Optimierung von Matrix-Abständen durch unitäre Ähnlichkeitstransformationen

Dissertation
zur Erlangung des Doktorgrades
der Fakultät für Mathematik der Universität Bielefeld

vorgelegt von
Ralf Siekmeyer

April 1999

1. Gutachter: Prof. Dr. Ludwig Elsner
2. Gutachter: Prof. Dr. Wolf-Jürgen Beyn
Datum der mündlichen Prüfung: 1. Juli 1999

Gedruckt auf alterungsbeständigem Papier ∞ ISO 9706

Inhaltsverzeichnis

Einleitung

In der vorliegenden Arbeit werden zwei Optimierungsprobleme behandelt, die in der Matrizentheorie von Bedeutung sind. Nachdem im ersten Teil dieser Arbeit beide Probleme zunächst theoretisch analysiert werden, werden im zweiten Teil numerische Verfahren zur Lösung entwickelt, die auf der iterativen Anwendung von unitären Ähnlichkeitstransformationen basieren.

Das erste Optimierungsproblem, das im folgenden als isospektrales Abstandsproblem bezeichnet wird, besteht darin, aus allen zu einer gegebenen komplexen Matrix A unitär ähnlichen Matrizen diejenigen zu bestimmen, deren Abstand zu einer komplexen Matrix B optimal ist. Wird mit $\Lambda(A)$ die isospektrale und kompakte Menge der zur Matrix A unitär ähnlichen Matrizen bezeichnet und mit

$$f_B(X) := \|B - X\|_F$$

der Abstand einer Matrix $X \in \Lambda(A)$ zur Matrix B in der Frobenius-Norm, so sollen also das Maximum und Minimum der Funktion f_B auf $\Lambda(A)$ bestimmt werden, sowie diejenigen Matrizen aus $\Lambda(A)$, für die die Funktion f_B ein Extremum annimmt.

Das zweite Optimierungsproblem ist das der besten normalen Approximation einer beliebigen Matrix in der Frobenius-Norm, das heißt zu einer gegebenen komplexen Matrix A soll die Abweichung

$$\min_{N \text{ ist normal}} \|A - N\|_F$$

der Matrix A von der Normalität bezüglich der Frobenius-Norm, sowie eine in der Frobenius-Norm nächstgelegene normale Matrix zu A bestimmt werden. Dieses und weitere Maße der Nichtnormalität einer Matrix wurden in [7] von Elsner und Paardekooper studiert. Da für eine nichtnormale Matrix A auch alle zur Matrix A unitär ähnlichen Matrizen nichtnormal sind, ist zunächst nicht ersichtlich, wie durch unitäre Ähnlichkeitstransformationen der Matrix A eine zu A nächstgelegene normale Matrix gewonnen werden kann. Darauf wird weiter unten eingegangen.

Diese Arbeit ist nicht die erste, in der das isospektrale Abstandsproblem behandelt wird. Bereits in [1] betrachtete Bhatia die Extrema der Funktion f_B^2 im Zusammenhang mit einer Verallgemeinerung des Satzes von Hoffman-Wielandt ([18, Theorem 6.3.5], eine allgemeinere Version ist in [6] zu finden). Wichtigstes Ergebnis seiner Untersuchung ist, daß alle Matrizen $X \in \Lambda(A)$, die die Funktion f_B^2 optimieren, dadurch gekennzeichnet sind, daß der Kommutator $[X, B^*]$ dieser Matrizen mit der Matrix B^* eine hermitesche Matrix ist.

Helmke und Hüper untersuchten in [15] ein Optimierungsproblem, das als Teilproblem des isospektralen Abstandsproblems angesehen werden kann. Sie studierten für reellsymmetrische Matrizen A und B die Extrema der Funktion f_B^2 auf der isospektralen Mannigfaltigkeit der zur Matrix A orthogonal ähnlichen Matrizen, die eine Teilmenge von $\Lambda(A)$ ist. Indem sie die Funktion f_B^2 auf eine einfachere Funktion zurückführten, lösten sie das Optimierungsproblem für den Spezialfall, daß B eine Diagonalmatrix mit paarweise verschiedenen und antiton angeordneten Diagonaleinträgen ist. In diesem Fall besitzt die Funktion f_B^2 auf der isospektralen Mannigfaltigkeit eine eindeutige Maximal- und Minimalstelle, bei der es sich gerade um die Diagonalmatrizen handelt, deren Diagonaleinträge die Eigenwerte der Matrix A in isotoner bzw. antitoner Anordnung sind. Ein Minimierungsverfahren für die Funktion f_B^2 diagonalisiert in diesem Fall also die Matrix A und sortiert die Eigenwerte der Matrix A der Größe nach. Helmke und Hüper konstruierten ein solches Verfahren, das darauf beruht, daß auf die Matrix A eine Folge von orthogonalen Ähnlichkeitstransformationen derart angewendet wird, daß jede Transformation die Funktion f_B^2 auf einer geschlossenen Kurve, die auf der Mannigfaltigkeit der zur Matrix A orthogonal ähnlichen Matrizen liegt, minimiert. Bei diesem Sortier-Jacobi-Algorithmus genannten Verfahren handelt es sich um ein iteratives zyklisches Verfahren, das wie das zyklische Jacobi-Verfahren zur Diagonalisierung reellsymmetrischer Matrizen, das etwa in Abschnitt 8.4 in [11] beschrieben wird, parallelisierbar ist, aber den Vorteil besitzt, daß es die Eigenwerte von A sortiert und die Konvergenz auch dann quadratisch ist, wenn A mehrfache Eigenwerte besitzt. Die Bezeichnung Sortier-Jacobi-Algorithmus wird dadurch gerechtfertigt, daß das zyklische Jacobi-Verfahren geometrisch so interpretiert werden kann, daß in jedem Schritt dieses Verfahrens die klassische Off-Norm Funktion über eine geschlossene Kurve auf der isospektralen Mannigfaltigkeit minimiert wird, wobei die Richtungen aller dieser Kurven mit den Richtungen der im Sortier-Jacobi-Algorithmus auftretenden Kurven übereinstimmen. Dies wurde von Helmke, Hüper und Moore in [16] gezeigt.

Auch das Problem der besten normalen Approximation einer komplexen Matrix wurde schon untersucht, etwa von Causey in [3]. Gabriel konnte es in [8] für 2×2 - Matrizen vollständig lösen. Ein wichtiges Ergebnis, das von Ruhe in [21] veröffentlicht wurde, beschreibt die Struktur einer Matrix A in den Koordinaten der Eigenvektoren einer zur Matrix A nächstgelegenen normalen Matrix: Wird A einer unitären Ähnlichkeitstransformation unterzogen, die eine zu A nächstgelegene normale Matrix diagonalisiert, so gehört die transformierte Matrix zur Klasse der sogenannten ΔH-Matrizen, und die Abweichung der Matrix A zur Normalität bezüglich der Frobenius-Norm ist identisch mit dem Abstand dieser transformierten Matrix zu ihrer Diagonale in der Frobenius-Norm. Eine Methode zur Bestimmung einer nächstgelegenen normalen Matrix besteht also darin, eine gegebene Matrix durch eine unitäre Ähnlichkeitstransformation in eine ΔH-Matrix zu überführen, die minimalen Abstand zu ihrer Diagonale besitzt, und anschließend den Diagonalteil dieser Matrix wieder zurückzutransformieren. Auf dieser Idee beruht das von Ruhe in [21] vorgestellte Verfahren, das durch eine Sequenz von unitären Ähnlichkeitstransformationen sukzessiv die 2×2 - Teilmatrizen einer Matrix derart in ΔH-Matrizen überführt, daß der Abstand der gesamten Matrix zu ihrer Diagonale verringert wird.

In dieser Arbeit werden zunächst einige neue Resultate zum isospektralen Abstandsproblem erarbeitet, sowie der Zusammenhang dieses Problems mit dem Problem der besten normalen Approximation einer Matrix erörtert. Anschließend werden jacobiartige Verfahren und Gradientenverfahren zur Lösung beider Optimierungsprobleme entwickelt. Dabei ist die vorliegende Arbeit wie folgt gegliedert:

In Kapitel 1 werden die in dieser Arbeit verwendeten Abkürzungen aufgelistet und einige später benötigte Resultate, hauptsächlich aus dem Bereich der Matrizentheorie, vorgestellt.

In Kapitel 2 wird im ersten Abschnitt das isospektrale Abstandsproblem untersucht. Wie in der Arbeit [15] von Helmke und Hüper wird zur Analyse der Funktion f_B eine einfachere Funktion g_B herangezogen, deren Maxima die Minima der Funktion f_B und deren Minima die Maxima von f_B sind. In Satz 2.6 wird für die Stationärpunkte von g_B eine Charakterisierung gegeben, die anschaulicher als die oben genannte Bedingung von Bhatia für das Vorliegen eines Extremums der Funktion f_B^2 ist. Mit Hilfe dieser Bedingung werden in Satz 2.9 für die Spezialfälle, daß die Matrix B hermitesch oder schiefhermitesch ist, sowie den Fall, daß sowohl A als auch B normale Matrizen sind, explizite Lösungen des isospektralen Abstandsproblems gegeben. Für den allgemeinen Fall wird mit Korollar 2.13 eine Abschätzung für die Extremalwerte von f_B und g_B vorgestellt. Im zweiten Abschnitt wird gezeigt, daß das oben angegebene Ergebnis von Ruhe eine direkte Folgerung aus einem im ersten Abschnitt gewonnenen Ergebnis zum isospektralen Abstandsproblem ist.

In Abschnitt 3.1 wird der in [15] vorgestellte Sortier-Jacobi-Algorithmus so verallgemeinert, daß Lösungsverfahren zur Bestimmung der Extrema der Funktion f_B auf der Menge $\Lambda(A)$, wobei A und B beliebige komplexe Matrizen sein können, erhalten werden. Diese jacobiartigen Verfahren beruhen darauf, daß die Funktion f_B in jedem Schritt der Verfahren über eine Kurve auf $\Lambda(A)$ mit vorgegebener Richtung optimiert wird, und sind erheblich komplizierter und aufwendiger als der Sortier-Jacobi-Algorithmus. Im Anschluß daran wird gezeigt, daß für diagonale Matrizen B einfachere und weniger aufwendige jacobiartige Verfahren zur Lösung des isospektralen Abstandsproblems existieren. In Abschnitt 3.2 wird aus dem Verfahren zur Minimierung der Funktion f_B mit diagonaler Matrix B ein Verfahren zur Berechnung einer nächstgelegenen normalen Matrix abgeleitet. Außerdem wird das von Ruhe in [21] vorgeschlagene Verfahren neu hergeleitet, da die Darstellung in [21] erhebliche Mängel aufweist.

In Kapitel 4 wird im ersten Teil zu jeder Matrix aus der Menge $\Lambda(A)$ eine Richtung bestimmt, in der die Funktion g_B den größten Anstieg verzeichnet, und darauf basierend ein Gradientenverfahren zur Lösung des isospektralen Abstandsproblems entwickelt. Im zweiten Teil wird analog zu Abschnitt 3.2 daraus ein Verfahren zur Bestimmung einer nächstgelegenen normalen Matrix abgeleitet.

In Kapitel 5 werden schließlich die in den Kapiteln 3 und 4 entwickelten Verfahren numerisch getestet, wobei sämtliche Berechnungen mit dem Programmpaket MATLAB durchgeführt werden. Im ersten Abschnitt wird die Konvergenz der jacobiartigen Verfahren und des Gradientenverfahrens zur Minimierung der Funktion f_B auf $\Lambda(A)$ anhand einiger Beispiele untersucht. Aufgrund experimenteller Beobachtungen werden Vorschläge gemacht, in welchen Fällen welche Verfahren verwendet werden sollten. Im zweiten Abschnitt werden die Verfahren zur Berechnung einer nächstgelegenen normalen Matrix auf Konvergenz getestet und die Konvergenzgeschwindigkeiten verglichen. Dabei zeigt sich, daß das Gradientenverfahren zur Bestimmung nächstgelegener normaler Matrizen weniger brauchbar ist als die jacobiartigen Verfahren.

An dieser Stelle möchte ich mich bei Prof. Dr. Ludwig Elsner für die Anregung zu diesem Thema und die Betreuung dieser Arbeit bedanken, sowie bei Prof. Dr. Wolf-Jürgen Beyn für einige Hinweise.

Kapitel 1

Grundlegende Definitionen und Sätze

In diesem Kapitel werden einige Definitionen und Sätze, vorwiegend aus dem Bereich der Matrizentheorie, aufgeführt, die wir in den folgenden Kapiteln benötigen werden. Zunächst einmal listen wir die in dieser Arbeit verwendeten Abkürzungen auf.

Definition 1.1 (Abkürzungsverzeichnis)
Sei im folgenden

$\mathbb{N}$	*die Menge der natürlichen Zahlen,*
$\mathbb{R}$	*die Menge der reellen Zahlen,*
$\mathbb{C}$	*die Menge der komplexen Zahlen,*
$\mathbb{C}^{m,n}$	*die Menge der komplexen $m \times n$ - Matrizen,*

und für $z \in \mathbb{C}$ sei

$\mathrm{re}(z)$	*der Realteil von z,*
$\mathrm{im}(z)$	*der Imaginärteil von z und*
$\arg(z)$	*das Argument von z,*

sowie für $x \in \mathbb{R}$

(1.1)
$$\mathrm{sign}(x) := \begin{cases} 1 & , \quad x > 0 \\ 0 & , \quad x = 0 \\ -1 & , \quad x < 0 \end{cases}$$

die Vorzeichenfunktion. Weiter sei

(1.2)
$$0_{m,n} = (0)_{\substack{k=1,\ldots,m \\ l=1,\ldots,n}} \in \mathbb{C}^{m,n}$$

die Nullmatrix, wobei auf eine Angabe der Indizes verzichtet wird, wenn die Größe der Matrix offensichtlich ist,

(1.3)
$$I_n = (\delta_{kl})_{k,l=1,\ldots,n} \quad mit \quad \delta_{kl} = \begin{cases} 1 & , \quad k = l \\ 0 & , \quad k \neq l \end{cases}$$

die n-reihige Einheitsmatrix und

$$P_n = (p_{kl})_{k,l=1,\dots,n} \quad \textit{mit} \quad p_{kl} = \begin{cases} 1 & , \quad k+l = n+1 \\ 0 & , \quad k+l \neq n+1 \end{cases} \tag{1.4}$$

die n-reihige Standard-Involution-Permutation Matrix.

Für $X = (x_{kl})_{k,l=1,\dots,n} \in \mathbb{C}^{n,n}$ *sei*

$$X^* := (\overline{x_{lk}})_{k,l=1,\dots,n} \tag{1.5}$$

die zur Matrix X *konjugiert transponierte Matrix,*

$$\operatorname{tr}(X) := \sum_{k=1}^{n} x_{kk} \tag{1.6}$$

die Spur der Matrix X*,*

$$\|X\|_F := \sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n} |x_{kl}|^2} \tag{1.7}$$

die Frobenius-Norm der Matrix X*,*

$$H(X) := \frac{1}{2}(X + X^*) \tag{1.8}$$

der hermitesche Teil der Matrix X *und*

$$S(X) := \frac{1}{2}(X - X^*) \tag{1.9}$$

der schiefhermitesche Teil der Matrix X*, sowie für* $X, Y \in \mathbb{C}^{n,n}$

$$[X, Y] := XY - YX \tag{1.10}$$

der Kommutator von X *und* Y*. Wir bezeichnen mit*

$$\mathcal{U}(n) := \{Y \in \mathbb{C}^{n,n} : YY^* = I_n\} \tag{1.11}$$

die Menge der unitären $n \times n$ *- Matrizen, sowie mit*

$$\mathcal{H}(n) := \{Y \in \mathbb{C}^{n,n} : Y^* = Y\} \tag{1.12}$$

die Menge der hermiteschen $n \times n$ *- Matrizen,*

$$\mathcal{S}(n) := \{Y \in \mathbb{C}^{n,n} : Y^* = -Y\} \tag{1.13}$$

die Menge der schiefhermiteschen $n \times n$ *- Matrizen,*

$$\mathcal{N}(n) := \{Y \in \mathbb{C}^{n,n} : YY^* = Y^*Y\} \tag{1.14}$$

die Menge der normalen $n \times n$ *- Matrizen und*

$$\Pi(n) := \left\{\pi : \{1, \dots, n\} \longrightarrow \{1, \dots, n\} : \pi \textit{ ist bijektiv}\right\} \tag{1.15}$$

die Gruppe der Permutationen von $\{1, \dots, n\}$*.*

Der nächste Satz zeigt, daß die in (1.7) definierte Funktion tatsächlich eine Matrixnorm ist, und beschreibt die wichtigsten Eigenschaften dieser Norm.

Satz 1.2 (Eigenschaften der Frobenius-Norm)

(i) $\|\cdot\|_F$ *definiert durch* (1.7) *ist eine Matrixnorm auf* $\mathbb{C}^{n,n}$.

(ii) $\|\cdot\|_F$ *ist unitär invariant, das heißt für alle* $X \in \mathbb{C}^{n,n}$ *und* $U, V \in \mathcal{U}(n)$ *ist*

$$\|X\|_F = \|UXV\|_F. \tag{1.16}$$

(iii) Für alle $X \in \mathbb{C}^{n,n}$ *gilt*

$$\|X\|_F^2 = \operatorname{tr}(XX^*). \tag{1.17}$$

Beweis:

(i) und (ii) werden in [18], Abschnitt 5.6, bewiesen.

(iii) folgt unmittelbar aus den Definitionen von $\|\cdot\|_F$ und tr. □

Im folgenden wird der Inhalt dieses Satzes als bekannt voraus gesetzt, das heißt wir werden die Eigenschaften der Frobenius-Norm verwenden ohne auf diesen Satz zu verweisen.

Eine wichtige Klasse von Matrizen sind die normalen Matrizen. Es gibt eine Vielzahl von Matrix-Eigenschaften, die zur Normalität äquivalent sind. Diejenigen dieser Eigenschaften, die wir im Laufe dieser Arbeit benötigen, werden im nächsten Satz zusammengefaßt.

Satz 1.3

Sei $X \in \mathbb{C}^{n,n}$ *und seien* λ_j, $j \in \{1, \ldots, n\}$, *die Eigenwerte von* X. *Dann sind die folgenden Aussagen äquivalent:*

(i) $X \in \mathcal{N}(n)$

(ii) Es existiert $U \in \mathcal{U}(n)$ *mit* $X = U \operatorname{diag}(\lambda_1, \ldots, \lambda_n) U^*$.

(iii) Es ist $H(X)S(X) = S(X)H(X)$.

(iv) Die Eigenwerte von $H(X)$ *sind* $\operatorname{re}(\lambda_j)$, $j \in \{1, \ldots, n\}$.

(v) Die Eigenwerte von $S(X)$ *sind* $\operatorname{im}(\lambda_j)i$, $j \in \{1, \ldots, n\}$.

Beweis:

Wir verweisen auf [13], wo 70 äquivalente Bedingungen dazu, daß eine Matrix normal ist, bewiesen werden.

(i) ist die Definition der Normalität, also "Condition 0" in [13].

(ii) entspricht "Condition 11" in [13].

(iii) entspricht "Condition 21" in [13].

(iv) entspricht "Condition 35" in [13].

(v) entspricht "Condition 36" in [13]. □

Eine andere wichtige Klasse von Matrizen sind die doppelt stochastischen Matrizen.

Definition 1.4
Sei $X = (x_{kl})_{k,l=1,\ldots,n} \in \mathbb{C}^{n,n}$. X heißt doppelt stochastisch, wenn $x_{kl} \geq 0$ und

$$\sum_{j=1}^{n} x_{kj} = \sum_{j=1}^{n} x_{jl} = 1$$

für alle $k, l \in \{1, \ldots, n\}$ gilt.

Prominenteste Vertreter der doppelt stochastischen Matrizen sind die Permutationsmatrizen, also diejenigen Matrizen, die in jeder Zeile und Spalte genau einmal den Eintrag 1 besitzen und ansonsten nur Nullen enthalten. Tatsächlich erzeugen sie alle doppelt stochastischen Matrizen in dem Sinn, daß jede doppelt stochastische Matrix eine Konvexkombination von Permutationsmatrizen ist, wie der Satz von Birkhoff zeigt.

Satz 1.5 (Birkhoff)
Eine Matrix $X \in \mathbb{C}^{n,n}$ ist genau dann doppelt stochastisch, wenn sie eine konvexe Kombination von Permutationsmatrizen ist, das heißt wenn es ein $m \in \mathbb{N}$ gibt, so daß für $j \in \{1, \ldots, m\}$ $\alpha_j > 0$ und Permutationsmatrizen Q_j existieren mit

$$\sum_{j=1}^{m} \alpha_j = 1 \qquad \text{und} \qquad X = \sum_{j=1}^{m} \alpha_j Q_j .$$

Beweis: [18, Theorem 8.7.1] □

Eine der bekanntesten Matrixfunktionen ist die Exponentialfunktion

$$\exp : \mathbb{C}^{n,n} \longrightarrow \mathbb{C}^{n,n},\ \exp(X) = e^X := \sum_{k=0}^{\infty} \frac{1}{k!} X^k . \tag{1.18}$$

Wir untersuchen im folgenden, wie sich die schiefhermiteschen Matrizen unter der Exponentialfunktion verhalten.

Lemma 1.6
Sei $Y \in \mathcal{S}(n)$. Dann ist $e^Y \in \mathcal{U}(n)$.

Beweis:
Für $Y \in \mathcal{S}(n)$ ist

$$e^Y \left(e^Y\right)^* = e^Y e^{Y^*} = e^Y e^{-Y} = e^{Y-Y} = e^{0_{n,n}} = I_n ,$$

also e^Y unitär. Die dritte Gleichung gilt, weil Y und $-Y$ kommutieren. □

Zum Abschluß dieses Kapitels erforschen wir noch, wie zwei endliche reelle Zahlenfolgen $(a_k)_{k=1,\ldots,n}$ und $(b_k)_{k=1,\ldots,n}$ angeordnet werden müssen, um die Summe $\sum_{k=1}^{n} a_k b_k$ zu optimieren.

Lemma 1.7
Sei $a_1 \leq \cdots \leq a_n$ und $b_1 \leq \cdots \leq b_n$. Dann gilt

$$\max_{\pi \in \Pi(n)} \sum_{k=1}^{n} a_{\pi(k)} b_k = \sum_{k=1}^{n} a_k b_k \tag{1.19}$$

und

$$\min_{\pi \in \Pi(n)} \sum_{k=1}^{n} a_{\pi(k)} b_k = \sum_{k=1}^{n} a_{n+1-k} b_k . \tag{1.20}$$

Seien weiter $0 = t_0 < \cdots < t_m = n$ *so, daß*

$$b_1 = b_{t_1} < b_{t_1+1} = b_{t_2} < \cdots < b_{t_{m-1}+1} = b_{t_m}.$$

Dann ist für $\rho \in \Pi(n)$

$$\sum_{k=1}^{n} a_{\rho(k)} b_k = \max_{\pi \in \Pi(n)} \sum_{k=1}^{n} a_{\pi(k)} b_k \tag{1.21}$$

genau dann, wenn für alle $j \in \{1, \ldots, m\}$

$$\{a_{\rho(k)} : k \in [t_{j-1}+1, t_j]\} = \{a_k : k \in [t_{j-1}+1, t_j]\} \tag{1.22}$$

gilt, und

$$\sum_{k=1}^{n} a_{\rho(k)} b_k = \min_{\pi \in \Pi(n)} \sum_{k=1}^{n} a_{\pi(k)} b_k \tag{1.23}$$

genau dann, wenn für alle $j \in \{1, \ldots, m\}$

$$\{a_{\rho(k)} : k \in [t_{j-1}+1, t_j]\} = \{a_{n+1-k} : k \in [t_{j-1}+1, t_j]\} \tag{1.24}$$

gilt.

