = 2n+2;\ k = 2n+1;\ l = 1, \ldots, n, \\ (D_2)_{il}\,, & j = k = 2n+2;\ l = 1, \ldots, n, \\ (D_2)_{i,l-n}\,, & j = 2n+2;\ k = 2n+1;\ l = n+1, \ldots, 2n, \\ 0\,, & \text{sonst}. \end{cases}$$

Dann ist $||r||_\infty = ||Cf(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2)||_\infty = \varphi$ und $||S||_\infty = ||I_{2n+2} - CJ(\tilde{x}, \tilde{\lambda})||_\infty = \sigma$. Weiter gilt

$$
\begin{aligned}
u_\infty &= \max_{1\le i\le 2n+2}\Big(\sum_{j=1}^{2n+2}\sum_{k=1}^{2n+2}\sum_{l=1}^{2n+2}|u_{ijkl}|\Big)\\
&= \max_{1\le i\le 2n+2}\Big(\sum_{l=1}^{n}(2|(D_2)_{il}|+2|(E_2)_{il}|)+\sum_{l=n+1}^{2n}(2|(D_2)_{i,l-n}|+2|(E_2)_{i,l-n}|)\Big)\\
&= \max_{1\le i\le 2n+2}\Big(\sum_{l=1}^{n}(4|(D_2)_{il}|+4|(E_2)_{il}|)\Big)\\
&\overset{\substack{\max \text{ für}\\ i\equiv i^*}}{=} \sum_{l=1}^{n}(4|(D_2)_{i^*l}|+4|(E_2)_{i^*l}|)\\
&= 4\sum_{l=1}^{n}\Big(\Big|\sum_{t=1}^{n}c_{i^*t}(A_2)_{tl}\Big|+\Big|\sum_{t=1}^{n}c_{i^*,n+t}(A_2)_{tl}\Big|\Big)\\
&\le 4\sum_{l=1}^{n}\Big(\sum_{t=1}^{n}|c_{i^*t}|\,|(A_2)_{tl}|+\sum_{t=1}^{n}|c_{i^*,n+t}|\,|(A_2)_{tl}|\Big)\\
&= 4\sum_{l=1}^{n}\sum_{t=1}^{n}(|c_{i^*t}|+|c_{i^*,n+t}|)|(A_2)_{tl}|\\
&= 4\sum_{t=1}^{n}\Big((|c_{i^*t}|+|c_{i^*,n+t}|)\sum_{l=1}^{n}|(A_2)_{tl}|\Big)\\
&\le 4||A_2||_\infty\sum_{t=1}^{n}(|c_{i^*t}|+|c_{i^*,n+t}|)\\
&\le 4||A_2||_\infty||C||_\infty = \gamma\,. \qquad (6.16)
\end{aligned}
$$

Weiterhin gilt für beliebiges $i \in \{1, \ldots, 2n+2\}$

$$
\begin{aligned}
&|t_{i,2n+1,2n+1}| + |t_{i,2n+2,2n+2}| + |t_{i,2n+1,2n+2}| + |t_{i,2n+2,2n+1}| \\
&= 2|(D_2\tilde{x}_1 + E_2\tilde{x}_2)_i| + 2|(-E_2\tilde{x}_1 + D_2\tilde{x}_2)_i| \\
&\leq 2\big(|(D_2\tilde{x}_1)_i| + |(E_2\tilde{x}_1)_i| + |(D_2\tilde{x}_2)_i| + |(E_2\tilde{x}_2)_i|\big) \\
&= 2\big(|\sum_{t=1}^{n} c_{it}(A_2\tilde{x}_1)_t| + |\sum_{t=1}^{n} c_{i,n+t}(A_2\tilde{x}_1)_t| + |\sum_{t=1}^{n} c_{it}(A_2\tilde{x}_2)_t| + |\sum_{t=1}^{n} c_{i,n+t}(A_2\tilde{x}_2)_t|\big) \\
&\leq 2\big(\sum_{t=1}^{n} |c_{it}|\,|(A_2\tilde{x}_1)_t| + \sum_{t=1}^{n} |c_{i,n+t}|\,|(A_2\tilde{x}_1)_t| + \sum_{t=1}^{n} |c_{it}|\,|(A_2\tilde{x}_2)_t| + \sum_{t=1}^{n} |c_{i,n+t}|\,|(A_2\tilde{x}_2)_t|\big) \\
&\leq 2\big(||A_2\tilde{x}_1||_\infty \sum_{t=1}^{n} |c_{it}| + ||A_2\tilde{x}_1||_\infty \sum_{t=1}^{n} |c_{i,n+t}| + ||A_2\tilde{x}_2||_\infty \sum_{t=1}^{n} |c_{it}| + ||A_2\tilde{x}_1||_\infty \sum_{t=1}^{n} |c_{i,n+t}|\big) \\
&= 2\big(||A_2\tilde{x}_1||_\infty \sum_{t=1}^{n} (|c_{it}| + |c_{i,n+t}|) + ||A_2\tilde{x}_2||_\infty \sum_{t=1}^{n} (|c_{it}| + |c_{i,n+t}|)\big) \\
&\leq 2||A_2||_\infty||\tilde{x}_1||_\infty \sum_{t=1}^{n} (|c_{it}| + |c_{i,n+t}|) + 2||A_2||_\infty||\tilde{x}_2||_\infty \sum_{t=1}^{n} (|c_{it}| + |c_{i,n+t}|) \\
&\leq 2||A_2||_\infty||\tilde{x}_1||_\infty||C||_\infty + 2||A_2||_\infty||\tilde{x}_2||_\infty||C||_\infty \\
&= 2||C||_\infty||A_2||_\infty(||\tilde{x}_1||_\infty + ||\tilde{x}_2||_\infty)\,. \qquad (6.17)
\end{aligned}
$$

Außerdem gilt

$$
\begin{aligned}
\sum_{k=1}^{n} |t_{i,2n+1,k}| &= \sum_{k=n+1}^{2n} |t_{i,2n+2,k}| = \sum_{k=1}^{n} |(2D_2\tilde{\lambda}_1 + 2E_2\tilde{\lambda}_2 + D_1)_{ik}| \\
&\leq \sum_{k=1}^{n} (2|\tilde{\lambda}_1|\,|(D_2)_{ik}| + 2|\tilde{\lambda}_2|\,|(E_2)_{ik}| + |(D_1)_{ik}|) \\
&\leq \sum_{k=1}^{n} (2|\tilde{\lambda}_1| \sum_{t=1}^{n} |c_{it}|\,|(A_2)_{tk}| + 2|\tilde{\lambda}_2| \sum_{t=1}^{n} |c_{i,n+t}|\,|(A_2)_{tk}| + \sum_{t=1}^{n} |c_{it}|\,|(A_1)_{tk}|) \\
&= 2|\tilde{\lambda}_1| \sum_{t=1}^{n} |c_{it}| (\sum_{k=1}^{n} |(A_2)_{tk}|) + 2|\tilde{\lambda}_2| \sum_{t=1}^{n} |c_{i,n+t}| (\sum_{k=1}^{n} |(A_2)_{tk}|) + \sum_{t=1}^{n} |c_{it}| (\sum_{k=1}^{n} |(A_1)_{tk}|) \\
&\leq 2|\tilde{\lambda}_1|\,||A_2||_\infty \sum_{t=1}^{n} |c_{it}| + 2|\tilde{\lambda}_2|\,||A_2||_\infty \sum_{t=1}^{n} |c_{i,n+t}| + ||A_1||_\infty \sum_{t=1}^{n} |c_{it}|
\end{aligned}
$$

und analog

$$
\begin{aligned}
\sum_{k=1}^{n} |t_{i,2n+2,k}| &= \sum_{k=n+1}^{2n} |t_{i,2n+1,k}| = \sum_{k=1}^{n} |(-2E_2\tilde{\lambda}_1 + 2D_2\tilde{\lambda}_2 - E_1)_{ik}| \\
&\leq 2|\tilde{\lambda}_1|\,||A_2||_\infty \sum_{t=1}^{n} |c_{i,n+t}| + 2|\tilde{\lambda}_2|\,||A_2||_\infty \sum_{t=1}^{n} |c_{it}| + ||A_1||_\infty \sum_{t=1}^{n} |c_{i,n+t}|\,.
\end{aligned}
$$

Somit ist

$$\sum_{k=1}^{n}|t_{i,2n+1,k}|+\sum_{k=n+1}^{2n}|t_{i,2n+2,k}|+\sum_{k=1}^{n}|t_{i,2n+2,k}|+\sum_{k=n+1}^{2n}|t_{i,2n+1,k}|$$

$$\begin{aligned}
&\leq 4|\tilde{\lambda}_1|\,||A_2||_\infty\sum_{t=1}^{n}(|c_{it}|+|c_{i,n+t}|)+4|\tilde{\lambda}_2|\,||A_2||_\infty\sum_{t=1}^{n}(|c_{it}|+|c_{i,n+t}|)\\
&\quad +2||A_1||_\infty\sum_{t=1}^{n}(|c_{it}|+|c_{i,n+t}|)\\
&\leq 4|\tilde{\lambda}_1|\,||A_2||_\infty||C||_\infty+4|\tilde{\lambda}_2|\,||A_2||_\infty||C||_\infty+2||A_1||_\infty||C||_\infty\\
&= 2||C||_\infty\{||A_2||_\infty(2|\tilde{\lambda}_1|+2|\tilde{\lambda}_2|)+||A_1||_\infty\}\,. \qquad (6.18)
\end{aligned}$$

Aus (6.17) und (6.18) folgt somit

$$\begin{aligned}
t_\infty &= \max_{1\leq i\leq 2n+2}\left(\sum_{j=1}^{2n+2}\sum_{k=1}^{2n+2}|t_{ijk}|\right)\\
&\leq 2||C||_\infty||A_2||_\infty(||\tilde{x}_1||_\infty+||\tilde{x}_2||_\infty)+2||C||_\infty\{||A_2||_\infty(2|\tilde{\lambda}_1|+2|\tilde{\lambda}_2|)+||A_1||_\infty\}\\
&= 2||C||_\infty\{||A_2||_\infty(2|\tilde{\lambda}_1|+2|\tilde{\lambda}_2|+||\tilde{x}_1||_\infty+||\tilde{x}_2||_\infty)+||A_1||_\infty\}=\tau\,.
\end{aligned}$$

C ist wegen $\sigma = ||I_{2n+2}-CJ(\tilde{x},\tilde{\lambda})||_\infty < 1$ nach Lemma 13 regulär. Weil auch A_2 nach Voraussetzung regulär ist, sind beide Matrizen ungleich der Nullmatrix, d.h. $\gamma\neq 0$. Außerdem ist der Funktionsausdruck (6.14) von $g:\mathbb{R}^{2n+2}\to\mathbb{R}^{2n+2}$ intervallmäßig auswertbar. Mit Satz 20 und Lemma 19 folgt dann der Satz. □

In Vorbereitung auf Satz 24 werden nun noch einige Begrifflichkeiten eingeführt und ein Lemma gezeigt.

Die Abbildung

$$R(A+iB)=\begin{pmatrix} A & -B\\ B & A\end{pmatrix},\quad A+Bi\in\mathbb{C}^{m\times m}, \qquad (6.19)$$

ist ein Ringisomorphismus zwischen $\mathbb{C}^{m\times m}$ und dem Unterring $\left\{\begin{pmatrix} A & -B\\ B & A\end{pmatrix}\,\middle|\, A,B\in\mathbb{R}^{m\times m}\right\}$ von $\mathbb{R}^{2m\times 2m}$, d.h. es gilt

$$R(C_1+C_2)=R(C_1)+R(C_2)\quad\text{und}\quad R(C_1\cdot C_2)=R(C_1)\cdot R(C_2)\,.$$

Dies ist leicht nachzurechnen.

Lemma 20

$C\in\mathbb{C}^{m\times m}$ ist genau dann nichtsingulär, wenn $R(C)\in\mathbb{R}^{2m\times 2m}$ nichtsingulär ist.

Beweis:

Die Matrix C sei nichtsingulär. Dann existiert ihre Inverse C^{-1}. Wegen

$$R(C)\cdot R(C^{-1})=R(C\cdot C^{-1})=R(I_m)=I_{2m}$$

ist dann also $R(C)$ invertierbar mit $R(C)^{-1} = R(C^{-1})$.

Wenn andererseits die Matrix $R(C) =: \begin{pmatrix} A & -B \\ B & A \end{pmatrix}$ mit $A, B \in \mathbb{R}^{m\times m}$ nichtsingulär ist, existiert ihre inverse Matrix $R(C)^{-1} =: \begin{pmatrix} D & -E \\ E & D \end{pmatrix}$ mit $D, E \in \mathbb{R}^{m\times m}$, d.h. es gilt

$$I_{2m} = R(C)\cdot R(C)^{-1} = \begin{pmatrix} A & -B \\ B & A \end{pmatrix}\begin{pmatrix} D & -E \\ E & D \end{pmatrix} = \begin{pmatrix} AD-BE & -(AE+BD) \\ AE+BD & AD-BE \end{pmatrix},$$

also $I_m = AD - BE = \mathrm{Re}\,((A+iB)(D+iE))$ und $O = AE + BD = \mathrm{Im}\,((A+iB)(D+iE))$ und damit $(A+iB)(D+iE) = I_m$. Damit ist $C = A + iB$ nichtsingulär mit $C^{-1} = D + iE$. □

Satz 24

Es sei (x^*, λ^*) ein komplexes Eigenpaar von $P(\lambda)$ mit $(x^*)_s \neq 0$. Dann gilt:
Der Eigenwert λ^* ist genau dann einfach, wenn $J(x^*, \lambda^*)$ aus (6.8) nichtsingulär ist.

Beweis:
Die Matrix

$$\begin{aligned} J(x^*,\lambda^*) &= \begin{pmatrix} R(P(\lambda^*)) & R(P'(\lambda^*)x^*) \\ {e_s}^T & 0^T \\ {e_{n+s}}^T & 0^T \end{pmatrix} \\ &= \begin{pmatrix} \mathrm{Re}\,(P(\lambda^*)) & -\mathrm{Im}\,(P(\lambda^*)) & \mathrm{Re}\,(P'(\lambda^*)x^*) & -\mathrm{Im}\,(P'(\lambda^*)x^*) \\ \mathrm{Im}\,(P(\lambda^*)) & \mathrm{Re}\,(P(\lambda^*)) & \mathrm{Im}\,(P'(\lambda^*)x^*) & \mathrm{Re}\,(P'(\lambda^*)x^*) \\ {e_s}^T & 0^T & 0 & 0 \\ 0^T & {e_s}^T & 0 & 0 \end{pmatrix} \in \mathbb{R}^{(2n+2)\times(2n+2)} \end{aligned}$$

(wobei $R(P'(\lambda^*)x^*) \in \mathbb{R}^{2n\times 2}$ definiert ist wie in (6.7)) kann man durch Spaltenvertauschungen in die Matrix

$$\begin{pmatrix} \mathrm{Re}\,(P(\lambda^*)) & \mathrm{Re}\,(P'(\lambda^*)x^*) & -\mathrm{Im}\,(P(\lambda^*)) & -\mathrm{Im}\,(P'(\lambda^*)x^*) \\ \mathrm{Im}\,(P(\lambda^*)) & \mathrm{Im}\,(P'(\lambda^*)x^*) & \mathrm{Re}\,(P(\lambda^*)) & \mathrm{Re}\,(P'(\lambda^*)x^*) \\ {e_s}^T & 0 & 0^T & 0 \\ 0^T & 0 & {e_s}^T & 0 \end{pmatrix}$$

und diese wiederum durch Zeilenvertauschungen in die Matrix

$$\begin{pmatrix} \mathrm{Re}\,(P(\lambda^*)) & \mathrm{Re}\,(P'(\lambda^*)x^*) & -\mathrm{Im}\,(P(\lambda^*)) & -\mathrm{Im}\,(P'(\lambda^*)x^*) \\ {e_s}^T & 0 & 0^T & 0 \\ \mathrm{Im}\,(P(\lambda^*)) & \mathrm{Im}\,(P'(\lambda^*)x^*) & \mathrm{Re}\,(P(\lambda^*)) & \mathrm{Re}\,(P'(\lambda^*)x^*) \\ 0^T & 0 & {e_s}^T & 0 \end{pmatrix} =: M$$

überführen. Es gilt dann $\det(M) = \pm\det(J(x^*, \lambda^*))$. Außerdem gilt für

$$B^* := \begin{pmatrix} P(\lambda^*) & P'(\lambda^*)x^* \\ {e_s}^T & 0 \end{pmatrix} \in \mathbb{C}^{(n+1)\times(n+1)}$$

die Beziehung $R(B^*) = M$ (siehe (6.19)), also ist $\det(R(B^*)) = \pm\det(J(x^*, \lambda^*))$. Damit ist $J(x^*, \lambda^*)$ genau dann nichtsingulär, wenn $R(B^*)$ nichtsingulär ist. Lemma 20 und Satz 14 liefern dann die Gültigkeit diese Satzes. □

Bemerkung 12

Die Voraussetzungen des Satzes 23 sind erfüllt, wenn alle folgenden Bedingungen gelten:

a) $(\tilde{x}, \tilde{\lambda})$ *ist eine hinreichend gute Näherung eines komplexen Eigenpaares* (x^*, λ^*) *von* $P(\lambda)$ *mit* $(x^*)_s = 1$ *; o.B.d.A. sei* $(\tilde{x}, \tilde{\lambda})$ *kein Eigenpaar von* $P(\lambda)$*.*

b) $J(\tilde{x}, \tilde{\lambda})$ *ist regulär.*

c) $C := J(\tilde{x}, \tilde{\lambda})^{-1}$ *oder* C *ist eine hinreichend gute Näherung von* $J(\tilde{x}, \tilde{\lambda})^{-1}$ *.*

Wenn a) gilt und der Eigenwert λ^* *zusätzlich einfach ist, impliziert dies b).*
Wenn a) gilt und λ^* *nicht einfach ist, dann ist* $J(\tilde{x}, \tilde{\lambda})$ *fast singulär.*

Wenn b) erfüllt ist, kann $C := J(\tilde{x}, \tilde{\lambda})^{-1}$ (oder zumindest eine hinreichend gute Näherung von $J(\tilde{x}, \tilde{\lambda})^{-1}$) gewählt werden (vgl. c)). Damit ist $\sigma = ||I_{n+1} - CJ(\tilde{x}, \tilde{\lambda})||_\infty = 0$ (oder zumindest $\sigma \approx 0$) und die Bedingung $\sigma < 1$ aus (6.12) erfüllt.

Da nach Annahme $(x^*, \lambda^*) \neq (\tilde{x}, \tilde{\lambda})$ gilt, ist $\varphi \neq 0$, d.h. die zweite Bedingung aus (6.12) ist erfüllt.

Für die hinreichend gute Eigenpaarnäherung aus a) ist $\varphi = ||Cf(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2)||_\infty$ hinreichend klein. Unter Verwendung des Beweises von Satz 23 kann man mit der Heuristik aus Bemerkung 8 folgern, dass $D < 0$ und damit die Bedingung (6.13) gilt. Damit sind alle Voraussetzungen von Satz 23 erfüllt.

Wenn a) gilt und λ^* einfach (bzw. nicht einfach) ist, ist die Matrix $J(x^*, \lambda^*)$ nach Satz 24 regulär (bzw. singulär). Für die hinreichend gute Näherung $(\tilde{x}, \tilde{\lambda})$ von (x^*, λ^*) ist dann aus Stetigkeitsgründen $J(\tilde{x}, \tilde{\lambda})$ regulär (bzw. fast singulär).

Kapitel 7

Einschließung von Eigenpaaren des komplexen QEP

Nun wird das komplexe quadratische Eigenwertproblem (QEP) betrachtet. Dabei ist das reguläre komplexe quadratische $n \times n$–Matrixpolynom

$$P(\lambda) = K_2\lambda^2 + K_1\lambda + K_0 \tag{7.1}$$

mit komplexen Koeffizientenmatrizen

$$K_2 = A_2 + iB_2\,, \quad K_1 = A_1 + iB_1\,, \quad K_0 = A_0 + iB_0 \in \mathbb{C}^{n\times n}$$

und $\det K_2 \neq 0$ gegeben. Gesucht werden komplexe Eigenwerte $\lambda^* = \lambda_1^* + i\lambda_2^*$ und zugehörige komplexe Eigenvektoren $x^* = x_1^* + ix_2^* \in \mathbb{C}^n$ von $P(\lambda)$, d.h. es soll

$$P(\lambda^*)x^* = (K_2\lambda^{*2} + K_1\lambda^* + K_0)x^* = 0$$

für $x^* \neq 0$ gelten.

Es wird jetzt analog zur Einschließung komplexer Eigenpaare des reellen QEP in Kapitel 6 vorgegangen: es werden reelle Intervalle bzw. Intervallvektoren mit hinreichend kleinem Radius gesucht, die Realteil und Imaginärteil der Eigenwerte bzw. der zugehörigen Eigenvektoren von $P(\lambda)$ einschließen.

Weil Lemma 17 mit (6.4) ebenfalls für ein komplexes quadratisches Matrixpolynom $P(\lambda)$ gilt, wird erneut die Funktion

$$f(x_1, x_2, \lambda_1, \lambda_2) := \begin{pmatrix} R(P(\lambda))\binom{x_1}{x_2} \\ {e_s}^T x_1 - 1 \\ {e_s}^T x_2 \end{pmatrix}, \quad x_1, x_2 \in \mathbb{R}^n,\ \lambda_1, \lambda_2 \in \mathbb{R}\ (x = x_1 + ix_2,\ \lambda = \lambda_1 + i\lambda_2),$$

betrachtet. Dabei ist

$$R(P(\lambda)) := \begin{pmatrix} \mathrm{Re}\,(P(\lambda)) & -\mathrm{Im}\,(P(\lambda)) \\ \mathrm{Im}\,(P(\lambda)) & \mathrm{Re}\,(P(\lambda)) \end{pmatrix} \in \mathbb{R}^{2n\times 2n},$$

wobei

$$\begin{aligned} \mathrm{Re}\,(P(\lambda)) &= (\lambda_1^2 - \lambda_2^2)A_2 + \lambda_1 A_1 + A_0 - 2\lambda_1\lambda_2 B_2 - \lambda_2 B_1\,, \\ \mathrm{Im}\,(P(\lambda)) &= 2\lambda_1\lambda_2 A_2 + \lambda_2 A_1 + (\lambda_1^2 - \lambda_2^2)B_2 + \lambda_1 B_1 + B_0\,. \end{aligned}$$

Für die Funktion f gilt dann Lemma 18 . Wir interessieren uns also wieder für Einschließungen der Nullstellen von f.