Beweis:
Sei $\rho \in \Pi(n)$. Erfüllt ρ (1.22) nicht, so existieren $1 \leq j_1 < j_2 \leq m$ und $k \in [t_{j_1-1}+1, t_{j_1}]$, $l \in [t_{j_2-1}+1, t_{j_2}]$ mit $a_{\rho(k)} > a_{\rho(l)}$. Dann ist

$$(a_{\rho(k)} - a_{\rho(l)})(b_l - b_k) > 0,$$

also

$$a_{\rho(k)} b_l + a_{\rho(l)} b_k > a_{\rho(k)} b_k + a_{\rho(l)} b_l.$$

Definiere $\sigma \in \Pi(n)$ durch

$$\sigma(r) := \begin{cases} \rho(r) & , \quad r \in \{1, \ldots, m\} \setminus \{k, l\} \\ \rho(l) & , \quad r = k \\ \rho(k) & , \quad r = l \end{cases} .$$

Dann gilt

$$\begin{aligned} \sum_{r=1}^{n} a_{\rho(r)} b_r &= \sum_{\substack{r=1 \\ r \notin \{k,l\}}}^{n} a_{\rho(r)} b_r + a_{\rho(k)} b_k + a_{\rho(l)} b_l \\ &< \sum_{\substack{r=1 \\ r \notin \{k,l\}}}^{n} a_{\rho(r)} b_r + a_{\rho(l)} b_k + a_{\rho(k)} b_l \\ &= \sum_{r=1}^{n} a_{\sigma(r)} b_r \end{aligned}$$

und ρ erfüllt (1.21) nicht. Also ist die Bedingung (1.22) notwendig für die Gültigkeit von (1.21). Erfüllt ρ (1.22), so ist

$$\begin{aligned} \sum_{r=1}^{n} a_{\rho(r)} b_r &= \sum_{r=1}^{m} \sum_{s=t_{r-1}+1}^{t_r} a_{\rho(s)} b_s = \sum_{r=1}^{m} b_{t_r} \sum_{s=t_{r-1}+1}^{t_r} a_{\rho(s)} \\ &= \sum_{r=1}^{m} b_{t_r} \sum_{s=t_{r-1}+1}^{t_r} a_s \quad = \sum_{r=1}^{m} \sum_{s=t_{r-1}+1}^{t_r} a_s b_s \\ &= \sum_{r=1}^{n} a_r b_r. \end{aligned}$$

Daraus folgt zum einen (1.19) und zum anderen, daß die Bedingung (1.22) hinreichend dafür ist, daß (1.21) gilt. Analog kann gezeigt werden, daß (1.20) gilt und die Bedingungen (1.23) und (1.24) äquivalent sind. $\square$

Damit sind die Definitionen und Sätze vorgestellt worden, die wir in den folgenden Kapiteln verwenden werden.

Kapitel 2

Optimierung von Matrix-Abständen

2.1 Das isospektrale Abstandsproblem

Im folgenden betrachten wir für gegebene $A, B \in \mathbb{C}^{n,n}$ die auf der kompakten Menge

$$\Lambda(A) := \{UAU^* : U \in \mathcal{U}(n)\} \tag{2.1}$$

der zu A unitär ähnlichen Matrizen definierte Funktion

$$f_B : \Lambda(A) \longrightarrow \mathbb{R},\ f_B(X) = \|B - X\|_F. \tag{2.2}$$

Wir interessieren uns für die Lösung des isospektralen Abstandsproblems, das heißt wir suchen das Maximum

$$F^+(A, B) := \max\{f_B(X) : X \in \Lambda(A)\} \tag{2.3}$$

und das Minimum

$$F^-(A, B) := \min\{f_B(X) : X \in \Lambda(A)\} \tag{2.4}$$

von f_B, sowie die Matrizen $X \in \Lambda(A)$ mit

$$f_B(X) = F^+(A, B) \tag{2.5}$$

und

$$f_B(X) = F^-(A, B). \tag{2.6}$$

Wir nennen (2.5) das isospektrale Maximierungsproblem und (2.6) das isospektrale Minimierungsproblem. Zur Untersuchung von f_B ziehen wir die Funktion

$$g_B : \Lambda(A) \longrightarrow \mathbb{R},\ g_B(X) = \mathrm{re}\{\mathrm{tr}(XB^*)\} \tag{2.7}$$

heran, sowie

$$G^+(A, B) := \max\{g_B(X) : X \in \Lambda(A)\} \tag{2.8}$$

und

$$G^-(A, B) := \min\{g_B(X) : X \in \Lambda(A)\}. \tag{2.9}$$

Zwischen f_B und g_B besteht der folgende Zusammenhang:

Lemma 2.1
Seien $A, B \in \mathbb{C}^{n,n}$ und $X \in \Lambda(A)$. Dann gilt

$$f_B(X)^2 = \|A\|_F^2 + \|B\|_F^2 - 2g_B(X). \tag{2.10}$$

Beweis:
Sei $X \in \Lambda(A)$. Dann ist $X = UAU^*$ mit $U \in \mathcal{U}(n)$ und es gilt

$$\begin{aligned} f_B(X)^2 &= \|B - X\|_F^2 \\ &= \operatorname{tr}\{(B-X)(B-X)^*\} \\ &= \operatorname{tr}(BB^* - BX^* - XB^* + XX^*) \\ &= \|B\|_F^2 + \|X\|_F^2 - \operatorname{tr}(BX^* + XB^*) \\ &= \|B\|_F^2 + \|UAU^*\|_F^2 - \operatorname{tr}\{(XB^*)^* + XB^*\} \\ &= \|B\|_F^2 + \|A\|_F^2 - 2\operatorname{re}\{\operatorname{tr}(XB^*)\} \\ &= \|A\|_F^2 + \|B\|_F^2 - 2g_B(X). \end{aligned}$$

□

Daraus folgt sofort:

Korollar 2.2
Seien $A, B \in \mathbb{C}^{n,n}$ und $X \in \Lambda(A)$. Dann gilt

$$F^+(A,B) = \sqrt{\|A\|_F^2 + \|B\|_F^2 - 2G^-(A,B)} \tag{2.11}$$

und

$$F^-(A,B) = \sqrt{\|A\|_F^2 + \|B\|_F^2 - 2G^+(A,B)}, \tag{2.12}$$

sowie

$$f_B(X) = F^+(A,B) \iff g_B(X) = G^-(A,B) \tag{2.13}$$

und

$$f_B(X) = F^-(A,B) \iff g_B(X) = G^+(A,B). \tag{2.14}$$

Wir bezeichnen mit

$$\begin{aligned} \mathcal{F}^+(A,B) :&= \{X \in \Lambda(A) : f_B(X) = F^+(A,B)\} \\ &= \{X \in \Lambda(A) : g_B(X) = G^-(A,B)\} \end{aligned} \tag{2.15}$$

die Menge der Lösungen des isospektralen Maximierungsproblems (2.5) und mit

$$\begin{aligned} \mathcal{F}^-(A,B) :&= \{X \in \Lambda(A) : f_B(X) = F^-(A,B)\} \\ &= \{X \in \Lambda(A) : g_B(X) = G^+(A,B)\} \end{aligned} \tag{2.16}$$

die Menge der Lösungen des isospektralen Minimierungsproblems (2.6).

$F^+(A,B)$, $F^-(A,B)$, $G^+(A,B)$ und $G^-(A,B)$ sind symmetrisch und unitär invariant, wie wir jetzt sehen werden.

Lemma 2.3
Seien $A, B \in \mathbb{C}^{n,n}$ und $U, V \in \mathcal{U}(n)$. Dann ist

$$F^+(UAU^*, VBV^*) = F^+(A,B) = F^+(B,A), \tag{2.17}$$
$$F^-(UAU^*, VBV^*) = F^-(A,B) = F^-(B,A), \tag{2.18}$$
$$G^+(UAU^*, VBV^*) = G^+(A,B) = G^+(B,A) \tag{2.19}$$

und

$$G^-(UAU^*, VBV^*) = G^-(A,B) = G^-(B,A). \tag{2.20}$$

Beweis:
Aus

$$\begin{aligned}
\{f_{VBV^*}(X) : X \in \Lambda(UAU^*)\} &= \{\|VBV^* - X\|_F : X \in \Lambda(UAU^*)\} \\
&= \{\|VBV^* - WUAU^*W^*\|_F : W \in \mathcal{U}(n)\} \\
&= \{\|B - V^*WUAU^*W^*V\|_F : W \in \mathcal{U}(n)\} \\
&= \{\|B - WAW^*\|_F : W \in \mathcal{U}(n)\} \\
&= \{f_B(X) : X \in \Lambda(A)\}
\end{aligned}$$

und

$$\begin{aligned}
\{f_A(X) : X \in \Lambda(B)\} &= \{\|A - WBW^*\|_F : W \in \mathcal{U}(n)\} \\
&= \{\|B - W^*AW\|_F : W \in \mathcal{U}(n)\} \\
&= \{f_B(X) : X \in \Lambda(A)\}
\end{aligned}$$

folgt

$$\begin{aligned}
F^+(A,B) &= \max\{f_B(X) : X \in \Lambda(A)\} \\
&= \max\{f_{VBV^*}(X) : X \in \Lambda(UAU^*)\} = F^+(UAU^*, VBV^*) \\
&= \max\{f_A(X) : X \in \Lambda(B)\} \qquad\quad = F^+(B,A)
\end{aligned}$$

und ebenso (2.18). Mit $\|UAU^*\|_F = \|A\|_F$ und $\|VBV^*\|_F = \|B\|_F$, sowie (2.11) und (2.12) folgen daraus (2.19) und (2.20). □

Ist also $UAU^* \in \Lambda(A)$ eine Lösung von (2.5), so löst U^*BU das duale Maximierungsproblem

"Finde $Y \in \Lambda(B)$ mit $f_A(Y) = F^+(B,A)$."

Ebenso ist für jede Lösung $VAV^* \in \Lambda(A)$ von (2.6) V^*BV eine Lösung des dualen Minimierungsproblems

"Finde $Y \in \Lambda(B)$ mit $f_A(Y) = F^-(B,A)$."

Sollen beim isospektralen Abstandsproblem nur $F^+(A,B)$ oder $F^-(A,B)$ bestimmt werden, so können A und B also vertauscht werden. Eine Lösung von (2.5) oder (2.6) kann dadurch gewonnen werden, daß man das duale Optimierungsproblem löst und anschließend eine Ähnlichkeitstransformation von A mit U^* durchführt, wobei $U \in \mathcal{U}(n)$ so gewählt wird, daß UBU^* eine Lösung des dualen Problems ist.

Um $F^+(A,B), F^-(A,B), G^+(A,B)$ und $G^-(A,B)$, sowie $\mathcal{F}^+(A,B)$ und $\mathcal{F}^-(A,B)$ zu bestimmen, untersuchen wir zunächst die Struktur der stationären Punkte von g_B, also der Punkte $X \in \Lambda(A)$, die die Eigenschaft besitzen, daß für jede differenzierbare Kurve $M : \mathbb{R} \longrightarrow \Lambda(A)$ mit $M(0) = X$

$$\frac{d}{dt} g_B(M)(0) = 0$$

ist. Da die Lösungen von (2.5) und (2.6) nach (2.13) und (2.14) auch für g_B extremal sind, sind sie insbesondere stationäre Punkte von g_B.

Sei $X \in \Lambda(A)$. Dann ist nach Lemma 1.6 für jedes $Y \in \mathcal{S}(n)$

$$\mathcal{K}_{X,Y} : \mathbb{R} \longrightarrow \Lambda(A),\ \mathcal{K}_{X,Y}(t) = e^{tY} X e^{-tY} \tag{2.21}$$

eine Kurve auf $\Lambda(A)$ durch X. Das folgende Lemma zeigt, daß zu jeder differenzierbaren Kurve auf $\Lambda(A)$, die durch X geht, eine Kurve $\mathcal{K}_{X,Y}$ existiert, die im Punkt X dieselbe Richtung besitzt.

Lemma 2.4
Sei $A \in \mathbb{C}^{n,n}$, $X \in \Lambda(A)$ und $M : \mathbb{R} \longrightarrow \Lambda(A)$ eine differenzierbare Kurve mit $M(0) = X$. Dann existiert $Y \in \mathcal{S}(n)$ mit

$$\left.\frac{d}{dt}\mathcal{K}_{X,Y}(t)\right|_{t=0} = \dot{M}(0). \tag{2.22}$$

Beweis:
Sei $U : \mathbb{R} \longrightarrow \mathcal{U}(n)$ mit $U(0) = I_n$ und $M(t) = U(t)XU^*(t)$ für alle $t \in \mathbb{R}$, sowie $Y := \dot{U}(0)$. Dann ist

$$Y + Y^* = \dot{U}(0)U^*(0) + U(0)\dot{U}^*(0) = \left.\frac{d}{dt}\big(U(t)U^*(t)\big)\right|_{t=0} = \left.\frac{d}{dt} I_n\right|_{t=0} = 0,$$

also $Y \in \mathcal{S}(n)$. Daraus folgt

$$\begin{aligned}
\dot{M}(0) &= \dot{U}(0)XU^*(0) + U(0)X\dot{U}^*(0)\\
&= YX + XY^*\\
&= YX - XY\\
&= Ye^{tY}Xe^{-tY} + e^{tY}Xe^{-tY}(-Y)\big|_{t=0}\\
&= \left.\frac{d}{dt}\big(e^{tY}Xe^{-tY}\big)\right|_{t=0}\\
&= \left.\frac{d}{dt}\mathcal{K}_{X,Y}(t)\right|_{t=0}.
\end{aligned}$$

□

Eine unmittelbare Folgerung aus diesem Lemma ist, daß $X \in \Lambda(A)$ genau dann ein stationärer Punkt von g_B ist, wenn für alle $Y \in \mathcal{S}(n)$

$$\frac{d}{dt} g_B(\mathcal{K}_{X,Y})(0) = 0 \tag{2.23}$$

gilt. Also berechnen wir als nächstes die Ableitung von $g_B(\mathcal{K}_{X,Y})$ im Nullpunkt.

Lemma 2.5
(i) *Seien $A, B \in \mathbb{C}^{n,n}$. Dann ist*

(2.24)
$$[H(A), H(B)] - [S(A), S(B)] = \frac{1}{2}(AB^* + A^*B - BA^* - B^*A) \in \mathcal{S}(n).$$

(ii) Seien $A, B \in \mathbb{C}^{n,n}$, $X \in \Lambda(A)$ und $Y \in \mathcal{S}(n)$. Dann ist

$$\frac{d}{dt} g_B(\mathcal{K}_{X,Y})(0) = \operatorname{tr}\big\{\big([H(X), H(B)] - [S(X), S(B)]\big)Y\big\}. \tag{2.25}$$

Beweis:

(i) Es ist

$$\begin{aligned}
&[H(A), H(B)] - [S(A), S(B)] \\
&= H(A)H(B) - H(B)H(A) - S(A)S(B) + S(B)S(A) \\
&= \frac{1}{2}(A + A^*)\frac{1}{2}(B + B^*) - \frac{1}{2}(B + B^*)\frac{1}{2}(A + A^*) \\
&\quad - \frac{1}{2}(A - A^*)\frac{1}{2}(B - B^*) + \frac{1}{2}(B - B^*)\frac{1}{2}(A - A^*) \\
&= \frac{1}{2}(AB^* + A^*B - BA^* - B^*A) \\
&= \frac{1}{2}\{(AB^* + A^*B) - (AB^* + A^*B)^*\} \\
&= S(AB^* + A^*B) \\
&\in \mathcal{S}(n).
\end{aligned}$$

(ii) Weiter ist

$$\begin{aligned}
&\frac{d}{dt} g_B(\mathcal{K}_{X,Y})(0) \\
&= \frac{d}{dt} g_B(e^{tY} X e^{-tY})\Big|_{t=0} \\
&= \frac{d}{dt} \operatorname{re}\{\operatorname{tr}(e^{tY} X e^{-tY} B^*)\}\Big|_{t=0} \\
&= \frac{d}{dt} \left\{\frac{1}{2} \operatorname{tr}\big(e^{tY} X e^{-tY} B^* + (e^{tY} X e^{-tY} B^*)^*\big)\right\}\Big|_{t=0} \\
&= \frac{1}{2} \operatorname{tr}\left\{\frac{d}{dt}(e^{tY} X e^{-tY} B^* + B e^{tY} X^* e^{-tY})\right\}\Big|_{t=0} \\
&= \frac{1}{2} \operatorname{tr}\{e^{tY}(YX - XY)e^{-tY} B^* + B e^{tY}(YX^* - X^*Y)e^{-tY}\}\Big|_{t=0} \\
&= \frac{1}{2} \operatorname{tr}(YXB^* - XYB^* + BYX^* - BX^*Y) \\
&= \operatorname{tr}\left\{\frac{1}{2}(XB^*Y + X^*BY - B^*XY - BX^*Y)\right\} \\
&= \operatorname{tr}\big\{\big([H(X), H(B)] - [S(X), S(B)]\big)Y\big\},
\end{aligned}$$

wobei die letzte Gleichheit aus (i) folgt. □

Mit Hilfe dieses Ergebnisses können wir die stationären Punkte von g_B charakterisieren.

Satz 2.6
Seien $A, B \in \mathbb{C}^{n,n}$. Dann ist $X \in \Lambda(A)$ genau dann ein stationärer Punkt von g_B, wenn

$$[H(X), H(B)] = [S(X), S(B)] \tag{2.26}$$

ist.

Beweis:
Sei $X \in \Lambda(A)$. Dann ist X genau dann stationärer Punkt von g_B, wenn für alle $Y \in \mathcal{S}(n)$ gilt

$$0 = \frac{d}{dt} g_B(\mathcal{K}_{X,Y})(0) \underset{(2.25)}{=} \operatorname{tr}\big\{\big([H(X), H(B)] - [S(X), S(B)]\big)Y\big\}.$$

Die Bedingung

$$[H(X), H(B)] = [S(X), S(B)]$$

ist sicherlich hinreichend dafür. Sie ist aber auch notwendig, da nach (2.24) $Y_0 := \big([H(X), H(B)] - [S(X), S(B)]\big)^* \in \mathcal{S}(n)$ ist, also insbesondere

$$\begin{aligned}
0 &= \frac{d}{dt} g_B(\mathcal{K}_{X,Y_0})(0) \\
&= \operatorname{tr}\big\{\big([H(X), H(B)] - [S(X), S(B)]\big)Y_0\big\} \\
&= \operatorname{tr}\big\{\big([H(X), H(B)] - [S(X), S(B)]\big)\big([H(X), H(B)] - [S(X), S(B)]\big)^*\big\} \\
&= \big\|[H(X), H(B)] - [S(X), S(B)]\big\|_F^2
\end{aligned}$$

gelten muß. □

Aus diesem Satz folgt sofort das Hauptergebnis der Untersuchung [1] von Bhatia, das wir jetzt mit Hilfe der in diesem Kapitel eingeführten Notation formulieren werden.

Korollar 2.7
Seien $A, B \in \mathbb{C}^{n,n}$ und $X \in \Lambda(A)$. Dann gilt

$$f_B^2(X) \in \left\{ \max_{Z \in \Lambda(A)} f_B^2(Z), \min_{Z \in \Lambda(A)} f_B^2(Z) \right\} \Longrightarrow [X, B^*] \in \mathcal{H}(n). \tag{2.27}$$

Beweis:
Sei $X \in \Lambda(A)$ eine Extremalstelle von f_B^2 auf $\Lambda(A)$. Dann ist X nach (2.10) auch eine Extremalstelle von g_B, also insbesondere ein stationärer Punkt. Somit folgt aus Satz 2.6, daß

$$[H(X), H(B)] = [S(X), S(B)]$$

ist, also nach (2.24)

$$\begin{aligned}
0 &= [H(X), H(B)] - [S(X), S(B)] \\
&= \frac{1}{2}(XB^* + X^*B - BX^* - B^*X) \\
&= \frac{1}{2}\big(XB^* - B^*X - (XB^* - B^*X)^*\big) \\
&= \frac{1}{2}\big([X, B^*] - [X, B^*]^*\big)
\end{aligned}$$

und $[X, B^*] \in \mathcal{H}(n)$. □

Wie aus dem Beweis dieses Korollars hervorgeht, ist die Bedingung (2.26) für das Vorliegen eines stationären Punktes von g_B sogar äquivalent zur Bedingung $[X, B^*] \in \mathcal{H}(n)$.

In den Spezialfällen, daß B hermitesch, schiefhermitesch oder diagonal ist, vereinfacht sich die Bedingung (2.26).

Korollar 2.8

(i) Sei $A \in \mathcal{H}(n)$ und $B \in \mathcal{S}(n)$ oder $A \in \mathcal{S}(n)$ und $B \in \mathcal{H}(n)$. Dann ist jedes $X \in \Lambda(A)$ ein stationärer Punkt von g_B.

(ii) Sei $A \in \mathbb{C}^{n,n}$ und $B \in \mathcal{H}(n)$. Dann ist $X \in \Lambda(A)$ genau dann ein stationärer Punkt von g_B, wenn

$$[H(X), B] = 0. \tag{2.28}$$

(iii) Sei $A \in \mathbb{C}^{n,n}$ und $B \in \mathcal{S}(n)$. Dann ist $X \in \Lambda(A)$ genau dann ein stationärer Punkt von g_B, wenn

$$[S(X), B] = 0. \tag{2.29}$$

(iv) Seien $A, B \in \mathcal{H}(n)$ oder $A, B \in \mathcal{S}(n)$. Dann ist $X \in \Lambda(A)$ genau dann ein stationärer Punkt von g_B, wenn

$$[X, B] = 0. \tag{2.30}$$

(v) Sei $A \in \mathbb{C}^{n,n}$ und $B = \mathrm{diag}(b_1, \ldots, b_n)$. Dann ist $X \in \Lambda(A)$ genau dann ein stationärer Punkt von g_B, wenn $H \in \mathcal{H}(n)$ und $Z \in \mathbb{C}^{n,n}$ existieren mit $z_{kl} = 0$ für $b_k \neq b_l$ und $X = Z + BH - HB$.

Beweis:
Nach dem letzten Satz ist $X \in \Lambda(A)$ genau dann ein stationärer Punkt von g_B, wenn $[H(X), H(B)] = [S(X), S(B)]$ ist.

(i) Sei $A \in \mathcal{H}(n)$ und $B \in \mathcal{S}(n)$. Dann ist $\Lambda(A) \subset \mathcal{H}(n)$ und für alle $X \in \Lambda(A)$ gilt

$$[H(X), H(B)] = [X, 0] = 0 = [0, B] = [S(X), S(B)].$$

Analog ist für $A \in \mathcal{S}(n)$ und $B \in \mathcal{H}(n)$ $\Lambda(A) \subset \mathcal{S}(n)$ und für alle $X \in \Lambda(A)$ ist

$$[H(X), H(B)] = [0, B] = 0 = [X, 0] = [S(X), S(B)].$$

(ii) Für $B \in \mathcal{H}(n)$ ist

$$[H(X), H(B)] = [H(X), B]$$

und

$$[S(X), S(B)] = [S(X), 0] = 0.$$

(iii) Für $B \in \mathcal{S}(n)$ ist

$$[H(X), H(B)] = [H(X), 0] = 0$$

und

$$[S(X), S(B)] = [S(X), B].$$

(iv) folgt unmittelbar aus (ii) und (iii).

(v) Es gilt

$$X = Z + BH - HB \iff x_{kl} = z_{kl} + h_{kl}(b_k - b_l) \text{ für alle } k,l \in \{1,\dots,n\}$$

und

$$\begin{aligned} & X \text{ ist stationärer Punkt von } g_B \\ \iff & 0 = [H(X),H(B)] - [S(X),S(B)] \\ & \underset{(2.24)}{=} \frac{1}{2}(XB^* + X^*B - BX^* - B^*X) \\ \iff & XB^* - B^*X = BX^* - X^*B \\ \iff & x_{kl}(\overline{b_l} - \overline{b_k}) = \overline{x_{lk}}(b_k - b_l) \text{ für alle } k,l \in \{1,\dots,n\}. \end{aligned}$$

Sei $X = Z + BH - HB$ mit $H \in \mathcal{H}(n)$ und $z_{kl} = 0$ für $b_k \neq b_l$. Dann ist

$$\begin{aligned} \overline{x_{lk}}(b_k - b_l) &= \overline{z_{lk}}(b_k - b_l) + \overline{h_{lk}}(\overline{b_l} - \overline{b_k})(b_k - b_l) \\ &= z_{kl}(\overline{b_l} - \overline{b_k}) + h_{kl}(b_k - b_l)(\overline{b_l} - \overline{b_k}) \\ &= x_{kl}(\overline{b_l} - \overline{b_k}), \end{aligned}$$

also X ein stationärer Punkt von g_B.
Sei andererseits X ein stationärer Punkt von g_B. Dann definiere H und Z durch

$$h_{kl} := \begin{cases} 0 & , \quad b_k = b_l \\ \frac{x_{kl}}{b_k - b_l} & , \quad b_k \neq b_l \end{cases} \qquad \text{und} \qquad z_{kl} := \begin{cases} x_{kl} & , \quad b_k = b_l \\ 0 & , \quad b_k \neq b_l \end{cases}.$$

Dann ist für $b_k = b_l$ $\overline{h_{lk}} = 0 = h_{kl}$ und für $b_k \neq b_l$

$$\overline{h_{lk}} = \frac{\overline{x_{lk}}}{\overline{b_l} - \overline{b_k}} = \frac{x_{kl}}{b_k - b_l} = h_{kl},$$

also $H \in \mathcal{H}(n)$, und

$$z_{kl} + h_{kl}(b_k - b_l) = \left\{ \begin{matrix} z_{kl} & , & b_k = b_l \\ 0 + \frac{x_{kl}}{b_k - b_l}(b_k - b_l) & , & b_k \neq b_l \end{matrix} \right\} = x_{kl},$$

also $X = Z + BH - HB$. □

Einige der im vorigen Korollar gefundenen Charakterisierungen der stationären Punkte von g_B erlauben es, diese explizit zu bestimmen. In den Fällen, in denen das möglich ist, können wir auch (2.5) und (2.6) lösen, da die Lösungen genau die stationären Punkte von g_B sind, für die g_B ein globales Optimum annimmt. Im folgenden Satz geben wir die Lösungen des isospektralen Abstandsproblems für diese Spezialfälle an.

Satz 2.9

(i) Sei $A \in \mathcal{H}(n)$ und $B \in \mathcal{S}(n)$ oder $A \in \mathcal{S}(n)$ und $B \in \mathcal{H}(n)$. Weiter seien λ_k, $k \in \{1,\dots,n\}$, die Eigenwerte von A und μ_k, $k \in \{1,\dots,n\}$, die Eigenwerte von B. Dann gilt

(2.31) $F^+(A,B) = F^-(A,B) = \sqrt{\|A\|_F^2 + \|B\|_F^2} = \sqrt{\sum_{k=1}^{n} |\lambda_k|^2 + \sum_{k=1}^{n} |\mu_k|^2}$,

(2.32) $G^+(A,B) = G^-(A,B) = 0$

und

(2.33) $\mathcal{F}^+(A,B) = \mathcal{F}^-(A,B) = \Lambda(A)$.

(ii) *Sei $A \in \mathbb{C}^{n,n}$ und $B \in \mathcal{H}(n)$. Weiter seien $\lambda_1 \leq \cdots \leq \lambda_n$ die Eigenwerte von $H(A)$, μ_k, $k \in \{1, \ldots, n\}$, die Eigenwerte von $S(A)$ und $\nu_1 \leq \cdots \leq \nu_n$ die Eigenwerte von B. Dann gilt*

$$F^+(A,B) = \sqrt{\sum_{k=1}^{n} (\lambda_{n+1-k} - \nu_k)^2 + \sum_{k=1}^{n} |\mu_k|^2}, \tag{2.34}$$

$$F^-(A,B) = \sqrt{\sum_{k=1}^{n} (\lambda_k - \nu_k)^2 + \sum_{k=1}^{n} |\mu_k|^2}, \tag{2.35}$$

$$G^+(A,B) = \sum_{k=1}^{n} \lambda_k \nu_k, \tag{2.36}$$

$$G^-(A,B) = \sum_{k=1}^{n} \lambda_{n+1-k} \nu_k. \tag{2.37}$$

Seien $V, W \in \mathcal{U}(n)$, $\rho_1 < \cdots < \rho_l$, $\sigma_1 < \cdots < \sigma_m$ und $r_1, \ldots, r_l, s_1, \ldots, s_m \in \mathbb{N}$ so, daß

$$H(A) = V \operatorname{diag}(\lambda_1, \ldots, \lambda_n) V^* = V \operatorname{diag}(\rho_1 I_{r_1}, \ldots, \rho_l I_{r_l}) V^*$$

und

$$B = W \operatorname{diag}(\nu_1, \ldots, \nu_n) W^* = W \operatorname{diag}(\sigma_1 I_{s_1}, \ldots, \sigma_m I_{s_m}) W^*.$$

Dann ist

$$\begin{aligned} \mathcal{F}^+(A,B) = \big\{ & W \operatorname{diag}(T_1, \ldots, T_m) P_n \operatorname{diag}(S_1, \ldots, S_l) V^* A V \\ & \cdot \operatorname{diag}(S_1, \ldots, S_l)^* P_n \operatorname{diag}(T_1, \ldots, T_m)^* W^* : \\ & S_j \in \mathcal{U}(r_j), T_k \in \mathcal{U}(s_k), j \in \{1, \ldots, l\}, k \in \{1, \ldots, m\} \big\} \end{aligned} \tag{2.38}$$

und

$$\begin{aligned} \mathcal{F}^-(A,B) = \big\{ & W \operatorname{diag}(T_1, \ldots, T_m) \operatorname{diag}(S_1, \ldots, S_l) V^* A V \\ & \cdot \operatorname{diag}(S_1, \ldots, S_l)^* \operatorname{diag}(T_1, \ldots, T_m)^* W^* : \\ & S_j \in \mathcal{U}(r_j), T_k \in \mathcal{U}(s_k), j \in \{1, \ldots, l\}, k \in \{1, \ldots, m\} \big\}. \end{aligned} \tag{2.39}$$

(iii) *Sei $A \in \mathbb{C}^{n,n}$ und $B \in \mathcal{S}(n)$. Weiter seien für $k \in \{1, \ldots, n\}$ λ_k die Eigenwerte von $H(A)$, μ_k die Eigenwerte von $S(A)$ und ν_k die Eigenwerte von B mit $\operatorname{im}(\mu_1) \leq \cdots \leq \operatorname{im}(\mu_n)$ und $\operatorname{im}(\nu_1) \leq \cdots \leq \operatorname{im}(\nu_n)$. Dann gilt*

$$F^+(A,B) = \sqrt{\sum_{k=1}^{n} \lambda_k^2 + \sum_{k=1}^{n} \big(\operatorname{im}(\mu_{n+1-k}) - \operatorname{im}(\nu_k)\big)^2}, \tag{2.40}$$

$$F^-(A,B) = \sqrt{\sum_{k=1}^{n} \lambda_k^2 + \sum_{k=1}^{n} \big(\operatorname{im}(\mu_k) - \operatorname{im}(\nu_k)\big)^2}, \tag{2.41}$$

$$G^+(A,B) = \sum_{k=1}^{n} \operatorname{im}(\mu_k) \operatorname{im}(\nu_k), \tag{2.42}$$

$$G^-(A,B) = \sum_{k=1}^{n} \operatorname{im}(\mu_{n+1-k}) \operatorname{im}(\nu_k). \tag{2.43}$$

Seien $V, W \in \mathcal{U}(n)$, $\rho_1 < \cdots < \rho_l$, $\sigma_1 < \cdots < \sigma_m$ *und* $r_1, \ldots, r_l, s_1, \ldots, s_m \in \mathbb{N}$ *so, daß*

$$S(A) = V \operatorname{diag}(\mu_1, \ldots, \mu_n) V^* = V \operatorname{diag}(\rho_1 i I_{r_1}, \ldots, \rho_l i I_{r_l}) V^*$$

und

$$B = W \operatorname{diag}(\nu_1, \ldots, \nu_n) W^* = W \operatorname{diag}(\sigma_1 i I_{s_1}, \ldots, \sigma_m i I_{s_m}) W^*.$$

Dann ist

$$\begin{aligned}(2.44)\quad \mathcal{F}^+(A,B) = \big\{ & W \operatorname{diag}(T_1, \ldots, T_m) P_n \operatorname{diag}(S_1, \ldots, S_l) V^* A V \\ & \cdot \operatorname{diag}(S_1, \ldots, S_l)^* P_n \operatorname{diag}(T_1, \ldots, T_m)^* W^* : \\ & S_j \in \mathcal{U}(r_j), T_k \in \mathcal{U}(s_k), j \in \{1, \ldots, l\}, k \in \{1, \ldots, m\} \big\}\end{aligned}$$

und

$$\begin{aligned}(2.45)\quad \mathcal{F}^-(A,B) = \big\{ & W \operatorname{diag}(T_1, \ldots, T_m) \operatorname{diag}(S_1, \ldots, S_l) V^* A V \\ & \cdot \operatorname{diag}(S_1, \ldots, S_l)^* \operatorname{diag}(T_1, \ldots, T_m)^* W^* : \\ & S_j \in \mathcal{U}(r_j), T_k \in \mathcal{U}(s_k), j \in \{1, \ldots, l\}, k \in \{1, \ldots, m\} \big\}.\end{aligned}$$

(iv) *Seien* $A, B \in \mathcal{H}(n)$ *und* $\lambda_1 \leq \cdots \leq \lambda_n$ *die Eigenwerte von* A *und* $\nu_1 \leq \cdots \leq \nu_n$ *die Eigenwerte von* B. *Dann gilt*

$$(2.46)\qquad F^+(A,B) = \sqrt{\sum_{k=1}^{n} (\lambda_{n+1-k} - \nu_k)^2},$$

$$(2.47)\qquad F^-(A,B) = \sqrt{\sum_{k=1}^{n} (\lambda_k - \nu_k)^2},$$

$$(2.48)\qquad G^+(A,B) = \sum_{k=1}^{n} \lambda_k \nu_k,$$

$$(2.49)\qquad G^-(A,B) = \sum_{k=1}^{n} \lambda_{n+1-k} \nu_k.$$

Sei $W \in \mathcal{U}(n)$, $\sigma_1 < \cdots < \sigma_m$ *und* $s_1, \ldots, s_m \in \mathbb{N}$ *so, daß*

$$B = W \operatorname{diag}(\nu_1, \ldots, \nu_n) W^* = W \operatorname{diag}(\sigma_1 I_{s_1}, \ldots, \sigma_m I_{s_m}) W^*.$$

Dann ist

(2.50)
$$\begin{aligned}\mathcal{F}^+(A,B) = \big\{ & W \operatorname{diag}(T_1, \ldots, T_m) \operatorname{diag}(\lambda_n, \ldots, \lambda_1) \operatorname{diag}(T_1, \ldots, T_m)^* W^* : \\ & T_k \in \mathcal{U}(s_k), k \in \{1, \ldots, m\} \big\}\end{aligned}$$

und

(2.51)
$$\begin{aligned}\mathcal{F}^-(A,B) = \big\{ & W \operatorname{diag}(T_1, \ldots, T_m) \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \operatorname{diag}(T_1, \ldots, T_m)^* W^* : \\ & T_k \in \mathcal{U}(s_k), k \in \{1, \ldots, m\} \big\}.\end{aligned}$$

(v) *Seien $A, B \in \mathcal{S}(n)$ und für $k \in \{1, \dots, n\}$ μ_k die Eigenwerte von A und ν_k die Eigenwerte von B mit $\operatorname{im}(\mu_1) \leq \dots \leq \operatorname{im}(\mu_n)$ und $\operatorname{im}(\nu_1) \leq \dots \leq \operatorname{im}(\nu_n)$. Dann gilt*

$$F^+(A, B) = \sqrt{\sum_{k=1}^{n} \big(\operatorname{im}(\mu_{n+1-k}) - \operatorname{im}(\nu_k)\big)^2}, \tag{2.52}$$

$$F^-(A, B) = \sqrt{\sum_{k=1}^{n} \big(\operatorname{im}(\mu_k) - \operatorname{im}(\nu_k)\big)^2}, \tag{2.53}$$

$$G^+(A, B) = \sum_{k=1}^{n} \operatorname{im}(\mu_k) \operatorname{im}(\nu_k), \tag{2.54}$$

$$G^-(A, B) = \sum_{k=1}^{n} \operatorname{im}(\mu_{n+1-k}) \operatorname{im}(\nu_k). \tag{2.55}$$

Sei $W \in \mathcal{U}(n)$, $\sigma_1 < \dots < \sigma_m$ und $s_1, \dots, s_m \in \mathbb{N}$ so, daß

$$B = W \operatorname{diag}(\nu_1, \dots, \nu_n) W^* = W \operatorname{diag}(\sigma_1 i I_{s_1}, \dots, \sigma_m i I_{s_m}) W^*.$$

Dann ist

(2.56)
$$\mathcal{F}^+(A, B) = \big\{W \operatorname{diag}(T_1, \dots, T_m) \operatorname{diag}(\mu_n, \dots, \mu_1) \operatorname{diag}(T_1, \dots, T_m)^* W^* : \\ T_k \in \mathcal{U}(s_k), k \in \{1, \dots, m\}\big\}$$

und

(2.57)
$$\mathcal{F}^-(A, B) = \{W \operatorname{diag}(T_1, \dots, T_m) \operatorname{diag}(\mu_1, \dots, \mu_n) \operatorname{diag}(T_1, \dots, T_m)^* W^* : \\ T_k \in \mathcal{U}(s_k), k \in \{1, \dots, m\}\}.$$

(vi) *Seien $A, B \in \mathcal{N}(n)$ und für $k \in \{1, \dots, n\}$ λ_k die Eigenwerte von A und μ_k die Eigenwerte von B. Dann gilt*

$$F^+(A, B) = \max_{\pi \in \Pi(n)} \sqrt{\sum_{k=1}^{n} |\lambda_{\pi(k)} - \mu_k|^2}, \tag{2.58}$$

$$F^-(A, B) = \min_{\pi \in \Pi(n)} \sqrt{\sum_{k=1}^{n} |\lambda_{\pi(k)} - \mu_k|^2}, \tag{2.59}$$

$$G^+(A, B) = \max_{\pi \in \Pi(n)} \sum_{k=1}^{n} \operatorname{re}(\lambda_{\pi(k)} \overline{\mu_k}), \tag{2.60}$$

$$G^-(A, B) = \min_{\pi \in \Pi(n)} \sum_{k=1}^{n} \operatorname{re}(\lambda_{\pi(k)} \overline{\mu_k}). \tag{2.61}$$

Beweis:

(i) Nach Korollar 2.8 (i) ist jedes $X \in \Lambda(A)$ ein stationärer Punkt von g_B, also ist g_B auf $\Lambda(A)$ konstant. Daraus folgt (2.33).
Sei $X_0 := \operatorname{diag}(\lambda_1, \dots, \lambda_n) \in \Lambda(A)$. Dann gilt für alle $X \in \Lambda(A)$

$$g_B(X) = g_B(X_0) = \operatorname{re}\{\operatorname{tr}(X_0 B^*)\} = \operatorname{re}\left(\sum_{k=1}^{n} \lambda_k \overline{b_{kk}}\right) = 0,$$

wobei sich die letzte Gleichheit daraus ergibt, daß für $A \in \mathcal{H}(n)$ und $B \in \mathcal{S}(n)$ $\lambda_k \in \mathbb{R}$ und $b_{kk} \in i\mathbb{R}$ ist und für $A \in \mathcal{S}(n)$ und $B \in \mathcal{H}(n)$ $\lambda_k \in i\mathbb{R}$ und $b_{kk} \in \mathbb{R}$. Daraus folgt (2.32). Weiter ist für alle $X \in \Lambda(A)$ nach (2.10)

$$\begin{aligned} f_B(X) &= \sqrt{\|A\|_F^2 + \|B\|_F^2 - 2g_B(X)} \\ &= \sqrt{\|A\|_F^2 + \|B\|_F^2} \\ &= \sqrt{\|\operatorname{diag}(\lambda_1, \ldots, \lambda_n)\|_F^2 + \|\operatorname{diag}(\mu_1, \ldots, \mu_n)\|_F^2} \\ &= \sqrt{\sum_{k=1}^n |\lambda_k|^2 + \sum_{k=1}^n |\mu_k|^2}, \end{aligned}$$

woraus (2.31) folgt. Insbesondere ist für $Y \in \mathbb{C}^{n,n}$

$$\begin{aligned} \|Y\|_F^2 &= \|H(Y) + S(Y)\|_F^2 = f_{H(Y)}\big(-S(Y)\big)^2 \\ &= \|H(Y)\|_F^2 + \|S(Y)\|_F^2. \end{aligned} \tag{2.62}$$

(ii) Sei

$$Y : \Lambda(A) \longrightarrow \Lambda\big(H(A)\big),\ Y(X) = W^*H(X)W.$$

Dann ist für $X \in \Lambda(A)$

$$\begin{aligned} Y(X)\operatorname{diag}(\nu_1, \ldots, \nu_n) &= W^*H(X)W\operatorname{diag}(\nu_1, \ldots, \nu_n)W^*W \\ &= W^*H(X)BW \end{aligned}$$

und

$$\begin{aligned} \operatorname{diag}(\nu_1, \ldots, \nu_n)Y(X) &= W^*W\operatorname{diag}(\nu_1, \ldots, \nu_n)W^*H(X)W \\ &= W^*BH(X)W. \end{aligned}$$

Also gilt nach Korollar 2.8 (ii)

$$\begin{aligned} & X \text{ ist stationärer Punkt von } g_B \\ \Longleftrightarrow\ & 0 = [H(X), B] = H(X)B - BH(X) \\ \Longleftrightarrow\ & 0 = W^*\{H(X)B - BH(X)\}W \\ & \quad = Y(X)\operatorname{diag}(\nu_1, \ldots, \nu_n) - \operatorname{diag}(\nu_1, \ldots, \nu_n)Y(X) \\ \Longleftrightarrow\ & Y(X)\operatorname{diag}(\nu_1, \ldots, \nu_n) = \operatorname{diag}(\nu_1, \ldots, \nu_n)Y(X) \\ \Longleftrightarrow\ & y_{jk}(X)\nu_k = \nu_j y_{jk}(X) \text{ für alle } j, k \in \{1, \ldots, n\} \\ \Longleftrightarrow\ & Y(X) = \operatorname{diag}(Y_1, \ldots, Y_m) \text{ mit } Y_j \in \mathcal{H}(s_j) \text{ für } j \in \{1, \ldots, m\} \\ \Longleftrightarrow\ & Y(X) = \operatorname{diag}(T_1, \ldots, T_m)\operatorname{diag}(\lambda_{\pi(1)}, \ldots, \lambda_{\pi(n)})\operatorname{diag}(T_1, \ldots, T_m)^* \\ & \text{mit } T_j \in \mathcal{U}(s_j) \text{ für } j \in \{1, \ldots, m\} \text{ und } \pi \in \Pi(n). \end{aligned}$$

Setze $t_j := \sum_{k=1}^j s_k$ für $j \in \{0, \ldots, m\}$. Dann ist für stationäre Punkte $X \in \Lambda(A)$ von g_B

$$\begin{aligned} g_B(X) &= \operatorname{re}\{\operatorname{tr}(XB^*)\} \\ &= \operatorname{re}\big\{\operatorname{tr}\big(H(X)B^*\big)\big\} + \underbrace{\operatorname{re}\big\{\operatorname{tr}\big(S(X)B^*\big)\big\}}_{=0 \text{ nach (i)}} \\ &= \operatorname{re}\big\{\operatorname{tr}\big(H(X)B\big)\big\} \end{aligned}$$

$$
\begin{aligned}
&= \operatorname{re}\bigl\{\operatorname{tr}\bigl(WY(X)W^*W \operatorname{diag}(\sigma_1 I_{s_1}, \dots, \sigma_m I_{s_m})W^*\bigr)\bigr\} \\
&= \operatorname{re}\bigl\{\operatorname{tr}\bigl(Y(X) \operatorname{diag}(\sigma_1 I_{s_1}, \dots, \sigma_m I_{s_m})\bigr)\bigr\} \\
&= \operatorname{re}\bigl\{\operatorname{tr}\bigl(\operatorname{diag}(T_1(X), \dots, T_m(X)) \operatorname{diag}(\lambda_{\pi(1)}, \dots, \lambda_{\pi(n)}) \\
&\qquad \cdot \operatorname{diag}(T_1(X), \dots, T_m(X))^* \operatorname{diag}(\sigma_1 I_{s_1}, \dots, \sigma_m I_{s_m})\bigr)\bigr\} \\
&= \operatorname{re}\left\{\sum_{j=1}^{m} \operatorname{tr}\bigl(T_j(X) \operatorname{diag}(\lambda_{\pi(t_{j-1}+1)}, \dots, \lambda_{\pi(t_j)})T_j^*(X)\sigma_j I_{s_j}\bigr)\right\} \\
&= \operatorname{re}\left\{\sum_{j=1}^{m} \operatorname{tr}\bigl(\sigma_j \operatorname{diag}(\lambda_{\pi(t_{j-1}+1)}, \dots, \lambda_{\pi(t_j)})\bigr)\right\} \\
&= \sum_{k=1}^{n} \lambda_{\pi(k)}\nu_k
\end{aligned}
$$

und nach (1.19) ist

$$
\begin{aligned}
G^+(A,B) &= \max\{g_B(X) : X \in \Lambda(A)\} \\
&= \max\{g_B(X) : X \text{ ist stationärer Punkt von } g_B\} \\
&= \max_{\pi \in \Pi(n)} \sum_{k=1}^{n} \lambda_{\pi(k)}\nu_k \\
&= \sum_{k=1}^{n} \lambda_k \nu_k,
\end{aligned}
$$

sowie nach (1.20)

$$
\begin{aligned}
G^-(A,B) &= \min\{g_B(X) : X \in \Lambda(A)\} \\
&= \min_{\pi \in \Pi(n)} \sum_{k=1}^{n} \lambda_{\pi(k)}\nu_k \\
&= \sum_{k=1}^{n} \lambda_{n+1-k}\nu_k.
\end{aligned}
$$