Für die Näherung $(\tilde{x}, \tilde{\lambda}) = (\tilde{x}_1 + i\tilde{x}_2, \tilde{\lambda}_1 + i\tilde{\lambda}_2)$ eines Eigenpaares (x^*, λ^*) mit $(x^*)_s = 1$ von $P(\lambda)$ sei nun wieder

$$\Delta x_1 := x_1 - \tilde{x}_1,\ \Delta x_2 := x_2 - \tilde{x}_2,\ \Delta\lambda_1 := \lambda_1 - \tilde{\lambda}_1,\ \Delta\lambda_2 := \lambda_2 - \tilde{\lambda}_2\,.$$

Analog zum Kapitel 6 gilt

$$\begin{aligned} f(x_1, x_2, \lambda_1, \lambda_2) &= f(\tilde{x}_1 + \Delta x_1, \tilde{x}_2 + \Delta x_2, \tilde{\lambda}_1 + \Delta\lambda_1, \tilde{\lambda}_2 + \Delta\lambda_2) \\ &= f(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2) + \qquad (7.2) \\ &\quad \begin{pmatrix} R(P(\tilde{\lambda})) & R(P'(\tilde{\lambda})\tilde{x}) + R(P'(\tilde{\lambda})\Delta x) + M(\tilde{x}, \Delta x, \Delta\lambda) \\ {e_s}^T & 0^T \\ {e_{n+s}}^T & 0^T \end{pmatrix} \begin{pmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta\lambda_1 \\ \Delta\lambda_2 \end{pmatrix}, \end{aligned}$$

wobei

$$\begin{aligned} R(P'(\lambda)x) &:= \begin{pmatrix} \mathrm{Re}\,(P'(\lambda)x) & -\mathrm{Im}\,(P'(\lambda)x) \\ \mathrm{Im}\,(P'(\lambda)x) & \mathrm{Re}\,(P'(\lambda)x) \end{pmatrix} \in \mathbb{R}^{2n\times 2}\,, \\ \mathrm{Re}\,(P'(\lambda)x) &= (2A_2\lambda_1 + A_1 - 2B_2\lambda_2)x_1 - (2A_2\lambda_2 + 2B_2\lambda_1 + B_1)x_2 \in \mathbb{R}^n\,, \\ \mathrm{Im}\,(P'(\lambda)x) &= (2A_2\lambda_2 + 2B_2\lambda_1 + B_1)x_1 + (2A_2\lambda_1 + A_1 - 2B_2\lambda_2)x_2 \in \mathbb{R}^n\,, \\ M(\tilde{x}, \Delta x, \Delta\lambda) &:= \begin{pmatrix} M_{1,\tilde{x},\Delta x,\Delta\lambda} & -M_{2,\tilde{x},\Delta x,\Delta\lambda} \\ M_{2,\tilde{x},\Delta x,\Delta\lambda} & M_{1,\tilde{x},\Delta x,\Delta\lambda} \end{pmatrix} \in \mathbb{R}^{2n\times 2}, \\ M_{1,\tilde{x},\Delta x,\Delta\lambda} &:= (A_2\tilde{x}_1)\Delta\lambda_1 + A_2\Delta x_1\Delta\lambda_1 - (A_2\tilde{x}_2)\Delta\lambda_2 - A_2\Delta x_2\Delta\lambda_2 \\ &\quad -(B_2\tilde{x}_2)\Delta\lambda_1 - B_2\Delta x_2\Delta\lambda_1 - (B_2\tilde{x}_1)\Delta\lambda_2 - B_2\Delta x_1\Delta\lambda_2\,, \\ M_{2,\tilde{x},\Delta x,\Delta\lambda} &:= (A_2\tilde{x}_2)\Delta\lambda_1 + A_2\Delta x_2\Delta\lambda_1 + (A_2\tilde{x}_1)\Delta\lambda_2 + A_2\Delta x_1\Delta\lambda_2 \\ &\quad +(B_2\tilde{x}_1)\Delta\lambda_1 + B_2\Delta x_1\Delta\lambda_1 - (B_2\tilde{x}_2)\Delta\lambda_2 - B_2\Delta x_2\Delta\lambda_2 \end{aligned}$$

mit $P'(\lambda) = 2K_2\lambda + K_1 = \big(2A_2\lambda_1 + A_1 - 2B_2\lambda_2\big) + i\cdot\big(2A_2\lambda_2 + 2B_2\lambda_1 + B_1\big)$, siehe (2.22).

(7.2) ist wieder die Taylor-Entwicklung von f an der Stelle $(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2)$. Deswegen hat die Jacobi–Matrix von f an der Stelle $(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2)$ die Gestalt

$$J(\tilde{x}, \tilde{\lambda}) := \begin{pmatrix} R(P(\tilde{\lambda})) & R(P'(\tilde{\lambda})\tilde{x}) \\ {e_s}^T & 0^T \\ {e_{n+s}}^T & 0^T \end{pmatrix}. \qquad (7.3)$$

Für die Funktion

$$\begin{aligned} g(\Delta x_1, \Delta x_2, \Delta\lambda_1, \Delta\lambda_2) := -Cf(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2) + \Bigg\{ I_{2n+2} - \qquad (7.4) \\ C \begin{pmatrix} R(P(\tilde{\lambda})) & R(P'(\tilde{\lambda})\tilde{x}) + R(P'(\tilde{\lambda})\Delta x) + M(\tilde{x}, \Delta x, \Delta\lambda) \\ {e_s}^T & 0^T \\ {e_{n+s}}^T & 0^T \end{pmatrix} \Bigg\} \begin{pmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta\lambda_1 \\ \Delta\lambda_2 \end{pmatrix} \end{aligned}$$

gilt dann Lemma 19. Nun wird die Funktion g wieder in eine für Intervall-Fixpunktiteration günstigere Gestalt gebracht.

$C^* := (c_{ij})_{\substack{i=1,\dots,2n+2\\ j=1,\dots,n}}$ bzw. $C^{**} := (c_{ij})_{\substack{i=1,\dots,2n+2\\ j=n+1,\dots,2n}}$ sei die Matrix mit den ersten n bzw. zweiten n Spalten von C.

Weiterhin seien die reellen $(2n+2) \times n$–Matrizen

$$\begin{aligned} D_1 &:= C^*A_1\,, \qquad D_2 := C^*A_2\,, \qquad E_1 := C^{**}A_1\,, \qquad E_2 := C^{**}A_2\,, \\ F_1 &:= C^*B_1\,, \qquad F_2 := C^*B_2\,, \qquad G_1 := C^{**}B_1\,, \qquad G_2 := C^{**}B_2 \end{aligned}$$

definiert.

(7.4) kann man unter Verwendung von (7.3) und (6.10) folgendermaßen umformen:

$$\begin{aligned} g(\Delta x_1, \Delta x_2, \Delta\lambda_1, \Delta\lambda_2) = -Cf(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2) + \Bigg\{ & I_{2n+2} - CJ(\tilde{x}, \tilde{\lambda}) \\ & -C \begin{pmatrix} O & R(P'(\tilde{\lambda})\Delta x) + M(\tilde{x}, \Delta x, \Delta\lambda) \\ 0^T & 0^T \\ 0^T & 0^T \end{pmatrix} \Bigg\} \begin{pmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta\lambda_1 \\ \Delta\lambda_2 \end{pmatrix} \\ = -Cf(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2) + \Big\{ & I_{2n+2} - CJ(\tilde{x}, \tilde{\lambda}) \\ & - \begin{pmatrix} O & N_1(\tilde{x}, \tilde{\lambda}, \Delta x, \Delta\lambda) & N_2(\tilde{x}, \tilde{\lambda}, \Delta x, \Delta\lambda) \end{pmatrix} \Big\} \begin{pmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta\lambda_1 \\ \Delta\lambda_2 \end{pmatrix} \end{aligned}$$

mit

$$\begin{aligned} N_1(\tilde{x}, \tilde{\lambda}, \Delta x, \Delta\lambda) &= C^*\big(\mathrm{Re}\,(P'(\tilde{\lambda})\Delta x) + M_{1,\tilde{x},\Delta x,\Delta\lambda}\big) + C^{**}\big(\mathrm{Im}\,(P'(\tilde{\lambda})\Delta x) + M_{2,\tilde{x},\Delta x,\Delta\lambda}\big) \\ &= (2D_2\tilde{\lambda}_1 + D_1 - 2F_2\tilde{\lambda}_2)\Delta x_1 - (2D_2\tilde{\lambda}_2 + 2F_2\tilde{\lambda}_1 + F_1)\Delta x_2 \\ &\quad +(D_2\tilde{x}_1)\Delta\lambda_1 + D_2\Delta x_1\Delta\lambda_1 - (D_2\tilde{x}_2)\Delta\lambda_2 - D_2\Delta x_2\Delta\lambda_2 \\ &\quad -(F_2\tilde{x}_2)\Delta\lambda_1 - F_2\Delta x_2\Delta\lambda_1 - (F_2\tilde{x}_1)\Delta\lambda_2 - F_2\Delta x_1\Delta\lambda_2 \\ &\quad +(2E_2\tilde{\lambda}_2 + 2G_2\tilde{\lambda}_1 + G_1)\Delta x_1 + (2E_2\tilde{\lambda}_1 + E_1 - 2G_2\tilde{\lambda}_2)\Delta x_2 \\ &\quad +(E_2\tilde{x}_2)\Delta\lambda_1 + E_2\Delta x_2\Delta\lambda_1 + (E_2\tilde{x}_1)\Delta\lambda_2 + E_2\Delta x_1\Delta\lambda_2 \\ &\quad +(G_2\tilde{x}_1)\Delta\lambda_1 + G_2\Delta x_1\Delta\lambda_1 - (G_2\tilde{x}_2)\Delta\lambda_2 - G_2\Delta x_2\Delta\lambda_2 \\ &= (D_2\tilde{x}_1 + E_2\tilde{x}_2 - F_2\tilde{x}_2 + G_2\tilde{x}_1)\Delta\lambda_1 - (D_2\tilde{x}_2 - E_2\tilde{x}_1 + F_2\tilde{x}_1 + G_2\tilde{x}_2)\Delta\lambda_2 \\ &\quad +(2D_2\tilde{\lambda}_1 + D_1 - 2F_2\tilde{\lambda}_2 + 2E_2\tilde{\lambda}_2 + 2G_2\tilde{\lambda}_1 + G_1)\Delta x_1 \\ &\quad -(2D_2\tilde{\lambda}_2 + 2F_2\tilde{\lambda}_1 + F_1 - 2E_2\tilde{\lambda}_1 - E_1 + 2G_2\tilde{\lambda}_2)\Delta x_2 \\ &\quad +\big((D_2 + G_2)\Delta x_1\big)\Delta\lambda_1 - \big((D_2 + G_2)\Delta x_2\big)\Delta\lambda_2 \\ &\quad +\big((E_2 - F_2)\Delta x_2\big)\Delta\lambda_1 + \big((E_2 - F_2)\Delta x_1\big)\Delta\lambda_2\,, \end{aligned}$$

$$
\begin{aligned}
N_2(\tilde{x},\tilde{\lambda},\Delta x,\Delta\lambda) &= -C^*\big(\mathrm{Im}\,(P'(\tilde{\lambda})\Delta x) + M_{2,\tilde{x},\Delta x,\Delta\lambda}\big) + C^{**}\big(\mathrm{Re}\,(P'(\tilde{\lambda})\Delta x) + M_{1,\tilde{x},\Delta x,\Delta\lambda}\big)\\
&= -(2D_2\tilde{\lambda}_2 + 2F_2\tilde{\lambda}_1 + F_1)\Delta x_1 - (2D_2\tilde{\lambda}_1 + D_1 - 2F_2\tilde{\lambda}_2)\Delta x_2\\
&\quad -(D_2\tilde{x}_2)\Delta\lambda_1 - D_2\Delta x_2\Delta\lambda_1 - (D_2\tilde{x}_1)\Delta\lambda_2 - D_2\Delta x_1\Delta\lambda_2\\
&\quad -(F_2\tilde{x}_1)\Delta\lambda_1 - F_2\Delta x_1\Delta\lambda_1 + (F_2\tilde{x}_2)\Delta\lambda_2 + F_2\Delta x_2\Delta\lambda_2\\
&\quad +(2E_2\tilde{\lambda}_1 + E_1 - 2G_2\tilde{\lambda}_2)\Delta x_1 - (2E_2\tilde{\lambda}_2 + 2G_2\tilde{\lambda}_1 + G_1)\Delta x_2\\
&\quad +(E_2\tilde{x}_1)\Delta\lambda_1 + E_2\Delta x_1\Delta\lambda_1 - (E_2\tilde{x}_2)\Delta\lambda_2 - E_2\Delta x_2\Delta\lambda_2\\
&\quad -(G_2\tilde{x}_2)\Delta\lambda_1 - G_2\Delta x_2\Delta\lambda_1 - (G_2\tilde{x}_1)\Delta\lambda_2 - G_2\Delta x_1\Delta\lambda_2\\
&= -(D_2\tilde{x}_2 - E_2\tilde{x}_1 + F_2\tilde{x}_1 + G_2\tilde{x}_2)\Delta\lambda_1 - (D_2\tilde{x}_1 + E_2\tilde{x}_2 - F_2\tilde{x}_2 + G_2\tilde{x}_1)\Delta\lambda_2\\
&\quad -(2D_2\tilde{\lambda}_2 + 2F_2\tilde{\lambda}_1 + F_1 - 2E_2\tilde{\lambda}_1 - E_1 + 2G_2\tilde{\lambda}_2)\Delta x_1\\
&\quad -(2D_2\tilde{\lambda}_1 + D_1 - 2F_2\tilde{\lambda}_2 + 2E_2\tilde{\lambda}_2 + 2G_2\tilde{\lambda}_1 + G_1)\Delta x_2\\
&\quad +\big((E_2 - F_2)\Delta x_1\big)\Delta\lambda_1 - \big((E_2 - F_2)\Delta x_2\big)\Delta\lambda_2\\
&\quad -\big((D_2 + G_2)\Delta x_2\big)\Delta\lambda_1 - \big((D_2 + G_2)\Delta x_1\big)\Delta\lambda_2\,.
\end{aligned}
$$

Satz 25

Es sei $P(\lambda)$ *wie in* (7.1), $\tilde{\lambda} = \tilde{\lambda}_1 + i\tilde{\lambda}_2 \in \mathbb{C}$, $\tilde{x} = \tilde{x}_1 + i\tilde{x}_2 \in \mathbb{C}^n$ *und* $C \in \mathbb{R}^{(2n+2)\times(2n+2)}$.
Für die wie folgt definierten Ausdrücke

$$\varphi := ||Cf(\tilde{x}_1,\tilde{x}_2,\tilde{\lambda}_1,\tilde{\lambda}_2)||_\infty\,, \quad \sigma := ||I_{2n+2} - CJ(\tilde{x},\tilde{\lambda})||_\infty\,,$$

$$\tau := 2||C||_\infty\{(||A_2||_\infty + ||B_2||_\infty)(2|\tilde{\lambda}_1| + 2|\tilde{\lambda}_2| + ||\tilde{x}_1||_\infty + ||\tilde{x}_2||_\infty) + ||A_1||_\infty + ||B_1||_\infty\}\,,$$

$$\gamma := 4||C||_\infty(||A_2||_\infty + ||B_2||_\infty)\,,$$

$$p := \frac{\sigma-1}{\gamma} - \frac{\tau^2}{3\,\gamma^2} \;\; \textit{und} \;\; q := \frac{2\,\tau^3}{27\,\gamma^3} - \frac{(\sigma-1)\tau}{3\,\gamma^2} + \frac{\varphi}{\gamma}$$

seien die Bedingungen

$$\varphi > 0\,,\ \sigma < 1 \tag{7.5}$$

und

$$D := \frac{q^2}{4} + \frac{p^3}{27} < 0 \tag{7.6}$$

erfüllt.

Mit $\alpha := \arccos\left(-\dfrac{\sqrt{27}q}{2\sqrt{-p^3}}\right)$ *gilt dann für*

$$\beta^- := 2\sqrt{\frac{-p}{3}}\cos\left(\frac{\alpha+4\pi}{3}\right) - \frac{\tau}{3\,\gamma} \qquad \textit{und} \qquad \beta^+ := 2\sqrt{\frac{-p}{3}}\cos\left(\frac{\alpha}{3}\right) - \frac{\tau}{3\,\gamma}$$

die Eigenschaft

$$0 < \beta^- < \beta^+\,.$$

Weiter sei

$$g(\Delta x_1, \Delta x_2, \Delta\lambda_1, \Delta\lambda_2) = -Cf(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2) + \Big\{ I_{2n+2} - CJ(\tilde{x}, \tilde{\lambda}) \\ - \begin{pmatrix} O & N_1(\tilde{x}, \tilde{\lambda}, \Delta x, \Delta\lambda) & N_2(\tilde{x}, \tilde{\lambda}, \Delta x, \Delta\lambda) \end{pmatrix} \Big\} \begin{pmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta\lambda_1 \\ \Delta\lambda_2 \end{pmatrix} \qquad (7.7)$$

mit

$$\begin{aligned} N_1(\tilde{x}, \tilde{\lambda}, \Delta x, \Delta\lambda) &= (D_2\tilde{x}_1 + E_2\tilde{x}_2 - F_2\tilde{x}_2 + G_2\tilde{x}_1)\Delta\lambda_1 - (D_2\tilde{x}_2 - E_2\tilde{x}_1 + F_2\tilde{x}_1 + G_2\tilde{x}_2)\Delta\lambda_2 \\ &\quad +(2D_2\tilde{\lambda}_1 + D_1 - 2F_2\tilde{\lambda}_2 + 2E_2\tilde{\lambda}_2 + 2G_2\tilde{\lambda}_1 + G_1)\Delta x_1 \\ &\quad -(2D_2\tilde{\lambda}_2 + 2F_2\tilde{\lambda}_1 + F_1 - 2E_2\tilde{\lambda}_1 - E_1 + 2G_2\tilde{\lambda}_2)\Delta x_2 \\ &\quad +\big((D_2 + G_2)\Delta x_1\big)\Delta\lambda_1 - \big((D_2 + G_2)\Delta x_2\big)\Delta\lambda_2 \\ &\quad +\big((E_2 - F_2)\Delta x_2\big)\Delta\lambda_1 + \big((E_2 - F_2)\Delta x_1\big)\Delta\lambda_2\,, \\ N_2(\tilde{x}, \tilde{\lambda}, \Delta x, \Delta\lambda) &= -(D_2\tilde{x}_2 - E_2\tilde{x}_1 + F_2\tilde{x}_1 + G_2\tilde{x}_2)\Delta\lambda_1 - (D_2\tilde{x}_1 + E_2\tilde{x}_2 - F_2\tilde{x}_2 + G_2\tilde{x}_1)\Delta\lambda_2 \\ &\quad -(2D_2\tilde{\lambda}_2 + 2F_2\tilde{\lambda}_1 + F_1 - 2E_2\tilde{\lambda}_1 - E_1 + 2G_2\tilde{\lambda}_2)\Delta x_1 \\ &\quad -(2D_2\tilde{\lambda}_1 + D_1 - 2F_2\tilde{\lambda}_2 + 2E_2\tilde{\lambda}_2 + 2G_2\tilde{\lambda}_1 + G_1)\Delta x_2 \\ &\quad +\big((E_2 - F_2)\Delta x_1\big)\Delta\lambda_1 - \big((E_2 - F_2)\Delta x_2\big)\Delta\lambda_2 \\ &\quad -\big((D_2 + G_2)\Delta x_2\big)\Delta\lambda_1 - \big((D_2 + G_2)\Delta x_1\big)\Delta\lambda_2 \end{aligned}$$

und

$$\begin{aligned} D_1 &:= C^*A_1\,, \qquad D_2 := C^*A_2\,, \qquad E_1 := C^{**}A_1\,, \qquad E_2 := C^{**}A_2\,, \\ F_1 &:= C^*B_1\,, \qquad F_2 := C^*B_2\,, \qquad G_1 := C^{**}B_1\,, \qquad G_2 := C^{**}B_2\,. \end{aligned}$$

Außerdem sei für beliebiges $\beta \in (\beta^-, \beta^+)$

$$\begin{pmatrix} [\Delta x_1]^{(0)} \\ [\Delta x_2]^{(0)} \\ [\Delta\lambda_1]^{(0)} \\ [\Delta\lambda_2]^{(0)} \end{pmatrix} := [-\beta, \beta]e \in \mathbf{IR}^{2n+2}.$$

Dann konvergiert die Iteriertenfolge

$$\begin{pmatrix} [\Delta x_1]^{(k+1)} \\ [\Delta x_2]^{(k+1)} \\ [\Delta\lambda_1]^{(k+1)} \\ [\Delta\lambda_2]^{(k+1)} \end{pmatrix} := g([\Delta x_1]^{(k)}, [\Delta x_2]^{(k)}, [\Delta\lambda_1]^{(k)}, [\Delta\lambda_2]^{(k)})\,, \quad k \in \mathbb{N}_0,$$

gegen einen Intervallvektor

$$\begin{pmatrix} [\Delta x_1]^* \\ [\Delta x_2]^* \\ [\Delta\lambda_1]^* \\ [\Delta\lambda_2]^* \end{pmatrix} \subseteq \begin{pmatrix} [\Delta x_1]^{(k+1)} \\ [\Delta x_2]^{(k+1)} \\ [\Delta\lambda_1]^{(k+1)} \\ [\Delta\lambda_2]^{(k+1)} \end{pmatrix} \subseteq \begin{pmatrix} [\Delta x_1]^{(k)} \\ [\Delta x_2]^{(k)} \\ [\Delta\lambda_1]^{(k)} \\ [\Delta\lambda_2]^{(k)} \end{pmatrix} \subseteq \ldots \subseteq \begin{pmatrix} [\Delta x_1]^{(0)} \\ [\Delta x_2]^{(0)} \\ [\Delta\lambda_1]^{(0)} \\ [\Delta\lambda_2]^{(0)} \end{pmatrix}, \quad k \in \mathbb{N},$$

und es existiert mindestens ein komplexes Eigenpaar $(x_1^* + ix_2^*, \lambda_1^* + i\lambda_2^*)$ *von* $P(\lambda)$ *mit* $(x_1^* + ix_2^*)_s = 1$ *und*

$$\begin{pmatrix} x_1^* \\ x_2^* \\ \lambda_1^* \\ \lambda_2^* \end{pmatrix} \in \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{\lambda}_1 \\ \tilde{\lambda}_2 \end{pmatrix} + \begin{pmatrix} [\Delta x_1]^{(k)} \\ [\Delta x_2]^{(k)} \\ [\Delta \lambda_1]^{(k)} \\ [\Delta \lambda_2]^{(k)} \end{pmatrix}, \quad k \in \mathbb{N}_0.$$

Unter der Einschränkung $\beta \in (\beta^-, \sqrt{\frac{-p}{3}} - \frac{\tau}{3\gamma})$ *ist dieses Eigenpaar eindeutig, und die Iterierten* $\begin{pmatrix} [\Delta x_1]^{(k)} \\ [\Delta x_2]^{(k)} \\ [\Delta \lambda_1]^{(k)} \\ [\Delta \lambda_2]^{(k)} \end{pmatrix}$ *konvergieren gegen dessen Approximationsfehler* $\begin{pmatrix} \Delta x_1^* \\ \Delta x_2^* \\ \Delta \lambda_1^* \\ \Delta \lambda_2^* \end{pmatrix}$.

Beweis:
Es sei $([\Delta x_1]^T, [\Delta x_2]^T, [\Delta \lambda_1], [\Delta \lambda_2])^T \in \mathbf{IR}^{2n+2}$, $[\Delta x] := [\Delta x_1] + i[\Delta x_2]$
und $[\Delta \lambda] := [\Delta \lambda_1] + i[\Delta \lambda_2]$. Dann gilt analog zu (6.15)

$$\begin{aligned} g([\Delta x_1], [\Delta x_2], [\Delta \lambda_1], [\Delta \lambda_2]) \quad &\subseteq \quad r + S \begin{pmatrix} [\Delta x_1] \\ [\Delta x_2] \\ [\Delta \lambda_1] \\ [\Delta \lambda_2] \end{pmatrix} + T \begin{pmatrix} [\Delta x_1] \\ [\Delta x_2] \\ [\Delta \lambda_1] \\ [\Delta \lambda_2] \end{pmatrix}^2 + U \begin{pmatrix} [\Delta x_1] \\ [\Delta x_2] \\ [\Delta \lambda_1] \\ [\Delta \lambda_2] \end{pmatrix}^3 \\ &=: \quad h([\Delta x_1], [\Delta x_2], [\Delta \lambda_1], [\Delta \lambda_2]), \end{aligned}$$

wobei

$$h(x) = r + Sx + Tx^2 + Ux^3, \quad x \in \mathbb{R}^{2n+2},$$

ein kubisches System (4.1) mit $m = 2n+2$, $r = -Cf(\tilde{x}_1, \tilde{x}_2, \tilde{\lambda}_1, \tilde{\lambda}_2)$, $S := I_{2n+2} - CJ(\tilde{x}, \tilde{\lambda})$, $T = (t_{ijk}) \in \mathbb{R}^{(2n+2)\times(2n+2)\times(2n+2)}$ und $U = (u_{ijkl}) \in \mathbb{R}^{(2n+2)\times(2n+2)\times(2n+2)\times(2n+2)}$ ist.
Für $i = 1, \ldots, 2n+2$ ist dabei

$$t_{ijk} = \begin{cases} -(D_2\tilde{x}_1 + E_2\tilde{x}_2 - F_2\tilde{x}_2 + G_2\tilde{x}_1)_i, & j = k = 2n+1, \\ (D_2\tilde{x}_2 - E_2\tilde{x}_1 + F_2\tilde{x}_1 + G_2\tilde{x}_2)_i, & j = 2n+1;\ k = 2n+2, \\ (D_2\tilde{x}_2 - E_2\tilde{x}_1 + F_2\tilde{x}_1 + G_2\tilde{x}_2)_i, & j = 2n+2;\ k = 2n+1, \\ (D_2\tilde{x}_1 + E_2\tilde{x}_2 - F_2\tilde{x}_2 + G_2\tilde{x}_1)_i, & j = k = 2n+2, \\ -(2D_2\tilde{\lambda}_1 + D_1 - 2F_2\tilde{\lambda}_2 + 2E_2\tilde{\lambda}_2 + 2G_2\tilde{\lambda}_1 + G_1)_{ik}, & j = 2n+1;\ k = 1, \ldots, n, \\ (2D_2\tilde{\lambda}_2 + 2F_2\tilde{\lambda}_1 + F_1 - 2E_2\tilde{\lambda}_1 - E_1 + 2G_2\tilde{\lambda}_2)_{i,k-n}, & j = 2n+1;\ k = n+1, \ldots, 2n, \\ (2D_2\tilde{\lambda}_2 + 2F_2\tilde{\lambda}_1 + F_1 - 2E_2\tilde{\lambda}_1 - E_1 + 2G_2\tilde{\lambda}_2)_{ik}, & j = 2n+2;\ k = 1, \ldots, n, \\ (2D_2\tilde{\lambda}_1 + D_1 - 2F_2\tilde{\lambda}_2 + 2E_2\tilde{\lambda}_2 + 2G_2\tilde{\lambda}_1 + G_1)_{i,k-n}, & j = 2n+2;\ k = n+1, \ldots, 2n, \\ 0, & \text{sonst}, \end{cases}$$

$$u_{ijkl} = \begin{cases} -(D_2+G_2)_{il}\,, & j=k=2n+1;\ l=1,\dots,n,\\ (D_2+G_2)_{i,l-n}\,, & j=2n+1;\ k=2n+2;\ l=n+1,\dots,2n,\\ -(E_2-F_2)_{i,l-n}\,, & j=k=2n+1;\ l=n+1,\dots,2n,\\ -(E_2-F_2)_{il}\,, & j=2n+1;\ k=2n+2;\ l=1,\dots,n,\\ (E_2-F_2)_{i,l-n}\,, & j=k=2n+2;\ l=n+1,\dots,2n,\\ -(E_2-F_2)_{il}\,, & j=2n+2;\ k=2n+1;\ l=1,\dots,n,\\ (D_2+G_2)_{il}\,, & j=k=2n+2;\ l=1,\dots,n,\\ (D_2+G_2)_{i,l-n}\,, & j=2n+2;\ k=2n+1;\ l=n+1,\dots,2n,\\ 0\,, & \text{sonst.} \end{cases}$$

Es gilt $||r||_\infty = ||Cf(\tilde{x}_1,\tilde{x}_2,\tilde{\lambda}_1,\tilde{\lambda}_2)||_\infty = \varphi$ und $||S||_\infty = ||I_{2n+2} - CJ(\tilde{x},\tilde{\lambda})||_\infty = \sigma$.
Weiter gilt

$$\begin{aligned}
u_\infty &= \max_{1\le i\le 2n+2}\left(\sum_{j=1}^{2n+2}\sum_{k=1}^{2n+2}\sum_{l=1}^{2n+2}|u_{ijkl}|\right)\\
&= \max_{1\le i\le 2n+2}\left(\sum_{l=1}^{n}(4|(D_2+G_2)_{il}| + 4|(E_2-F_2)_{il}|)\right)\\
&\overset{\substack{\max \text{ für}\\ i\equiv i^*}}{=} \sum_{l=1}^{n}(4|(D_2+G_2)_{i^*l}| + 4|(E_2-F_2)_{i^*l}|)\\
&\le 4\sum_{l=1}^{n}(|(D_2)_{i^*l}| + |(G_2)_{i^*l}| + |(E_2)_{i^*l}| + |(F_2)_{i^*l}|)\\
&= 4\sum_{l=1}^{n}(|(D_2)_{i^*l}| + |(E_2)_{i^*l}|) + 4\sum_{l=1}^{n}(|(F_2)_{i^*l}| + |(G_2)_{i^*l}|)\\
&= 4\sum_{l=1}^{n}\left(\Big|\sum_{t=1}^{n}c_{i^*t}(A_2)_{tl}\Big| + \Big|\sum_{t=1}^{n}c_{i^*,n+t}(A_2)_{tl}\Big|\right)\\
&\quad + 4\sum_{l=1}^{n}\left(\Big|\sum_{t=1}^{n}c_{i^*t}(B_2)_{tl}\Big| + \Big|\sum_{t=1}^{n}c_{i^*,n+t}(B_2)_{tl}\Big|\right)\\
&\overset{(6.16)}{\le} 4||A_2||_\infty||C||_\infty + 4||B_2||_\infty||C||_\infty\\
&= 4||C||_\infty(||A_2||_\infty + ||B_2||_\infty) = \gamma\,.
\end{aligned}$$

Für beliebiges $i \in \{1, \ldots, 2n+2\}$ ist

$$\begin{aligned}
&|t_{i,2n+1,2n+1}| + |t_{i,2n+2,2n+2}| + |t_{i,2n+1,2n+2}| + |t_{i,2n+2,2n+1}| \\
&\quad = \quad 2|(D_2\tilde{x}_1 + E_2\tilde{x}_2 - F_2\tilde{x}_2 + G_2\tilde{x}_1)_i| + 2|(D_2\tilde{x}_2 - E_2\tilde{x}_1 + F_2\tilde{x}_1 + G_2\tilde{x}_2)_i| \\
&\quad \leq \quad 2\big(|(D_2\tilde{x}_1)_i| + |(E_2\tilde{x}_1)_i| + |(D_2\tilde{x}_2)_i| + |(E_2\tilde{x}_2)_i| \\
&\qquad + |(F_2\tilde{x}_1)_i| + |(G_2\tilde{x}_1)_i| + |(F_2\tilde{x}_2)_i| + |(G_2\tilde{x}_2)_i|\big) \\
&\quad \overset{(6.17)}{\leq} \quad 2||C||_\infty||A_2||_\infty(||\tilde{x}_1||_\infty + ||\tilde{x}_2||_\infty) + 2||C||_\infty||B_2||_\infty(||\tilde{x}_1||_\infty + ||\tilde{x}_2||_\infty) \\
&\quad = \quad 2||C||_\infty(||A_2||_\infty + ||B_2||_\infty)(||\tilde{x}_1||_\infty + ||\tilde{x}_2||_\infty)\,. \qquad (7.8)
\end{aligned}$$

Desweiteren gilt

$$\begin{aligned}
&\sum_{k=1}^{n} |t_{i,2n+1,k}| + \sum_{k=n+1}^{2n} |t_{i,2n+2,k}| + \sum_{k=1}^{n} |t_{i,2n+2,k}| + \sum_{k=n+1}^{2n} |t_{i,2n+1,k}| \\
&\quad = \quad 2\sum_{k=1}^{n} |(2D_2\tilde{\lambda}_1 + D_1 - 2F_2\tilde{\lambda}_2 + 2E_2\tilde{\lambda}_2 + 2G_2\tilde{\lambda}_1 + G_1)_{ik}| \\
&\qquad + 2\sum_{k=1}^{n} |(2D_2\tilde{\lambda}_2 + 2F_2\tilde{\lambda}_1 + F_1 - 2E_2\tilde{\lambda}_1 - E_1 + 2G_2\tilde{\lambda}_2)_{ik}| \\
&\quad \leq \quad 2\sum_{k=1}^{n} \Big(2|\tilde{\lambda}_1||(D_2)_{ik}| + 2|\tilde{\lambda}_1||(E_2)_{ik}| + 2|\tilde{\lambda}_2||(D_2)_{ik}| + 2|\tilde{\lambda}_2||(E_2)_{ik}| + |(D_1)_{ik}| + |(E_1)_{ik}| \\
&\qquad + 2|\tilde{\lambda}_1||(F_2)_{ik}| + 2|\tilde{\lambda}_1||(G_2)_{ik}| + 2|\tilde{\lambda}_2||(F_2)_{ik}| + 2|\tilde{\lambda}_2||(G_2)_{ik}| + |(F_1)_{ik}| + |(G_1)_{ik}|\Big) \\
&\quad \overset{(6.18)}{\leq} \quad 2||C||_\infty\{||A_2||_\infty(2|\tilde{\lambda}_1| + 2|\tilde{\lambda}_2|) + ||A_1||_\infty\} + 2||C||_\infty\{||B_2||_\infty(2|\tilde{\lambda}_1| + 2|\tilde{\lambda}_2|) + ||B_1||_\infty\} \\
&\quad = \quad 2||C||_\infty\{(||A_2||_\infty + ||B_2||_\infty)(2|\tilde{\lambda}_1| + 2|\tilde{\lambda}_2|) + ||A_1||_\infty + ||B_1||_\infty\}\,. \qquad (7.9)
\end{aligned}$$

Aus (7.8) und (7.9) folgt somit

$$\begin{aligned}
t_\infty \quad &= \quad \max_{1\leq i\leq 2n+2}\Big(\sum_{j=1}^{2n+2}\sum_{k=1}^{2n+2} |t_{ijk}|\Big) \\
&\leq \quad 2||C||_\infty(||A_2||_\infty + ||B_2||_\infty)(||\tilde{x}_1||_\infty + ||\tilde{x}_2||_\infty) \\
&\qquad + 2||C||_\infty\{(||A_2||_\infty + ||B_2||_\infty)(2|\tilde{\lambda}_1| + 2|\tilde{\lambda}_2|) + ||A_1||_\infty + ||B_1||_\infty\} \\
&= \quad 2||C||_\infty\{(||A_2||_\infty + ||B_2||_\infty)(2|\tilde{\lambda}_1| + 2|\tilde{\lambda}_2| + ||\tilde{x}_1||_\infty + ||\tilde{x}_2||_\infty) + ||A_1||_\infty + ||B_1||_\infty\} \\
&= \quad \tau\,.
\end{aligned}$$

C ist wegen $\sigma = ||I_{2n+2} - CJ(\tilde{x}, \tilde{\lambda})||_\infty < 1$ nach Lemma 13 regulär, also ungleich der Nullmatrix. Weil nach Voraussetzung $K_2 = A_2 + iB_2$ ebenfalls regulär ist, sind A_2 und B_2 nicht beide die Nullmatrix, d.h. $\gamma \neq 0$. Außerdem ist der Funktionsausdruck (7.7) von $g : \mathbb{R}^{2n+2} \to \mathbb{R}^{2n+2}$ intervallmäßig auswertbar. Mit Satz 20 und Lemma 19 folgt dann der Satz. □

Satz 24 und Bemerkung 12 gelten genauso im hier betrachteten Fall des komplexen QEP. Wenn man Satz 23 und Satz 25 vergleicht, sieht man, dass es sich bei Satz 23 um einen Spezialfall

von Satz 25 handelt (Spezialfall $B_0 = B_1 = B_2 = O \in \mathbb{R}^{n\times n}$). In diesem Kapitel 7 wurde also eine Verallgemeinerung von Kapitel 6 vorgenommen.

Kapitel 8

Numerische Ergebnisse

8.1 Numerische Verfahren zur Berechnung von Eigenpaarnäherungen beim QEP

Einen sehr guten Überblick über numerische Verfahren zur Berechnung von Eigenpaarnäherungen (x^*, λ^*) des regulären QEPs $P(\lambda)x = (A_2\lambda^2 + A_1\lambda + A_0)x = 0$ der Dimension n findet man in dem Übersichtsartikel „The Quadratic Eigenvalue Problem“ von Tisseur und Meerbergen [27]. Es werden zwei Arten von Verfahren unterschieden: solche, die eine Linearisierung $\tilde{A} \in \mathbb{C}^{2n\times 2n}$ (2.1) bzw. $A - \lambda B$ (2.45) (mit $A, B \in \mathbb{C}^{2n\times 2n}$) des QEPs verwenden, und solche, die das QEP direkt angehen.

Außerdem kann man das QEP als einen Spezialfall des nichtlinearen Eigenwertproblems

$$F(\lambda)x = F(\lambda; M_0, \ldots, M_l, v)x = 0$$

betrachten, wobei $F : \mathbb{C} \to \mathbb{C}^{m\times n}$ eine Funktion ist, die von Matrizenkoeffizienten $M_0, \ldots, M_l \in \mathbb{C}^{m\times n}$ und einem Parameter–Vektor $v \in \mathbb{C}^r$ abhängen kann. Numerische Verfahren zur Lösung des nichtlinearen Eigenwertproblems werden in [20] beschrieben.

8.1.1 Verfahren, die Linearisierungen des QEPs verwenden

Wenn $\tilde{A}$ eine Linearisierung (2.1) von $P(\lambda)$ ist, kann man Eigenpaare des speziellen Eigenwertproblems (SEP) $\tilde{A}y = \lambda y$ der Dimension $2n$ näherungsweise bestimmen und erhält daraus mit Hilfe von Satz 15 bzw. Bemerkung 4 Näherungen der Eigenpaare von $P(\lambda)$. Numerische Verfahren zur Eigenpaarbestimmung beim SEP sind z.B. der QR–Algorithmus (wenn $2n$ relativ klein ist), der Lanczos–Algorithmus (wenn $\tilde{A}$ hermitesch und $2n$ groß ist) oder der Arnoldi–Algorithmus (wenn $\tilde{A}$ nicht hermitesch und $2n$ groß ist), siehe [6]. Wenn man aber z.B. die Begleitmatrix

$$C_P = \begin{pmatrix} O & I_n \\ -A_2^{-1}A_0 & -A_2^{-1}A_1 \end{pmatrix} \in \mathbb{C}^{2n\times 2n}$$

als Linearisierung verwendet, muss A_2^{-1} explizit berechnet werden, was schlecht für die Konditionierung des Problems ist.

Deswegen verwendet man besser eine Linearisierung der Form $A - \lambda B$, z.B. mit

$$A = \begin{pmatrix} O & cI_n \\ -A_0 & -A_1 \end{pmatrix}, \quad B = \begin{pmatrix} cI_n & O \\ O & A_2 \end{pmatrix} \tag{8.1}$$

für $c \neq 0$. Aus den Eigenpaaren des verallgemeinerten Eigenwertproblems (GEP) $Ay = \lambda By$ der Dimension $2n$ erhält man dann mit Hilfe von Satz 16 Eigenpaare von $P(\lambda)$. Für die Eigenpaarapproximation beim GEP gibt es eine Reihe bekannter numerischer Verfahren. Das QZ–Verfahren (für $2n$ relativ klein) wird im Anschluss genauer beschrieben. Die Eigenpaarnäherungen in den später folgenden numerischen Experimenten wurden mit der MATLAB-Prozedur `polyeig(`A_0, A_1, A_2`)` berechnet. Diese wiederum verwendet Linearisierungen der Form $A - \lambda B$ von $P(\lambda)$ und löst das so erhaltene GEP näherungsweise mit dem QZ–Algorithmus.

$A - \lambda B$ mit $A, B \in \mathbb{C}^{2n \times 2n}$ sei nun stets eine Linearisierung von $P(\lambda)$. Für das Matrizentupel (A, B) existieren unitäre Matrizen Q und Z (d.h. $Q^H Q = I_{2n} = Z^H Z$), so dass

$$Q^H A Z = S \qquad \text{und} \qquad Q^H B Z = T \tag{8.2}$$

obere Dreiecksmatrizen sind. (8.2) heißt verallgemeinerte Schur–Zerlegung von (A, B). Der QZ–Algorithmus nun bringt A und B simultan mit Hilfe von Ähnlichkeitstransformationen (Householder-Spiegelungen und Givens-Rotationen) auf obere Dreiecksgestalt (siehe [15]). Seien nun $S = (s_{ij})$ und $T = (t_{ij})$ die so erhaltenen oberen Dreiecksmatrizen. Wegen

$$A - \lambda B = Q(S - \lambda T)Z^H$$

sind dann $\lambda_i^* = s_{ii}/t_{ii}$, $i = 1, \ldots, 2n$, die Eigenwerte des GEPs $Ay = \lambda By$ und damit auch die Eigenwerte von $P(\lambda)$.

Wegen $S - \lambda T = Q^H(A - \lambda B)Z$ und Satz 16 gilt dann für eine Linearisierung vom Typ (8.1): wenn z^* Eigenvektor zum Eigenwert λ^* bzgl. des GEPs (S, T) ist, dann ist $y^* = Zz^*$ Eigenvektor zu λ^* bzgl. des GEPs (A, B), also folglich $(y_1^*, \ldots, y_n^*)^T$ und $(y_{n+1}^*, \ldots, y_{2n}^*)^T$ (linear abhängige) Eigenvektoren zu λ^* bzgl. des QEPs. Die Lösung z^* des homogenen linearen Gleichungssystems oberer Dreiecksgestalt $(S - \lambda^* T)z^* = 0$ ist durch Rückwärtseinsetzen leicht zu berechnen.

Für $A, B \in \mathbb{R}^{2n \times 2n}$ existiert eine reelle verallgemeinerte Schur-Zerlegung der Gestalt (8.2) mit $Q, Z \in \mathbb{R}^{2n \times 2n}$ orthogonal und S von oberer Quasi-Dreiecksgestalt (d.h. es können bei S 2×2–Diagonalblöcke auftreten, welche den konjugiert komplexen Eigenwert-Paaren von (A, B) entsprechen).

Der QZ–Algorithmus ist numerisch stabil bzgl. der Lösung des GEPs, jedoch nicht numerisch stabil bzgl. der Lösung des QEPs [27]. Wenn A_2, A_1 und A_0 keine spezielle Struktur besitzen und n nicht zu groß ist, wird der QZ-Algorithmus in der Praxis dennoch gerne zur Berechnung von Eigenpaaren des QEPs verwendet. Wenn z.B. A_2, A_1 und A_0 symmetrisch sind, kann man eine symmetrische Linearisierung $A - \lambda B$ (d.h. A und B symmetrisch) des QEPs verwenden und Verfahren zur Bestimmung von Eigenpaarnäherungen des zugehörigen GEPs $Ay = \lambda By$ nutzen, welche die Symmetrie von A und B berücksichtigen.

Bei großer Dimension n des QEPs sind der Speicheraufwand und die Rechenzeit bei der Verwendung des QZ–Algorithmus sehr groß. In den Anwendungen solcher großen QEPs werden zudem oft nur Näherungen von einigen Eigenpaaren benötigt. Es werden dann projektive Verfahren zur Eigenpaarnäherung beim GEP $Ay = \lambda By$ verwendet, zumeist Krylow–Unterraum–Verfahren.

$$K_m(S, v) = \text{span}\{v, Sv, S^2 v, \ldots, S^{m-1} v\}$$

ist der m–te Krylow–Unterraum ($m \leq 2n$) für einen festen Vektor $v \in \mathbb{C}^{2n}$ und eine feste Matrix $S \in \mathbb{C}^{2n \times 2n}$ (z.B. $S = B^{-1}A$), für die man einen Eigenwert μ^* am äußeren Rand ihres Spektrums approximieren will (x^* sei ein zugehöriger Eigenvektor). Für $S = B^{-1}A$ ist dann (x^*, μ^*) Eigenpaar des QEPs. Eigenwerte λ^* des QEPs in der Nähe von σ mit zugehörigem Eigenvektor x^* kann man approximieren, wenn man z.B. $S = (A - \sigma B)^{-1}B$ (mit Eigenpaaren $(x^*, \mu^* = \frac{1}{\lambda^* - \sigma})$) verwendet. Mit dem Lanczos–Verfahren für S hermitesch bzw. dem Arnoldi–Verfahren für S nicht hermitesch konstruiert man sukzessive mit Hilfe von Gram-Schmidt-Orthogonalisierung eine Matrix $V_m \in \mathbb{C}^{2n \times m}$, deren Spalten eine Orthonormalbasis von $K_m(S, v)$ bilden und für die $H_m = V_m^H S V_m \in \mathbb{C}^{m \times m}$ eine hermitesche Tridiagonalmatrix bzw. obere Hessenbergmatrix ist. Für relativ kleines m ist dann H_m eine Matrix von relativ kleiner Dimension, deren Eigenpaare (u^*, τ^*) z.B. mit dem QR–Algorithmus approximiert werden können. $(V_m u^*, \tau^*)$ sind dann Eigenpaare von S.

Für ein reelles bzw. symmetrisches QEP kann man diese Algorithmen adaptieren. Außerdem gibt es noch weitere Krylow–Unterraum–Verfahren, z.B. das nicht–hermitesche Lanczos–Verfahren [27].

8.1.2 Verfahren, die das QEP direkt angehen

Eine Möglichkeit, reelle Eigenpaare des reellen QEPs zu approximieren, besteht darin, Nullstellen der nichtlinearen Funktion $f : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ aus (5.2)

$$f(x, \lambda) = \begin{pmatrix} (A_2\lambda^2 + A_1\lambda + A_0)x \\ {e_s}^T x - 1 \end{pmatrix}, \qquad x \in \mathbb{R}^n,\ \lambda \in \mathbb{R},$$

mit dem mehrdimensionalen Newton–Verfahren zu berechnen. Es sei daran erinnert, dass nach Satz 14 die Jacobi-Matrix von f in einem Eigenpaar (x^*, λ^*) von $P(\lambda)$ mit λ^* einfach und $(x^*)_s = 1$ invertierbar ist. Für komplexe Eigenpaare des reellen QEPs kann man das mehrdimensionale Newton–Verfahren auf die Funktion f aus (6.5) anwenden. Weitere Newton–artige Verfahren für das QEP als Spezialfall des nichtlinearen Eigenwertproblems werden in [20] beschrieben.

Eine weitere Klasse von Verfahren sind projektive Verfahren, die direkt auf das QEP angewandt werden. Es wird wieder sukzessive eine Orthononormalbasis $V_m = (v_1, \ldots, v_m) \in \mathbb{C}^{n \times m}$ eines Unterraumes K_m der Dimension $m \leq n$ berechnet, dessen Eigenraum Teilmenge des Eigenraumes von $P(\lambda)$ ist. Die Wahl von $v_m \in \mathbb{C}^n$, mit dem man V_{m-1} zu V_m ergänzt, führt auf die verschiedenen Verfahren, z.B. das Residuum–Verfahren oder das Jacobi–Davidson–Verfahren für das QEP (siehe [27]). Die Eigenpaare (u^*, λ^*) des QEP $V_m^H P(\lambda) V_m u = 0$ kleinerer Dimension m kann man z.B. mit dem QZ–Verfahren näherungsweise bestimmen. So erhält man Näherungen einiger Eigenpaare von $P(\lambda)$.

8.2 Numerische Ergebnisse

Alle Rechnungen wurden in MATLAB 7.4.0 durchgeführt. Für alle Intervallrechnungen wurde die MATLAB–Toolbox INTLAB (Version 5.3) von Siegfried Rump verwendet, welche eine schnelle und effiziente Software für Intervallarithmetik ist (siehe [23]). Die jeweils aktuelle INTLAB Version für Unix und Windows steht zur freien Verfügung auf der Internetseite

`http://www.ti3.tu-hamburg.de/rump/intlab/`

zusammen mit einigen Installations– und Nutzungsinformationen (z.B. [11]).

Es wurde in einem 64 Bit–Gleitkommasystem (double precision) gerechnet, also mit Gleitkommazahlen mit 15–16 Stellen. Die für meine Verfahren benötigten Eigenpaarnäherungen des QEP lieferte der in Kapitel 8.1.1 erläuterte MATLAB–Befehl `polyeig`.

Mit dem MATLAB–Befehl

```
rand('state', i);   Ai = -10+20*rand(n);                    (8.3)
```

wurden $n \times n$–Matrizen $A_i = A_i^{(n)}$, $i = 0, 1, \ldots, 5$, erzeugt, deren Einträge reelle Zufallszahlen aus dem Intervall $[-10, 10]$ sind. Dabei war $A_i^{(n)} \neq A_j^{(n)}$ für $i \neq j$. Für $n \in \mathbb{N}$ erhält man auf diese Weise ein quadratisches reelles $n \times n$–Matrixpolynom $P_n(\lambda) = A_2^{(n)}\lambda^2 + A_1^{(n)}\lambda + A_0^{(n)}$ und kann das entsprechende reelle QEP $P_n(\lambda)x = 0$ der Dimension n betrachten. Für $n = 10$, 50, 100, 200 war $\det A_2^{(n)} \neq 0$ also $P_n(\lambda)$ regulär. Die Eigenpaare von $P_n(\lambda)$ sind entweder reell oder treten in konjugiert komplexen Paaren auf (siehe Satz 1). Es sollen nun einfache Eigenwerte λ^* und jeweils zugehörige Eigenvektoren x^* mit dem Verfahren aus Satz 21 (bei reellen Eigenpaaren) bzw. Satz 23 (bei komplexen Eigenpaaren) verifiziert und möglichst enge Schranken für sie gefunden werden.