Weiter ist nach (2.62)

$$
\begin{aligned}
F^+(A,B) &= \sqrt{\|A\|_F^2 + \|B\|_F^2 - 2G^-(A,B)} \\
&= \sqrt{\|H(A)\|_F^2 + \|S(A)\|_F^2 + \|B\|_F^2 - 2\sum_{k=1}^{n} \lambda_{n+1-k}\nu_k} \\
&= \Bigl(\|\operatorname{diag}(\lambda_1, \dots, \lambda_n)\|_F^2 + \|\operatorname{diag}(\mu_1, \dots, \mu_n)\|_F^2 \\
&\qquad + \|\operatorname{diag}(\nu_1, \dots, \nu_n)\|_F^2 - 2\sum_{k=1}^{n} \lambda_{n+1-k}\nu_k\Bigr)^{\frac{1}{2}} \\
&= \sqrt{\sum_{k=1}^{n} \lambda_{n+1-k}^2 + \sum_{k=1}^{n} |\mu_k|^2 + \sum_{k=1}^{n} \nu_k^2 - 2\sum_{k=1}^{n} \lambda_{n+1-k}\nu_k} \\
&= \sqrt{\sum_{k=1}^{n} (\lambda_{n+1-k} - \nu_k)^2 + \sum_{k=1}^{n} |\mu_k|^2}
\end{aligned}
$$

und nach (2.12)

$$\begin{aligned}
F^-(A,B) &= \sqrt{\|A\|_F^2 + \|B\|_F^2 - 2G^+(A,B)} \\
&= \sqrt{\sum_{k=1}^n \lambda_k^2 + \sum_{k=1}^n |\mu_k|^2 + \sum_{k=1}^n \nu_k^2 - 2\sum_{k=1}^n \lambda_k\nu_k} \\
&= \sqrt{\sum_{k=1}^n (\lambda_k - \nu_k)^2 + \sum_{k=1}^n |\mu_k|^2}.
\end{aligned}$$

Für stationäre Punkte $X \in \Lambda(A)$ gilt nach Lemma 1.7 außerdem

$$\begin{aligned}
& g_B(X) = \sum_{k=1}^n \lambda_{\pi(k)}\nu_k = \min_{\tau\in\Pi(n)} \sum_{k=1}^n \lambda_{\tau(k)}\nu_k = G^-(A,B) \\
\Longleftrightarrow\ & \text{für alle } j \in \{1,\dots,m\} \text{ ist} \\
& \{\lambda_{\pi(k)} : k \in [t_{j-1}+1, t_j]\} = \{\lambda_{n+1-k} : k \in [t_{j-1}+1, t_j]\} \\
\Longleftrightarrow\ & Y(X) = \operatorname{diag}(T_1,\dots,T_m)\operatorname{diag}(\lambda_n,\dots,\lambda_1)\operatorname{diag}(T_1,\dots,T_m)^* \\
& \text{mit } T_j \in \mathcal{U}(s_j) \text{ für } j \in \{1,\dots,m\} \\
\Longleftrightarrow\ & H(X) = W\operatorname{diag}(T_1,\dots,T_m)P_n\operatorname{diag}(\lambda_1,\dots,\lambda_n)P_n\operatorname{diag}(T_1,\dots,T_m)^*W^* \\
& \text{mit } T_j \in \mathcal{U}(s_j) \text{ für } j \in \{1,\dots,m\}.
\end{aligned}$$

Jetzt müssen die $X = UAU^* \in \Lambda(A)$ bestimmt werden, für die $H(X)$ die passende Form hat. Es gilt

$$\begin{aligned}
& UV\operatorname{diag}(\rho_1 I_{r_1},\dots,\rho_l I_{r_l})V^*U^* \\
&= UH(A)U^* \\
&= H(UAU^*) \\
&= H(X) \\
&= W\operatorname{diag}(T_1,\dots,T_m)P_n\operatorname{diag}(\rho_1 I_{r_1},\dots,\rho_l I_{r_l})P_n\operatorname{diag}(T_1,\dots,T_m)^*W^* \\
&\quad \text{mit } T_j \in \mathcal{U}(s_j) \text{ für } j \in \{1,\dots,m\} \\
\Longleftrightarrow\ & P_n\operatorname{diag}(T_1,\dots,T_m)^*W^*UV\operatorname{diag}(\rho_1 I_{r_1},\dots,\rho_l I_{r_l}) \\
& = \operatorname{diag}(\rho_1 I_{r_1},\dots,\rho_l I_{r_l})P_n\operatorname{diag}(T_1,\dots,T_m)^*W^*UV \\
\Longleftrightarrow\ & P_n\operatorname{diag}(T_1,\dots,T_m)^*W^*UV = \operatorname{diag}(S_1,\dots,S_l) \\
& \text{mit } S_k \in \mathcal{U}(r_k) \text{ für } k \in \{1,\dots,l\} \\
\Longleftrightarrow\ & U = W\operatorname{diag}(T_1,\dots,T_m)P_n\operatorname{diag}(S_1,\dots,S_l)V^* \\
& \text{mit } S_k \in \mathcal{U}(r_k) \text{ für } k \in \{1,\dots,l\}.
\end{aligned}$$

Daraus folgt (2.38). Ebenso gilt nach Lemma 1.7

$$\begin{aligned}
& g_B(X) = \sum_{k=1}^n \lambda_{\pi(k)}\nu_k = \max_{\tau\in\Pi(n)} \sum_{k=1}^n \lambda_{\tau(k)}\nu_k = G^+(A,B) \\
\Longleftrightarrow\ & \text{für alle } j \in \{1,\dots,m\} \text{ ist} \\
& \{\lambda_{\pi(k)} : k \in [t_{j-1}+1, t_j]\} = \{\lambda_k : k \in [t_{j-1}+1, t_j]\} \\
\Longleftrightarrow\ & H(X) = W\operatorname{diag}(T_1,\dots,T_m)\operatorname{diag}(\lambda_1,\dots,\lambda_n)\operatorname{diag}(T_1,\dots,T_m)^*W^* \\
& \text{mit } T_j \in \mathcal{U}(s_j) \text{ für } j \in \{1,\dots,m\},
\end{aligned}$$

woraus (2.39) folgt.

(iii) Analog zu (ii). Sei

$$Y : \Lambda(A) \longrightarrow \Lambda\big(S(A)\big),\ Y(X) = W^* S(X) W.$$

Dann ist für $X \in \Lambda(A)$

$$Y(X)\operatorname{diag}(\nu_1, \ldots, \nu_n) = W^* S(X) B W$$

und

$$\operatorname{diag}(\nu_1, \ldots, \nu_n) Y(X) = W^* B S(X) W.$$

Also gilt nach Korollar 2.8 (iii)

$$\begin{aligned}
& X \text{ ist stationärer Punkt von } g_B \\
\Longleftrightarrow\ & 0 = [S(X), B] = S(X)B - BS(X) \\
\Longleftrightarrow\ & Y(X)\operatorname{diag}(\nu_1, \ldots, \nu_n) = \operatorname{diag}(\nu_1, \ldots, \nu_n) Y(X) \\
\Longleftrightarrow\ & Y(X) = \operatorname{diag}(Y_1, \ldots, Y_m) \\
& \text{mit } Y_j \in \mathcal{S}(s_j) \text{ für } j \in \{1, \ldots, m\} \\
\Longleftrightarrow\ & \mathrm{Y}(X) = \operatorname{diag}(T_1, \ldots, T_m)\operatorname{diag}(\mu_{\pi(1)}, \ldots, \mu_{\pi(n)})\operatorname{diag}(T_1, \ldots, T_m)^* \\
& \text{mit } T_j \in \mathcal{U}(s_j) \text{ für } j \in \{1, \ldots, m\} \text{ und } \pi \in \Pi(n),
\end{aligned}$$

und mit $t_j := \sum_{k=1}^{j} s_k$ für $j \in \{0, \ldots, m\}$ ist für stationäre Punkte $X \in \Lambda(A)$ von g_B

$$\begin{aligned}
g_B(X) &= \operatorname{re}\{\operatorname{tr}(XB^*)\} \\
&= \underbrace{\operatorname{re}\big\{\operatorname{tr}\big(H(X)B^*\big)\big\}}_{=0 \text{ nach (i)}} + \operatorname{re}\big\{\operatorname{tr}\big(S(X)B^*\big)\big\} \\
&= \operatorname{re}\big\{\operatorname{tr}\big(-S(X)B\big)\big\} \\
&= -\operatorname{re}\big\{\operatorname{tr}\big(Y(X)\operatorname{diag}(\sigma_1 i I_{s_1}, \ldots, \sigma_m i I_{s_m})\big)\big\} \\
&= -\operatorname{re}\left\{\sum_{j=1}^{m} \operatorname{tr}\big(\sigma_j i \operatorname{diag}(\mu_{\pi(t_{j-1}+1)}, \ldots, \mu_{\pi(t_j)})\big)\right\} \\
&= -\operatorname{re}\left\{\sum_{k=1}^{n} \big(-\operatorname{im}(\mu_{\pi(k)})\operatorname{im}(\nu_k)\big)\right\} \\
&= \sum_{k=1}^{n} \operatorname{im}(\mu_{\pi(k)})\operatorname{im}(\nu_k).
\end{aligned}$$

Der Rest ist vollkommen analog zu (ii).

(iv) In diesem Fall ist $H(A) = A$ und $S(A) = 0$, also folgen (2.46) bis (2.49) direkt aus (ii). Weiter ist mit der Notation aus (ii) für $S_j \in \mathcal{U}(r_j)$, $j \in \{1, \ldots, l\}$,

$$\begin{aligned}
& \operatorname{diag}(S_1, \ldots, S_l) V^* A V \operatorname{diag}(S_1, \ldots, S_l)^* \\
=\ & \operatorname{diag}(S_1, \ldots, S_l) V^* H(A) V \operatorname{diag}(S_1, \ldots, S_l)^* \\
=\ & \operatorname{diag}(S_1, \ldots, S_l)\operatorname{diag}(\rho_1 I_{r_1}, \ldots, \rho_l I_{r_l})\operatorname{diag}(S_1, \ldots, S_l)^* \\
=\ & \operatorname{diag}(S_1 \rho_1 I_{r_1} S_1^*, \ldots, S_l \rho_l I_{r_l} S_l^*) \\
=\ & \operatorname{diag}(\lambda_1, \ldots, \lambda_n),
\end{aligned}$$

woraus mit Hilfe von (2.39) (2.51) folgt, und

$$\begin{aligned}
P_n \operatorname{diag}(S_1, \ldots, S_l) V^* A V \operatorname{diag}(S_1, \ldots, S_l)^* P_n &= P_n \operatorname{diag}(\lambda_1, \ldots, \lambda_n) P_n \\
&= \operatorname{diag}(\lambda_n, \ldots, \lambda_1),
\end{aligned}$$

woraus mit Hilfe von (2.38) (2.50) folgt.

(v) folgt aus (iii) ebenso, wie (iv) aus (ii) folgt.

(vi) Ein Großteil dieser Aussagen wurde bereits bei verschiedenen Beweisen des Satzes von Hoffman-Wielandt gezeigt, so etwa (2.58) und (2.59) in [6] und (2.60) in [18], Abschnitt 6.3.5. Da diese Beweise auf einer vollkommen anderen Idee beruhen als die Beweise der übrigen Teile dieses Satzes, werden wir sie an dieser Stelle noch einmal ausführlich darstellen.
Sei $W \in \mathcal{U}(n)$ mit $B = W \operatorname{diag}(\mu_1, \ldots, \mu_n) W^*$ und $Y : \Lambda(A) \longrightarrow \mathcal{U}(n)$ so, daß $Y(X) \operatorname{diag}(\lambda_1, \ldots, \lambda_n) Y(X)^* = W^* X W$ für alle $X \in \Lambda(A)$ ist. Dann ist

$$\begin{aligned} g_B(X) &= \operatorname{re}\{\operatorname{tr}(XB^*)\} \\ &= \operatorname{re}\big\{\operatorname{tr}\big(W^*XW \operatorname{diag}(\mu_1, \ldots, \mu_n)^*\big)\big\} \\ &= \operatorname{re}\big\{\operatorname{tr}\big(Y(X) \operatorname{diag}(\lambda_1, \ldots, \lambda_n) Y(X)^* \operatorname{diag}(\overline{\mu_1}, \ldots, \overline{\mu_n})\big)\big\} \\ &= \operatorname{re}\left\{\sum_{k=1}^{n} [Y(X) \operatorname{diag}(\lambda_1, \ldots, \lambda_n) Y(X)^*]_{kk} \overline{\mu_k}\right\} \\ &= \sum_{k=1}^{n} \operatorname{re}\left(\sum_{l=1}^{n} y_{kl}(X) \lambda_l \overline{y_{kl}(X) \mu_k}\right) \\ &= \sum_{k=1}^{n} \sum_{l=1}^{n} |y_{kl}(X)|^2 \operatorname{re}(\lambda_l \overline{\mu_k}). \end{aligned}$$

Wegen $Y(X)Y(X)^* = I_n = Y(X)^*Y(X)$ ist für alle $k, l \in \{1, \ldots, n\}$

$$\sum_{j=1}^{n} |y_{kj}(X)|^2 = \sum_{j=1}^{n} |y_{jl}(X)|^2 = 1,$$

also die Matrix $\big(|y_{kl}(X)|^2\big)_{k,l=1,\ldots,n}$ doppelt stochastisch. Somit ist

$$\begin{aligned} G^+(A, B) &= \max\{g_B(X) : X \in \Lambda(A)\} \\ &= \max\left\{\sum_{k=1}^{n} \sum_{l=1}^{n} |y_{kl}|^2 \operatorname{re}(\lambda_l \overline{\mu_k}) : Y \in \mathcal{U}(n)\right\} \\ &\leq \max\left\{\sum_{k=1}^{n} \sum_{l=1}^{n} c_{kl} \operatorname{re}(\lambda_l \overline{\mu_k}) : C \text{ ist doppelt stochastisch}\right\}. \end{aligned}$$

Nach Satz 1.5 bilden die doppelt stochastischen Matrizen eine konvexe Menge, deren Extremalpunkte die Permutationsmatrizen sind. Also nimmt die lineare Funktion

$$C \longmapsto \sum_{k=1}^{n} \sum_{l=1}^{n} c_{kl} \operatorname{re}(\lambda_l \overline{\mu_k})$$

ihre Extremalwerte für Permutationsmatrizen an. Da Permutationsmatrizen insbesondere unitär sind, folgt

$$\begin{aligned} G^+(A, B) &= \max\left\{\sum_{k=1}^{n} \sum_{l=1}^{n} p_{kl} \operatorname{re}(\lambda_l \overline{\mu_k}) : P \text{ ist Permutationsmatrix}\right\} \\ &= \max_{\pi \in \Pi(n)} \sum_{k=1}^{n} \operatorname{re}(\lambda_{\pi(k)} \overline{\mu_k}). \end{aligned}$$

Analog ist

$$
\begin{aligned}
G^-(A,B) &= \min\{g_B(X) : X \in \Lambda(A)\} \\
&= \min\left\{\sum_{k=1}^{n}\sum_{l=1}^{n} p_{kl}\,\mathrm{re}(\lambda_l\overline{\mu_k}) : P \text{ ist Permutationsmatrix}\right\} \\
&= \min_{\pi\in\Pi(n)} \sum_{k=1}^{n} \mathrm{re}(\lambda_{\pi(k)}\overline{\mu_k}).
\end{aligned}
$$

Weiter ist nach (2.11)

$$
\begin{aligned}
F^+(A,B)^2 &= \|A\|_F^2 + \|B\|_F^2 - 2G^-(A,B) \\
&= \|\operatorname{diag}(\lambda_1,\dots,\lambda_n)\|_F^2 + \|\operatorname{diag}(\mu_1,\dots,\mu_n)\|_F^2 \\
&\quad - 2\min_{\pi\in\Pi(n)} \sum_{k=1}^{n} \mathrm{re}(\lambda_{\pi(k)}\overline{\mu_k}) \\
&= \max_{\pi\in\Pi(n)} \sum_{k=1}^{n} \left\{|\lambda_{\pi(k)}|^2 + |\mu_k|^2 - 2\,\mathrm{re}(\lambda_{\pi(k)}\overline{\mu_k})\right\} \\
&= \max_{\pi\in\Pi(n)} \sum_{k=1}^{n} (\lambda_{\pi(k)} - \mu_k)(\overline{\lambda_{\pi(k)}} - \overline{\mu_k}) \\
&= \max_{\pi\in\Pi(n)} \sum_{k=1}^{n} |\lambda_{\pi(k)} - \mu_k|^2,
\end{aligned}
$$

woraus (2.58) folgt, und analog nach (2.12)

$$
\begin{aligned}
F^-(A,B)^2 &= \|A\|_F^2 + \|B\|_F^2 - 2G^+(A,B) \\
&= \min_{\pi\in\Pi(n)} \sum_{k=1}^{n} |\lambda_{\pi(k)} - \mu_k|^2,
\end{aligned}
$$

woraus (2.59) folgt. □

Wie wir im vorigen Satz gesehen haben, sind die Lösungen des isospektralen Maximierungsproblems (2.5) und des isospektralen Minimierungsproblems (2.6) im allgemeinen nicht eindeutig, das heißt $\mathcal{F}^+(A,B)$ und $\mathcal{F}^-(A,B)$ enthalten mehr als ein Element. Im folgenden geben wir für einige Spezialfälle Bedingungen an, die Eindeutigkeit garantieren.

Korollar 2.10

(i) Sei $A \in \mathcal{H}(n)$ und $B \in \mathcal{S}(n)$ oder $A \in \mathcal{S}(n)$ und $B \in \mathcal{H}(n)$. Dann sind (2.5) und (2.6) genau dann eindeutig lösbar, wenn A ein Vielfaches der Einheitsmatrix ist.

(ii) Sei $A \in \mathcal{N}(n)$ und $B \in \mathcal{H}(n)$. Gilt für die Eigenwerte κ_j, $j \in \{1,\dots,n\}$, von A $\mathrm{re}(\kappa_1) < \cdots < \mathrm{re}(\kappa_n)$ und sind $\nu_1 < \cdots < \nu_n$ die Eigenwerte von B, sowie $W \in \mathcal{U}(n)$ mit $B = W\operatorname{diag}(\nu_1,\dots,\nu_n)W^*$, so ist

(2.63) $$\mathcal{F}^+(A,B) = \left\{W\operatorname{diag}(\kappa_n,\dots,\kappa_1)W^*\right\}$$

und

(2.64) $$\mathcal{F}^-(A,B) = \left\{W\operatorname{diag}(\kappa_1,\dots,\kappa_n)W^*\right\}.$$

(iii) Sei $A \in \mathcal{N}(n)$ und $B \in \mathcal{S}(n)$. Gilt für die Eigenwerte κ_j, $j \in \{1,\dots,n\}$, von A $\mathrm{im}(\kappa_1) < \cdots < \mathrm{im}(\kappa_n)$ und für die Eigenwerte ν_j, $j \in \{1,\dots,n\}$, von B $\mathrm{im}(\nu_1) < \cdots < \mathrm{im}(\nu_n)$, so ist

$$\mathcal{F}^+(A,B) = \{W \operatorname{diag}(\kappa_n,\dots,\kappa_1)W^*\} \tag{2.65}$$

und

$$\mathcal{F}^-(A,B) = \{W \operatorname{diag}(\kappa_1,\dots,\kappa_n)W^*\}, \tag{2.66}$$

wobei $W \in \mathcal{U}(n)$ ist mit $B = W \operatorname{diag}(\nu_1,\dots,\nu_n)W^*$.

(iv) Seien $A, B \in \mathcal{H}(n)$ und $\lambda_1 \le \cdots \le \lambda_n$ die Eigenwerte von A. Sind die Eigenwerte von B $\nu_1 < \cdots < \nu_n$ und ist $W \in \mathcal{U}(n)$ mit $B = W \operatorname{diag}(\nu_1,\dots,\nu_n)W^*$, so ist

$$\mathcal{F}^+(A,B) = \{W \operatorname{diag}(\lambda_n,\dots,\lambda_1)W^*\} \tag{2.67}$$

und

$$\mathcal{F}^-(A,B) = \{W \operatorname{diag}(\lambda_1,\dots,\lambda_n)W^*\}. \tag{2.68}$$

(v) Seien $A, B \in \mathcal{S}(n)$ und μ_k, $k \in \{1,\dots,n\}$, mit $\mathrm{im}(\mu_1) \le \cdots \le \mathrm{im}(\mu_n)$ die Eigenwerte von A. Gilt für die Eigenwerte ν_k, $k \in \{1,\dots,n\}$, von B $\mathrm{im}(\nu_1) < \cdots < \mathrm{im}(\nu_n)$ und ist $W \in \mathcal{U}(n)$ mit $B = W \operatorname{diag}(\nu_1,\dots,\nu_n)W^*$, so ist

$$\mathcal{F}^+(A,B) = \{W \operatorname{diag}(\mu_n,\dots,\mu_1)W^*\} \tag{2.69}$$

und

$$\mathcal{F}^-(A,B) = \{W \operatorname{diag}(\mu_1,\dots,\mu_n)W^*\}. \tag{2.70}$$

Beweis:

(i) Nach (2.33) ist in diesem Fall $\mathcal{F}^+(A,B) = \mathcal{F}^-(A,B) = \Lambda(A)$. Seien λ_j, $j \in \{1,\dots,n\}$, die Eigenwerte von A. $\Lambda(A)$ ist genau dann einelementig, wenn für alle $U \in \mathcal{U}(n)$ $UAU^* = \operatorname{diag}(\lambda_1,\dots,\lambda_n) \in \Lambda(A)$ ist. Da Permutationsmatrizen unitär sind, muß insbesondere

$$\operatorname{diag}(\lambda_1,\dots,\lambda_n) = \operatorname{diag}(\lambda_2,\dots,\lambda_n,\lambda_1) = \cdots = \operatorname{diag}(\lambda_n,\lambda_1,\dots,\lambda_{n-1})$$

sein, also $\lambda_1 = \cdots = \lambda_n$ und A Vielfaches der Einheitsmatrix. Diese Bedingung ist auch hinreichend.