Für diese Verfahren wird jeweils eine hinreichend gute Näherung $(\tilde{x}, \tilde{\lambda})$ eines einfachen Eigenpaares von $P_n(\lambda)$ mit $(\tilde{x})_s \approx 1$ für ein $s \in \{1, \ldots, n\}$ benötigt (siehe Bemerkung 10 und 12). Dazu wurden in allen Beispielrechnungen dieses Kapitels mit dem MATLAB–Befehl `polyeig(A0,A1,A2)` Näherungen $(\hat{x}, \hat{\lambda})$ von einfachen Eigenpaaren von $P_n(\lambda)$ berechnet. $\hat{\lambda} \in \mathbb{R}$ wurde dann auf eine gewisse Anzahl $z_{\tilde{\lambda}}$ von Stellen im Gleitpunktsystem gerundet und dies als $\tilde{\lambda}$ gewählt (bei $\hat{\lambda} \in \mathbb{C}$ wurden Real- und Imaginärteil jeweils auf $z_{\tilde{\lambda}}$ Stellen gerundet). $\hat{x} \in \mathbb{R}$ wurde so normiert, dass der betragsmäßig größte (s–te) Eintrag und damit die Maximumsnorm gleich Eins ist; anschließend wurden die Vektoreinträge jeweils auf $z_{\tilde{x}}$ Gleikomma-Stellen gerundet. Bei $\hat{x} \in \mathbb{C}$ wurden nach Normierung auf $||\,|\hat{x}|\,||_\infty = 1$ mit $(\hat{x})_s = 1$ wiederum Real– und Imaginärteil jeweils auf $z_{\tilde{x}}$ Stellen gerundet und dies als Näherung $\tilde{x}$ verwendet. Damit gilt in beiden Fällen $(\tilde{x})_s = 1$ für ein $s \in \{1, \ldots, n\}$. Die Matrix C in Satz 21 wurde als die mit MATLAB berechnete Approximation der Inversen von $f'(\tilde{x}, \tilde{\lambda}) \in \mathbb{R}^{(n+1)\times(n+1)}$ (siehe (5.5) und Bemerkung 10) bzw. C in Satz 23 als die mit MATLAB berechnete Approximation der Inversen von $J(\tilde{x}, \tilde{\lambda}) \in \mathbb{R}^{(2n+2)\times(2n+2)}$ (siehe (6.8) und Bemerkung 12) gewählt. Die Anzahl $z_{\tilde{\lambda}}$ der Stellen der Eigenwertnäherung und die Anzahl $z_{\tilde{x}}$ der Stellen der Eigenvektornäherung wurden stets so gewählt, dass die Voraussetzungen von Satz 21 bzw. 23 erfüllt und $z_{\tilde{\lambda}}, z_{\tilde{x}}$ beide möglichst klein sind.

Wir untersuchen zunächst das reelle QEP $P_{10}(\lambda)x = 0$. `polyeig` deutete an, dass dessen Eigenwerte alle einfach sind. Es gibt 4 relle Eigenwerte und 8 Paare komplex konjugierter Eigenwerte. Betrachten wir zuerst die reellen Eigenpaare von $P_{10}(\lambda)$. Eine hinreichend gute Näherung eines reellen Eigenpaares, die wie eben beschrieben gewonnen wurde, ist

$$\tilde{\lambda} = 3.954\,, \quad \tilde{x} = \begin{pmatrix} 1.833E-01 \\ -1.371E-01 \\ -1.548E-01 \\ 3.178E-01 \\ 1.065E-01 \\ 4.334E-01 \\ -4.567E-02 \\ 1 \\ -3.269E-01 \\ 3.815E-01 \end{pmatrix} \tag{8.4}$$

$(z_{\tilde{\lambda}} = 4, z_{\tilde{x}} = 4, s = 8)$.

Wenn, wie bereits erwähnt, $C \approx f'(\tilde{x}, \tilde{\lambda})^{-1} \in \mathbb{R}^{11\times 11}$ gewählt ist, sind damit alle Voraussetzungen von Satz 21 erfüllt. Dies gilt nicht für $z_{\tilde{\lambda}} \leq 3$ und $z_{\tilde{x}} \leq 3$. Mit

$$\beta := 0.0005 \in (\beta^-, \beta^*) = (2.072898719434058E-04\,,\ 7.900156591471585E-04)\,,$$

wobei $\beta^* := \sqrt{\frac{-p}{3}} - \frac{\tau}{3\gamma}$, liefert das Verfahren aus Satz 21 nach 4 Iterationen die Existenz eines eindeutigen reellen Eigenpaares (x^*, λ^*) von $P_{10}(\lambda)$ mit

$$\lambda^* \in [3.95418009328405^{1}_{0}] := [3.954180093284050, 3.954180093284051] := [\lambda]^*,$$

$$x^* \in \begin{pmatrix} [1.83315056803542^{3}_{2}E-01] \\ [-1.37132703118297^{39}_{40}E-01] \\ [-1.54796119721177^{78}_{80}E-01] \\ [3.17831125509128^{81}_{79}E-01] \\ [1.06492320107227^{1}_{0}E-01] \\ [4.33407811445218^{80}_{78}E-01] \\ [-4.56694003548936^{6}_{9}E-02] \\ 1.0 + [-1,1]\cdot 10^{-15} \\ [-3.26911970099437^{18}_{20}E-01] \\ [3.81519419506171^{8}_{5}E-01] \end{pmatrix} = [x]^*\,.$$

Es sind damit die führenden 14 bzw. 15 Stellen in diesem Gleitpunktsystem garantiert. Wenn man den relativen Durchmesser eines Intervallvektors $[x] \in \mathbf{IR}^n$ als

$$w([x]) := \max_{\substack{1\leq i\leq n \\ [x]_i \neq 0}} \frac{2\,\mathrm{rad}([x]_i)}{|[x]_i|}$$

definiert, ergibt sich $w([x]^*) \leq 6.7E-16$.

Analog wurden Einschließungen der anderen drei reellen Eigenpaare berechnet. Die verwendeten Eigenpaarnäherungen, die benötigte Anzahl an Iterationen (mit $\beta := \frac{1}{2}(\beta^- + \beta^*)$) und die so mit dem Verfahren erhaltenen Eigenpaareinschließungen sind in der folgenden Tabelle festgehalten.

$\tilde{\lambda}$	$[\lambda]^*$	Stellenanzahl $z_{\tilde{x}}$	Schranke für $w([x]^*)$	Iterationen
$-2.78E-01$	$[-2.779711284051978_{81}E-01]$	3	$4.4E-16$	5
$7.2E-01$	$[7.191801964141538_{4}E-01]$	3	$7.6E-16$	7
$3.24E-01$	$[3.241928621154863_{1}E-01]$	3	$3.9E-16$	6

Für die Real– und Imaginärteile der komplexen Eigenpaare $(x^*, \lambda^*) = (x_1^* + ix_2^*, \lambda_1^* + i\lambda_2^*)$ von $P_{10}(\lambda)$ wurden Einschließungen

$$x_1^* \in [x_1]^*,\ x_2^* \in [x_2]^*,\ \lambda_1^* \in [\lambda_1]^*,\ \lambda_2^* \in [\lambda_2]^*$$

mit dem Verfahren aus Satz 23 unter Berücksichtigung von Bemerkung 12 gewonnen. Es sei nochmals an Satz 1 erinnert. Die Wahl der jeweiligen Eigenwertnäherung $\tilde{\lambda} = \tilde{\lambda}_1 + i\tilde{\lambda}_2$ und der dazugehörigen Eigenvektornäherung $\tilde{x} = \tilde{x}_1 + i\tilde{x}_2$ wurde bereits beschrieben. Mit
$C \approx J(\tilde{x}, \tilde{\lambda})^{-1} \in \mathbb{R}^{22\times 22}$ und $\beta := \frac{1}{2}(\beta^- + \beta^*)$ lieferte das Verfahren aus Satz 23 dann folgende Resultate:

$\tilde{\lambda}_1\,,\ \tilde{\lambda}_2$	$[\lambda_1]^*\,,\,[\lambda_2]^*$	Stellenanzahl $z_{\tilde{x}}$	Schranke für $w([x_1]^*)$ u. $w([x_2]^*)$	It.
$-6.1808E-01\,,$ 1.9984	$[-6.180841353542603_{3}E-01]\,,$ $[1.998396178439575_{3}]$	4	$4.1E-16$	5
$-1.3087\,,$ $4.9143E-01$	$[-1.308674944880659_{61}]\,,$ $[4.914285470117102_{0}E-01]$	4	$4.5E-16$	5
$-7.457E-01\,,$ 1.037	$[-7.456830715265740_{3}E-01]\,,$ $[1.037142949012460_{59}]$	4	$4.2E-16$	6
$1.177\,,$ $7.105E-01$	$[1.177195110815153_{1}]\,,$ $[7.104724786484783_{0}E-01]$	4	$4.2E-16$	5
$-1.195E-01\,,$ $9.039E-01$	$[-1.195136694142751_{2}E-01]\ ,$ $[9.039447324391151_{48}E-01]$	4	$4.3E-16$	6
$-2.321E-01\,,$ $5.836E-01$	$[-2.321447349622283_{5}E-01]\,,$ $[5.836331234319512_{0}E-01]$	4	$4.5E-16$	6
$5.036E-01\,,$ $4.374E-01$	$[5.035908959577196_{4}E-01]\,,$ $[4.374328933751938_{6}E-01]$	4	$4.4E-16$	5
$2.92E-01\,,$ $4.646E-01$	$[2.920461383787791_{89}E-01]\,,$ $[4.645619003014199_{7}E-01]$	4	$4.1E-16$	6

Weiterhin wurden die QEPs $P_n(\lambda)x = 0$ für $n = 50, 100$ und 200 untersucht. Dabei wurde vor allem auf den Aspekt ein Auge geworfen, wie viele Stellen $z_{\tilde{\lambda}}$ bei der Eigenwertnäherung $\tilde{\lambda}$ und wie viele Stellen $z_{\tilde{x}}$ bei der Eigenvektornäherung $\tilde{x}$, die auf die bereits beschriebene Art mit Hilfe von `polyeig` gewonnen wurden, nötig sind, damit die Voraussetzungen von Satz 21 bzw. Satz 23 erfüllt sind. Dazu wurden jeweils exemplarisch einige einfache Eigenpaare mit $|\tilde{\lambda}| \in [2.5, 4]$ bzw. $|\tilde{\lambda}_1|, |\tilde{\lambda}_2| \in [1, 10]$ betrachtet.

Mit der reellen Eigenwertnäherung $\tilde{\lambda} = 2.8718$ ($z_{\tilde{\lambda}} = 5$) und der zugehörigen auf $z_{\tilde{x}} = 4$ Stellen gerundeten Eigenvektornäherung $\tilde{x}$ von $P_{50}(\lambda)$ erhält man mit $\beta = \frac{1}{2}(\beta^- + \beta^*)$ nach

5 Iterationen des Verfahrens aus Satz 21 die Eigenwerteinschließung $[2.87178557641220_7^8]$ und eine Eigenvektoreinschließung $[x]^*$ mit $w([x]^*) \leq 4.4E-016$. Für die Eigenpaarnäherung $(\tilde{x}, \tilde{\lambda} = 3.46835)$ von $P_{100}(\lambda)$ mit $z_{\tilde{\lambda}} = 6$ und $z_{\tilde{x}} = 5$ erhält man nach 5 Iterationen eine sehr gute Einschließung eines einfachen Eigenpaares von $P_{100}(\lambda)$. Für $P_{200}(\lambda)$ waren ebenfalls $z_{\tilde{\lambda}} = 6$ Stellen bei der Eigenwertnäherung $\tilde{\lambda} = -3.73252$ und $z_{\tilde{x}} = 5$ Stellen bei der entsprechenden Eigenvektornäherung nötig, damit die Voraussetzungen von Satz 21 erfüllt sind. Es sieht also so aus, als ob man bei großer Dimension n von $P_n(\lambda)$ bessere Eigenpaarnäherungen als bei kleinerer Dimension benötigt, um die Voraussetzungen von Satz 21 zu erfüllen und das dazugehörige Einschließungsverfahren anwenden zu können.

Problematisch ist aber nicht direkt die Größe von n. In unserem Beispiel sind $A_1^{(n)}$ und $A_2^{(n)}$ für $n = 10, 50, 100, 200$ so gewählt (siehe (8.3)), dass jeweils

$$\begin{aligned} \{||A_2^{(n)}||_\infty\}_{n=10,50,100,200} &= \{64.23..., 310.93..., 558.18..., 1115.43...\} \quad \text{und} \\ \{||A_1^{(n)}||_\infty\}_{n=10,50,100,200} &= \{60.66..., 301.20..., 578.88..., 1108.36...\} \end{aligned}$$

stark monoton wachsende Folgen sind. Lemma 16 garantiert $||C||_\infty \geq 1$ für $C = f'(\tilde{x}, \tilde{\lambda})$. In unseren letzten drei Beispielen wie auch bei der Eigenpaarnäherung $(\tilde{x}, \tilde{\lambda} = 3.945)$ von $P_{10}(\lambda)$ aus (8.4) gilt $||C||_\infty \leq 1.125 = \frac{9}{8}$. Die Eigenvektornäherung wurde außerdem stets so gewählt, dass $||\tilde{x}||_\infty = 1$ gilt. Wegen $2.5 \leq |\tilde{\lambda}| \leq 4$ gilt dann

$$\begin{aligned} 6||A_2^{(n)}||_\infty + ||A_1^{(n)}||_\infty \leq \tau^{(n)} &:= ||C||_\infty\{(2|\tilde{\lambda}| + ||\tilde{x}||_\infty)||A_2^{(n)}||_\infty + ||A_1^{(n)}||_\infty\} \\ &\leq \frac{81}{8}||A_2^{(n)}||_\infty + \frac{9}{8}||A_1^{(n)}||_\infty\,, \\ ||A_2^{(n)}||_\infty \leq \gamma^{(n)} &:= ||C||_\infty||A_2^{(n)}||_\infty \leq \frac{9}{8}||A_2^{(n)}||_\infty\,. \end{aligned}$$

Weil $\{||A_2^{(n)}||_\infty\}_{n=10,50,100,200}$ und $\{||A_1^{(n)}||_\infty\}_{n=10,50,100,200}$ sehr stark monoton wachsend sind, sind dann auch $\{\tau^{(n)}\}_{n=10,50,100,200}$ und $\{\gamma^{(n)}\}_{n=10,50,100,200}$ monoton wachsend. Bemerkung 11 liefert nun eine Erklärung dafür, warum die Näherungen $(\tilde{x}, \tilde{\lambda})$ für größeres n besser sein müssen.

Bei der Einschließung von komplexen Eigenpaaren mit Satz 23 sieht die Sache ähnlich aus. Mit $\tilde{\lambda} = 1.02294 + i\,1.34267$ (d.h. $z_{\tilde{\lambda}} = 6$) und $z_{\tilde{x}} = 5$ bei $P_{50}(\lambda)$ ergibt sich nach 4 Iterationen eine gute Einschließung. Bei $P_{100}(\lambda)$ und $\tilde{\lambda} = 9.632125 + i\,4.772286$ bzw. $P_{200}(\lambda)$ und $\tilde{\lambda} = 2.877178 + i\,3.374934$ mussten sogar $z_{\tilde{\lambda}} = 7$ und $z_{\tilde{x}} = 6$ Stellen bei der Eigenpaarnäherung gewählt werden, damit die Voraussetzungen von Satz 23 erfüllt sind. Wenn man sich die Definition von τ und γ in Satz 23 anschaut, kann man nach ähnlichen Überlegungen wie im reellen Fall einen Zusammenhang zwischen (für wachsendes n) wachsenden Normen $||A_2^{(n)}||_\infty$ und $||A_1^{(n)}||_\infty$ auf der einen Seite und größerer Genauigkeit der benötigten Eigenpaarnäherungen auf der anderen Seite herstellen.

Abschließend wird das komplexe quadratische $n \times n$–Matrixpolynom

$$Q_n(\lambda) = K_2^{(n)}\lambda^2 + K_1^{(n)}\lambda + K_0^{(n)}$$

mit $K_2^{(n)} = A_2^{(n)} + iA_5^{(n)}$, $K_1^{(n)} = A_1^{(n)} + iA_4^{(n)}$ und $K_0^{(n)} = A_0^{(n)} + iA_3^{(n)}$ für $A_i^{(n)}$ aus (8.3) betrachtet. Es sollen komplexe einfache Eigenpaare $(x_1^* + ix_2^*, \lambda_1^* + i\lambda_2^*)$ von $Q_n(\lambda)$ mit Hilfe

von Satz 25 und Bemerkung 12 (d.h. $C \approx J(\tilde{x}, \tilde{\lambda})^{-1}$, siehe (7.3)) eingeschlossen werden. Dazu werden Eigenpaarnäherungen analog zum Fall komplexer Eigenpaare beim reellen QEP mit Hilfe des MATLAB–Befehls `polyeig(A0+i*A3,A1+i*A4,A2+i*A5)` erzeugt. Für $n = 10$ und die Näherung $(\tilde{x}, \tilde{\lambda} = 1.5174 - i\,2.7743)$ mit $z_{\tilde{x}} = 4$ und $z_{\tilde{\lambda}} = 5$ erhält man mit $\beta = \frac{1}{2}(\beta^- + \beta^*)$ so nach 4 Iterationen die Einschließungen

$$\lambda_1^* \in [1.51738676352439_{3}^{4}],\ \lambda_2^* \in [-2.77431355669888_{90}^{89}],\ \max\{w([x_1]^*), w([x_2]^*)\} \leq 3.8E-16\,.$$

Für $n = 100$ und $n = 200$ wurden ebenfalls numerische Experimente durchgeführt. So erhielt man für $n = 100$, $\tilde{\lambda} = 1.22875 - i\,2.55582$ (d.h. $z_{\tilde{\lambda}} = 6$) und $z_{\tilde{x}} = 5$ nach 5 Iterationen und für $n = 200$, $\tilde{\lambda} = 1.110027 - i\,2.222785$ (d.h. $z_{\tilde{\lambda}} = 7$) und $z_{\tilde{x}} = 6$ nach 4 Iterationen sehr gute Eigenpaareinschließungen.

Kapitel 9

Eine Anwendung des QEP

9.1 Die Millennium Bridge

Die Millennium Bridge ist eine 320 m lange Hängebrücke für Fußgänger über die Themse im Zentrum von London. Sie wurde von dem bekannten britischen Architekten Lord Norman Forster, der auch die Reichstagskuppel gestaltete, konzipiert. Als die Brücke am 10. Juni 2000 für den Publikumsverkehr geöffnet wurde, war die Begeisterung groß: ca. 100000 Personen überquerten sie an diesem Tag. Wenn große Personengruppen über die Brücke gingen, kam es zu viel stärkeren Seitwärtsbewegungen der Brücke als erwartet. Sie schwankte stark nach rechts und links (insgesamt bis zu 7 cm). So kam es 2 Tage später zur Schließung der Brücke und anschließend zu einer genauen Untersuchung dieses unerwarteten Phänomens, siehe [8] und `http://www.arup.com/MillenniumBridge/`.

Abbildung 9.1: Millennium Bridge in London.

Es gibt einen Zusammenhang zwischen diesem Vorfall und dem QEP. Es handelt sich bei der Brücke um ein schwingendes System mit einer gewissen Dämpfung. Es besitzt natürliche Frequenzen, mit denen es bevorzugt schwingt. Wir werden später darauf eingehen, dass es einen Zusammenhang zwischen diesen natürlichen Frequenzen und den Eigenwerten eines speziellen reellen quadratischen Matrixpolynoms gibt. Wenn das System von einer äußeren Kraft erregt wird, deren Frequenz nahe der Eigenfrequenz ist, werden die Schwingungen des Systems verstärkt und es wird instabil. Dieses Verhalten wird als Resonanz bezeichnet.

Bei der Millennium Bridge wurden diese äußeren Kräfte durch Bewegungen der Fußgänger verursacht. Eine Person erzeugt beim Gehen nicht nur eine vertikale Kraft, die wegen der Auf- und Abwärtsbewegung der Körpermasse bei jedem Schritt fluktuiert, sondern auch eine kleine seitwärts gerichtete Kraft. Denn die leicht auseinanderstehenden Beine verursachen beim Gehen ein Verlagern der Körpermasse abwechselnd nach rechts und links. Die dabei entstehende Kraft ist nach links gerichtet, wenn man mit dem linken Fuß auftritt, und nach rechts gerichtet, wenn man mit dem rechten Fuß auftritt. Das Ingenieurteam der Millennium Bridge fand nach einigen Tests und Nachforschungen heraus, dass diese geringe seitwärts gerichtete Kraft die Ursache für das starke Schwanken der Brücke war.

Das Gleichgewicht einer Person beim Gehen wird kaum beeinflusst von vertikalen Bewegungen des Untergrunds. Ganz anders ist das bei Seitwärtsbewegungen. Wenn der Untergrund seitwärts gerichtet ein wenig oszilliert (bei einer Hängebrücke z.B. verursacht durch seitlichen Wind), setzt man beim Gehen die Beine weiter auseinander, um sich zu stabilisieren. Damit vergrößert sich die seitwärts gerichtete Kraft der Person. Außerdem ist es für die Person bequemer, im Einklang mit der Bewegung des Untergrunds zu gehen. Diese Neigung des Fußgängers zur Synchronisation hat den Effekt, dass jeder Schritt die Oszillation des Untergrunds verstärkt. Wegen der sich verstärkenden Bewegung des Untergrunds versucht der Fußgänger wiederum, sein Gehen wie eben beschrieben mit dieser Bewegung zu synchronisieren.

Bei einer Menschenmenge verstärkt sich dieser Effekt. Die individuellen Reaktionsmuster unterscheiden sich zwar, aber die Mehrheit der Menschenmenge läuft synchron zur Brückenbewegung, weil es so am bequemsten ist. Dieses instinktive Verhalten führt dazu, dass sich die von einer Menschenmenge erzeugten seitwärts gerichteten Kräfte der Eigenfrequenz und Phase der Seitwärtsbewegung der Brücke anpassen und diese Bewegung so verstärken. Dieses Phänomen wird als Synchrone Laterale Erregung (Synchronous Lateral Excitation) bezeichnet.

Jede Struktur besitzt eine gewisse Dämpfung, welche die Auslenkung der Schwingung der Struktur verringert. Die Dämpfungskraft wächst mit wachsender Bewegung der Struktur. Mit wachsender Anzahl der Personen auf der Brücke steigt auch die durch sie verursachte Erregungskraft an, die Dämpfung der Brücke bleibt aber unverändert. Solange die Dämpfungskraft größer als die Erregungskraft ist, ist die Seitwärtsbewegung der Brücke gering. Wenn aber so viele Personen auf der Brücke gehen, dass deren Erregungskraft größer als die Dämpfungskraft ist, tritt Synchrone Laterale Erregung auf, und die Seitwärtsbewegung der Brücke verstärkt sich drastisch.

Nach einigen Tests stellte sich heraus, dass die durchschnittliche seitwärts gerichtete Kraft, die eine Person beim Gehen über eine Brücke auf diese ausübt, propotional zur Geschwindigkeit der Seitwärtsbewegung der Brücke ist. Somit kann man nun die Größe der Menschenmenge berechnen, die Synchrone Laterale Erregung auf der Brücke verursacht. Synchrone Laterale Erregung ist kein gradueller Effekt, sondern tritt plötzlich auf: bei einem Test auf einem Teilstück der Brücke war die Seitwärtsbewegung der Brücke ziemlich gleichbleibend bei bis zu 156 Fußgängern auf der Brücke. Als nur 10 weitere Personen dazu kamen, verstärkte sich die Be-

wegung plötzlich sehr stark und der Test musste abgebrochen werden. Dieser Test bestätigte die Berechnungen der Ingenieure.

Das Phänomen der Synchronen Lateralen Erregung trat natürlich nicht nur bei der Millennium Bridge im Speziellen auf. Es kann theoretisch bei jeder Brücke dazu kommen, und nach einigen Nachforschungen konnten diverse Fälle von vermuteter Synchroner Lateraler Erregung bei anderen Brücken ausgemacht werden. Neben der Anzahl der Personen, die Synchrone Laterale Erregung auf einer speziellen Brücke verursacht, kann nun auch die nötige Dämpfung einer Brücke berechnet werden, die ein Auftreten dieses Phänomens verhindert. Dies kann bei der Konzeption zukünftiger Brücken berücksichtigt werden.