(ii) Unter Verwendung der Notation aus Satz 2.9 (ii) gilt für $k \in \{1,\dots,n\}$ nach Satz 1.3 (iv) $\lambda_k = \mathrm{re}(\kappa_k)$, also ist $\lambda_1 < \cdots < \lambda_n$ und

$$\begin{aligned}
[\operatorname{diag}(\lambda_1,\dots,\lambda_n), V^*S(A)V] &= [V^*H(A)V, V^*S(A)V] \\
&= V^*[H(A),S(A)]V \\
&= 0,
\end{aligned}$$

wobei die letzte Gleichheit aus Satz 1.3 (iii) folgt. Somit ist

$$\operatorname{diag}(\lambda_1,\dots,\lambda_n)V^*S(A)V = V^*S(A)V \operatorname{diag}(\lambda_1,\dots,\lambda_n)$$

und für alle $j,k \in \{1,\dots,n\}$

$$\lambda_j[V^*S(A)V]_{jk} = [V^*S(A)V]_{jk}\lambda_k.$$

Daraus folgt, daß $V^*S(A)V = \operatorname{diag}\bigl(\mu_{\pi(1)},\dots,\mu_{\pi(n)}\bigr)$ mit $\pi \in \Pi(n)$ ist und

$$\begin{aligned}
V^*AV &= V^*H(A)V + V^*S(A)V \\
&= \operatorname{diag}(\lambda_1,\dots,\lambda_n) + \operatorname{diag}\bigl(\mu_{\pi(1)},\dots,\mu_{\pi(n)}\bigr) \\
&= \operatorname{diag}\bigl(\lambda_1 + \mu_{\pi(1)},\dots,\lambda_n + \mu_{\pi(n)}\bigr) \\
&= \operatorname{diag}(\kappa_1,\dots,\kappa_n).
\end{aligned}$$

Also ist nach (2.38)

$$\begin{aligned}
\mathcal{F}^+(A,B) = \bigl\{&W \operatorname{diag}(T_1,\dots,T_n) P_n \operatorname{diag}(S_1,\dots,S_n) \operatorname{diag}(\kappa_1,\dots,\kappa_n) \\
&\cdot \operatorname{diag}(S_1,\dots,S_n)^* P_n \operatorname{diag}(T_1,\dots,T_n)^* W^* : \\
&S_j, T_k \in \mathcal{U}(1), j,k \in \{1,\dots,n\}\bigr\} \\
= \bigl\{&W \operatorname{diag}(\kappa_n,\dots,\kappa_1) W^*\bigr\}.
\end{aligned}$$

Ebenso folgt aus (2.39)

$$\mathcal{F}^-(A,B) = \bigl\{W \operatorname{diag}(\kappa_1,\dots,\kappa_n) W^*\bigr\}.$$

(iii) folgt analog zu (ii) mit Hilfe von Satz 1.3 (iii) und (v) aus (2.44) und (2.45).

(iv) folgt direkt aus (2.50) und (2.51).

(v) folgt direkt aus (2.56) und (2.57). □

Also sind für $A, B \in \mathcal{H}(n)$ oder $A, B \in \mathcal{S}(n)$ die Lösungen von (2.5) und (2.6) eindeutig, wenn die Eigenwerte von B einfach sind. Für $A \in \mathcal{N}(n)$ und $B \in \mathcal{H}(n)$ liegt ebenfalls Eindeutigkeit vor, wenn zusätzlich noch die Realteile der Eigenwerte von A paarweise verschieden sind, und für $A \in \mathcal{N}(n)$ und $B \in \mathcal{S}(n)$, wenn zusätzlich zu den Eigenwerten von B noch die Imaginärteile der Eigenwerte von A paarweise verschieden sind. Man könnte vermuten, daß für $A, B \in \mathcal{N}(n)$ die Lösungen von (2.5) und (2.6) eindeutig sind, wenn die Real- und die Imaginärteile der Eigenwerte von A und B paarweise verschieden sind. Daß dies jedoch nicht zutreffend ist, zeigt das folgende Beispiel.

Beispiel 2.11
Seien $\alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathbb{C}$ mit $\operatorname{re}\{(\alpha_1 - \alpha_2)(\overline{\beta_1} - \overline{\beta_2})\} = 0$, sowie $A = \operatorname{diag}(\alpha_1, \alpha_2)$, $B = \operatorname{diag}(\beta_1, \beta_2)$ und $U \in \mathcal{U}(n)$. Dann ist $|u_{11}|^2 = |u_{22}|^2$ und $|u_{12}|^2 = |u_{21}|^2 = 1 - |u_{11}|^2$, also

$$\begin{aligned}
g_B(UAU^*) &= \operatorname{re}\left(\operatorname{tr}\left\{\begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22}\end{pmatrix}\begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2\end{pmatrix}\begin{pmatrix} \overline{u_{11}} & \overline{u_{21}} \\ \overline{u_{12}} & \overline{u_{22}}\end{pmatrix}\begin{pmatrix} \overline{\beta_1} & 0 \\ 0 & \overline{\beta_2}\end{pmatrix}\right\}\right) \\
&= \operatorname{re}\left(\operatorname{tr}\left\{\begin{pmatrix} u_{11}\alpha_1 & u_{12}\alpha_2 \\ u_{21}\alpha_1 & u_{22}\alpha_2\end{pmatrix}\begin{pmatrix} \overline{u_{11}\beta_1} & \overline{u_{21}\beta_2} \\ \overline{u_{12}\beta_1} & \overline{u_{22}\beta_2}\end{pmatrix}\right\}\right) \\
&= \operatorname{re}\{\alpha_1\overline{\beta_1}|u_{11}|^2 + \alpha_2\overline{\beta_1}|u_{12}|^2 + \alpha_1\overline{\beta_2}|u_{21}|^2 + \alpha_2\overline{\beta_2}|u_{22}|^2\} \\
&= \operatorname{re}\{(\alpha_1\overline{\beta_1} + \alpha_2\overline{\beta_2})|u_{11}|^2 + (\alpha_1\overline{\beta_2} + \alpha_2\overline{\beta_1})(1 - |u_{11}|^2)\} \\
&= \operatorname{re}\{(\alpha_1\overline{\beta_1} + \alpha_2\overline{\beta_2} - \alpha_1\overline{\beta_2} - \alpha_2\overline{\beta_1})|u_{11}|^2 + (\alpha_1\overline{\beta_2} + \alpha_2\overline{\beta_1})\} \\
&= \operatorname{re}\{(\alpha_1 - \alpha_2)(\overline{\beta_1} - \overline{\beta_2})\}|u_{11}|^2 + \operatorname{re}\{\alpha_1\overline{\beta_2} + \alpha_2\overline{\beta_1}\} \\
&= \operatorname{re}\{\alpha_1\overline{\beta_2} + \alpha_2\overline{\beta_1}\}
\end{aligned}$$

unabhängig von U und damit

$$\mathcal{F}^+(A,B) = \mathcal{F}^-(A,B) = \Lambda(A).$$

Für $\alpha_1 = 1+i$, $\alpha_2 = 4+3i$, $\beta_1 = -3+4i$ und $\beta_2 = 1-2i$ sind zwar die Real- und Imaginärteile der Eigenwerte von A und B verschieden, aber wegen

$$\begin{aligned}(\alpha_1-\alpha_2)(\overline{\beta_1}-\overline{\beta_2}) &= (1+i-4-3i)(-3-4i-1-2i)\\ &= (-3-2i)(-4-6i) = 26i\end{aligned}$$

ist die Lösung wie oben gezeigt nicht eindeutig.

Auch für hermitesches oder schiefhermitesches B und eine allgemeine Matrix A garantiert die Einfachheit der Eigenwerte von $H(A)$, $S(A)$ und B keine Eindeutigkeit der Lösungen von (2.5) und (2.6), wie uns das nächsten Beispiel zeigt.

Beispiel 2.12
Sei $H(A) = \mathrm{diag}(\alpha_1, \alpha_2)$, $B = \mathrm{diag}(\beta_1, \beta_2)$ und

$$S(A) = \begin{pmatrix} \gamma_1 & 1 \\ -1 & \gamma_2 \end{pmatrix}$$

mit $\alpha_1 < \alpha_2$, $\beta_1 < \beta_2$, $\mathrm{re}(\gamma_1) = \mathrm{re}(\gamma_2) = 0$ und $\mathrm{im}(\gamma_1) < \mathrm{im}(\gamma_2)$. Dann sind die Eigenwerte von $S(A)$

$$\frac{1}{2}\left\{\mathrm{im}(\gamma_1) + \mathrm{im}(\gamma_2) \pm \sqrt{\left(\mathrm{im}(\gamma_1) - \mathrm{im}(\gamma_2)\right)^2 + 4}\right\} i$$

ebenso verschieden wie die Eigenwerte von $H(A)$ und B. Jedoch ist für $\phi_1, \phi_2 \in \mathbb{C}$ mit $|\phi_1| = |\phi_2| = 1$ und $\phi = \phi_1\overline{\phi_2}$

$$\begin{aligned}\mathrm{diag}(\phi_1,\phi_2)A\,\mathrm{diag}(\phi_1,\phi_2)^* &= \begin{pmatrix} \phi_1 & 0 \\ 0 & \phi_2 \end{pmatrix}\left\{\begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} + \begin{pmatrix} \gamma_1 & 1 \\ -1 & \gamma_2 \end{pmatrix}\right\}\begin{pmatrix} \overline{\phi_1} & 0 \\ 0 & \overline{\phi_2} \end{pmatrix}\\ &= \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} + \begin{pmatrix} \phi_1\gamma_1\overline{\phi_1} & \phi_1\overline{\phi_2} \\ -\phi_2\overline{\phi_1} & \phi_2\gamma_2\overline{\phi_2} \end{pmatrix}\\ &= \begin{pmatrix} \alpha_1+\gamma_1 & \phi \\ -\overline{\phi} & \alpha_2+\gamma_2 \end{pmatrix},\end{aligned}$$

also nach (2.39)

$$\begin{aligned}\mathcal{F}^-(A,B) &= \{\mathrm{diag}(T_1,T_2)\,\mathrm{diag}(S_1,S_2)A\,\mathrm{diag}(S_1,S_2)^*\,\mathrm{diag}(T_1,T_2)^* : \\ &\qquad S_1,S_2,T_1,T_2 \in \mathcal{U}(1)\}\\ &= \{\mathrm{diag}(\phi_1,\phi_2)A\,\mathrm{diag}(\overline{\phi_1},\overline{\phi_2}) : \phi_1,\phi_2 \in \mathbb{C}, |\phi_1| = |\phi_2| = 1\}\\ &= \left\{\begin{pmatrix} \alpha_1+\gamma_1 & \phi \\ -\overline{\phi} & \alpha_2+\gamma_2 \end{pmatrix} : \phi \in \mathbb{C}, |\phi| = 1\right\},\end{aligned}$$

und für $\psi_1, \psi_2 \in \mathbb{C}$ mit $|\psi_1| = |\psi_2| = 1$ und $\psi = -\psi_1\overline{\phi\psi_2}$

$$\begin{aligned}&\mathrm{diag}(\psi_1,\psi_2)P_n\,\mathrm{diag}(\phi_1,\phi_2)A\,\mathrm{diag}(\phi_1,\phi_2)^*P_n\,\mathrm{diag}(\psi_1,\psi_2)^*\\ &= \begin{pmatrix} \psi_1 & 0 \\ 0 & \psi_2 \end{pmatrix}\begin{pmatrix} \alpha_2+\gamma_2 & -\overline{\phi} \\ \phi & \alpha_1+\gamma_1 \end{pmatrix}\begin{pmatrix} \overline{\psi_1} & 0 \\ 0 & \overline{\psi_2} \end{pmatrix}\\ &= \begin{pmatrix} \alpha_2+\gamma_2 & -\psi_1\overline{\phi\psi_2} \\ \psi_2\phi\overline{\psi_1} & \alpha_1+\gamma_1 \end{pmatrix}\\ &= \begin{pmatrix} \alpha_2+\gamma_2 & \psi \\ -\overline{\psi} & \alpha_1+\gamma_1 \end{pmatrix},\end{aligned}$$

also nach (2.38)

$$\begin{aligned}\mathcal{F}^+(A,B) &= \{\operatorname{diag}(T_1,T_2)P_n \operatorname{diag}(S_1,S_2)A \operatorname{diag}(S_1,S_2)^* P_n \operatorname{diag}(T_1,T_2)^* : \\ &\qquad S_1,S_2,T_1,T_2 \in \mathcal{U}(1)\} \\ &= \left\{ \begin{pmatrix} \alpha_2+\gamma_2 & \psi \\ -\overline{\psi} & \alpha_1+\gamma_1 \end{pmatrix} : \psi \in \mathbb{C}, |\psi| = 1 \right\}.\end{aligned}$$

Somit sind die Lösungen von (2.5) und (2.6) nicht eindeutig.

In anderen Fällen als den in Satz 2.9 behandelten konnten bislang noch keine expliziten Lösungen erzielt werden. Wir können aber mit Hilfe der Resultate aus Satz 2.9 eine Abschätzung für $F^+(A,B), F^-(A,B), G^+(A,B)$ und $G^-(A,B)$ herleiten, die für alle $A,B \in \mathbb{C}^{n,n}$ gilt.

Korollar 2.13
Seien $A,B \in \mathbb{C}^{n,n}$, $\lambda_1 \le \cdots \le \lambda_n$ die Eigenwerte von $H(A)$, $\nu_1 \le \cdots \le \nu_n$ die Eigenwerte von $H(B)$ und für $k \in \{1,\dots,n\}$ μ_k die Eigenwerte von $S(A)$ und ξ_k die Eigenwerte von $S(B)$ mit $\operatorname{im}(\mu_1) \le \cdots \le \operatorname{im}(\mu_n)$ und $\operatorname{im}(\xi_1) \le \cdots \le \operatorname{im}(\xi_n)$. Setze

$$K_1(A,B) := \sqrt{\sum_{k=1}^n (\lambda_k - \nu_k)^2 + \sum_{k=1}^n \big(\operatorname{im}(\mu_k) - \operatorname{im}(\xi_k)\big)^2}, \tag{2.71}$$

$$\begin{aligned}K_2(A,B) := \min \Bigg\{ &\sqrt{\sum_{k=1}^n (\lambda_k - \nu_k)^2 + \sum_{k=1}^n \big(\operatorname{im}(\mu_{n+1-k}) - \operatorname{im}(\xi_k)\big)^2}\ , \\ &\sqrt{\sum_{k=1}^n (\lambda_{n+1-k} - \nu_k)^2 + \sum_{k=1}^n \big(\operatorname{im}(\mu_k) - \operatorname{im}(\xi_k)\big)^2} \Bigg\},\end{aligned} \tag{2.72}$$

$$\begin{aligned}K_3(A,B) := \max \Bigg\{ &\sqrt{\sum_{k=1}^n (\lambda_k - \nu_k)^2 + \sum_{k=1}^n \big(\operatorname{im}(\mu_{n+1-k}) - \operatorname{im}(\xi_k)\big)^2}\ , \\ &\sqrt{\sum_{k=1}^n (\lambda_{n+1-k} - \nu_k)^2 + \sum_{k=1}^n \big(\operatorname{im}(\mu_k) - \operatorname{im}(\xi_k)\big)^2} \Bigg\},\end{aligned} \tag{2.73}$$

$$K_4(A,B) := \sqrt{\sum_{k=1}^n (\lambda_{n+1-k} - \nu_k)^2 + \sum_{k=1}^n \big(\operatorname{im}(\mu_{n+1-k}) - \operatorname{im}(\xi_k)\big)^2}, \tag{2.74}$$

$$L_1(A,B) := \sum_{k=1}^n \lambda_{n+1-k}\nu_k + \sum_{k=1}^n \operatorname{im}(\mu_{n+1-k}) \operatorname{im}(\xi_k), \tag{2.75}$$

$$\begin{aligned}L_2(A,B) := \min \Bigg\{ &\sum_{k=1}^n \lambda_k \nu_k + \sum_{k=1}^n \operatorname{im}(\mu_{n+1-k}) \operatorname{im}(\xi_k)\ , \\ &\sum_{k=1}^n \lambda_{n+1-k}\nu_k + \sum_{k=1}^n \operatorname{im}(\mu_k) \operatorname{im}(\xi_k) \Bigg\},\end{aligned} \tag{2.76}$$

$$\begin{aligned}L_3(A,B) := \max \Bigg\{ &\sum_{k=1}^n \lambda_k \nu_k + \sum_{k=1}^n \operatorname{im}(\mu_{n+1-k}) \operatorname{im}(\xi_k)\ , \\ &\sum_{k=1}^n \lambda_{n+1-k}\nu_k + \sum_{k=1}^n \operatorname{im}(\mu_k) \operatorname{im}(\xi_k) \Bigg\},\end{aligned} \tag{2.77}$$

$$(2.78) \qquad L_4(A,B) := \sum_{k=1}^{n} \lambda_k \nu_k + \sum_{k=1}^{n} \operatorname{im}(\mu_k)\operatorname{im}(\xi_k).$$

Dann gilt

$$(2.79) \quad K_1(A,B) \le F^-(A,B) \le K_2(A,B) \le K_3(A,B) \le F^+(A,B) \le K_4(A,B)$$

und

$$(2.80) \quad L_1(A,B) \le G^-(A,B) \le L_2(A,B) \le L_3(A,B) \le G^+(A,B) \le L_4(A,B).$$

Beweis:
Sei $X \in \Lambda(A)$. Dann ist

$$\begin{aligned}
g_B(X) &= \operatorname{re}\{\operatorname{tr}(XB^*)\} \\
&= \operatorname{re}\Big\{\operatorname{tr}\Big(\big(H(X)+S(X)\big)\big(H(B)+S(B)\big)^*\Big)\Big\} \\
&= \operatorname{re}\big\{\operatorname{tr}\big(H(X)H(B)^*\big)\big\} + \underbrace{\operatorname{re}\big\{\operatorname{tr}\big(H(X)S(B)^*\big)\big\}}_{=0 \text{ nach } (2.32)} \\
&\quad + \underbrace{\operatorname{re}\big\{\operatorname{tr}\big(S(X)H(B)^*\big)\big\}}_{=0 \text{ nach } (2.32)} + \operatorname{re}\big\{\operatorname{tr}\big(S(X)S(B)^*\big)\big\} \\
&= \operatorname{re}\big\{\operatorname{tr}\big(H(X)H(B)^*\big)\big\} + \operatorname{re}\big\{\operatorname{tr}\big(S(X)S(B)^*\big)\big\}.
\end{aligned}$$

Weiter ist nach (2.48) und (2.49)

$$\begin{aligned}
G^+\big(H(A),H(B)\big) &= \sum_{k=1}^{n} \lambda_k \nu_k, \\
G^-\big(H(A),H(B)\big) &= \sum_{k=1}^{n} \lambda_{n+1-k} \nu_k,
\end{aligned}$$

sowie nach (2.54) und (2.55)

$$\begin{aligned}
G^+\big(S(A),S(B)\big) &= \sum_{k=1}^{n} \operatorname{im}(\mu_k)\operatorname{im}(\xi_k), \\
G^-\big(S(A),S(B)\big) &= \sum_{k=1}^{n} \operatorname{im}(\mu_{n+1-k})\operatorname{im}(\xi_k),
\end{aligned}$$

also

$$\begin{aligned}
L_1(A,B) &= G^-\big(H(A),H(B)\big) + G^-\big(S(A),S(B)\big) \\
L_2(A,B) &= \min\big\{G^+\big(H(A),H(B)\big) + G^-\big(S(A),S(B)\big), \\
&\qquad\quad G^-\big(H(A),H(B)\big) + G^+\big(S(A),S(B)\big)\big\}, \\
L_3(A,B) &= \max\big\{G^+\big(H(A),H(B)\big) + G^-\big(S(A),S(B)\big), \\
&\qquad\quad G^-\big(H(A),H(B)\big) + G^+\big(S(A),S(B)\big)\big\},
\end{aligned}$$

und

$$L_4(A,B) = G^+\big(H(A),H(B)\big) + G^+\big(S(A),S(B)\big),$$

woraus (2.80) folgt. Ebenso ist nach (2.62)

$$f_B(X) = \sqrt{\|B-X\|_F^2} = \sqrt{\|H(B)-H(X)\|_F^2 + \|S(B)-S(X)\|_F^2}$$

und nach (2.46) und (2.47)

$$F^+\big(H(A),H(B)\big) = \sqrt{\sum_{k=1}^{n}(\lambda_{n+1-k}-\nu_k)^2},$$

$$F^-\big(H(A),H(B)\big) = \sqrt{\sum_{k=1}^{n}(\lambda_{k}-\nu_k)^2},$$

sowie nach (2.52) und (2.53)

$$F^+\big(S(A),S(B)\big) = \sqrt{\sum_{k=1}^{n}\big(\mathrm{im}(\mu_{n+1-k})-\mathrm{im}(\xi_k)\big)^2},$$

$$F^-\big(S(A),S(B)\big) = \sqrt{\sum_{k=1}^{n}\big(\mathrm{im}(\mu_{k})-\mathrm{im}(\xi_k)\big)^2},$$

woraus ebenso (2.79) folgt. □

Wir werden in den Abschnitten 3.1 und 4.1 numerische Verfahren zur Lösung des isospektralen Maximierungsproblems (2.5) und des isospektralen Minimierungsproblems (2.6) entwickeln.

2.2 Nächstgelegene normale Matrizen

Ein bislang nur teilweise gelöstes Problem ist das der besten normalen Approximation einer beliebigen komplexen Matrix in der Frobenius-Norm, das heißt für eine gegebene Matrix $A \in \mathbb{C}^{n,n}$ wird die Abweichung

$$\min_{M\in\mathcal{N}(n)} \|A-M\|_F$$

der Matrix A von der Normalität gesucht, sowie $N \in \mathcal{N}(n)$ mit

$$\|A-N\|_F = \min_{M\in\mathcal{N}(n)} \|A-M\|_F, \tag{2.81}$$

also eine in der Frobenius-Norm nächstgelegene normale Matrix zu A. Untersucht wurde das Problem unter anderem von Causey in [3] und von Ruhe in [21]. In [8] konnte Gabriel (2.81) für den Fall $A \in \mathbb{C}^{2,2}$ vollständig lösen. Dabei zeigte sich auch, daß die Lösung von (2.81) im allgemeinen nicht eindeutig ist. Eine wichtige Rolle bei diesem Problem spielen die sogenannten ΔH-Matrizen.

Definition 2.14
Sei $G \in \mathbb{C}^{n,n}$. G heißt ΔH-Matrix, wenn $D = \mathrm{diag}(d_1,\dots,d_n) \in \mathbb{C}^{n,n}$ und $H \in \mathcal{H}(n)$ existieren mit $G = D + DH - HD$.

Der folgende Satz ist eine leichte Modifikation von [21, Theorem 1] und beschreibt die Struktur der Matrix A in den Koordinaten der Eigenvektoren einer nächstgelegenen normalen Matrix N. Mit Hilfe von Korollar 2.8 (v) geben wir einen Alternativbeweis zu [21]. Wir verwenden im folgenden die Funktion

$$\Delta:\ \mathbb{C}^{n,n} \longrightarrow \mathbb{C}^{n,n},\ \Delta(Z) = \mathrm{diag}(z_{11},\dots,z_{nn}). \tag{2.82}$$

Satz 2.15
Sei $A \in \mathbb{C}^{n,n}$ gegeben und $N \in \mathcal{N}(n)$ erfülle (2.81). Sind $b_1, \dots, b_n$ die Eigenwerte von N, $B := \operatorname{diag}(b_1, \dots, b_n)$ und $V \in \mathcal{U}(n)$ mit $N = VBV^*$, so existiert $H \in \mathcal{H}(n)$ mit

$$V^*AV = B + BH - HB. \tag{2.83}$$

Insbesondere ist V^*AV eine ΔH-Matrix.