Im Falle der Millennnium Bridge wurden viskose Dämpfer unter dem Brückendeck, an den Brückenpfeilern und an der Verbindung der Brücke zum Ufergrund angebracht. Sie dämpfen die Seitwärtsbewegung der Brücke. Jeder einzelne Dämpfer baut die Energie der Seitwärtsbewegung dadurch ab, dass sich ein Kolben innerhalb einer Flüssigkeit vor– und zurückbewegt.

Die Kosten für den Umbau der Brücke betrugen ca. 5 Mio. britische Pfund. Die Millennium Bridge wurde im Februar 2002 wiedereröffnet und bereitet seitdem keine Probleme.

9.2 Ideales gedämpftes Masse-Feder-System

Ein ideales gedämpftes Masse-Feder-System kann man sich folgendermaßen vorstellen: Ein starrer Block der Masse M (in kg) befindet sich auf Rollen auf einer ebenen Fläche. Der Block ist durch eine Feder und einen Dämpfer mit einer Wand verbunden. Die Feder besitzt die Feder-Konstante K (in $\frac{N}{m}$) und der Dämpfer die Dämpfer-Konstante C (in $\frac{Ns}{m}$). Es wird der Einfachheit halber angenommen, dass keine äußere Kraft auf den Block einwirkt.

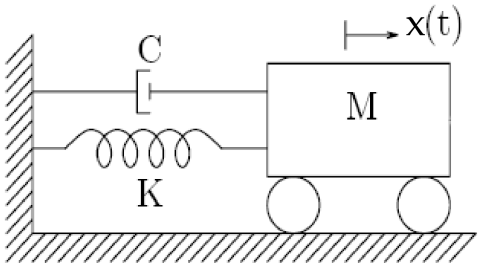

Abbildung 9.2: ideales gedämpftes Masse-Feder-System

Angenommen der Block wurde aus seiner Gleichgewichtslage $x = 0$ gebracht. Dann kann man den Ort des Blocks x in Abhängigkeit von der Zeit t mit den folgenden Gleichungen beschreiben:

$$F_{\text{Feder}} = -Kx \tag{9.1}$$

$$F_{\text{Dämpfer}} = -Cv = -C\dot{x} \tag{9.2}$$

$$F_{\text{gesamt}} = F_{\text{Feder}} + F_{\text{Dämpfer}}$$

(9.1) ist das Hookesche Gesetz. Es besagt, dass die Federkraft proportional zur Auslenkung des Blocks aus seiner Gleichgewichtslage ist. (9.2) sagt aus, dass die Dämpfungskraft proportional

zur Geschwindigkeit des Blocks ist. Wegen des 2. Newtonschen Gesetzes $F_{\text{gesamt}} = Ma = M\ddot{x}$ liefert dies die homogene Differentialgleichung 2. Ordnung

$$M\ddot{x} + C\dot{x} + Kx = 0\,,$$

welche äquivalent ist zu

$$\ddot{x} + \frac{C}{M}\dot{x} + \frac{K}{M}x = 0\,.$$

Diese Differentialgleichung kann man nun folgendermaßen umschreiben:

$$\ddot{x} + 2\zeta\omega\,\dot{x} + \omega^2\,x = 0\,, \qquad \omega = \sqrt{\frac{K}{M}}\,, \quad \zeta = \frac{C}{2\sqrt{MK}}\,. \tag{9.3}$$

In den Ingenieurwissenschaften wird ω natürliche Frequenz und das dimensionslose ζ viskoser Dämpfungsfaktor genannt. Das zur homogenen reellen Differentialgleichung 2. Ordnung mit konstanten Koeffizienten (9.3) gehörige charakteristische Polynom

$$\lambda^2 + 2\zeta\omega\,\lambda + \omega^2$$

besitzt die Nullstellen

$$\lambda_{1,2} = (-\zeta \pm \sqrt{\zeta^2 - 1})\omega\,.$$

Also ist die allgemeine Lösung von (9.3)

$$\begin{aligned}
x(t) &= \alpha e^{\mathrm{Re}\,(\lambda_1)t}\cos(\mathrm{Im}\,(\lambda_1)t) + \beta e^{\mathrm{Re}\,(\lambda_1)t}\sin(\mathrm{Im}\,(\lambda_1)t) \qquad \text{für } \lambda_1, \lambda_2 = \overline{\lambda_1} \in \mathbb{C}\,,\\
x(t) &= \alpha e^{\lambda_1 t} + \beta e^{\lambda_2 t} \qquad \text{für } \lambda_1, \lambda_2 \in \mathbb{R} \text{ mit } \lambda_1 \neq \lambda_2\,,\\
x(t) &= \alpha e^{\lambda_1 t} + \beta t e^{\lambda_1 t} \qquad \text{für } \lambda_1 = \lambda_2 \in \mathbb{R}\,,
\end{aligned}$$

wobei α und β reelle Konstanten sind, die durch die Anfangsbedingungen $x(t_0)$, $\dot{x}(t_0)$ festgelegt sind.

Die Lösung $x(t)$ verhält sich für verschiedene Werte des viskosen Dämpfungsfaktors ζ also unterschiedlich.

Für $\zeta = 0$ gilt $\lambda_{1,2} = \pm i\omega$. Nach einer Transformation erhält man

$$x(t) = \gamma\sin(\omega t + \varphi)\,,$$

wobei die reellen Konstanten γ und φ durch die Anfangsbedingungen festgelegt sind. Das System schwingt also harmonisch mit der natürlichen Frequenz ω. Dies ist der ungedämpfte Fall.

Für $\zeta \in (0,1)$ gilt $\lambda_{1,2} = (-\zeta \pm i\eta)\omega$ mit $\eta = \sqrt{1-\zeta^2} \in (0,1)$. Damit hat die Lösung die Gestalt

$$x(t) = \gamma e^{-\zeta\omega t}\sin(\eta\omega t + \varphi)\,,$$

ist also eine Sinuskurve mit abnehmender Amplitude. Dies ist der unterdämpfte Fall. Hier tritt eine schwache Dämpfung auf.

Für $\zeta = 1$ tritt die doppelte Nullstelle $\lambda_{1,2} = -\omega$ auf. Dann gilt

$$x(t) = (\alpha + \beta t)e^{-\omega t}$$

mit α, $\beta \in \mathbb{R}$. Dies ist die geringste Dämpfung, bei der keine Schwingung in $x(t)$ auftritt (kritische Dämpfung).

Letztlich gibt es noch den Fall $\zeta > 1$ mit den reellen negativen Nullstellen $\lambda_1 \neq \lambda_2$. Dann ist also

$$x(t) = \alpha e^{\lambda_1 t} + \beta e^{\lambda_2 t}$$

exponentiell fallend. Dies ist der überdämpfte Fall (starke Dämpfung).

9.3 Gedämpftes Masse-Feder-System mit n Freiheitsgraden

Wir betrachten nun ein zusammenhängendes gedämpftes Masse-Feder-System, siehe [27]. Man kann sich so stark vereinfacht eine Brücke vorstellen. Es gibt n Masseblöcke. Der i–te Masseblock besitzt das Gewicht m_i und ist mit dem $(i+1)$–ten Masseblock durch eine Feder mit Federkonstante k_i und einen Dämpfer mit Dämpfungskonstante c_i verbunden. Der i–te Masseblock ist ebenfalls mit dem Untergrund durch eine Feder (Konstante κ_i) und einen Dämpfer (Konstante τ_i) verbunden. Es wirken keine äußeren Kräfte auf das System ein.

Den Ort $x_i(t)$ des i–ten Blocks $(i = 1, \ldots, n)$ kann man mit den folgenden Gleichungen beschreiben:

$$\begin{aligned}
F_{\mathrm{F},i-} &= -k_{i-1}(x_i - x_{i-1}) \\
F_{\mathrm{D},i-} &= -c_{i-1}(\dot{x}_i - \dot{x}_{i-1}) \\
F_{\mathrm{F},i+} &= k_i(x_{i+1} - x_i) \\
F_{\mathrm{D},i+} &= c_i(\dot{x}_{i+1} - \dot{x}_i) \\
F_{\mathrm{F},g} &= -\kappa_i\, x_i \\
F_{\mathrm{D},g} &= -\tau_i\, \dot{x}_i \\
F_{\mathrm{ges},i} &= F_{\mathrm{F},i-} + F_{\mathrm{D},i-} + F_{\mathrm{F},i+} + F_{\mathrm{D},i+} + F_{\mathrm{F},g} + F_{\mathrm{D},g}\,,
\end{aligned}$$

wobei $k_0 = c_0 = k_n = c_n = 0$ definiert wird. Wegen $F_{\mathrm{ges},i} = m_i\, \ddot{x}_i$ ergibt sich dann für $i = 1, \ldots, n$ die Differentialgleichung

$$m_i\ddot{x}_i - c_{i-1}\dot{x}_{i-1} + (c_{i-1} + c_i + \tau_i)\dot{x}_i - c_i\dot{x}_{i+1} - k_{i-1}x_{i-1} + (k_{i-1} + k_i + \kappa_i)x_i - k_i x_{i+1} = 0\,.$$

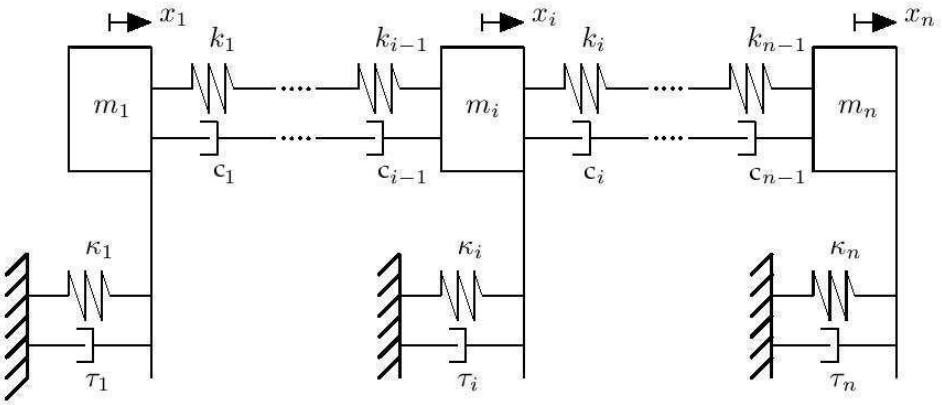

Abbildung 9.3: gedämpftes Masse-Feder-System mit n Freiheitsgraden

Mit $x = (x_1, x_2, \ldots, x_n)^T$ kann man dann das gedämpfte Masse-Feder-System mit n Freiheitsgraden also als homogene Differentialgleichung 2. Ordnung

$$M\ddot{x} + C\dot{x} + K = 0$$

mit Massematrix $M = \operatorname{diag}(m_1, \ldots, m_n)$, Steifheitsmatrix K und Dämpfungsmatrix C beschreiben. K und C sind reelle symmetrische $n \times n$–Tridiagonalmatrizen, die folgende Gestalt besitzen:

$$K = \begin{pmatrix} k_1+\kappa_1 & -k_1 & 0 & \ldots & & \ldots & 0 \\ -k_1 & k_1+k_2+\kappa_2 & -k_2 & & & & \vdots \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ \vdots & & & -k_{n-2} & k_{n-2}+k_{n-1}+\kappa_{n-1} & -k_{n-1} & \\ 0 & \ldots & & \ldots & 0 & -k_{n-1} & k_{n-1}+\kappa_n \end{pmatrix},$$

$$C = \begin{pmatrix} c_1+\tau_1 & -c_1 & 0 & \ldots & & \ldots & 0 \\ -c_1 & c_1+c_2+\tau_2 & -c_2 & & & & \vdots \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ \vdots & & & -c_{n-2} & c_{n-2}+c_{n-1}+\tau_{n-1} & -c_{n-1} & \\ 0 & \ldots & & \ldots & 0 & -c_{n-1} & c_{n-1}+\tau_n \end{pmatrix}.$$

Mit $P = \operatorname{tridiag}(-1, 1, 0)$ kann man aber auch eleganter

$$\begin{aligned} K &= P \operatorname{diag}(k_1, \ldots, k_{n-1}, 0)\, P^T + \operatorname{diag}(\kappa_1, \ldots, \kappa_n), \\ C &= P \operatorname{diag}(c_1, \ldots, c_{n-1}, 0)\, P^T + \operatorname{diag}(\tau_1, \ldots, \tau_n) \end{aligned}$$

schreiben.

Vereinfachender Weise nehmen wir nun an, dass $k_i = \kappa_j = \kappa$ und $c_i = \tau_j = \tau$ für $i = 1, \ldots, n-1$ und $j = 2, \ldots, n-1$. Außerdem sei $\kappa_1 = \kappa_n = 2\kappa$, $\tau_1 = \tau_n = 2\tau$ und $m_i \equiv 1$. Das heißt alle Federn (bzw. Dämpfer) außer der ersten und der letzten besitzen dieselbe Konstante $\kappa > 0$ (bzw. $\tau > 0$) und alle Masseblöcke haben dasselbe Gewicht. Dann ergibt sich

$$M = I\,, \qquad C = \tau \operatorname{tridiag}(-1, 3, -1) =: C_\tau\,, \qquad K = \kappa \operatorname{tridiag}(-1, 3, -1) =: K_\kappa\,.$$

Angenommen das reelle reguläre quadratische Eigenwertproblem $(A_2\lambda^2 + A_1\lambda + A_0)v = 0$ der Dimension $n \times n$ besitzt $2n$ verschiedene (d.h. einfache) Eigenwerte $\lambda_i \in \mathbb{C}$, $i = 1, \ldots, 2n$. Außerdem sei $v_i \in \mathbb{C}^n$ der Eigenvektor zu λ_i, $i = 1, \ldots, 2n$. Nach Satz 9 ist dann $x(t) = e^{\lambda_i t} v_i$ für $i = 1, \ldots, 2n$ Lösung der homogenen Differentialgleichung $A_2\ddot{x} + A_1\dot{x} + A_0 = 0$. Nach Theorem 6.3 aus [16] lautet die allgemeine (komplexe) Lösung dieser Differentialgleichung

$$x(t) = \sum_{i=1}^{2n} \alpha_i e^{\lambda_i t} v_i$$

mit Konstanten $\alpha_i \in \mathbb{C}\,,\ i = 1, \ldots, 2n$.

Wir betrachten nun die Differentialgleichung

$$\ddot{x} + C_\tau \dot{x} + K_\kappa x = 0$$

bzw. das zugehörige quadratische Eigenwertproblem

$$(I\lambda^2 + C_\tau \lambda + K_\kappa)v = 0 \tag{9.4}$$

für $n = 50$.

Für $\kappa = 5$ und $\tau = 8$ wurden mit dem MATLAB–Befehl `polyeig` die Eigenpaare von (9.4) approximiert. Es gibt $2n = 100$ verschiedene Eigenwerte λ_i, die reell und negativ sind (siehe Abbildung 9.4). Auch die zugehörigen Eigenvektoren v_i sind folglich reell. Es gibt eine große Lücke zwischen den n betragsmäßig größten ($-39.4 < \lambda_i < -7.3$ für $i = 1, \ldots, 50$) und den n betragsmäßig kleinsten Eigenwerten ($-0.684 < \lambda_i < -0.635$ für $i = 51, \ldots, 100$). Zur Verdeutlichung: die mit dem Verfahren aus Satz 21 unter Berücksichtigung von Bemerkung 10 erhaltenen engen Einschließungen der beiden betragsmäßig kleinsten Eigenwerte sind
$\lambda_{100} \in [-6.35091175899986^1_4E-01],\ \lambda_{99} \in [-6.351149411512900^0_4E-01]\,.$

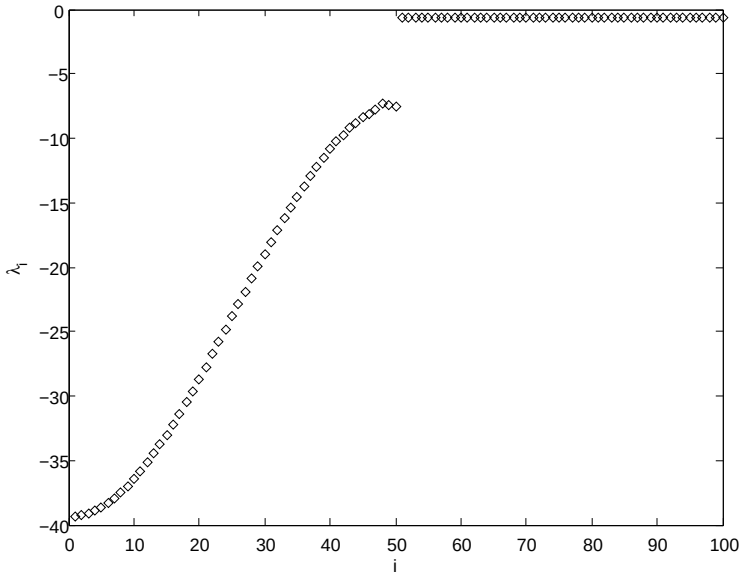

Abbildung 9.4: Eigenwertverteilung des QEPs mit $\kappa = 5$, $\tau = 8$ und $n = 50$

Die allgemeine (reelle) Lösung der Differentialgleichung hat dann die Gestalt

$$x(t) = \sum_{i=1}^{100} \alpha_i e^{\lambda_i t} v_i$$

mit

$$0 > \lambda_i \in \mathbb{R}\,, \quad i = 1, \ldots, 100 \tag{9.5}$$

und beliebigen Konstanten $\alpha_i \in \mathbb{R}$, d.h. in der Lösung treten ausschließlich exponentiell fallende Funktionen, jedoch keine Oszillation auf. Das System ist also überdämpft (sehr stark gedämpft). Dies ist im Fall einer Brücke ja gewünscht.

Für $\kappa = 5$ und $\tau = 3$ besitzt (9.4) nach `polyeig` ebenfalls $2n = 100$ verschiedene Eigenwerte λ_i. Diese sind entweder reell und negativ oder treten in komplex konjugierten Paaren mit negativem Realteil auf (siehe Abbildung 9.5). Die Eigenvektoren können alle reell gewählt werden. Konjugiert komplexe Eigenwerte besitzen denselben Eigenvektor (siehe Satz 1). Damit hat die allgemeine (reelle) Lösung dieser Differentialgleichung die Gestalt

$$x(t) = \sum_{\substack{i \text{ mit} \\ \mathrm{Im}\,(\lambda_i) = 0}} \alpha_i e^{\lambda_i t} v_i + \sum_{\substack{i \text{ mit} \\ \mathrm{Im}\,(\lambda_i) > 0}} \alpha_i e^{\mathrm{Re}\,(\lambda_i) t} \sin(\mathrm{Im}\,(\lambda_i) t + \varphi_i) v_i$$

mit beliebigen Konstanten $\alpha_i \in \mathbb{R}$. Wegen

$$\mathrm{Re}\,(\lambda_i) < 0\,, \quad i = 1, \ldots, 100\,, \tag{9.6}$$

gibt es also Anteile der Lösung ohne Oszillation und Anteile, die Sinuskurven mit abnehmender Amplitude entsprechen. Das System ist damit stabil. Es tritt Dämpfung auf.

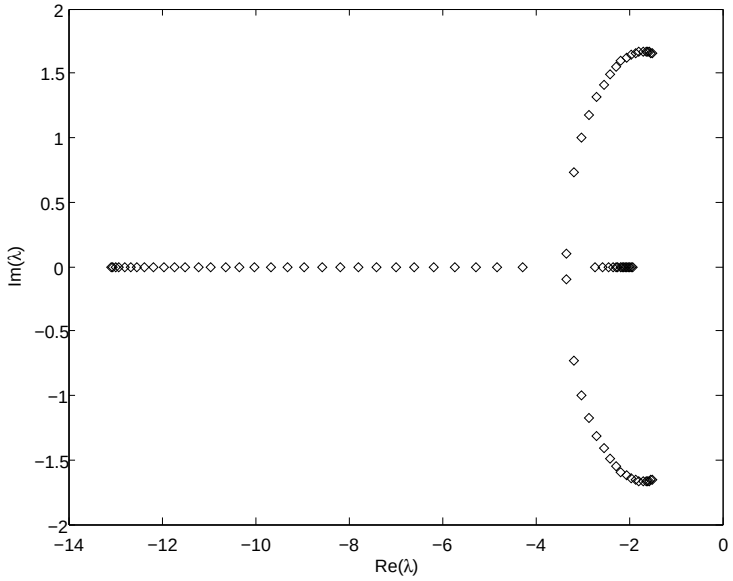

Abbildung 9.5: Eigenwertverteilung des QEPs mit $\kappa = 5$, $\tau = 3$ und $n = 50$

Es sei nun das Differentialgleichungssystem $A_2\ddot{x} + A_1\dot{x} + A_0 x = 0$ mit $A_0, A_1, A_2 \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit betrachtet. Lancaster hat schon in [16] (Kapitel 7.5) gezeigt, dass für die Eigenwerte des zugehörigen QEP die Eigenschaft (9.6) gilt. Das System ist dann stabil. Im Kapitel 7.6 von [16] wurde Folgendes bewiesen: Wenn zusätzlich die Überdämpfungsbedingung

$$(x^T A_1 x)^2 - 4(x^T A_0 x)(x^T A_2 x) > 0 \qquad \text{für } x \neq 0$$

erfüllt ist, gilt (9.5) und das Differentialgleichungssystem ist überdämpft; außerdem existiert dann eine Lücke zwischen den n größten und den n kleinsten Eigenwerten des QEP. Die eben betrachteten Beispiele gehören in diese Klasse von Differentialgleichungen.

Kapitel 10

Einschließung von reellen Eigenpaaren des reellen polynomialen Eigenwertproblems

In diesem Kapitel wird eine Verallgemeinerung von Kapitel 5 vorgenommen: reelle einfache Eigenpaare (x^*, λ^*) des reellen polynomialen Eigenwertproblems $P(\lambda)x = 0$ für das reguläre $n \times n$–Matrixpolynom vom Grad l

$$P(\lambda) = A_l\lambda^l + \ldots + A_1\lambda + A_0 \tag{10.1}$$

(wobei $A_0, \ldots, A_l \in \mathbb{R}^{n\times n}$ und $\det A_l \neq 0$) sollen eingeschlossen werden.

Als Hilsmittel dazu definieren wir in Analogie zu (4.1) ein System vom Grad r.

Definition 10

Eine Funktion $h : \mathbb{R}^m \to \mathbb{R}^m$ *der Gestalt*

$$h(x) = m_0 + M_1x + M_2x^2 + \ldots + M_rx^r\,,$$

wobei $m_0 \in \mathbb{R}^m$, $M_1 \in \mathbb{R}^{m\times m}$ *und*

$$M_t : \begin{cases} \overbrace{\mathbb{R}^m \times \ldots \times \mathbb{R}^m}^{t-mal} & \to & \mathbb{R}^m \\ (x^{(1)}, x^{(2)}, \ldots, x^{(t)}) & \mapsto & \left(\displaystyle\sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\ldots\sum_{i_t=1}^{m} m^{(t)}_{i,i_1,i_2,\ldots,i_t}x^{(1)}_{i_t}\cdot\ldots\cdot x^{(t-1)}_{i_2}x^{(t)}_{i_1}\right)_{i=1,\ldots,m} \end{cases}$$

mit

$$M_tx^t := M_t(\underbrace{x, \ldots, x}_{t-mal}) \qquad \textit{für} \quad 2 \leq t \leq r\,,$$

heißt System vom Grad r *der Dimension* m.

Für $2 \leq t \leq r$ *wird die Abbildung* M_t *mit* $(m^{(t)}_{i,i_1,\ldots,i_t}) \in \overbrace{\mathbb{R}^{m\times\ldots\times m}}^{(t+1)-mal}$ *identifiziert. Außerdem sei*

$$||M_t||_\infty := \max_{i=1,\ldots,m}\left\{\sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\ldots\sum_{i_t=1}^{m}|m^{(t)}_{i,i_1,i_2,\ldots,i_t}|\right\}$$

definiert. Für $D = (d_{ij}) \in \mathbb{R}^{m\times m}$ *ist*

$$D \cdot M_t := \Big(\sum_{j=1}^{m} d_{ij}\, m^{(t)}_{j,i_1,i_2,\ldots,i_t}\Big) \in \mathbb{R}^{\overbrace{m\times\ldots\times m}^{(t+1)-mal}}.$$

Lemma 21

Für $D \in \mathbb{R}^{m\times m}$ *und* $M_t \in \mathbb{R}^{\overbrace{m\times\ldots\times m}^{(t+1)-mal}}$ *gilt*

$$||D \cdot M_t||_\infty \le ||D||_\infty \cdot ||M_t||_\infty .$$

Beweis:
Mit der Bezeichnung $D = (d_{ij})$ und $M_t = (m^{(t)}_{j,i_1,\ldots,i_t})$ gilt für ein $i^* \in \{1,\ldots,m\}$

$$\begin{aligned}
||DM_t||_\infty &= \max_{i=1,\ldots,m}\Big\{\sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\ldots\sum_{i_t=1}^{m}\Big|\sum_{j=1}^{m} d_{ij} m^{(t)}_{j,i_1,i_2,\ldots,i_t}\Big|\Big\}\\
&= \sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\ldots\sum_{i_t=1}^{m}\Big|\sum_{j=1}^{m} d_{i^*j} m^{(t)}_{j,i_1,i_2,\ldots,i_t}\Big|\\
&\le \sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\ldots\sum_{i_t=1}^{m}\sum_{j=1}^{m}|d_{i^*j}|\,|m^{(t)}_{j,i_1,i_2,\ldots,i_t}|\\
&= \sum_{j=1}^{m}|d_{i^*j}|\Big(\sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\ldots\sum_{i_t=1}^{m}|m^{(t)}_{j,i_1,i_2,\ldots,i_t}|\Big)\\
&\le \sum_{j=1}^{m}|d_{i^*j}|\cdot||M_t||_\infty = ||M_t||_\infty\cdot\Big(\sum_{j=1}^{m}|d_{i^*j}|\Big)\\
&\le ||M_t||_\infty\cdot||D||_\infty .
\end{aligned}$$

□

Es wird nun wieder die Funktion $f = (f_1,\ldots,f_{n+1})^T : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ durch

$$f(x,\lambda) := \begin{pmatrix} P(\lambda)x \\ {e_s}^T x - 1 \end{pmatrix}, \qquad x \in \mathbb{R}^n,\ \lambda \in \mathbb{R}, \tag{10.2}$$

definiert, deren Nullstellen (x^*,λ^*) genau die Eigenpaare von $P(\lambda)$ mit $(x^*)_s = 1$ sind.
$(\tilde{x},\tilde{\lambda})$ sei eine Näherung eines Eigenpaares (x^*,λ^*) mit $(x^*)_s = 1$ und

$$(\Delta x, \Delta\lambda) := (x - \tilde{x}, \lambda - \tilde{\lambda}).$$

Die Taylor–Entwicklung von $f(x,\lambda)$ zur Entwicklungsstelle $(\tilde{x},\tilde{\lambda})$ lautet dann

$$\begin{aligned}
f(x,\lambda) &= f(\tilde{x},\tilde{\lambda}) + f'(\tilde{x},\tilde{\lambda})\begin{pmatrix}\Delta x\\ \Delta\lambda\end{pmatrix} + \frac{1}{2}f''(\tilde{x},\tilde{\lambda})\begin{pmatrix}\Delta x\\ \Delta\lambda\end{pmatrix}^2 + \ldots\\
&\quad + \frac{1}{(l+1)!}f^{(l+1)}(\tilde{x},\tilde{\lambda})\begin{pmatrix}\Delta x\\ \Delta\lambda\end{pmatrix}^{l+1},
\end{aligned} \tag{10.3}$$

wobei

$$f^{(t)}(\tilde{x},\tilde{\lambda}) := \left(\frac{\partial^t f_i}{\partial x_{i_t}\dots\partial x_{i_1}}(\tilde{x},\tilde{\lambda})\right)_{i,i_1,\dots,i_t=1,\dots,n+1} \in \mathbb{R}^{\overbrace{(n+1)\times\dots\times(n+1)}^{(t+1)\text{–mal}}}$$

(mit $x=(x_i)_{i=1,\dots,n}$ und $x_{n+1}=\lambda$) die t–te Fréchet–Ableitung des Operators f im Punkt $(\tilde{x},\tilde{\lambda})$ ist (siehe [21]). (10.3) ist ein System vom Grad $l+1$ der Dimension $n+1$ in $(\Delta x^T,\Delta\lambda)^T$. Es wird im folgenden Lemma u.a. gezeigt, dass $f^{(t)}(x,\lambda)=O$ für $t\geq l+2$ gilt. Deswegen ist die Taylor-Entwicklung (10.3) endlich.

Lemma 22

Für die Fréchet–Ableitungen von (10.2) *gilt*

$$f'(x,\lambda) = \begin{pmatrix} P(\lambda) & P'(\lambda)x \\ e_s^T & 0 \end{pmatrix}, \tag{10.4}$$

$$(f^{(t)}(x,\lambda))_{i,i_1,\dots,i_t} = \begin{cases} (P^{(t-1)}(\lambda))_{ii_1}\,, & i,i_1=1,\dots,n;\ i_2=i_3=\dots=i_t=n+1 \\ (P^{(t-1)}(\lambda))_{ii_2}\,, & i,i_2=1,\dots,n;\ i_1=i_3=\dots=i_t=n+1 \\ \quad\vdots & \\ (P^{(t-1)}(\lambda))_{ii_t}\,, & i,i_t=1,\dots,n;\ i_1=i_2=\dots=i_{t-1}=n+1 \\ (P^{(t)}(\lambda)x)_i\,, & i=1,\dots,n;\ i_1=i_2=\dots=i_t=n+1 \\ 0\,, & \textit{sonst} \end{cases} \tag{10.5}$$

für $2\leq t\leq l+1$ *und* $f^{(t)}(x,\lambda)=O$ *für* $t\geq l+2$.

Beweis:

Wegen

$$(f(x,\lambda))_i = \begin{cases} \sum_{r=1}^n (P(\lambda))_{ir}x_r\,, & i=1,\dots,n \\ x_s-1\,, & i=n+1 \end{cases}$$

gilt für die erste Fréchet–Ableitung von f nach Definition

$$(f'(x,\lambda))_{ij} = \begin{cases} (P(\lambda))_{ij}\,, & i,j=1,\dots,n \\ \sum_{r=1}^n (P'(\lambda))_{ir}x_r = (P'(\lambda)x)_i\,, & i=1,\dots,n;\ j=n+1 \\ 1\,, & i=n+1;\ j=s \\ 0\,, & \text{sonst} \end{cases},$$

es gilt also (10.4). Wenn man davon ausgehend die zweite Fréchet–Ableitung von f bildet, erhält man

$$(f''(x,\lambda))_{ijk} = \begin{cases} (P'(\lambda))_{ij}\,, & i,j=1,\dots,n;\ k=n+1 \\ (P'(\lambda))_{ik}\,, & i=1,\dots,n;\ j=n+1;\ k=1,\dots,n \\ \sum_{r=1}^n (P''(\lambda))_{ir}x_r = (P''(\lambda)x)_i\,, & i=1,\dots,n;\ j=k=n+1 \\ 0\,, & \text{sonst} \end{cases},$$

d.h. (10.5) gilt für $t = 2$. (10.5) für $t > 2$ zeigt man dann ganz analog per vollständiger Induktion. Wegen $P^{(t-1)}(\lambda) = O$ für $t - 1 \geq l + 1$ gilt dann $f^{(t)}(x, \lambda) = O$ für $t \geq l + 2$. $\square$

Man kann aber auch die Taylor–Entwicklung von $P(\lambda)$ zur Entwicklungsstelle $\tilde{\lambda}$

$$P(\lambda) = P(\tilde{\lambda} + \Delta\lambda) = P(\tilde{\lambda}) + \Delta\lambda P'(\tilde{\lambda}) + \frac{1}{2}(\Delta\lambda)^2 P''(\tilde{\lambda}) + \ldots + \frac{1}{l!}(\Delta\lambda)^l P^{(l)}(\tilde{\lambda})$$

bilden ($P^{(t)}(\lambda) = O$ für $t \geq l + 1$) und erhält damit

$$\begin{aligned}
&f(x, \lambda) \\
&\quad = f(\tilde{x} + \Delta x, \tilde{\lambda} + \Delta\lambda) = \begin{pmatrix} P(\tilde{\lambda} + \Delta\lambda)(\tilde{x} + \Delta x) \\ {e_s}^T(\tilde{x} + \Delta x) - 1 \end{pmatrix} \\
&\quad = \begin{pmatrix} \big(P(\tilde{\lambda}) + \Delta\lambda P'(\tilde{\lambda}) + \ldots + \frac{1}{l!}(\Delta\lambda)^l P^{(l)}(\tilde{\lambda})\big)(\tilde{x} + \Delta x) \\ {e_s}^T\tilde{x} - 1 + {e_s}^T\Delta x \end{pmatrix} \\
&\quad = f(\tilde{x}, \tilde{\lambda}) + \begin{pmatrix} P(\tilde{\lambda}) & \big(P'(\tilde{\lambda}) + \ldots + \frac{1}{l!}(\Delta\lambda)^{l-1} P^{(l)}(\tilde{\lambda})\big)(\tilde{x} + \Delta x) \\ e_s^T & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} \\
&\quad = f(\tilde{x}, \tilde{\lambda}) + \left\{ \begin{pmatrix} P(\tilde{\lambda}) & P'(\tilde{\lambda})\tilde{x} \\ e_s^T & 0 \end{pmatrix} + \right. \\
&\qquad \left. \begin{pmatrix} O & P'(\tilde{\lambda})\Delta x + \ldots + \frac{1}{l!}(\Delta\lambda)^{l-1} P^{(l)}(\tilde{\lambda})\tilde{x} + \frac{1}{l!}(\Delta\lambda)^{l-1} P^{(l)}(\tilde{\lambda})\Delta x \\ 0^T & 0 \end{pmatrix} \right\} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} \\
&\quad = f(\tilde{x}, \tilde{\lambda}) + \Bigg\{ f'(\tilde{x}, \tilde{\lambda}) + \\
&\begin{pmatrix} O & \displaystyle\sum_{t=2}^{l} \Big\{ \tfrac{1}{(t-1)!}(P^{(t-1)}(\tilde{\lambda})\Delta x)(\Delta\lambda)^{t-2} + \tfrac{1}{t!}(P^{(t)}(\tilde{\lambda})\tilde{x})(\Delta\lambda)^{t-1} \Big\} + \tfrac{1}{l!}(P^{(l)}(\tilde{\lambda})\Delta x)(\Delta\lambda)^{l-1} \\ 0^T & 0 \end{pmatrix} \Bigg\} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix}.
\end{aligned} \tag{10.6}$$

Die rechte Seite von (10.6) wird mit einer geeigneten Matrix $-C = -(c_{ij}) \in \mathbb{R}^{(n+1)\times(n+1)}$ multipliziert und anschließend $(\Delta x^T, \Delta\lambda)^T$ addiert. So erhält man für

$$C^* := (c_{ij})_{\substack{i = 1, \ldots, n+1 \\ j = 1, \ldots, n}}, \quad \text{und} \quad Q^{(t)}(\tilde{\lambda}) := C^* P^{(t)}(\tilde{\lambda}) \in \mathbb{R}^{(n+1)\times n}$$

die Funktion $g : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$

$$\begin{aligned}
&-Cf(\tilde{x}, \tilde{\lambda}) + \Big\{ I_{n+1} - Cf'(\tilde{x}, \tilde{\lambda}) \\
&-C \begin{pmatrix} O & \displaystyle\sum_{t=2}^{l} \Big\{ \tfrac{1}{(t-1)!}(P^{(t-1)}(\tilde{\lambda})\Delta x)(\Delta\lambda)^{t-2} + \tfrac{1}{t!}(P^{(t)}(\tilde{\lambda})\tilde{x})(\Delta\lambda)^{t-1} \Big\} + \tfrac{1}{l!}(P^{(l)}(\tilde{\lambda})\Delta x)(\Delta\lambda)^{l-1} \\ 0^T & 0 \end{pmatrix} \Bigg\} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} \\
&\quad = -Cf(\tilde{x}, \tilde{\lambda}) + \Big\{ I_{n+1} - Cf'(\tilde{x}, \tilde{\lambda}) \\
&\quad -\left(O \ \displaystyle\sum_{t=2}^{l} \Big\{ \tfrac{1}{(t-1)!}(Q^{(t-1)}(\tilde{\lambda})\Delta x)(\Delta\lambda)^{t-2} + \tfrac{1}{t!}(Q^{(t)}(\tilde{\lambda})\tilde{x})(\Delta\lambda)^{t-1} \Big\} + \tfrac{1}{l!}(Q^{(l)}(\tilde{\lambda})\Delta x)(\Delta\lambda)^{l-1} \right) \Bigg\} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} \\
&\quad =: \ g(\Delta x, \Delta\lambda).
\end{aligned}$$

Für eine reguläre Matrix C ist $(\Delta x^*, \Delta\lambda^*) = (x^* - \tilde{x}, \lambda^* - \tilde{\lambda})$ genau dann ein Fixpunkt von g, wenn (x^*, λ^*) ein Eigenpaar von $P(\lambda)$ mit $(x^*)_s = 1$ ist.

Andererseits kann man auch (10.3) mit $-C$ multiplizieren und $(\Delta x^T, \Delta\lambda)^T$ dazu addieren. Dies liefert die Funktion

$$\begin{aligned} h(\Delta x, \Delta\lambda) &= -Cf(\tilde{x}, \tilde{\lambda}) + \{I_{n+1} - Cf'(\tilde{x}, \tilde{\lambda})\} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} - \frac{1}{2}(Cf''(\tilde{x}, \tilde{\lambda})) \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix}^2 - \dots \\ &\quad - \frac{1}{(l+1)!}(Cf^{(l+1)}(\tilde{x}, \tilde{\lambda})) \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix}^{l+1} \end{aligned} \tag{10.7}$$

mit

$$C \cdot f^{(t)}(\tilde{x}, \tilde{\lambda}) \in \mathbb{R}^{\overbrace{(n+1) \times \dots \times (n+1)}^{(t+1)\text{–mal}}}$$

für $2 \le t \le l+1$ (siehe Definition 10).

Es gilt im restlichen Kapitel die Festlegung

$$[\Delta\lambda]^t := \prod_{j=1}^{t} [\Delta\lambda] \quad \text{für } [\Delta\lambda] \in \mathbf{IR} \text{ und } 2 \le t \le l.$$

Lemma 23

Für alle $([\Delta x]^T, [\Delta\lambda])^T \in \mathbf{IR}^{n+1}$ *gilt*

$$g([\Delta x], [\Delta\lambda]) \subseteq h([\Delta x], [\Delta\lambda])\,.$$

Beweis:

Wegen der Subdistributivität in **IR** gilt

$$\begin{aligned} g([\Delta x], [\Delta\lambda]) \subseteq\ & -Cf(\tilde{x}, \tilde{\lambda}) + (I_{n+1} - Cf'(\tilde{x}, \tilde{\lambda})) \begin{pmatrix} [\Delta x] \\ [\Delta\lambda] \end{pmatrix} \\ & - \sum_{t=2}^{l} \Big\{ \frac{1}{(t-1)!}(Q^{(t-1)}(\tilde{\lambda})[\Delta x])[\Delta\lambda]^{t-2} + \frac{1}{t!}(Q^{(t)}(\tilde{\lambda})\tilde{x})[\Delta\lambda]^{t-1} \Big\} [\Delta\lambda] \\ & - \frac{1}{l!}(Q^{(l)}(\tilde{\lambda})[\Delta x])[\Delta\lambda]^l\,. \end{aligned}$$

Es bleibt zu zeigen, dass für $2 \le t \le l+1$

$$\Big\{ \frac{1}{(t-1)!}(Q^{(t-1)}(\tilde{\lambda})[\Delta x])[\Delta\lambda]^{t-2} + \frac{1}{t!}(Q^{(t)}(\tilde{\lambda})\tilde{x})[\Delta\lambda]^{t-1} \Big\} [\Delta\lambda] \subseteq \frac{1}{t!}(Cf^{(t)}(\tilde{x}, \tilde{\lambda})) \begin{pmatrix} [\Delta x] \\ [\Delta\lambda] \end{pmatrix}^t$$

gilt (wegen $Q^{(l+1)}(\lambda) = C^* P^{(l+1)}(\lambda) = O$).

Wir haben

$$\begin{aligned}
&\Big\{\tfrac{1}{(t-1)!}(Q^{(t-1)}(\tilde\lambda)[\Delta x])[\Delta\lambda]^{t-2} + \tfrac{1}{t!}(Q^{(t)}(\tilde\lambda)\tilde x)[\Delta\lambda]^{t-1}\Big\}[\Delta\lambda] \\
&\quad\subseteq \frac{1}{(t-1)!}(Q^{(t-1)}(\tilde\lambda)[\Delta x])[\Delta\lambda]^{t-1} + \frac{1}{t!}(Q^{(t)}(\tilde\lambda)\tilde x)[\Delta\lambda]^{t} \\
&\quad= \frac{1}{(t-1)!}\big((C^*P^{(t-1)}(\tilde\lambda))[\Delta x]\big)[\Delta\lambda]^{t-1} + \frac{1}{t!}(C^*P^{(t)}(\tilde\lambda)\tilde x)[\Delta\lambda]^{t} \\
&\quad= \frac{1}{(t-1)!}\Big(\sum_{j=1}^{n}(C^*P^{(t-1)}(\tilde\lambda))_{ij}[\Delta x]_j\Big)_{i=1,\dots,n+1}[\Delta\lambda]^{t-1} + \frac{1}{t!}\Big((C^*P^{(t)}(\tilde\lambda)\tilde x)_i[\Delta\lambda]^{t}\Big)_{i=1,\dots,n+1} \\
&\quad\subseteq \frac{1}{(t-1)!}\Big(\sum_{j=1}^{n}(C^*P^{(t-1)}(\tilde\lambda))_{ij}[\Delta x]_j[\Delta\lambda]^{t-1}\Big)_{i=1,\dots,n+1} + \frac{1}{t!}\Big((C^*P^{(t)}(\tilde\lambda)\tilde x)_i[\Delta\lambda]^{t}\Big)_{i=1,\dots,n+1} \\
&\quad= \frac{1}{t!}\Big(t\sum_{j=1}^{n}(C^*P^{(t-1)}(\tilde\lambda))_{ij}[\Delta x]_j[\Delta\lambda]^{t-1} + (C^*P^{(t)}(\tilde\lambda)\tilde x)_i[\Delta\lambda]^{t}\Big)_{i=1,\dots,n+1}. \qquad (10.8)
\end{aligned}$$

Andererseits gilt aufgrund von Lemma 22 für $2 \le t \le l+1$

$$\begin{aligned}
(Cf^{(t)}(\tilde x,\tilde\lambda))_{i,i_1,\dots,i_t} &= \sum_{j=1}^{n+1} c_{ij}(f^{(t)}(\tilde x,\tilde\lambda))_{j,i_1,\dots,i_t} \\
&= \begin{cases}
(C^*P^{(t-1)}(\tilde\lambda))_{ii_1}, & i=1,\dots,n+1;\ i_1=1,\dots,n;\\ & i_2=i_3=\dots=i_t=n+1\\
\vdots & \\
(C^*P^{(t-1)}(\tilde\lambda))_{ii_t}, & i=1,\dots,n+1;\ i_t=1,\dots,n;\\ & i_1=i_2=\dots=i_{t-1}=n+1\\
(C^*P^{(t)}(\tilde\lambda)\tilde x)_i, & i=1,\dots,n+1;\\ & i_1=i_2=\dots=i_t=n+1\\
0, & \text{sonst}.
\end{cases}
\end{aligned}$$

Dies liefert nach Definition

$$\begin{aligned}
&\frac{1}{t!}(Cf^{(t)}(\tilde x,\tilde\lambda))\begin{pmatrix}[\Delta x]\\ [\Delta\lambda]\end{pmatrix}^t \\
&\quad= \frac{1}{t!}\Big(\sum_{i_1=1}^{n}(C^*P^{(t-1)}(\tilde\lambda))_{ii_1}[\Delta\lambda]^{t-1}[\Delta x]_{i_1} + \dots + \sum_{i_t=1}^{n}(C^*P^{(t-1)}(\tilde\lambda))_{ii_t}[\Delta x]_{i_t}[\Delta\lambda]^{t-1} \\
&\qquad +(C^*P^{(t)}(\tilde\lambda)\tilde x)_i[\Delta\lambda]^{t}\Big)_{i=1,\dots,n+1} \\
&\quad= \frac{1}{t!}\Big(\sum_{i_1=1}^{n}(C^*P^{(t-1)}(\tilde\lambda))_{ii_1}[\Delta x]_{i_1}[\Delta\lambda]^{t-1} + \dots + \sum_{i_t=1}^{n}(C^*P^{(t-1)}(\tilde\lambda))_{ii_t}[\Delta x]_{i_t}[\Delta\lambda]^{t-1} \\
&\qquad +(C^*P^{(t)}(\tilde\lambda)\tilde x)_i[\Delta\lambda]^{t}\Big)_{i=1,\dots,n+1} \\
&\quad= \frac{1}{t!}\Big(t\sum_{j=1}^{n}(C^*P^{(t-1)}(\tilde\lambda))_{ij}[\Delta x]_j[\Delta\lambda]^{t-1} + (C^*P^{(t)}(\tilde\lambda)\tilde x)_i[\Delta\lambda]^{t}\Big)_{i=1,\dots,n+1}. \qquad (10.9)
\end{aligned}$$

Aus (10.8) und (10.9) folgt die Behauptung. □

Lemma 24

Für alle $2 \leq t \leq l+1$ *gilt*

$$\frac{1}{t!}||Cf^{(t)}(\tilde{x},\tilde{\lambda})||_\infty \leq ||C||_\infty \left(\frac{1}{(t-1)!}\,(P^{(t-1)}(\tilde{\lambda}))_\infty + \frac{1}{t!}(P^{(t)}(\tilde{\lambda}))_\infty ||\tilde{x}||_\infty \right).$$

Dabei gilt die Bezeichnung

$$(\sum_{i=0}^{r} B_i \tilde{\lambda}^i)_\infty := \sum_{i=0}^{r} ||B_i||_\infty |\tilde{\lambda}|^i$$

$(B_0 \ldots, B_r \in \mathbb{R}^{n\times n})$.

Beweis:
Nach Lemma 21 gilt

$$\frac{1}{t!}||Cf^{(t)}(\tilde{x},\tilde{\lambda})||_\infty \leq \frac{1}{t!}||C||_\infty \cdot ||f^{(t)}(\tilde{x},\tilde{\lambda})||_\infty .$$

Es bleibt also zu zeigen, dass

$$||f^{(t)}(\tilde{x},\tilde{\lambda})||_\infty \leq t\,(P^{(t-1)}(\tilde{\lambda}))_\infty + (P^{(t)}(\tilde{\lambda}))_\infty ||\tilde{x}||_\infty$$

ist. Nach Definition 10 und Lemma 22 gilt

$$\begin{aligned}
||f^{(t)}(\tilde{x},\tilde{\lambda})||_\infty &= \max_{i=1,\ldots,n+1} \left\{ \sum_{i_1=1}^{n+1} \cdots \sum_{i_t=1}^{n+1} |f^{(t)}(\tilde{x},\tilde{\lambda})_{i,i_1,\ldots,i_t}| \right\} \\
&= \sum_{i_1=1}^{n+1} \cdots \sum_{i_t=1}^{n+1} |f^{(t)}(\tilde{x},\tilde{\lambda})_{i^*,i_1,\ldots,i_t}| \\
&= t\sum_{j=1}^{n} |P^{(t-1)}(\tilde{\lambda})_{i^*j}| + |(P^{(t)}(\tilde{\lambda})\tilde{x})_{i^*}| \\
&= t\sum_{j=1}^{n} |P^{(t-1)}(\tilde{\lambda})_{i^*j}| + \left| \sum_{j=1}^{n} P^{(t)}(\tilde{\lambda})_{i^*j}\tilde{x}_j \right| \\
&\leq t\sum_{j=1}^{n} |P^{(t-1)}(\tilde{\lambda})_{i^*j}| + \sum_{j=1}^{n} |P^{(t)}(\tilde{\lambda})_{i^*j}|\,|\tilde{x}_j| \\
&\leq t\sum_{j=1}^{n} |P^{(t-1)}(\tilde{\lambda})_{i^*j}| + ||\tilde{x}||_\infty \sum_{j=1}^{n} |P^{(t)}(\tilde{\lambda})_{i^*j}| ,
\end{aligned}$$

wobei $i^* \in \{1,\ldots,n+1\}$ geeignet gewählt ist.