Beweis:
Der Abstand zwischen A und N ist minimal für alle normalen Matrizen, also ist insbesondere

$$\begin{aligned}\|B - V^*AV\|_F &= \|N - A\|_F \\ &= \min_{U \in \mathcal{U}(n)} \|VU^*BUV^* - A\|_F \\ &= \min_{U \in \mathcal{U}(n)} \|B - UV^*AVU^*\|_F.\end{aligned}$$

Somit ist V^*AV eine Extremalstelle von f_B und damit ein stationärer Punkt von g_B, und aus Korollar 2.8 (v) folgt, daß $V^*AV = Z + BH - HB$ ist mit $H \in \mathcal{H}(n)$ und $z_{rs} = 0$ für $b_r \neq b_s$. Wir wollen $Z = B$ zeigen. Aus

$$\begin{aligned}\|N - A\|_F^2 &= \|B - V^*AV\|_F^2 \\ &= \|B - Z - BH + HB\|_F^2 \\ &= \|B - \Delta(Z)\|_F^2 + \|Z - \Delta(Z)\|_F^2 + \|HB - BH\|_F^2\end{aligned}$$

folgt $B = \Delta(Z)$, da sonst

$$\begin{aligned}\|V\Delta(Z)V^* - A\|_F^2 &= \|\Delta(Z) - V^*AV\|_F^2 \\ &= \|Z - \Delta(Z)\|_F^2 + \|HB - BH\|_F^2 \\ &< \|N - A\|_F^2\end{aligned}$$

wäre, also die normale Matrix $V\Delta(Z)V^*$ näher an A liegen würde als N. Angenommen, es ist Z nicht diagonal, also $Z \neq \Delta(Z) = B$. Dann existiert $1 \leq k < l \leq n$ mit $b_k = b_l$ und $(z_{kl}, z_{lk}) \neq (0, 0)$. Daraus konstruieren wir jetzt eine Matrix C so, daß VCV^* näher an A liegt als N. Setze $\omega := \arg(z_{kl} z_{lk})$ und definiere $C \in \mathbb{C}^{n,n}$ durch

$$c_{rs} := \begin{cases} z_{rr} & , \quad r = s \\ \frac{1}{2}(z_{kl} + e^{i\omega}\overline{z_{lk}}) & , \quad r = k, s = l \\ \frac{1}{2}(z_{lk} + e^{i\omega}\overline{z_{kl}}) & , \quad r = l, s = k \\ 0 & , \quad \text{sonst} \end{cases}.$$

Dann ist

$$\begin{aligned}&|c_{kl} - z_{kl}|^2 + |c_{lk} - z_{lk}|^2 \\ &= \left|\frac{1}{2}(z_{kl} + e^{i\omega}\overline{z_{lk}}) - z_{kl}\right|^2 + \left|\frac{1}{2}(z_{lk} + e^{i\omega}\overline{z_{kl}}) - z_{lk}\right|^2 \\ &= \frac{1}{4}\left\{(e^{i\omega}\overline{z_{lk}} - z_{kl})(e^{-i\omega}z_{lk} - \overline{z_{kl}}) + (e^{i\omega}\overline{z_{kl}} - z_{lk})(e^{-i\omega}z_{kl} - \overline{z_{lk}})\right\} \\ &= \frac{1}{2}\left\{|z_{kl}|^2 + |z_{lk}|^2 - z_{kl}z_{lk}e^{-i\arg(z_{kl}z_{lk})} - \overline{z_{kl}z_{lk}}e^{i\arg(z_{kl}z_{lk})}\right\} \\ &= \frac{1}{2}\left\{|z_{kl}|^2 + |z_{lk}|^2 - 2|z_{kl}z_{lk}|\right\},\end{aligned}$$

also

$$
\begin{aligned}
&|c_{kl} - z_{kl}|^2 + |c_{lk} - z_{lk}|^2 - |z_{kl}|^2 - |z_{lk}|^2 \\
&= \frac{1}{2}\big\{|z_{kl}|^2 + |z_{lk}|^2 - 2|z_{kl} z_{lk}|\big\} - |z_{kl}|^2 - |z_{lk}|^2 \\
&= -\frac{1}{2}\big(|z_{kl}| + |z_{lk}|\big)^2 \\
&< 0
\end{aligned}
$$

und

$$
\begin{aligned}
\|VCV^* - A\|_F^2 &= \|C - V^*AV\|_F^2 \\
&= \|C - Z - BH + HB\|_F^2 \\
&= \|\Delta(C) - \Delta(Z)\|_F^2 + |c_{kl} - z_{kl}|^2 + |c_{lk} - z_{lk}|^2 \\
&\quad + \|Z - \Delta(Z)\|_F^2 - |z_{kl}|^2 - |z_{lk}|^2 + \|HB - BH\|_F^2 \\
&< \|Z - \Delta(Z)\|_F^2 + \|HB - BH\|_F^2 \\
&= \|N - A\|_F^2,
\end{aligned}
$$

das heißt VCV^* liegt näher an A als N. Es bleibt zu zeigen, daß VCV^* normal ist. Wegen

$$c_{kk} = z_{kk} = b_k = b_l = z_{ll} = c_{ll}$$

und

$$|c_{kl}| = \left|\frac{1}{2}(z_{kl} + e^{i\omega}\overline{z_{lk}})\right| = \left|\frac{1}{2}(\overline{z_{kl}} + e^{-i\omega} z_{lk})e^{i\omega}\right| = \left|\frac{1}{2}(e^{i\omega}\overline{z_{kl}} + z_{lk})\right| = |c_{lk}|$$

ist

$$
\begin{aligned}
[CC^*]_{kk} &= |c_{kk}|^2 + |c_{kl}|^2 = |c_{kk}|^2 + |c_{lk}|^2 = [C^*C]_{kk}, \\
[CC^*]_{kl} &= c_{kk}\overline{c_{lk}} + c_{kl}\overline{c_{ll}} = \overline{c_{kk}}c_{kl} + \overline{c_{lk}}c_{ll} = [C^*C]_{kl}, \\
[CC^*]_{lk} &= c_{lk}\overline{c_{kk}} + c_{ll}\overline{c_{kl}} = \overline{c_{kl}}c_{kk} + \overline{c_{ll}}c_{lk} = [C^*C]_{lk}, \\
[CC^*]_{ll} &= |c_{lk}|^2 + |c_{ll}|^2 = |c_{kl}|^2 + |c_{ll}|^2 = [C^*C]_{ll},
\end{aligned}
$$

und für $(r,s) \notin \{(k,k),(k,l),(l,k),(l,l)\}$ ist

$$[CC^*]_{rs} = \left\{\begin{matrix} |c_{rr}|^2 & , & r = s \\ 0 & , & r \neq s \end{matrix}\right\} = [C^*C]_{rs}.$$

Somit ist C normal. Damit ist aber auch VCV^* normal, was im Widerspruch dazu steht, daß N die nächste normale Matrix zu A ist. Also ist $Z = \Delta(Z) = B$ und $V^*AV = B + BH - HB$ eine ΔH-Matrix. □

In den Abschnitten 3.2 und 4.2 werden wir diese Eigenschaft benutzen, um numerische Verfahren zur Berechnung einer nächstgelegenen normalen Matrix zu entwickeln.

Kapitel 3

Jacobiartige Verfahren

3.1 Verfahren zur Lösung des isospektralen Abstandsproblems

Das klassische Jacobi-Verfahren zur Berechnung der Eigenwerte einer reellsymmetrischen Matrix wurde bereits Mitte des vorigen Jahrhunderts von Jacobi entwickelt und erstmalig in [19] publiziert. Bei diesem Verfahren werden durch eine Folge von orthogonalen Ähnlichkeitstransformationen sukzessiv die Nichtdiagonalelemente einer reellsymmetrischen Matrix annihiliert. Da sich dabei bereits vorher annihilierte Elemente wieder verändern, handelt es sich um ein nicht-endliches Verfahren. In jedem Schritt des Verfahrens wird die sogenannte Off-Norm

$$\text{off} : \mathbb{C}^{n,n} \longrightarrow \mathbb{R},\ \text{off}(X) = \sum_{k=1}^{n} \sum_{\substack{l=1 \\ l \neq k}}^{n} |x_{kl}|^2 \tag{3.1}$$

verringert, was zur Folge hat, daß die durch Iteration erzeugte Folge von Matrizen in der Frobenius-Norm gegen eine Diagonalmatrix konvergiert. Wird beim klassischen Jacobi-Verfahren stets ein betragsgrößtes Nichtdiagonalelement annihiliert, so werden beim zyklischen Jacobi-Verfahren die Nichtdiagonalelemente nacheinander in der Reihenfolge ihres Auftretens in der Matrix annihiliert. Dadurch wird die Suche nach dem betragsgrößten Nichtdiagonalelement vermieden, die bei großen Matrizen weit mehr Computer-Rechenzeit in Anspruch nimmt als die Bestimmung der Ähnlichkeitstransformationen. Dies hat zur Folge, daß das zyklische Jacobi-Verfahren erheblich schneller ist als das klassische Jacobi-Verfahren. Beide Verfahren werden ausführlich in [11], Abschnitt 8.4, beschrieben, wo auch ein Konvergenzbeweis für das klassische Jacobi-Verfahren gegeben wird. Die Konvergenz des zyklischen Jacobi-Verfahrens ist in einer ganzen Reihe von Publikationen ([14],[17],[22],[23],[26],[27]) untersucht worden.

Basierend auf dem Jacobi-Verfahren wurden in den letzten 40 Jahren eine Reihe ähnlicher Verfahren zur Eigenwertbestimmung entwickelt, etwa das Verfahren [9] von Goldstine und Horwitz zur Diagonalisierung normaler Matrizen oder das Verfahren [25] von Stewart zur Transformation einer beliebigen Matrix in eine obere Dreiecksmatrix, die beide auf unitären Ähnlichkeitstransformationen beruhen, sowie das Verfahren [5] von Eberlein, das mit nichtunitären Ähnlichkeitstransformationen arbeitet. Es entstand eine ganze Klasse solcher Verfahren, die unter dem Oberbegriff jacobiartige Verfahren oder jacobiähnliche Verfahren zusammengefaßt werden. Gerade in jüngerer Zeit gibt es ein vermehrtes Interesse an jacobiartigen Verfahren, da sie zum einen leicht parallelisierbar sind ([2],[12]) und zum anderen die Eigenwerte mit einer höheren Genauigkeit berechnen ([4],[20]) als andere Eigenwertmethoden, wie etwa der QR-Algorithmus oder das Bisektionsverfahren.

Ein Ergebnis dieses neuerwachten Interesses an jacobiartigen Verfahren ist die Arbeit [16], in der Helmke, Hüper und Moore eine geometrische Interpretation des zyklischen Jacobi-Verfahrens geben. Sie zeigen, daß für eine reellsymmetrische Matrix A die von dem Verfahren erzeugte Folge der zu A orthogonal ähnlichen Iterierten $A^{(q)}$ dadurch zustande kommt, daß jede Iterierte $A^{(q)}$ die Off-Norm Funktion über eine Kurve $\mathcal{K}_{A^{(q-1)},Y^{(q-1)}}$ minimiert, wobei $Y^{(q-1)}$ ein Element der Standard-Basis des reellen Vektorraumes der schiefsymmetrischen Matrizen ist. In einem Zyklus des Verfahrens kommt jedes Element dieser Basis genau einmal vor. Ändert man das Verfahren derart, daß in jedem Schritt nicht mehr die Off-Norm minimiert, sondern die im letzten Kapitel eingeführte Funktion g_B maximiert wird, wobei B eine beliebige reelle Diagonalmatrix mit der Größe nach geordneten, paarweise verschiedenen Einträgen ist, so erhält man ein Sortier-Jacobi-Algorithmus genanntes Verfahren, das von Helmke und Hüper in [15] vorgestellt wurde. Bei diesem Verfahren konvergiert die Folge $A^{(q)}$ gegen eine Diagonalmatrix, deren Einträge ebenfalls der Größe nach geordnet sind. Gegenüber dem zyklischen Jacobi-Verfahren hat es den Vorteil, daß es die Eigenwerte der Matrix A der Größe nach sortiert und die Konvergenz auch dann quadratisch ist, wenn A mehrfache Eigenwerte besitzt.

Der Sortier-Jacobi-Algorithmus diagonalisiert also reellsymmetrische A, indem er in jedem Schritt die Funktion g_B für eine reelle Diagonalmatrix B über eine Kurve mit vorgegebener Richtung maximiert. Wir wollen dieses Verfahren in diesem Abschnitt verallgemeinern, um ein jacobiartiges Verfahren zu bekommen, das für beliebige $A, B \in \mathbb{C}^{n,n}$ die Funktion g_B auf $\Lambda(A)$ optimiert. Dazu verallgemeinern wir zunächst [15, Algorithm 3.3] so, daß wir ein allgemeines Verfahren zur Maximierung einer glatten Funktion $y : \Lambda(A) \longrightarrow \mathbb{R}$ auf $\Lambda(A)$ erhalten.

Algorithmus 3.1 (Jacobiartiges Maximierungsverfahren)
Eingabe: $A \in \mathbb{C}^{n,n}$, $y : \Lambda(A) \longrightarrow \mathbb{R}$

- Wähle eine Basis $Y_1, \ldots, Y_{n^2}$ von $\mathcal{S}(n)$.
- Setze $A_0^{(0)} := A$.
- Für $q = 0, 1, \ldots$
  - ∘ für $k = 1, \ldots, n^2$
    - ∗ falls $y\big(\mathcal{K}_{A_{k-1}^{(q)},Y_k}\big)$ konstant ist, so
      - · setze $t_k := 0$,
    - ∗ sonst
      - · berechne $t_k \in \mathbb{R}$ mit $y\big(\mathcal{K}_{A_{k-1}^{(q)},Y_k}\big)(t_k) = \max_{t\in\mathbb{R}} y\big(\mathcal{K}_{A_{k-1}^{(q)},Y_k}\big)(t)$,
    - ∗ setze $A_k^{(q)} := \mathcal{K}_{A_{k-1}^{(q)},Y_k}(t_k)$,
  - ∘ setze $A_0^{(q+1)} := A_{n^2}^{(q)}$.

Ausgabe: $A^{(q)} := A_0^{(q)}$, $q \in \mathbb{N}$

Analog sieht ein allgemeines Verfahren zur Minimierung von $y : \Lambda(A) \longrightarrow \mathbb{R}$ auf $\Lambda(A)$ folgendermaßen aus:

Algorithmus 3.2 (Jacobiartiges Minimierungsverfahren)
Ersetze in Algorithmus 3.1

· berechne $t_k \in \mathbb{R}$ mit $y\big(\mathcal{K}_{A^{(q)}_{k-1},Y_k}\big)(t_k) = \max_{t\in\mathbb{R}} y\big(\mathcal{K}_{A^{(q)}_{k-1},Y_k}\big)(t)$

durch

· berechne $t_k \in \mathbb{R}$ mit $y\big(\mathcal{K}_{A^{(q)}_{k-1},Y_k}\big)(t_k) = \min_{t\in\mathbb{R}} y\big(\mathcal{K}_{A^{(q)}_{k-1},Y_k}\big)(t)$.

Ein Optimierungszyklus, in dem jedes Basiselement von $\mathcal{S}(n)$ als Richtung genau einmal vorkommt, wird üblicherweise als Sweep bezeichnet. Wir bemerken an dieser Stelle, daß in den Algorithmen 3.1 und 3.2 nichts dazu gesagt wird, wie die Optimierung von y auf den durch Y_k gegebenen Kurven praktisch durchzuführen ist. Ob sie sich überhaupt bewerkstelligen läßt, hängt sicherlich von der Komplexität der Funktion y und von der Wahl der Basis von $\mathcal{S}(n)$ ab. Insbesondere, wenn die durch die Basiselemente gegebenen Kurven auf $\Lambda(A)$ nicht geschlossen sind, werden wir das Maximum und Minimum von y auf diesen Kurven im allgemeinen nicht explizit bestimmen können. Wie sieht es mit der Konvergenz der beiden Verfahren aus, wenn sie durchführbar sind? Der nächste Satz gibt ein allgemeines Konvergenzresultat.

Satz 3.1 (Konvergenz der jacobiartigen Algorithmen 3.1 und 3.2)
Sei $y : \Lambda(A) \longrightarrow \mathbb{R}$ eine glatte Funktion und $A^{(q)}$, $q \in \mathbb{N}$, die von Algorithmus 3.1 (Algorithmus 3.2) generierte Folge. Dann konvergiert $A^{(q)}$ gegen einen stationären Punkt $A_{\max} \in \Lambda(A)$ ($A_{\min} \in \Lambda(A)$) von y, dessen Funktionswert maximal (minimal) über alle Kurven $\mathcal{K}_{A_{\max},Y_k}$ ($\mathcal{K}_{A_{\min},Y_k}$), $k \in \{1,\ldots,n^2\}$, ist.

Beweis: Vollkommen analog zum Beweis von [15, Theorem 3.1]. □

Im folgenden entwickeln wir aus den allgemeinen Algorithmen 3.1 und 3.2 explizite Verfahren zur Optimierung von g_B auf $\Lambda(A)$. Dazu wählen wir zunächst eine Basis von $\mathcal{S}(n)$.

Sei für $1 \le k < l \le n$

$$R_{kl} : \mathbb{C} \longrightarrow \mathcal{S}(n)$$

$$R_{kl}(x) = \begin{pmatrix} 0_{k-1,k-1} & 0_{k-1,1} & 0_{k-1,l-k-1} & 0_{k-1,1} & 0_{k-1,n-l} \\ 0_{1,k-1} & 0 & 0_{1,l-k-1} & x & 0_{1,n-l} \\ 0_{l-k-1,k-1} & 0_{l-k-1,1} & 0_{l-k-1,l-k-1} & 0_{l-k-1,1} & 0_{l-k-1,n-l} \\ 0_{1,k-1} & -\overline{x} & 0_{1,l-k-1} & 0 & 0_{1,n-l} \\ 0_{n-l,k-1} & 0_{n-l,1} & 0_{n-l,l-k-1} & 0_{n-l,1} & 0_{n-l,n-l} \end{pmatrix} \tag{3.2}$$

und für $1 \le m \le n$

$$S_m : i\mathbb{R} \longrightarrow \mathcal{S}(n)$$

$$S_m(y) = \begin{pmatrix} 0_{m-1,m-1} & 0_{m-1,1} & 0_{m-1,n-m} \\ 0_{1,m-1} & y & 0_{1,n-m} \\ 0_{n-m,m-1} & 0_{n-m,1} & 0_{n-m,n-m} \end{pmatrix}. \tag{3.3}$$

Dann gilt:

Satz 3.2
$\{R_{kl}(1) : 1 \le k < l \le n\} \cup \{R_{kl}(i) : 1 \le k < l \le n\} \cup \{S_m(i) : 1 \le m \le n\}$ ist eine Basis des reellen Vektorraumes $\mathcal{S}(n)$.

Wir wollen g_B nacheinander über die Kurven, deren Richtungen $R_{kl}(1)$, $R_{kl}(i)$ und $S_m(i)$ sind, optimieren. Um den Verlauf von g_B auf diesen Kurven festzustellen, verschaffen wir uns erst einmal eine einfachere Darstellung der Kurven, in der die Matrix-Exponentialfunktion nicht mehr vorkommt. Dazu untersuchen wir, wie die Exponentialfunktion die Basisrichtungen transformiert.

Sei für $1 \le k < l \le n$

$$U_{kl} : \{(x,y) \in \mathbb{R} \times \mathbb{C} : x^2 + |y|^2 = 1\} \longrightarrow \mathcal{U}(n)$$

$$U_{kl}(x,y) = \begin{pmatrix} I_{k-1} & 0_{k-1,1} & 0_{k-1,l-k-1} & 0_{k-1,1} & 0_{k-1,n-l} \\ 0_{1,k-1} & x & 0_{1,l-k-1} & y & 0_{1,n-l} \\ 0_{l-k-1,k-1} & 0_{l-k-1,1} & I_{l-k-1} & 0_{l-k-1,1} & 0_{l-k-1,n-l} \\ 0_{1,k-1} & -\overline{y} & 0_{1,l-k-1} & x & 0_{1,n-l} \\ 0_{n-l,k-1} & 0_{n-l,1} & 0_{n-l,l-k-1} & 0_{n-l,1} & I_{n-l} \end{pmatrix} \tag{3.4}$$

und für $1 \le m \le n$

$$V_m : \{z \in \mathbb{C} : |z| = 1\} \longrightarrow \mathcal{U}(n)$$

$$V_m(z) = \begin{pmatrix} I_{m-1} & 0_{m-1,1} & 0_{m-1,n-m} \\ 0_{1,m-1} & z & 0_{1,n-m} \\ 0_{n-m,m-1} & 0_{n-m,1} & I_{n-m} \end{pmatrix}. \tag{3.5}$$

Dann gilt:

Lemma 3.3
Sei $1 \le k < l \le n$ und $1 \le m \le n$. Dann ist für $x \in \mathbb{C}$

$$e^{R_{kl}(x)} = U_{kl}\left(\cos|x|, \frac{x}{|x|}\sin|x|\right) \tag{3.6}$$

und für $y \in i\mathbb{R}$

$$e^{S_m(y)} = V_m(e^y). \tag{3.7}$$

Beweis:
Sei für $1 \le k < l \le n$

$$W_{kl} : \mathbb{R} \times \mathbb{C} \longrightarrow \mathbb{C}^{n,n}$$

$$W_{kl}(x,y) = \begin{pmatrix} 0_{k-1,k-1} & 0_{k-1,1} & 0_{k-1,l-k-1} & 0_{k-1,1} & 0_{k-1,n-l} \\ 0_{1,k-1} & x & 0_{1,l-k-1} & y & 0_{1,n-l} \\ 0_{l-k-1,k-1} & 0_{l-k-1,1} & 0_{l-k-1,l-k-1} & 0_{l-k-1,1} & 0_{l-k-1,n-l} \\ 0_{1,k-1} & -\overline{y} & 0_{1,l-k-1} & x & 0_{1,n-l} \\ 0_{n-l,k-1} & 0_{n-l,1} & 0_{n-l,l-k-1} & 0_{n-l,1} & 0_{n-l,n-l} \end{pmatrix}.$$

Dann ist

$$R_{kl}(x) = W_{kl}(0,x),$$

und wegen

$$\begin{pmatrix} 0 & x \\ -\overline{x} & 0 \end{pmatrix}\begin{pmatrix} 0 & x \\ -\overline{x} & 0 \end{pmatrix} = \begin{pmatrix} -x\overline{x} & 0 \\ 0 & -x\overline{x} \end{pmatrix}$$

ist

$$R_{kl}(x)^2 = W_{kl}(-x\overline{x}, 0) = -|x|^2 W_{kl}(1,0),$$

also für $j \in \mathbb{N}$

$$R_{kl}(x)^{2j} = \left(R_{kl}(x)^2\right)^j = (-1)^j |x|^{2j} W_{kl}(1,0)$$

und

$$\begin{aligned} R_{kl}(x)^{2j+1} &= R_{kl}(x)^{2j} R_{kl}(x) \\ &= (-1)^j |x|^{2j} W_{kl}(1,0) R_{kl}(x) \\ &= (-1)^j |x|^{2j} R_{kl}(x) \\ &= (-1)^j |x|^{2j} W_{kl}(0,x). \end{aligned}$$