Da für ein beliebiges reelles $n \times n$–Matrixpolynom $\hat{P}(\lambda) = B_r\lambda^r + \ldots + B_1\lambda + B_0$ vom Grad r

$$\begin{aligned}
\sum_{j=1}^{n} |\hat{P}(\tilde{\lambda})_{i^*j}| &= \sum_{j=1}^{n} |(B_r)_{i^*j}\tilde{\lambda}^r + \ldots + (B_1)_{i^*j}\tilde{\lambda} + (B_0)_{i^*j}| \\
&\leq \sum_{j=1}^{n} \left(|(B_r)_{i^*j}|\,|\tilde{\lambda}|^r + \ldots + |(B_1)_{i^*j}|\,|\tilde{\lambda}| + |(B_0)_{i^*j}| \right) \\
&= |\tilde{\lambda}|^r \sum_{j=1}^{n} |(B_r)_{i^*j}| + \ldots + |\tilde{\lambda}| \sum_{j=1}^{n} |(B_1)_{i^*j}| + \sum_{j=1}^{n} |(B_0)_{i^*j}| \\
&\leq |\tilde{\lambda}|^r\, ||B_r||_\infty + \ldots + |\tilde{\lambda}|\, ||B_1||_\infty + ||B_0||_\infty \\
&= (\hat{P}(\tilde{\lambda}))_\infty
\end{aligned}$$

gilt, ist die Aussage des Lemmas gezeigt. □

Satz 26

Es sei $P(\lambda) = A_l\lambda^l + \ldots + A_1\lambda + A_0$ *wie in* (10.1), $\tilde{\lambda} \in \mathbb{R}$, $\tilde{x} \in \mathbb{R}^n$ *und* $C \in \mathbb{R}^{(n+1)\times(n+1)}$ *regulär.*
Für die wie folgt definierten Ausdrücke

$$\varphi := ||Cf(\tilde{x}, \tilde{\lambda})||_\infty\,, \quad \sigma := ||I_{n+1} - Cf'(\tilde{x}, \tilde{\lambda})||_\infty\,,$$

$$\tau_t := ||C||_\infty \left(\tfrac{1}{(t-1)!}\,(P^{(t-1)}(\tilde{\lambda}))_\infty + \tfrac{1}{t!}(P^{(t)}(\tilde{\lambda}))_\infty ||\tilde{x}||_\infty \right), \quad 2 \leq t \leq l+1,$$

besitze das skalare Polynom $(l+1)$*–ten Grades*

$$p(\beta) := \varphi + (\sigma - 1)\beta + \sum_{t=2}^{l+1} \tau_t \beta^t \tag{10.10}$$

zwei reelle Nullstellen $0 < \beta^- < \beta^+$ *mit*

$$p(\beta) < 0 \quad \textit{für } \beta \in (\beta^-, \beta^+)\,. \tag{10.11}$$

Weiter sei

$$g(\Delta x, \Delta\lambda) := -Cf(\tilde{x}, \tilde{\lambda}) + \Bigg\{ I_{n+1} - Cf'(\tilde{x}, \tilde{\lambda})$$

$$-\left(O \;\sum_{t=2}^{l} \left\{ \tfrac{1}{(t-1)!}(Q^{(t-1)}(\tilde{\lambda})\Delta x)(\Delta\lambda)^{t-2} + \tfrac{1}{t!}(Q^{(t)}(\tilde{\lambda})\tilde{x})(\Delta\lambda)^{t-1} \right\} + \tfrac{1}{l!}(Q^{(l)}(\tilde{\lambda})\Delta x)(\Delta\lambda)^{l-1} \right) \Bigg\} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix}$$

mit

$$Q^{(t)}(\tilde{\lambda}) := C^* P^{(t)}(\tilde{\lambda})$$

und

$$\begin{pmatrix} [\Delta x]^{(0)} \\ [\Delta\lambda]^{(0)} \end{pmatrix} := [-\beta, \beta]e \in \mathbf{IR}^{n+1}$$

für beliebiges $\beta \in (\beta^-, \beta^+)$.
Dann konvergiert die Iteriertenfolge

$$\begin{pmatrix} [\Delta x]^{(k+1)} \\ [\Delta\lambda]^{(k+1)} \end{pmatrix} := g([\Delta x]^{(k)}, [\Delta\lambda]^{(k)}), \quad k \in \mathbb{N}_0, \tag{10.12}$$

gegen einen Intervallvektor $\begin{pmatrix} [\Delta x]^* \\ [\Delta\lambda]^* \end{pmatrix}$ *mit*

$$\begin{pmatrix} [\Delta x]^* \\ [\Delta\lambda]^* \end{pmatrix} \subseteq \begin{pmatrix} [\Delta x]^{(k+1)} \\ [\Delta\lambda]^{(k+1)} \end{pmatrix} \subseteq \begin{pmatrix} [\Delta x]^{(k)} \\ [\Delta\lambda]^{(k)} \end{pmatrix} \subseteq \ldots \subseteq \begin{pmatrix} [\Delta x]^{(0)} \\ [\Delta\lambda]^{(0)} \end{pmatrix}, \quad k \in \mathbb{N},$$

und es existiert mindestens ein reelles Eigenpaar (x^*, λ^*) *von* $P(\lambda)$ *mit* $(x^*)_s = 1$ *und*

$$\begin{pmatrix} x^* \\ \lambda^* \end{pmatrix} \in \begin{pmatrix} \tilde{x} \\ \tilde{\lambda} \end{pmatrix} + \begin{pmatrix} [\Delta x]^{(k)} \\ [\Delta\lambda]^{(k)} \end{pmatrix}, \quad k \in \mathbb{N}_0 . \tag{10.13}$$

Sei des Weiteren β^* *eine Nullstelle von* $p'(\beta)$ *mit* $\beta^- < \beta^* < \beta^+$ *und*

$$p'(\beta) = (\sigma - 1) + \sum_{t=2}^{l+1} t\tau_t \beta^{t-1} < 0 \quad \textit{für} \quad \beta \in (\beta^-, \beta^*) . \tag{10.14}$$

Unter der Einschränkung $\beta \in (\beta^-, \beta^*)$ *ist das Eigenpaar* (x^*, λ^*) *mit* $(x^*)_s = 1$ *und* (10.13) *eindeutig, und die Iterierten* $\begin{pmatrix} [\Delta x]^{(k)} \\ [\Delta\lambda]^{(k)} \end{pmatrix}$ *konvergieren gegen dessen Approximationsfehler* $\begin{pmatrix} \Delta x^* \\ \Delta\lambda^* \end{pmatrix}$.

Beweis:
i)

Die Funktion h sei definiert wie in (10.7). Außerdem gelten die Bezeichnungen $m_0 := -Cf(\tilde{x}, \tilde{\lambda})$,

$$M_1 = (m_{ij}^{(1)}) := I_{n+1} - Cf'(\tilde{x}, \tilde{\lambda}), \qquad M_t = (m_{i,i_1,\ldots,i_t}^{(t)}) := -\frac{1}{t!} Cf^{(t)}(\tilde{x}, \tilde{\lambda}) \quad (2 \le t \le l+1) .$$

Es wird gleich gezeigt, dass

$$\begin{pmatrix} [\Delta x] \\ [\Delta\lambda] \end{pmatrix} := \begin{pmatrix} [\Delta x]^{(0)} \\ [\Delta\lambda]^{(0)} \end{pmatrix} = [-\beta, \beta]e$$

für $\beta > 0$ die Inklusion

$$\begin{aligned} h([\Delta x], [\Delta\lambda]) &= m_0 + M_1 \begin{pmatrix} [\Delta x] \\ [\Delta\lambda] \end{pmatrix} + M_2 \begin{pmatrix} [\Delta x] \\ [\Delta\lambda] \end{pmatrix}^2 + \ldots + M_{l+1} \begin{pmatrix} [\Delta x] \\ [\Delta\lambda] \end{pmatrix}^{l+1} \\ &\subseteq \operatorname{int} \begin{pmatrix} [\Delta x] \\ [\Delta\lambda] \end{pmatrix} \end{aligned} \tag{10.15}$$

erfüllt, wenn $\beta \in (\beta^-, \beta^+)$. Dann gilt nach Lemma 23 auch

$$g([\Delta x], [\Delta\lambda]) \subseteq \operatorname{int} \begin{pmatrix} [\Delta x] \\ [\Delta\lambda] \end{pmatrix},$$

d.h. $\begin{pmatrix}[\Delta x]^{(1)}\\ [\Delta\lambda]^{(1)}\end{pmatrix} \subseteq \begin{pmatrix}[\Delta x]^{(0)}\\ [\Delta\lambda]^{(0)}\end{pmatrix}$. Die Existenz mindestens eines Fixpunktes $\begin{pmatrix}\Delta x^*\\ \Delta\lambda^*\end{pmatrix} \in \begin{pmatrix}[\Delta x]\\ [\Delta\lambda]\end{pmatrix}$ von g ist dann garantiert durch den Brouwerschen Fixpunktsatz (Satz 19), da trivialerweise $g(x,\lambda) \in g([\Delta x],[\Delta\lambda])\ \forall\,(x,\lambda) \in ([\Delta x],[\Delta\lambda]) \in \mathbf{IR}^n \times \mathbf{IR}$ gilt.

Da der Funktionsausdruck von g intervallmäßig auswertbar ist, liefert uns Satz 17 die Konvergenz der Iteriertenfolge $\begin{pmatrix}[\Delta x]^{(k)}\\ [\Delta\lambda]^{(k)}\end{pmatrix}_{k\geq 0}$ gegen einen Grenzwert $\begin{pmatrix}[\Delta x]^*\\ [\Delta\lambda]^*\end{pmatrix} \in \mathbf{IR}^{n+1}$ und $\begin{pmatrix}\Delta x^*\\ \Delta\lambda^*\end{pmatrix} \in \begin{pmatrix}[\Delta x]^{(k)}\\ [\Delta\lambda]^{(k)}\end{pmatrix}$, $k \in \mathbb{N}_0$. Wegen der Regularität von C ist dann (x^*,λ^*) ein Eigenpaar von $P(\lambda)$ mit $(x^*)_s = 1$ und (10.13), d.h. der erste Teil des Satzes ist gezeigt.

Zum Beweis, dass (10.15) für $\beta \in (\beta^-,\beta^+)$ gilt: die Inklusion (10.15) ist äquivalent zu

$$\begin{aligned}
\operatorname{int}([-\beta,\beta]e) \quad &\supseteq \quad m_0 + \left(\sum_{j=1}^{n+1} m_{ij}^{(1)}[-\beta,\beta]\right) + \left(\sum_{j=1}^{n+1}\sum_{k=1}^{n+1} m_{ijk}^{(2)}[-\beta,\beta][-\beta,\beta]\right) + \ldots \\
&\quad + \left(\sum_{i_1=1}^{n+1}\sum_{i_2=1}^{n+1}\cdots\sum_{i_{l+1}=1}^{n+1} m_{i,i_1,i_2,\ldots,i_{l+1}}^{(l+1)}[-\beta,\beta]^{l+1}\right) \\
&\overset{\text{L. 8b)}}{=} \quad m_0 + \left(\sum_{j=1}^{n+1} |m_{ij}^{(1)}|[-\beta,\beta]\right) + \left(\sum_{j=1}^{n+1}\sum_{k=1}^{n+1} |m_{ijk}^{(2)}|[-\beta^2,\beta^2]\right) + \ldots \\
&\quad + \left(\sum_{i_1=1}^{n+1}\sum_{i_2=1}^{n+1}\cdots\sum_{i_{l+1}=1}^{n+1} |m_{i,i_1,i_2,\ldots,i_{l+1}}^{(l+1)}|[-\beta^{l+1},\beta^{l+1}]\right) \\
&\overset{\text{L. 8a) \& Ind.}}{=} \quad m_0 + \left([-\beta,\beta]\sum_{j=1}^{n+1} |m_{ij}^{(1)}|\right) + \left([-\beta^2,\beta^2]\sum_{j=1}^{n+1}\sum_{k=1}^{n+1} |m_{ijk}^{(2)}|\right) + \ldots \\
&\quad + \left([-\beta^{l+1},\beta^{l+1}]\sum_{i_1=1}^{n+1}\sum_{i_2=1}^{n+1}\cdots\sum_{i_{l+1}=1}^{n+1} |m_{i,i_1,i_2,\ldots,i_{l+1}}^{(l+1)}|\right) \\
&= \quad m_0 + [-\beta,\beta]|M_1|e + [-\beta^2,\beta^2]|M_2|e^2 + \ldots + [-\beta^{l+1},\beta^{l+1}]|M_{l+1}|e^{l+1}
\end{aligned} \tag{10.16}$$

mit $|M_t| := (|m_{i,i_1,\ldots,i_t}^{(t)}|)$, $2 \leq t \leq l+1$.

Lemma 9 b) angewandt auf jede Komponente von (10.16) liefert, dass (10.16) äquivalent ist zu

$$|m_0| + \beta|M_1|e + \beta^2|M_2|e^2 + \ldots + \beta^{l+1}|M_{l+1}|e^{l+1} < \beta e\,. \tag{10.17}$$

Die Ungleichung (10.17) ist sicherlich erfüllt, wenn

$$||m_0||_\infty + \beta||M_1||_\infty + \beta^2||M_2||_\infty + \ldots + \beta^{l+1}||M_{l+1}||_\infty < \beta\,,$$

d.h. wenn

$$||m_0||_\infty + \beta(||M_1||_\infty - 1) + \beta^2||M_2||_\infty + \ldots + \beta^{l+1}||M_{l+1}||_\infty < 0\,. \tag{10.18}$$

Es gilt $||m_0||_\infty = \varphi$ und $||M_1||_\infty = \sigma$. Außerdem liefert Lemma 24 die Abschätzungen $||M_t||_\infty \leq \tau_t \quad (2 \leq t \leq l+1)$. Also ist

$$p(\beta) := \varphi + \beta(\sigma - 1) + \beta^2\tau_2 + \ldots + \beta^{l+1}\tau_{l+1} < 0 \tag{10.19}$$

für $\beta > 0$ eine für (10.18) und damit für (10.15) hinreichende Bedingung. Da (10.19) nach Voraussetzung für $\beta \in (\beta^-, \beta^+)$ erfüllt ist, ist der erste Teil des Satzes bewiesen.

ii)

Sei nun $\beta \in (\beta^-, \beta^*) \subseteq (\beta^-, \beta^+)$. Damit gelten die in i) hergeleiteten Aussagen für β.
$[y]^* := \begin{pmatrix} [\Delta x]^* \\ [\Delta\lambda]^* \end{pmatrix}$ sei der Grenzwert der Iterierten $\begin{pmatrix} [\Delta x]^{(k)} \\ [\Delta\lambda]^{(k)} \end{pmatrix}$ aus (10.12). Aus (10.12) und Lemma 23 folgt dann

$$\begin{aligned} \begin{pmatrix} [\Delta x]^{(k+1)} \\ [\Delta\lambda]^{(k+1)} \end{pmatrix} &= g([\Delta x]^{(k)}, [\Delta\lambda]^{(k)}) \subseteq h([\Delta x]^{(k)}, [\Delta\lambda]^{(k)}) \\ &= m_0 + M_1 \begin{pmatrix} [\Delta x]^{(k)} \\ [\Delta\lambda]^{(k)} \end{pmatrix} + M_2 \begin{pmatrix} [\Delta x]^{(k)} \\ [\Delta\lambda]^{(k)} \end{pmatrix}^2 + \ldots + M_{l+1} \begin{pmatrix} [\Delta x]^{(k)} \\ [\Delta\lambda]^{(k)} \end{pmatrix}^{l+1} \end{aligned}$$

für $k \geq 0$. Für $k \to \infty$ erhält man

$$[y]^* \subseteq m_0 + M_1[y]^* + M_2([y]^*)^2 + \ldots + M_{l+1}([y]^*)^{l+1}.$$

Für die Radien liefert dies mit Lemma 10 a), b) und d)

$$\begin{aligned} \mathrm{rad}([y]^*) &\leq \mathrm{rad}(m_0 + M_1[y]^* + M_2([y]^*)^2 + \ldots + M_{l+1}([y]^*)^{l+1}) \\ &= \mathrm{rad}(M_1[y]^*) + \mathrm{rad}(M_2([y]^*)^2) + \ldots + \mathrm{rad}(M_{l+1}([y]^*)^{l+1}) \\ &= \left(\sum_{j=1}^{n+1} |m_{ij}^{(1)}| \mathrm{rad}([y]_j^*)\right) + \left(\sum_{j=1}^{n+1}\sum_{k=1}^{n+1} |m_{ijk}^{(2)}| \mathrm{rad}([y]_k^*[y]_j^*)\right) + \ldots \\ &\quad + \left(\sum_{i_1=1}^{n+1}\sum_{i_2=1}^{n+1}\cdots\sum_{i_{l+1}=1}^{n+1} |m_{i,i_1,i_2,\ldots,i_{l+1}}^{(l+1)}| \,\mathrm{rad}([y]_{i_{l+1}}^* \ldots [y]_{i_2}^*[y]_{i_1}^*)\right). \end{aligned} \tag{10.20}$$

Es sei definiert

$$r_\infty := ||\mathrm{rad}([y]^*)||_\infty = \max_{i=1}^{n+1} \mathrm{rad}([y]_i^*)\,.$$

Weil $[y]^* \subseteq \begin{pmatrix} [\Delta x]^{(0)} \\ [\Delta\lambda]^{(0)} \end{pmatrix} = [-\beta, \beta]e\,,$ gilt $|[y]_j^*| \leq \beta$ für $j = 1, \ldots, n+1$.

Es wird nun per vollständiger Induktion gezeigt, dass für $t \in \mathbb{N}$ die Aussage

$$\mathrm{rad}([y]_{i_t}^* \ldots [y]_{i_2}^*[y]_{i_1}^*) \leq \beta^{t-1} t\, r_\infty \quad (\text{wobei } i_1, \ldots, i_t \in \{1, \ldots, n+1\}) \tag{10.21}$$

gilt. Für $t = 1$ gilt (10.21) wegen der Definition von r_∞. Angenommen (10.21) ist richtig für ein $t \in \mathbb{N}$. Wegen Lemma 10 c) und e) gilt dann

$$\begin{aligned} \mathrm{rad}([y]_{i_{t+1}}^*[y]_{i_t}^* \ldots [y]_{i_1}^*) &\leq \mathrm{rad}([y]_{i_{t+1}}^*)\,|[y]_{i_t}^* \ldots [y]_{i_1}^*| + |[y]_{i_{t+1}}^*|\,\mathrm{rad}([y]_{i_t}^* \ldots [y]_{i_1}^*) \\ &= \mathrm{rad}([y]_{i_{t+1}}^*)\,|[y]_{i_t}^*| \ldots |[y]_{i_1}^*| + |[y]_{i_{t+1}}^*|\,\mathrm{rad}([y]_{i_t}^* \ldots [y]_{i_1}^*) \\ &\leq r_\infty \beta^t + \beta\{\beta^{t-1} t\, r_\infty\} \\ &= (t+1) r_\infty \beta^t\,, \end{aligned}$$

d.h. (10.21) hat auch Gültigkeit für $t+1$. Damit ist (10.21) für alle $t \in \mathbb{N}$ gezeigt. Aus (10.20) erhält man dann

$$\begin{aligned} \mathrm{rad}([y]^*) \;\leq\; & r_\infty \left(\sum_{j=1}^{n+1} |m_{ij}^{(1)}| \right) + \beta 2 r_\infty \left(\sum_{j=1}^{n+1} \sum_{k=1}^{n+1} |m_{ijk}^{(2)}| \right) + \ldots \\ & + \beta^l (l+1) r_\infty \left(\sum_{i_1=1}^{n+1} \sum_{i_2=1}^{n+1} \ldots \sum_{i_{l+1}=1}^{n+1} |m_{i,i_1,i_2,\ldots,i_{l+1}}^{(l+1)}| \right). \end{aligned} \tag{10.22}$$

Sei nun $\mathrm{rad}([y]_{i'}^*) = r_\infty$. Es gilt dann mit (10.22)

$$\begin{aligned} r_\infty &= \mathrm{rad}([y]_{i'}^*) \\ &\leq r_\infty \sum_{j=1}^{n+1} |m_{i'j}^{(1)}| + \beta 2 r_\infty \sum_{j=1}^{n+1} \sum_{k=1}^{n+1} |m_{i'jk}^{(2)}| + \ldots + \beta^l (l+1) r_\infty \sum_{i_1=1}^{n+1} \sum_{i_2=1}^{n+1} \ldots \sum_{i_{l+1}=1}^{n+1} |m_{i',i_1,i_2,\ldots,i_{l+1}}^{(l+1)}| \\ &\leq r_\infty ||M_1||_\infty + \beta 2 r_\infty ||M_2||_\infty + \ldots + \beta^l (l+1) r_\infty ||M_{l+1}||_\infty \\ &\leq r_\infty \sigma + 2 r_\infty \tau_2 \beta + \ldots + (l+1) r_\infty \tau_{l+1} \beta^l \,. \end{aligned} \tag{10.23}$$

Angenommen $r_\infty > 0$. Dann ist die Aussage (10.23) äquivalent zu

$$1 \leq \sigma + 2\tau_2\beta + \ldots + (l+1)\tau_{l+1}\beta^l \quad \Longleftrightarrow \quad 0 \leq (\sigma - 1) + 2\tau_2\beta + \ldots + (l+1)\tau_{l+1}\beta^l = p'(\beta)\,,$$

was ein Widerspruch zur Voraussetzung (10.14) ist, da $\beta \in (\beta^-, \beta^*)$. Also gilt $r_\infty = 0$.

Da jeder Fixpunkt von g, der in $\begin{pmatrix} [\Delta x]^{(0)} \\ [\Delta\lambda]^{(0)} \end{pmatrix}$ enthalten ist, in $\begin{pmatrix} [\Delta x]^{(k)} \\ [\Delta\lambda]^{(k)} \end{pmatrix}$ bleibt (für $k \in \mathbb{N}$), liefert uns $r_\infty = 0$ die Eindeutigkeit des Fixpunkts $\begin{pmatrix} \Delta x^* \\ \Delta\lambda^* \end{pmatrix}$ von g in $\begin{pmatrix} [\Delta x]^{(0)} \\ [\Delta\lambda]^{(0)} \end{pmatrix}$.

Wegen $\begin{pmatrix} \Delta x^* \\ \Delta\lambda^* \end{pmatrix} \in [y]^*$ (nach i)) und $r_\infty = ||\mathrm{rad}([y]^*)||_\infty = 0$ gilt dann

$$\begin{pmatrix} \Delta x^* \\ \Delta\lambda^* \end{pmatrix} = [y]^* = \begin{pmatrix} [\Delta x]^* \\ [\Delta\lambda]^* \end{pmatrix} \in \mathbb{R}^{n+1}\,.$$

Damit ist auch der letzte Teil des Satzes bewiesen. □

Bemerkung 13

Die Voraussetzungen des Satzes 26 sind erfüllt, wenn alle folgenden Bedingungen gelten:

a) $(\tilde{x}, \tilde{\lambda})$ *ist eine hinreichend gute Näherung eines reellen Eigenpaares* (x^*, λ^*) *von* $P(\lambda)$ *mit* $(x^*)_s = 1$ *; o.B.d.A. sei* $(\tilde{x}, \tilde{\lambda})$ *kein Eigenpaar von* $P(\lambda)$ *.*

b) $f'(\tilde{x}, \tilde{\lambda})$ *ist regulär.*

c) $C := f'(\tilde{x}, \tilde{\lambda})^{-1}$ *oder* C *ist eine hinreichend gute Näherung von* $f'(\tilde{x}, \tilde{\lambda})^{-1}$ *.*

β^- und β^+ sind dann die beiden größten reellen Nullstellen von $p(\beta)$ und β^* die größte reelle Nullstelle von $p'(\beta)$.

Wenn a) gilt und der Eigenwert λ^* zusätzlich einfach ist, impliziert dies b).
Wenn a) gilt und λ^* nicht einfach ist, dann ist $f'(\tilde{x}, \tilde{\lambda})$ fast singulär.