Somit ist

$$\begin{aligned} e^{R_{kl}(x)} &= \sum_{j=0}^{\infty} \frac{1}{j!} R_{kl}(x)^j \\ &= I_n + \sum_{j=0}^{\infty} \frac{1}{(2j+1)!} R_{kl}(x)^{2j+1} + \sum_{j=1}^{\infty} \frac{1}{(2j)!} R_{kl}(x)^{2j} \\ &= I_n + \sum_{j=0}^{\infty} \frac{1}{(2j+1)!} (-1)^j |x|^{2j} W_{kl}(0,x) + \sum_{j=1}^{\infty} \frac{1}{(2j)!} (-1)^j |x|^{2j} W_{kl}(1,0) \\ &= \operatorname{diag}(I_{k-1}, 0, I_{l-k-1}, 0, I_{n-l}) \\ &\quad + \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j+1)!} |x|^{2j+1} W_{kl}\left(0, \frac{x}{|x|}\right) + \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j)!} |x|^{2j} W_{kl}(1,0) \\ &= \operatorname{diag}(I_{k-1}, 0, I_{l-k-1}, 0, I_{n-l}) + \sin|x| W_{kl}\left(0, \frac{x}{|x|}\right) + \cos|x| W_{kl}(1,0) \\ &= U_{kl}\left(\cos|x|, \frac{x}{|x|}\sin|x|\right). \end{aligned}$$

(3.7) ist wegen

$$e^{S_m(y)} = e^{\operatorname{diag}(0_{m-1,m-1}, y, 0_{n-m,n-m})} = \operatorname{diag}(I_{m-1}, e^y, I_{n-m})$$

trivial. □

Daraus ergibt sich für die Kurven mit den Richtungen $R_{kl}(1)$, $R_{kl}(i)$ und $S_m(i)$ die folgende einfache Darstellung:

Korollar 3.4
Sei $X \in \mathbb{C}^{n,n}$, $1 \le k < l \le n$ und $1 \le m \le n$. Dann ist für $t \in \mathbb{R}$

(3.8) $$\mathcal{K}_{X,R_{kl}(1)}(t) = U_{kl}(\cos t, \sin t) X U_{kl}^*(\cos t, \sin t),$$

(3.9) $$\mathcal{K}_{X,R_{kl}(i)}(t) = U_{kl}(\cos t, i\sin t) X U_{kl}^*(\cos t, i\sin t)$$

und

(3.10) $$\mathcal{K}_{X,S_m(i)}(t) = V_m(e^{it}) X V_m^*(e^{it}).$$

Beweis:
Nach (3.6) ist

$$\begin{aligned} \mathcal{K}_{X,R_{kl}(1)}(t) &= e^{tR_{kl}(1)} X e^{-tR_{kl}(1)} \\ &= e^{R_{kl}(t)} X e^{R_{kl}(-t)} \\ &= U_{kl}\left(\cos|t|, \frac{t}{|t|}\sin|t|\right) X U_{kl}\left(\cos|-t|, \frac{-t}{|-t|}\sin|-t|\right) \\ &= U_{kl}\left(\cos t, \operatorname{sign}(t)\sin|t|\right) X U_{kl}\left(\cos t, -\operatorname{sign}(t)\sin|t|\right) \\ &= U_{kl}(\cos t, \sin t) X U_{kl}^*(\cos t, \sin t) \end{aligned}$$

und

$$\begin{aligned}
\mathcal{K}_{X,R_{kl}(i)}(t) &= e^{tR_{kl}(i)}Xe^{-tR_{kl}(i)} \\
&= e^{R_{kl}(it)}Xe^{R_{kl}(-it)} \\
&= U_{kl}\left(\cos|it|, \frac{it}{|it|}\sin|it|\right)XU_{kl}\left(\cos|-it|, \frac{-it}{|-it|}\sin|-it|\right) \\
&= U_{kl}(\cos t, i\sin t)XU_{kl}(\cos t, -i\sin t) \\
&= U_{kl}(\cos t, i\sin t)XU_{kl}^*(\cos t, i\sin t),
\end{aligned}$$

sowie nach (3.7)

$$\begin{aligned}
\mathcal{K}_{X,S_m(i)}(t) &= e^{tS_m(i)}Xe^{-tS_m(i)} \\
&= e^{S_m(it)}Xe^{S_m(-it)} \\
&= V_m(e^{it})XV_m(e^{-it}) \\
&= V_m(e^{it})XV_m^*(e^{it}).
\end{aligned}$$

□

Eine Folgerung aus dem letzten Korollar ist, daß eine Kurve, die die Richtung $R_{kl}(1)$, $R_{kl}(i)$ oder $S_m(i)$ besitzt, geschlossen ist. Deshalb verläuft g_B auf ihr periodisch, wobei die Periode nach den obigen Resultaten höchstens 2π beträgt. Also nimmt g_B auf diesen Kurven seine Extrema in jedem Intervall der Länge 2π an.

Da $g_B(X) = \mathrm{re}\{\mathrm{tr}(XB^*)\}$ ist, studieren wir, um g_B auf den Kurven berechnen zu können, zunächst einmal die beiden Funktionen $\mathrm{tr}\{U_{kl}(x,y)XU_{kl}^*(x,y)B^*\}$ und $\mathrm{tr}\{V_m(z)XV_m^*(z)B^*\}$. Dazu führen wir die folgende Matrix-Partionierung ein:

Seien für $Z = (z_{rs})_{r,s=1,\dots,n} \in \mathbb{C}^{n,n}$ und $1 \le k < l \le n$

$$Z_{klrs}, \quad r,s \in \{1,\dots,5\}, (r,s) \notin \{(2,2),(2,4),(4,2),(4,4)\},$$

sowie für $1 \le m \le n$

$$Z_{mrs}, \quad r,s \in \{1,2,3\}, (r,s) \ne (2,2),$$

so gewählt, daß

$$Z = \begin{pmatrix} Z_{kl11} & Z_{kl12} & Z_{kl13} & Z_{kl14} & Z_{kl15} \\ Z_{kl21} & z_{kk} & Z_{kl23} & z_{kl} & Z_{kl25} \\ Z_{kl31} & Z_{kl32} & Z_{kl33} & Z_{kl34} & Z_{kl35} \\ Z_{kl41} & z_{lk} & Z_{kl43} & z_{ll} & Z_{kl45} \\ Z_{kl51} & Z_{kl52} & Z_{kl53} & Z_{kl54} & Z_{kl55} \end{pmatrix} = \begin{pmatrix} Z_{m11} & Z_{m12} & Z_{m13} \\ Z_{m21} & z_{mm} & Z_{m23} \\ Z_{m31} & Z_{m32} & Z_{m33} \end{pmatrix}.$$

Wir bemerken an dieser Stelle, daß mit Z_{klrs}^* und Z_{mrs}^* im folgenden stets $(Z_{klrs})^*$ und $(Z_{mrs})^*$ gemeint sind.

Mit Hilfe dieser Partitionierung können wir die Konstanten angeben, die bei der Darstellung der Funktionen $\mathrm{tr}\{U_{kl}(x,y)XU_{kl}^*(x,y)B^*\}$ und $\mathrm{tr}\{V_m(z)XV_m^*(z)B^*\}$ auftreten.

Wir definieren für $1 \le k < l \le n$

$$\eta_{klr} : \mathbb{C}^{n,n} \times \mathbb{C}^{n,n} \longrightarrow \mathbb{C}, \quad r \in \{0,\dots,9\},$$

durch

(3.11)

$$\left|\begin{aligned}
\eta_{kl0}(X,Y) &:= \mathrm{tr}(X_{kl11}Y^*_{kl11} + X_{kl13}Y^*_{kl13} + X_{kl15}Y^*_{kl15})\\
&\quad + \mathrm{tr}(X_{kl31}Y^*_{kl31} + X_{kl33}Y^*_{kl33} + X_{kl35}Y^*_{kl35})\\
&\quad + \mathrm{tr}(X_{kl51}Y^*_{kl51} + X_{kl53}Y^*_{kl53} + X_{kl55}Y^*_{kl55})\\
\eta_{kl1}(X,Y) &:= \mathrm{tr}(X_{kl12}Y^*_{kl12} + X_{kl14}Y^*_{kl14}) + X_{kl21}Y^*_{kl21} + X_{kl23}Y^*_{kl23}\\
&\quad + X_{kl25}Y^*_{kl25} + \mathrm{tr}(X_{kl32}Y^*_{kl32} + X_{kl34}Y^*_{kl34}) + X_{kl41}Y^*_{kl41}\\
&\quad + X_{kl43}Y^*_{kl43} + X_{kl45}Y^*_{kl45} + \mathrm{tr}(X_{kl52}Y^*_{kl52} + X_{kl54}Y^*_{kl54})\\
\eta_{kl2}(X,Y) &:= -\,\mathrm{tr}(X_{kl12}Y^*_{kl14}) + X_{kl41}Y^*_{kl21} + X_{kl43}Y^*_{kl23}\\
&\quad + X_{kl45}Y^*_{kl25} - \mathrm{tr}(X_{kl32}Y^*_{kl34}) - \mathrm{tr}(X_{kl52}Y^*_{kl54})\\
\eta_{kl3}(X,Y) &:= \mathrm{tr}(X_{kl14}Y^*_{kl12}) + \mathrm{tr}(X_{kl34}Y^*_{kl32}) - X_{kl21}Y^*_{kl41}\\
&\quad - X_{kl23}Y^*_{kl43} - X_{kl25}Y^*_{kl45} + \mathrm{tr}(X_{kl54}Y^*_{kl52})\\
\eta_{kl4}(X,Y) &:= x_{kk}\overline{y_{kk}} + x_{kl}\overline{y_{kl}} + x_{lk}\overline{y_{lk}} + x_{ll}\overline{y_{ll}}\\
\eta_{kl5}(X,Y) &:= x_{lk}\overline{y_{kk}} - x_{kk}\overline{y_{kl}} + x_{ll}\overline{y_{kl}} - x_{lk}\overline{y_{ll}}\\
\eta_{kl6}(X,Y) &:= x_{kl}\overline{y_{kk}} - x_{kk}\overline{y_{lk}} + x_{ll}\overline{y_{lk}} - x_{kl}\overline{y_{ll}}\\
\eta_{kl7}(X,Y) &:= -x_{lk}\overline{y_{kl}}\\
\eta_{kl8}(X,Y) &:= -x_{kl}\overline{y_{lk}}\\
\eta_{kl9}(X,Y) &:= x_{ll}\overline{y_{kk}} + x_{kk}\overline{y_{ll}}
\end{aligned}\right|$$

und für $1 \le m \le n$

$$\theta_{ms}:\ \mathbb{C}^{n,n} \times \mathbb{C}^{n,n} \longrightarrow \mathbb{C}, \quad s \in \{0,1,2\},$$

durch

$$\left|\begin{aligned}
\theta_{m0}(X,Y) &:= \mathrm{tr}(X_{m11}Y^*_{m11} + X_{m13}Y^*_{m13}) + x_{mm}\overline{y_{mm}}\\
&\quad + \mathrm{tr}(X_{m31}Y^*_{m31} + X_{m33}Y^*_{m33})\\
\theta_{m1}(X,Y) &:= X_{m21}Y^*_{m21} + X_{m23}Y^*_{m23}\\
\theta_{m2}(X,Y) &:= \mathrm{tr}(X_{m12}Y^*_{m12}) + \mathrm{tr}(X_{m32}Y^*_{m32})
\end{aligned}\right|. \tag{3.12}$$

Dann gilt:

Lemma 3.5
Sei $1 \le k < l \le n$, $x \in \mathbb{R}$ und $y \in \mathbb{C}$ mit $x^2 + |y|^2 = 1$. Dann ist

(3.13)

$$\begin{aligned}
\mathrm{tr}\{U_{kl}(x,y)XU^*_{kl}(x,y)B^*\} = \eta_{kl0}(X,B) &+ \eta_{kl1}(X,B)x + \eta_{kl2}(X,B)y\\
&+ \eta_{kl3}(X,B)\overline{y} + \eta_{kl4}(X,B)x^2 + \eta_{kl5}(X,B)xy\\
&+ \eta_{kl6}(X,B)x\overline{y} + \eta_{kl7}(X,B)y^2 + \eta_{kl8}(X,B)\overline{y}^2\\
&+ \eta_{kl9}(X,B)|y|^2.
\end{aligned}$$

Weiter ist für $1 \le m \le n$ und $z \in \mathbb{C}$ mit $|z| = 1$

$$\mathrm{tr}\{V_m(z)XV^*_m(z)B^*\} = \theta_{m0}(X,B) + \theta_{m1}(X,B)z + \theta_{m2}(X,B)\overline{z}. \tag{3.14}$$

Beweis:

$$
\begin{aligned}
&\operatorname{tr}\{U_{kl}(x,y)XU^*_{kl}(x,y)B^*\}\\
&= \operatorname{tr}\left\{\begin{pmatrix} I_{k-1} & 0 & 0 & 0 & 0\\ 0 & x & 0 & y & 0\\ 0 & 0 & I_{l-k-1} & 0 & 0\\ 0 & -\overline{y} & 0 & x & 0\\ 0 & 0 & 0 & 0 & I_{n-l}\end{pmatrix}\begin{pmatrix} X_{kl11} & X_{kl12} & X_{kl13} & X_{kl14} & X_{kl15}\\ X_{kl21} & x_{kk} & X_{kl23} & x_{kl} & X_{kl25}\\ X_{kl31} & X_{kl32} & X_{kl33} & X_{kl34} & X_{kl35}\\ X_{kl41} & x_{lk} & X_{kl43} & x_{ll} & X_{kl45}\\ X_{kl51} & X_{kl52} & X_{kl53} & X_{kl54} & X_{kl55}\end{pmatrix}\right.\\
&\qquad \left.\cdot R^*_{kl}(x,y)B^*\right\}\\
&= \operatorname{tr}\left\{\left(\begin{matrix} X_{kl11} & X_{kl12} & X_{kl13} & X_{kl14}\\ xX_{kl21}+yX_{kl41} & xx_{kk}+yx_{lk} & xX_{kl23}+yX_{kl43} & xx_{kl}+yx_{ll}\\ X_{kl31} & X_{kl32} & X_{kl33} & X_{kl34}\\ -\overline{y}X_{kl21}+xX_{kl41} & -\overline{y}x_{kk}+xx_{lk} & -\overline{y}X_{kl23}+xX_{kl43} & -\overline{y}x_{kl}+xx_{ll}\\ X_{kl51} & X_{kl52} & X_{kl53} & X_{kl54}\end{matrix}\right.\right.\\
&\qquad \left.\left.\begin{matrix} X_{kl15}\\ xX_{kl25}+yX_{kl45}\\ X_{kl35}\\ -\overline{y}X_{kl25}+xX_{kl45}\\ X_{kl55}\end{matrix}\right)\begin{pmatrix} I_{k-1} & 0 & 0 & 0 & 0\\ 0 & x & 0 & -y & 0\\ 0 & 0 & I_{l-k-1} & 0 & 0\\ 0 & \overline{y} & 0 & x & 0\\ 0 & 0 & 0 & 0 & I_{n-l}\end{pmatrix}B^*\right\}\\
&= \operatorname{tr}\left\{\left(\begin{matrix} X_{kl11} & xX_{kl12}+\overline{y}X_{kl14} & X_{kl13}\\ xX_{kl21}+yX_{kl41} & x^2x_{kk}+xyx_{lk}+x\overline{y}x_{kl}+y\overline{y}x_{ll} & xX_{kl23}+yX_{kl43}\\ X_{kl31} & xX_{kl32}+\overline{y}X_{kl34} & X_{kl33}\\ -\overline{y}X_{kl21}+xX_{kl41} & -x\overline{y}x_{kk}+x^2x_{lk}-\overline{y}^2x_{kl}+x\overline{y}x_{ll} & -\overline{y}X_{kl23}+xX_{kl43}\\ X_{kl51} & xX_{kl52}+\overline{y}X_{kl54} & X_{kl53}\end{matrix}\right.\right.\\
&\qquad \left.\left.\begin{matrix} -yX_{kl12}+xX_{kl14} & X_{kl15}\\ -xyx_{kk}-y^2x_{lk}+x^2x_{kl}+xyx_{ll} & xX_{kl25}+yX_{kl45}\\ -yX_{kl32}+xX_{kl34} & X_{kl35}\\ y\overline{y}x_{kk}-xyx_{lk}-x\overline{y}x_{kl}+x^2x_{ll} & -\overline{y}X_{kl25}+xX_{kl45}\\ -yX_{kl52}+xX_{kl54} & X_{kl55}\end{matrix}\right)B^*\right\}\\
&= \operatorname{tr}\{X_{kl11}B^*_{kl11}+(xX_{kl12}+\overline{y}X_{kl14})B^*_{kl12}+X_{kl13}B^*_{kl13}\\
&\qquad +(-yX_{kl12}+xX_{kl14})B^*_{kl14}+X_{kl15}B^*_{kl15}\}\\
&\quad +(xX_{kl21}+yX_{kl41})B^*_{kl21}+(x^2x_{kk}+xyx_{lk}+x\overline{y}x_{kl}+y\overline{y}x_{ll})\overline{b_{kk}}\\
&\quad +(xX_{kl23}+yX_{kl43})B^*_{kl23}+(-xyx_{kk}-y^2x_{lk}+x^2x_{kl}+xyx_{ll})\overline{b_{kl}}\\
&\quad +(xX_{kl25}+yX_{kl45})B^*_{kl25}\\
&\quad +\operatorname{tr}\{X_{kl31}B^*_{kl31}+(xX_{kl32}+\overline{y}X_{kl34})B^*_{kl32}+X_{kl33}B^*_{kl33}\\
&\qquad +(-yX_{kl32}+xX_{kl34})B^*_{kl34}+X_{kl35}B^*_{kl35}\}\\
&\quad +(-\overline{y}X_{kl21}+xX_{kl41})B^*_{kl41}+(-x\overline{y}x_{kk}+x^2x_{lk}-\overline{y}^2x_{kl}+x\overline{y}x_{ll})\overline{b_{lk}}\\
&\quad +(-\overline{y}X_{kl23}+xX_{kl43})B^*_{kl43}+(y\overline{y}x_{kk}-xyx_{lk}-x\overline{y}x_{kl}+x^2x_{ll})\overline{b_{ll}}\\
&\quad +(-\overline{y}X_{kl25}+xX_{kl45})B^*_{kl45}\\
&\quad +\operatorname{tr}\{X_{kl51}B^*_{kl51}+(xX_{kl52}+\overline{y}X_{kl54})B^*_{kl52}+X_{kl53}B^*_{kl53}\\
&\qquad +(-yX_{kl52}+xX_{kl54})B^*_{kl54}+X_{kl55}B^*_{kl55}\}\\
&= \operatorname{tr}(X_{kl11}B^*_{kl11}+X_{kl13}B^*_{kl13}+X_{kl15}B^*_{kl15})\\
&\quad +\operatorname{tr}(X_{kl31}B^*_{kl31}+X_{kl33}B^*_{kl33}+X_{kl35}B^*_{kl35})\\
&\quad +\operatorname{tr}(X_{kl51}B^*_{kl51}+X_{kl53}B^*_{kl53}+X_{kl55}B^*_{kl55})
\end{aligned}
$$

$$
\begin{aligned}
&+ x\{\operatorname{tr}(X_{kl12}B^*_{kl12} + X_{kl14}B^*_{kl14}) + X_{kl21}B^*_{kl21} + X_{kl23}B^*_{kl23} + X_{kl25}B^*_{kl25} \\
&\qquad + \operatorname{tr}(X_{kl32}B^*_{kl32} + X_{kl34}B^*_{kl34}) + X_{kl41}B^*_{kl41} + X_{kl43}B^*_{kl43} + X_{kl45}B^*_{kl45} \\
&\qquad + \operatorname{tr}(X_{kl52}B^*_{kl52} + X_{kl54}B^*_{kl54})\} \\
&+ y\{-\operatorname{tr}(X_{kl12}B^*_{kl14}) + X_{kl41}B^*_{kl21} + X_{kl43}B^*_{kl23} + X_{kl45}B^*_{kl25} - \operatorname{tr}(X_{kl32}B^*_{kl34}) \\
&\qquad - \operatorname{tr}(X_{kl52}B^*_{kl54})\} \\
&+ \overline{y}\{\operatorname{tr}(X_{kl14}B^*_{kl12}) + \operatorname{tr}(X_{kl34}B^*_{kl32}) - X_{kl21}B^*_{kl41} - X_{kl23}B^*_{kl43} - X_{kl25}B^*_{kl45} \\
&\qquad + \operatorname{tr}(X_{kl54}B^*_{kl52})\} \\
&+ x^2\{x_{kk}\overline{b_{kk}} + x_{kl}\overline{b_{kl}} + x_{lk}\overline{b_{lk}} + x_{ll}\overline{b_{ll}}\} \\
&+ xy\{x_{lk}\overline{b_{kk}} - x_{kk}\overline{b_{kl}} + x_{ll}\overline{b_{kl}} - x_{lk}\overline{b_{ll}}\} \\
&+ x\overline{y}\{x_{kl}\overline{b_{kk}} - x_{kk}\overline{b_{lk}} + x_{ll}\overline{b_{lk}} - x_{kl}\overline{b_{ll}}\} \\
&- y^2 x_{lk}\overline{b_{kl}} - \overline{y}^2 x_{kl}\overline{b_{lk}} + |y|^2\{x_{ll}\overline{b_{kk}} + x_{kk}\overline{b_{ll}}\} \\
= {}& \eta_{kl0}(X,B) + \eta_{kl1}(X,B)x + \eta_{kl2}(X,B)y + \eta_{kl3}(X,B)\overline{y} + \eta_{kl4}(X,B)x^2 \\
&+ \eta_{kl5}(X,B)xy + \eta_{kl6}(X,B)x\overline{y} + \eta_{kl7}(X,B)y^2 + \eta_{kl8}(X,B)\overline{y}^2 + \eta_{kl9}(X,B)|y|^2
\end{aligned}
$$

und

$$
\begin{aligned}
&\operatorname{tr}\{V_m(z)XV_m^*(z)B^*\} \\
&= \operatorname{tr}\left\{\begin{pmatrix} I_{m-1} & 0 & 0 \\ 0 & z & 0 \\ 0 & 0 & I_{n-m}\end{pmatrix}\begin{pmatrix} X_{m11} & X_{m12} & X_{m13} \\ X_{m21} & x_{mm} & X_{m23} \\ X_{m31} & X_{m32} & X_{m33}\end{pmatrix}\begin{pmatrix} I_{m-1} & 0 & 0 \\ 0 & \overline{z} & 0 \\ 0 & 0 & I_{n-m}\end{pmatrix} B^*\right\} \\
&= \operatorname{tr}\left\{\begin{pmatrix} X_{m11} & \overline{z}X_{m12} & X_{m13} \\ zX_{m21} & z\overline{z}x_{mm} & zX_{m23} \\ X_{m31} & \overline{z}X_{m32} & X_{m33}\end{pmatrix} B^*\right\} \\
&= \operatorname{tr}(X_{m11}B^*_{m11} + \overline{z}X_{m12}B^*_{m12} + X_{m13}B^*_{m13}) \\
&\quad + zX_{m21}B_{m21} + |z|^2 x_{mm}\overline{b_{mm}} + zX_{m23}B_{m23} \\
&\quad + \operatorname{tr}(X_{m31}B^*_{m31} + \overline{z}X_{m32}B^*_{m32} + X_{m33}B^*_{m33}) \\
&= \operatorname{tr}(X_{m11}B^*_{m11} + X_{m13}B^*_{m13}) + x_{mm}\overline{b_{mm}} + \operatorname{tr}(X_{m31}B^*_{m31} + X_{m33}B^*_{m33}) \\
&\quad + z\{X_{m21}B^*_{m21} + X_{m23}B^*_{m23}\} + \overline{z}\{\operatorname{tr}(X_{m12}B^*_{m12}) + \operatorname{tr}(X_{m32}B^*_{m32})\} \\
&= \theta_{m0}(X,B) + \theta_{m1}(X,B)z + \theta_{m2}(X,B)\overline{z}.
\end{aligned}
$$

□

Unter Zuhilfenahme dieser Ergebnisse können wir die Funktionswerte von g_B auf den Kurven mit den Richtungen $R_{kl}(1)$, $R_{kl}(i)$ und $S_m(i)$ berechnen.