Beweis:

Wenn b) erfüllt ist, kann $C := f'(\tilde{x}, \tilde{\lambda})^{-1}$ (oder zumindest eine hinreichend gute Näherung von $f'(\tilde{x}, \tilde{\lambda})^{-1}$) gewählt werden (siehe c)).

Damit ist $\sigma = ||I_{n+1} - Cf'(\tilde{x}, \tilde{\lambda})||_\infty = 0$ (oder zumindest $\sigma \approx 0$).

Wegen $P^{(l)}(\tilde{\lambda}) = l!A_l$ und $P^{(l+1)}(\tilde{\lambda}) = O$ gilt nach Definition $\tau_{l+1} = ||C||_\infty ||A_l||_\infty$. Wegen der Regularität von C und A_l ist dann also $\tau_{l+1} > 0$.

Da nach Annahme $(x^*, \lambda^*) \neq (\tilde{x}, \tilde{\lambda})$ gilt, ist $\varphi = ||Cf(\tilde{x}, \tilde{\lambda})||_\infty > 0$. Für die hinreichend gute Eigenpaarnäherung aus a) ist φ allerdings hinreichend klein.

Nun wollen wir das Polynom

$$p(\beta) = \varphi + (\sigma - 1)\beta + \sum_{t=2}^{l+1} \tau_t \beta^t$$

aus (10.10) heuristisch betrachten. Für $\varphi > 0$ hinreichend klein und $\beta' > 0$ sehr klein ist dann wegen $\sigma - 1 \approx -1$

$$p(\beta') = \varphi + \beta'\{(\sigma - 1) + \sum_{t=2}^{l+1} \tau_t (\beta')^{t-1}\} < 0 \,.$$

Wegen $\tau_{l+1} > 0$ gilt außerdem $\lim_{\beta \to \infty} p(\beta) = \infty$, d.h. p besitzt eine positive reelle Nullstelle $\beta^+ > \beta'$. Da die Vorzeichenregel von Descartes (Satz 18) besagt, dass p entweder zwei oder keine positive reelle Nullstelle besitzt ($p(\beta)$ hat genau zwei Vorzeichenwechsel bei seinen Koeffizienten: $\varphi > 0$, $\sigma - 1 < 0$, $\tau_t \geq 0$ $(2 \leq t \leq l)$, $\tau_{l+1} > 0$), muss p also genau zwei reelle positive Nullstellen $0 < \beta^- \leq \beta^+$ besitzen. Wegen $p(0) = \varphi > 0$, $p(\beta') < 0$ und $p(\beta) > 0$ für $\beta >> \beta'$ gilt $\beta^- < \beta^+$ und $p(\beta) < 0$ für $\beta^- < \beta < \beta^+$.

Dann muss an einer Stelle $\beta^* \in (\beta^-, \beta^+)$ ein relatives Minimum von $p(\beta)$ also eine (positive reelle) Nullstelle von $p'(\beta)$ existieren. Weil $p'(\beta)$ genau einen Vorzeichenwechsel bei seinen Koeffizienten besitzt, besagt Satz 18, dass p' genau eine positive reelle Nullstelle besitzt (nämlich β^*). Das Polynom p ist in (β^-, β^*) monoton fallend. Also gilt $p'(\beta) < 0$ für $\beta \in (\beta^-, \beta^*)$.

Damit sind alle Voraussetzungen von Satz 26 erfüllt.

Wenn a) erfüllt und λ^* einfach ist, ist nach Satz 14 die Matrix $f'(x^*, \lambda^*) = \begin{pmatrix} P(\lambda^*) & P'(\lambda^*)x^* \\ e_s^T & 0 \end{pmatrix}$ regulär. Für die hinreichend gute Näherung $(\tilde{x}, \tilde{\lambda})$ von (x^*, λ^*) ist dann aus Stetigkeitsgründen auch $f'(\tilde{x}, \tilde{\lambda})$ regulär, d.h. b) gilt. □

Bemerkung 14

Die Nullstellen des Polynoms $(l + 1)$–ten Grades $p(\beta)$ aus Satz 26 kann man in MATLAB näherungsweise mit dem Befehl `roots` *bestimmen, der die Eigenwerte der Begleitmatrix von $p(\beta)$ berechnet. Angenommen, dies liefert Näherungen zweier reeller positiver Nullstellen*

$\beta^- < \beta^+$ von $p(\beta)$ mit $p(\beta) < 0$ für $\beta \in (\beta^-, \beta^+)$. Diese reellen Nullstellen eines skalaren Polynoms können dann mit Hilfe des INTLAB–Befehls `verifypoly` *eng eingeschlossen werden: $\beta^- \in [\beta^-]$, $\beta^+ \in [\beta^+]$. `verifypoly` benötigt Nullstellennäherungen und seine Verfahren basieren auf [25]. Für $\beta \in [\sup([\beta^-]), \inf([\beta^+])]$ gilt dann auf jeden Fall $p(\beta) < 0$. Analog kann man bei der positiven reellen Nullstelle β^* von $p'(\beta)$, für die $p'(\beta) < 0$ für $\beta \in (\beta^-, \beta^*)$ gilt, vorgehen.*

Bemerkung 15

Der MATLAB–Befehl `polyeig(A0,A1,...,Al)` *liefert Näherungen aller Eigenpaare des PEPs $P(\lambda)x = 0$ mit $P(\lambda)$ wie in* (10.1)*. Der Befehl verwendet eine Linearisierung der Form*

$$A - \lambda B = M_1(\lambda) \begin{pmatrix} P(\lambda) & O \\ O & I_{(l-1)n} \end{pmatrix} M_2(\lambda)$$

von $P(\lambda)$ (mit $\det(M_1(\lambda)) = c_1 \neq 0$, $\det(M_2(\lambda)) = c_2 \neq 0$) und löst das so erhaltene verallgemeinerte Eigenwertproblem (GEP) $Ay = \lambda By$ ($A, B \in \mathbb{R}^{ln \times ln}$) näherungsweise mit dem QZ–Algorithmus. Die Eigenwerte des PEPs sind dann genau die Eigenwerte des GEPs, und aus den Eigenvektoren des GEPs erhält man die Eigenvektoren des PEPs.

Beispiel 7

Für das reelle quadratische Eigenwertproblem $(A_2\lambda^2 + A_1\lambda + A_0)x = 0$ mit $\det A_2 \neq 0$ (d.h. $l = 2$) lautet in Satz 26

$$\varphi = ||Cf(\tilde{x}, \tilde{\lambda})||_\infty, \quad \sigma = ||I_{n+1} - Cf'(\tilde{x}, \tilde{\lambda})||_\infty,$$

$$\tau_2 = ||C||_\infty \big(2|\tilde{\lambda}|\, ||A_2||_\infty + ||A_1||_\infty + ||A_2||_\infty ||\tilde{x}||_\infty\big), \quad \tau_3 = ||C||_\infty ||A_2||_\infty$$

und

$$\begin{aligned} g(\Delta x, \Delta\lambda) &= -Cf(\tilde{x}, \tilde{\lambda}) + \Big\{ I_{n+1} - Cf'(\tilde{x}, \tilde{\lambda}) \\ &\quad - \big(O \;\; (Q'(\tilde{\lambda})\Delta x) + \tfrac{1}{2}(Q''(\tilde{\lambda})\tilde{x})\Delta\lambda + \tfrac{1}{2}(Q''(\tilde{\lambda})\Delta x)\Delta\lambda\big)\Big\} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} \\ &= -Cf(\tilde{x}, \tilde{\lambda}) + \Big\{ I_{n+1} - Cf'(\tilde{x}, \tilde{\lambda}) \\ &\quad - \big(O \;\; (2C^*A_2\tilde{\lambda} + C^*A_1)\Delta x + \tfrac{1}{2}(2C^*A_2\tilde{x})\Delta\lambda + \tfrac{1}{2}(2C^*A_2\Delta x)\Delta\lambda\big)\Big\} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} \\ &= -Cf(\tilde{x}, \tilde{\lambda}) + \Big\{ I_{n+1} - Cf'(\tilde{x}, \tilde{\lambda}) \\ &\quad - \big(O \;\; (2\tilde{\lambda}C^*A_2 + C^*A_1)\Delta x + (C^*A_2\tilde{x})\Delta\lambda + (C^*A_2\Delta x)\Delta\lambda\big)\Big\} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix}. \end{aligned}$$

Mit $\tau = \tau_2$ und $\gamma = \tau_3$ stimmt dies also mit den entsprechenden Bezeichnungen in Satz 21 überein.

Wenn die Voraussetzungen (5.7) *und* (5.8) *aus Satz 21 erfüllt sind, dann erfüllen genau β^- und β^+ aus* (5.9) *die Bedingung* (10.11) *in Satz 26. Wenn man $\beta^* = \sqrt{\frac{-p}{3}} - \frac{\tau}{3\gamma}$ wählt, erfüllt*

genau dies die Bedingung (10.14) *in Satz 26. Unter den Voraussetzungen* (5.7) *und* (5.8) *sind die Sätze 21 und 26 also identisch.*

Beispiel 8

Für das reelle kubische Eigenwertproblem $(A_3\lambda^3 + A_2\lambda^2 + A_1\lambda + A_0)x = 0$ *mit* $\det A_3 \neq 0$ *(d.h.* $l = 3$*) lautet wegen*

$$P'(\lambda) = 3A_3\lambda^2 + 2A_2\lambda + A_1\,, \quad P''(\lambda) = 6A_3\lambda + 2A_2\,, \quad P^{(3)}(\lambda) = 6A_3$$

in Satz 26

$$\varphi = ||Cf(\tilde{x},\tilde{\lambda})||_\infty\,, \quad \sigma = ||I_{n+1} - Cf'(\tilde{x},\tilde{\lambda})||_\infty\,,$$

$$\tau_2 = ||C||_\infty\big(3||A_3||_\infty|\tilde{\lambda}|^2 + 2||A_2||_\infty|\tilde{\lambda}| + ||A_1||_\infty + 3||A_3||_\infty|\tilde{\lambda}|\,||\tilde{x}||_\infty + ||A_2||_\infty||\tilde{x}||_\infty\big)\,,$$

$$\tau_3 = ||C||_\infty\big(3||A_3||_\infty|\tilde{\lambda}| + ||A_2||_\infty + ||A_3||_\infty||\tilde{x}||_\infty\big)\,, \quad \tau_4 = ||C||_\infty||A_3||_\infty$$

und

$$g(\Delta x,\Delta\lambda) = -Cf(\tilde{x},\tilde{\lambda}) + \Big\{ I_{n+1} - Cf'(\tilde{x},\tilde{\lambda})$$

$$-\big(O \;\; Q'(\tilde{\lambda})\Delta x + \tfrac{1}{2}(Q''(\tilde{\lambda})\tilde{x})\Delta\lambda + \tfrac{1}{2}(Q''(\tilde{\lambda})\Delta x)\Delta\lambda + \tfrac{1}{6}(Q^{(3)}(\tilde{\lambda})\tilde{x})(\Delta\lambda)^2 + \tfrac{1}{6}(Q^{(3)}(\tilde{\lambda})\Delta x)(\Delta\lambda)^2\big)\Big\}\begin{pmatrix}\Delta x\\ \Delta\lambda\end{pmatrix}$$

$$= -Cf(\tilde{x},\tilde{\lambda}) + \Big\{ I_{n+1} - Cf'(\tilde{x},\tilde{\lambda})$$

$$-\big(O \;\; (3D_3\tilde{\lambda}^2 + 2D_2\tilde{\lambda} + D_1)\Delta x + ((3D_3\tilde{\lambda} + D_2)\tilde{x})\Delta\lambda + ((3D_3\tilde{\lambda} + D_2)\Delta x)\Delta\lambda + (D_3\tilde{x})(\Delta\lambda)^2 + (D_3\Delta x)(\Delta\lambda)^2\big)\Big\}\begin{pmatrix}\Delta x\\ \Delta\lambda\end{pmatrix}$$

mit $D_3 := C^*A_3\,,\; D_2 := C^*A_2$ *und* $D_1 := C^*A_1\,.$
Erinnert sei außerdem an die Festlegung $[\Delta\lambda]^2 := [\Delta\lambda]\cdot[\Delta\lambda]\,.$

Beispiel 9

Es sollen mit dem Verfahren aus Beispiel 8 unter Berücksichtigung von Bemerkung 13 ein einfacher reeller Eigenwert λ^* *und zugehöriger Eigenvektor* x^* *des kubischen reellen* 10×10*–Matrixpolynoms* $P(\lambda) = A_3\lambda^3 + A_2\lambda^2 + A_1\lambda + A_0$ *(mit* A_i *wie in* (8.3) *für* $n = 10$*) eingeschlossen werden. Wie in Kapitel 8.2 wurde mit MATLAB und INTLAB in einem 64 Bit–Gleitkommasystem gerechnet. Eine hinreichend gute Eigenpaarnäherung* $(\tilde{x},\tilde{\lambda})$ *wurde mit Hilfe des MATLAB–Befehls* `polyeig(A0,A1,A2,A3)` *gewonnen. Dazu wurde die von* `polyeig` *berechnete Eigenvektornäherung so normiert, dass der betragsmäßig größte Eintrag und damit die Maximumsnorm gleich Eins ist, und dann die führenden (gerundeten)* $z_{\tilde{\lambda}}$ *bzw.* $z_{\tilde{x}}$ *Stellen der Eigenwertnäherung bzw. normierten Eigenvektornäherung als* $\tilde{\lambda}$ *bzw.* $\tilde{x}$ *verwendet.* C *wurde wiederum als die mit MATLAB berechnete Inverse von* $f'(\tilde{x},\tilde{\lambda}) \in \mathbb{R}^{11\times 11}$ *gewählt (siehe* (10.4)*).*

Mit der einfachen Eigenpaarnäherung

$$\tilde{\lambda} = 1.356, \quad \tilde{x} = \begin{pmatrix} 6.965E-02 \\ 1 \\ -1.621E-01 \\ -1.939E-01 \\ 8.471E-01 \\ 4.544E-01 \\ 1.942E-01 \\ -1.49E-01 \\ -6.133E-01 \\ 9.276E-02 \end{pmatrix}$$

(d.h. $z_{\tilde{\lambda}} = z_{\tilde{x}} = 4$*) sind dann alle Voraussetzungen des Verfahrens aus Beispiel* 8 *erfüllt (für* $z_{\tilde{\lambda}} \leq 3$ *und* $z_{\tilde{x}} \leq 3$ *ist dies nicht der Fall). Wie in Bemerkung 14 beschrieben, wurden* $\sup([\beta^-])$ *und* $\inf([\beta^*])$ *berechnet. Mit* $\beta = 0.0004 \in [\sup([\beta^-]), \inf([\beta^*])]$ *wurden nach 5 Iterationen die eindeutigen Einschließungen*

$$\lambda^* \in [1.356212914130808^{9}_{8}], \quad x^* \in \begin{pmatrix} [6.965013011297347^{7}_{3}E-002] \\ 1 \\ [-1.620586353762361^{1}_{3}E-001] \\ [-1.939234377827459^{59}_{61}E-001] \\ [8.470688289142049^{9}_{5}E-001] \\ [4.543655689653972^{2}_{0}E-001] \\ [1.941825722627263^{3}_{1}E-001] \\ [-1.489767367321^{899}_{901}E-001] \\ [-6.133353216358955^{55}_{60}E-001] \\ [9.275702305434314^{14}_{09}E-002] \end{pmatrix}$$

gewonnen.

Kapitel 11

Zusammenfassung und Ausblick

In dieser Dissertation wurden Einschließungsverfahren für einfache Eigenwerte λ^* und zugehörige Eigenvektoren x^* des quadratischen Eigenwertproblems

$$(A_2\lambda^2 + A_1\lambda + A_0)x = 0 \quad (A_0, A_1, A_2 \in \mathbb{C}^{n\times n};\, \det A_2 \neq 0)$$

entwickelt (reelle Eigenpaare des reellen QEP: Satz 21; komplexe Eigenpaare des reellen QEP: Satz 23; komplexe Eigenpaare des komplexen QEP: Satz 25). Für das reelle polynomiale Eigenwertproblem $(A_l\lambda^l + A_{l-1}\lambda^{l-1} + \ldots + A_1\lambda + A_0)x = 0$ $(A_0, \ldots, A_l \in \mathbb{R}^{n\times n};\, \det A_l \neq 0)$ beliebigen Grades l wurde ein Verfahren zur Einschließung von reellen einfachen Eigenpaaren hergeleitet (Satz 26).

Alle diese Verfahren benötigen bereits sehr gute Eigenpaarnäherungen, wie es in Kapitel 8.2 an Beispielen sowie in Beispiel 9 verdeutlicht wurde. Numerische Verfahren zur Berechnung von Eigenpaarnäherungen beim QEP wurden in Kapitel 8.1 vorgestellt. Man kann die Einschließungsverfahren ausschließlich für Näherungen von einfachen Eigenpaaren anwenden. Nur dann kann man die in den Verfahren verwendete Matrix C wie in Bemerkung 10, 12 und 13 wählen. Wenn $(\tilde{x}, \tilde{\lambda})$ eine sehr gute Näherung eines mehrfachen Eigenpaares ist, dann ist die Matrix $f'(\tilde{x}, \tilde{\lambda})$ (siehe (5.5)) bzw. $J(\tilde{x}, \tilde{\lambda})$ (siehe (6.8)) fast singulär, und somit kann man nicht deren Inverse als C wählen.

In [18] und [2] wurden Einschließungsverfahren für reelle Eigenpaare des reellen einfachen Eigenwertproblems $Ax = \lambda x$ entwickelt. Wenn die Voraussetzungen dieser Verfahren für eine beliebige Matrix C erfüllt sind, dann impliziert dies (bei hinreichend guter Eigenpaarnäherung) die Einfachheit des eindeutigen eingeschlossenen Eigenpaares. Dieses Resulat konnte ich auf meine Verfahren leider nicht übertragen.

Die in dieser Dissertation vorgestellten Verfahren liefern nach wenigen Iterationen sehr enge Schranken für das eingeschlossene Eigenpaar und garantieren dann die meisten führenden Stellen im gegebenen Gleitpunktsystem. Auch bei relativ großer Dimension n des QEP ($n = 200$ siehe Kapitel 8.2) sind die Verfahren aus Satz 21, 23 und 25 praktisch durchführbar.

Die Verfahren zur Einschließung von reellen einfachen Eigenpaaren beim reellen PEP werden mit wachsendem Grad l des PEP natürlich immer komplexer. Dazu betrachte man sich allein das Verfahren für das kubische PEP in Beispiel 8.

Einschließungsverfahren für komplexe einfache Eigenpaare beim reellen bzw. komplexen PEP lassen sich in Analogie zu Kapitel 6 und 7 herleiten. Wie man doppelte Eigenwerte und zugehörige Eigenvektoren des QEP bzw. PEP verifiziert, ist natürlich ein weiteres Thema, das sich

zu betrachten lohnt. Ein entsprechendes Verfahren für reelle Eigenpaare des reellen einfachen Eigenwertproblems ist in [4] zu finden.

Literaturverzeichnis

[1] G. Alefeld, J. Herzberger: Einführung in die Intervallrechnung, Bibliographisches Institut AG, Zürich, 1974.

[2] G. Alefeld: Berechenbare Fehlerschranken für ein Eigenpaar unter Einschluß von Rundungsfehlern bei Verwendung des genauen Skalarprodukts, ZAMM, 67 (1987), S. 145 – 152.

[3] G. Alefeld: Berechenbare Fehlerschranken für ein Eigenpaar beim verallgemeinerten Eigenwertproblem, ZAMM, 68 (1988), S. 181 – 184.

[4] G. Alefeld, H. Spreuer: Iterative improvement of componentwise errorbounds for invariant subspaces belonging to a double or nearly double eigenvalue, Computing, 36 (1986), S. 321 – 334.

[5] B. Anderson, J. Jackson, M. Sitharam: Descartes' Rule of Signs Revisited, Amer. Math. Monthly, 105 (1998), S. 447 – 451.

[6] F. Chatelin: Eigenvalues of Matrices, John Wiley & Sons, Chichester, 1993.

[7] A. Duschek: Vorlesungen über höhere Mathematik I, 3. Aufl., Springer Verlag, Wien, 1960.

[8] T. Fitzpatrick, P. Dallard, S. Le Bourva, A. Low, R. Smith, M. Willford: Linking London: The Millennium Bridge, Royal Academy of Engineering, Paper ISBN 1 871634 99 7, London, 2001.

[9] F.R. Gantmacher: Matrizenrechnung I, 2. Aufl., Deutscher Verlag der Wissenschaften, Berlin, 1965.

[10] I. Gohberg, P. Lancaster, L. Rodman: Matrix Polynomials, Academic Press, New York, 1982.

[11] G.I. Hargreaves: Interval Analysis in MATLAB, Numerical Analysis Report, 416 (2002), Manchester Centre for Computational Mathematics, Manchester (England).

[12] H. Heuser: Lehrbuch der Analysis, Teil 1, 9. Aufl., B.G. Teubner, Stuttgart, 1991.

[13] H. Heuser: Lehrbuch der Analysis, Teil 2, 6. Aufl., B.G. Teubner, Stuttgart, 1991.

[14] R. Krawczyk: Iterative Verbesserung von Schranken für Eigenwerte und Eigenvektoren reeller Matrizen, ZAMM, 48 (1968), S. T80 – T83.

[15] D. Kressner: Numerical Methods for General and Structured Eigenvalues Problems, Springer Verlag, Berlin, 2005.

[16] P. Lancaster: Lambda-Matrices and Vibrating Systems, Pergamon Press, Oxford, UK, 1966.

[17] P. Lancaster, M. Tismenetsky: The Theory of Matrices, 2nd ed., Academic Press, London, 1985.

[18] G. Mayer: Result verification for eigenvectors and eigenvalues, in:
J. Herzberger (ed.): Topics in validated computations, North–Holland (ELSEVIER), 1994, S. 209 – 276.

[19] G. Mayer: A unified approach to enclosure methods for eigenpairs, ZAMM, 74 (1994), S. 115 – 128.

[20] V. Mehrmann, H. Voss: Nonlinear eigenvalue problems: a challenge for modern eigenvalue methods, GAMM Mitteilungen, 27 (2004), S. 121 – 152.

[21] L.B. Rall: Computational Solution of Nonlinear Operator Equations, John Wiley & Sons, New York, 1969.

[22] S.M. Rump: Exakte Fehlerschranken für Eigenwerte und Eigenvektoren, ZAMM, 61 (1981), S. T311 – T313.

[23] S.M. Rump: INTLAB – INTerval LABoratory, in:
T. Csendes (ed.), Developments in Reliable Computing, Kluwer, Dordrecht, 1999, S. 77 – 105.

[24] S.M. Rump: Solving Algebraic Problems with High Accuracy, in:
U.W. Kulisch, W.L. Miranker (eds.): A New Approach to Scientific Computation, Academic Press, New York, 1983, S. 53 – 120.

[25] S.M. Rump: Ten methods to bound multiple roots of polynomials, Journal of Computational and Applied Mathematics, 156 (2003), S. 403 – 432.

[26] H.J. Symm, J.H. Wilkinson: Realisitic Error Bounds for a Simple Eigenvalue and its Associated Eigenvector, Numer. Math., 35 (1980), S. 113 – 126.

[27] F. Tisseur, K. Meerbergen: The Quadratic Eigenvalue Problem, SIAM Rev., 43 (2001), S. 235 – 286.

Lebenslauf

Name: Friederike Decker

Geburtsdatum: 8.12.1979

Geburtsort: Rostock

Eltern: Lutz Decker

Cornelia Decker, geb. Pietsch

Schulbildung: Sept. 1986 – Juni 1991:

Polytechnische Oberschule "Gerhart Hauptmann" in Ribnitz

Sept. 1991 – Juni 1998:

Gymnasium "Richard Wossidlo" in Ribnitz
Abiturprüfung am 3. Juli 1998

Studium: Okt. 1998 – Juni 2004:

Diplom-Mathematik mit Nebenfach Informatik
an der Universität Rostock
Diplomprüfung am 29. Juni 2004

Okt. 2002 – März 2003:

Auslandssemester an der University of Glasgow (GB)

Berufstätigkeit: Okt. 2000 – Juli 2002:

Studentische Hilfskraft
am Institut für Mathematik der Universität Rostock

seit Juli 2004:

Wissenschaftliche Angestellte
am Institut für Angewandte und Numerische Mathematik (Lehrstuhl 1)
der Universität Karlsruhe