3.1.1 Verfahren für allgemeine B

Sei

$$
\begin{aligned}
&h_1 : \mathbb{R}^5 \times \mathbb{R} \longrightarrow \mathbb{R}, \\
&h_1(\alpha, t) = \alpha_1 \cos^2 t + \alpha_2 \sin^2 t + \alpha_3 \sin t \cos t + \alpha_4 \cos t + \alpha_5 \sin t
\end{aligned}
\tag{3.15}
$$

und

$$
h_2 : \mathbb{R}^2 \times \mathbb{R} \longrightarrow \mathbb{R},\ h_2(\beta, t) = \beta_1 \cos t + \beta_2 \sin t. \tag{3.16}
$$

Definiere für $1 \le k < l \le n$

$$
\rho_{klr} :\ \mathbb{C}^{n,n} \times\ \mathbb{C}^{n,n} \longrightarrow \mathbb{R}, \quad r \in \{0, \ldots, 5\},
$$

durch

$$
(3.17) \qquad \left|\begin{array}{l}
\rho_{kl0}(X,Y) := \mathrm{re}\{\eta_{kl0}(X,Y)\} \\
\rho_{kl1}(X,Y) := \mathrm{re}\{\eta_{kl4}(X,Y)\} \\
\rho_{kl2}(X,Y) := \mathrm{re}\{\eta_{kl7}(X,Y) + \eta_{kl8}(X,Y) + \eta_{kl9}(X,Y)\} \\
\rho_{kl3}(X,Y) := \mathrm{re}\{\eta_{kl5}(X,Y) + \eta_{kl6}(X,Y)\} \\
\rho_{kl4}(X,Y) := \mathrm{re}\{\eta_{kl1}(X,Y)\} \\
\rho_{kl5}(X,Y) := \mathrm{re}\{\eta_{kl2}(X,Y) + \eta_{kl3}(X,Y)\}
\end{array}\right|
$$

und

$$
\sigma_{klr} : \mathbb{C}^{n,n} \times \mathbb{C}^{n,n} \longrightarrow \mathbb{R}, \quad r \in \{0,\ldots,5\},
$$

durch

$$
(3.18) \qquad \left|\begin{array}{l}
\sigma_{kl0}(X,Y) := \mathrm{re}\{\eta_{kl0}(X,Y)\} \\
\sigma_{kl1}(X,Y) := \mathrm{re}\{\eta_{kl4}(X,Y)\} \\
\sigma_{kl2}(X,Y) := \mathrm{re}\{\eta_{kl9}(X,Y) - \eta_{kl7}(X,Y) - \eta_{kl8}(X,Y)\} \\
\sigma_{kl3}(X,Y) := \mathrm{im}\{\eta_{kl6}(X,Y) - \eta_{kl5}(X,Y)\} \\
\sigma_{kl4}(X,Y) := \mathrm{re}\{\eta_{kl1}(X,Y)\} \\
\sigma_{kl5}(X,Y) := \mathrm{im}\{\eta_{kl3}(X,Y) - \eta_{kl2}(X,Y)\}
\end{array}\right| ,
$$

sowie für $1 \leq m \leq n$

$$
\tau_{ms} : \mathbb{C}^{n,n} \times \mathbb{C}^{n,n} \longrightarrow \mathbb{R}, \quad s \in \{0,1,2\},
$$

durch

$$
(3.19) \qquad \left|\begin{array}{l}
\tau_{m0}(X,Y) := \mathrm{re}\{\theta_{m0}(X,Y)\} \\
\tau_{m1}(X,Y) := \mathrm{re}\{\theta_{m1}(X,Y) + \theta_{m2}(X,Y)\} \\
\tau_{m2}(X,Y) := \mathrm{im}\{\theta_{m2}(X,Y) - \theta_{m1}(X,Y)\}
\end{array}\right| .
$$

Dann gilt:

Satz 3.6
Seien $X, B \in \mathbb{C}^{n,n}$, $1 \leq k < l \leq n$ und $1 \leq m \leq n$. Dann ist für $t \in \mathbb{R}$

$$
(3.20) \qquad g_B\big(\mathcal{K}_{X,R_{kl}(1)}(t)\big) = h_1\big(\rho_{kl1}(X,B),\ldots,\rho_{kl5}(X,B),t\big) + \rho_{kl0}(X,B),
$$

$$
(3.21) \qquad g_B\big(\mathcal{K}_{X,R_{kl}(i)}(t)\big) = h_1\big(\sigma_{kl1}(X,B),\ldots,\sigma_{kl5}(X,B),t\big) + \sigma_{kl0}(X,B)
$$

und

$$
(3.22) \qquad g_B\big(\mathcal{K}_{X,S_m(i)}(t)\big) = h_2\big(\tau_{m1}(X,B),\tau_{m2}(X,B),t\big) + \tau_{m0}(X,B).
$$

Beweis:
Nach (3.8) und (3.13) ist

$$
\begin{aligned}
&g_B\big(\mathcal{K}_{X,R_{kl}(1)}(t)\big) \\
&= \mathrm{re}\big\{\mathrm{tr}\big(U_{kl}(\cos t,\sin t)XU_{kl}^*(\cos t,\sin t)B^*\big)\big\} \\
&= \mathrm{re}\{\eta_{kl0}(X,B) + \eta_{kl1}(X,B)\cos t + \eta_{kl2}(X,B)\sin t + \eta_{kl3}(X,B)\sin t \\
&\qquad + \eta_{kl4}(X,B)\cos^2 t + \eta_{kl5}(X,B)\sin t\cos t + \eta_{kl6}(X,B)\sin t\cos t \\
&\qquad + \eta_{kl7}(X,B)\sin^2 t + \eta_{kl8}(X,B)\sin^2 t + \eta_{kl9}(X,B)\sin^2 t\}
\end{aligned}
$$

$$
\begin{aligned}
&= \mathrm{re}\{\eta_{kl4}(X,B)\}\cos^2 t + \mathrm{re}\{\eta_{kl7}(X,B)+\eta_{kl8}(X,B)+\eta_{kl9}(X,B)\}\sin^2 t\\
&\quad + \mathrm{re}\{\eta_{kl5}(X,B)+\eta_{kl6}(X,B)\}\sin t\cos t + \mathrm{re}\{\eta_{kl1}(X,B)\}\cos t\\
&\quad + \mathrm{re}\{\eta_{kl2}(X,B)+\eta_{kl3}(X,B)\}\sin t + \mathrm{re}\{\eta_{kl0}(X,B)\}\\
&= \rho_{kl1}(X,B)\cos^2 t + \rho_{kl2}(X,B)\sin^2 t + \rho_{kl3}(X,B)\sin t\cos t + \rho_{kl4}(X,B)\cos t\\
&\quad + \rho_{kl5}(X,B)\sin t + \rho_{kl0}(X,B)\\
&= h_1\big(\rho_{kl1}(X,B),\rho_{kl2}(X,B),\rho_{kl3}(X,B),\rho_{kl4}(X,B),\rho_{kl5}(X,B),t\big) + \rho_{kl0}(X,B)
\end{aligned}
$$

und nach (3.9) und (3.13)

$$
\begin{aligned}
&g_B\big(\mathcal{K}_{X,R_{kl}(i)}(t)\big)\\
&= \mathrm{re}\big\{\mathrm{tr}\big(U_{kl}(\cos t, i\sin t)XU_{kl}^*(\cos t, i\sin t)B^*\big)\big\}\\
&= \mathrm{re}\{\eta_{kl0}(X,B)+\eta_{kl1}(X,B)\cos t+\eta_{kl2}(X,B)i\sin t-\eta_{kl3}(X,B)i\sin t\\
&\qquad + \eta_{kl4}(X,B)\cos^2 t+\eta_{kl5}(X,B)i\sin t\cos t-\eta_{kl6}(X,B)i\sin t\cos t\\
&\qquad - \eta_{kl7}(X,B)\sin^2 t-\eta_{kl8}(X,B)\sin^2 t+\eta_{kl9}(X,B)\sin^2 t\}\\
&= \mathrm{re}\{\eta_{kl4}(X,B)\}\cos^2 t + \mathrm{re}\{\eta_{kl9}(X,B)-\eta_{kl7}(X,B)-\eta_{kl8}(X,B)\}\sin^2 t\\
&\quad + \mathrm{im}\{\eta_{kl6}(X,B)-\eta_{kl5}(X,B)\}\sin t\cos t + \mathrm{re}\{\eta_{kl1}(X,B)\}\cos t\\
&\quad + \mathrm{im}\{\eta_{kl3}(X,B)-\eta_{kl2}(X,B)\}\sin t + \mathrm{re}\{\eta_{kl0}(X,B)\}\\
&= \sigma_{kl1}(X,B)\cos^2 t + \sigma_{kl2}(X,B)\sin^2 t + \sigma_{kl3}(X,B)\sin t\cos t + \sigma_{kl4}(X,B)\cos t\\
&\quad + \sigma_{kl5}(X,B)\sin t + \sigma_{kl0}(X,B)\\
&= h_1\big(\sigma_{kl1}(X,B),\sigma_{kl2}(X,B),\sigma_{kl3}(X,B),\sigma_{kl4}(X,B),\sigma_{kl5}(X,B),t\big) + \sigma_{kl0}(X,B),
\end{aligned}
$$

sowie nach (3.10) und (3.14)

$$
\begin{aligned}
&g_B\big(\mathcal{K}_{X,S_m(i)}(t)\big)\\
&= \mathrm{re}\big\{\mathrm{tr}\big(V_m(e^{it})XV_m^*(e^{it})B^*\big)\big\}\\
&= \mathrm{re}\{\theta_{m0}(X,B)+\theta_{m1}(X,B)e^{it}+\theta_{m2}(X,B)e^{-it}\}\\
&= \mathrm{re}\{\theta_{m0}(X,B)+\theta_{m1}(X,B)(\cos t+i\sin t)+\theta_{m2}(X,B)(\cos t-i\sin t\}\\
&= \mathrm{re}\big\{\theta_{m0}(X,B)+\big(\theta_{m1}(X,B)+\theta_{m2}(X,B)\big)\cos t\\
&\qquad + i\big(\theta_{m1}(X,B)-\theta_{m2}(X,B)\big)\sin t\big\}\\
&= \mathrm{re}\{\theta_{m0}(X,B)\} + \mathrm{re}\{\theta_{m1}(X,B)+\theta_{m2}(X,B)\}\cos t\\
&\qquad + \mathrm{im}\{\theta_{m2}(X,B)-\theta_{m1}(X,B)\}\sin t\\
&= h_2\big(\tau_{m1}(X,B),\tau_{m2}(X,B),t\big) + \tau_{m0}(X,B).
\end{aligned}
$$

□

Die Extremalstellen von g_B auf den Kurven mit den Richtungen $R_{kl}(1)$ bzw. $R_{kl}(i)$ sind also die Extremalstellen der Funktion h_1 mit den Parametern $\rho_{kl1},\ldots,\rho_{kl5}$ bzw. $\sigma_{kl1},\ldots,\sigma_{kl5}$. Ebenso sind die Extremalstellen von g_B auf den Kurven, die die Richtung $S_m(i)$ besitzen, die Extremalstellen der Funktion h_2 mit den Parametern τ_{m1},τ_{m2}. Zur Charakterisierung der Extremalstellen von h_1 verwenden wir das Polynom

$$
\begin{aligned}
&p : \mathbb{R}^4\times\mathbb{R}\longrightarrow\mathbb{R},\\
(3.23)\qquad &p(\gamma,x) = 4(\gamma_1^2+\gamma_2^2)x^4 + 4(\gamma_1\gamma_3+\gamma_2\gamma_4)x^3 + (-4\gamma_1^2-4\gamma_2^2+\gamma_3^2+\gamma_4^2)x^2\\
&\qquad - (4\gamma_1\gamma_3+2\gamma_2\gamma_4)x + \gamma_2^2-\gamma_3^2.
\end{aligned}
$$

Satz 3.7
Sei $\alpha \in \mathbb{R}^5$. Dann gilt

(3.24)
$$h_1(\alpha, t) \in \left\{\max_{T\in\mathbb{R}} h_1(\alpha, T), \min_{T\in\mathbb{R}} h_1(\alpha, T)\right\} \Longrightarrow p(\alpha_1 - \alpha_2, \alpha_3, \alpha_4, \alpha_5, \cos t) = 0.$$

Beweis:
Sei $t \in \mathbb{R}$ eine Maximal- oder Minimalstelle von h_1. Dann ist

$$\begin{aligned} 0 &= \frac{\partial h_1}{\partial t}(\alpha, t) \\ &= \alpha_3(\cos^2 t - \sin^2 t) + 2(\alpha_2 - \alpha_1)\sin t \cos t + \alpha_5 \cos t - \alpha_4 \sin t \\ &= \alpha_3(2\cos^2 t - 1) + \alpha_5 \cos t - \{2(\alpha_1 - \alpha_2)\cos t + \alpha_4\}\sin t, \end{aligned}$$

also

$$\alpha_3(2\cos^2 t - 1) + \alpha_5 \cos t = \{2(\alpha_1 - \alpha_2)\cos t + \alpha_4\}\sin t,$$

woraus

$$\begin{aligned} &4\alpha_3^2 \cos^4 t + 4\alpha_3\alpha_5 \cos^3 t + (-4\alpha_3^2 + \alpha_5^2)\cos^2 t - 2\alpha_3\alpha_5 \cos t + \alpha_3^2 \\ = {}& \alpha_3^2(4\cos^4 t - 4\cos^2 t + 1) + 2\alpha_3\alpha_5(2\cos^3 t - \cos t) + \alpha_5^2 \cos^2 t \\ = {}& \{\alpha_3(2\cos^2 t - 1) + \alpha_5 \cos t\}^2 \\ = {}& \{2(\alpha_1 - \alpha_2)\cos t + \alpha_4\}^2(1 - \cos^2 t) \\ = {}& \{4(\alpha_1 - \alpha_2)^2 \cos^2 t + 4(\alpha_1 - \alpha_2)\alpha_4 \cos t + \alpha_4^2\}(1 - \cos^2 t) \\ = {}& -4(\alpha_1 - \alpha_2)^2 \cos^4 t - 4(\alpha_1 - \alpha_2)\alpha_4 \cos^3 t + \{4(\alpha_1 - \alpha_2)^2 - \alpha_4^2\}\cos^2 t \\ & + 4(\alpha_1 - \alpha_2)\alpha_4 \cos t + \alpha_4^2 \end{aligned}$$

folgt, also

$$\begin{aligned} 0 &= 4\{(\alpha_1 - \alpha_2)^2 + \alpha_3^2\}\cos^4 t + 4\{(\alpha_1 - \alpha_2)\alpha_4 + \alpha_3\alpha_5\}\cos^3 t \\ &\quad + \{-4(\alpha_1 - \alpha_2)^2 - 4\alpha_3^2 + \alpha_4^2 + \alpha_5^2\}\cos^2 t \\ &\quad - \{4(\alpha_1 - \alpha_2)\alpha_4 + 2\alpha_3\alpha_5\}\cos t + \alpha_3^2 - \alpha_4^2 \\ &= p(\alpha_1 - \alpha_2, \alpha_3, \alpha_4, \alpha_5, \cos t). \end{aligned}$$

□

Eine Maximal- oder Minimalstelle von h_1 ist also eine Nullstelle des Polynoms p und damit eine Lösung einer Gleichung 4. Grades. Da in den Darstellungen (3.8) und (3.9) der Kurven mit den Richtungen $R_{kl}(1)$ und $R_{kl}(i)$ die Kurvenparameter nicht direkt erscheinen, sondern ihre Sinus- und Kosinuswerte, wollen wir die Extremalstellen von g_B nicht direkt berechnen, sondern die Werte der Sinus- und Kosinusfunktion an diesen Stellen. Dazu benutzen wir den folgenden Algorithmus, der das Paar $(\cos t_+, \sin t_+)$ für eine Maximalstelle t_+ von h_1 berechnet. Zur Vereinfachung der Notation führen wir die Funktion

(3.25)
$$\begin{aligned} \widetilde{h_1} &: \mathbb{R}^5 \times \{(x, y) \in \mathbb{R} \times \mathbb{R} : x^2 + y^2 = 1\} \longrightarrow \mathbb{R}, \\ \widetilde{h_1}(\alpha, c, s) &= \alpha_1 c^2 + \alpha_2 s^2 + \alpha_3 sc + \alpha_4 c + \alpha_5 s \end{aligned}$$

ein. Für alle $t \in \mathbb{R}$ ist dann

(3.26)
$$h_1(\alpha, t) = \widetilde{h_1}(\alpha, \cos t, \sin t).$$

Algorithmus 3.3 (Maximierungsverfahren für h_1)
Eingabe: $\alpha \in \mathbb{R}^5$

- Falls $\alpha_1 - \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$, so
    - setze $c_+(\alpha) := 1$ und $s_+(\alpha) := 0$ und beende das Verfahren,
- sonst
    - berechne die Lösungen $x_1, \dots, x_r$, $r \in \{2,4\}$, von $p(\alpha_1 - \alpha_2, \alpha_3, \alpha_4, \alpha_5, x) = 0$,
    - bestimme $\{c_1, \dots, c_j\} := \{x_1, \dots, x_r\} \cap [-1,1]$,
    - für $l = 1, \dots, j$
        - berechne $y_{1l} := \sqrt{1 - c_l^2}$ und $y_{2l} := -\sqrt{1 - c_l^2}$,
    - bestimme $h_1^+ := \max\left\{\widetilde{h_1}(\alpha, c_l, y_{kl}) : k \in \{1,2\}, l \in \{1, \dots, j\}\right\}$,
    - falls $\widetilde{h_1}(\alpha, 1, 0) = h_1^+$, so
        - setze $c_+(\alpha) := 1$ und $s_+(\alpha) := 0$ und beende das Verfahren,
    - sonst
        - wähle $k \in \{1,2\}$ und $l \in \{1, \dots, j\}$ mit $\widetilde{h_1}(\alpha, c_l, y_{kl}) = h_1^+$ beliebig und setze $c_+(\alpha) := c_l$ und $s_+(\alpha) := y_{kl}$.

Ausgabe: $c_+(\alpha), s_+(\alpha)$

Ersetzen wir im obigen Algorithmus h_1^+ durch den Wert

$$h_1^- := \min\left\{\widetilde{h_1}(\alpha, c_l, y_{kl}) : k \in \{1,2\}, l \in \{1, \dots, j\}\right\}$$

und $c_+(\alpha)$ durch $c_-(\alpha)$, sowie $s_+(\alpha)$ durch $s_-(\alpha)$, so erhalten wir einen Algorithmus, der ein Paar

$$\big(c_-(\alpha), s_-(\alpha)\big) = (\cos t_-, \sin t_-)$$

für eine Minimalstelle t_- von $h_1(\alpha, t)$ berechnet.

Im Gegensatz zur Funktion h_1 lassen sich die Extremalstellen der Funktion h_2 direkt bestimmen.

Satz 3.8
Sei $\beta \in \mathbb{R}^2$, sowie

$$t_+(\beta) := \begin{cases} \operatorname{arccot} \frac{\beta_1}{\beta_2} & , \quad \beta_2 > 0 \\ \operatorname{arccot} \frac{\beta_1}{\beta_2} + \pi & , \quad \beta_2 < 0 \\ 0 & , \quad \beta_1 \geq 0 = \beta_2 \\ \pi & , \quad \beta_1 < 0 = \beta_2 \end{cases} \tag{3.27}$$

und

$$t_-(\beta) := \begin{cases} \operatorname{arccot} \frac{\beta_1}{\beta_2} + \pi & , \quad \beta_2 > 0 \\ \operatorname{arccot} \frac{\beta_1}{\beta_2} & , \quad \beta_2 < 0 \\ \pi & , \quad \beta_1 > 0 = \beta_2 \\ 0 & , \quad \beta_1 \leq 0 = \beta_2 \end{cases} . \tag{3.28}$$

Dann gilt

$$\max_{t\in\mathbb{R}} h_2(\beta,t) = h_2\big(\beta,t_+(\beta)\big) = \sqrt{\beta_1^2+\beta_2^2} \tag{3.29}$$

und

$$\min_{t\in\mathbb{R}} h_2(\beta,t) = h_2\big(\beta,t_-(\beta)\big) = -\sqrt{\beta_1^2+\beta_2^2}. \tag{3.30}$$

Beweis:
Es ist

$$\frac{\partial h_2}{\partial t}(\beta,t) = -\beta_1 \sin t + \beta_2 \cos t = 0$$

genau dann, wenn $\beta_1 \sin t = \beta_2 \cos t$ ist.

<u>Fall 1: $\beta_2 = 0$</u>
Für $\beta_1 = 0$ sind (3.29) und (3.30) trivial. Für $\beta_1 \neq 0$ ist $\frac{\partial h_2}{\partial t}(\beta,t) = 0$ für $t_1 = 0$ und $t_2 = \pi$. Aus $h_2(0) = \beta_1$ und $h_2(\pi) = -\beta_1$ folgen damit (3.29) und (3.30).

<u>Fall 2: $\beta_2 \neq 0$</u>
In diesem Fall ist $\frac{\partial h_2}{\partial t}(\beta,t) = 0$ genau dann, wenn $\cot t = \frac{\beta_1}{\beta_2}$, also für $t_1 = \operatorname{arccot}\frac{\beta_1}{\beta_2}$ und $t_2 = \operatorname{arccot}\frac{\beta_1}{\beta_2} + \pi$. Aus

$$\begin{aligned}
h_2(\beta,t_1) &= \beta_1 \cos\left(\operatorname{arccot}\frac{\beta_1}{\beta_2}\right) + \beta_2 \sin\left(\operatorname{arccot}\frac{\beta_1}{\beta_2}\right) \\
&= \beta_1 \frac{\frac{\beta_1}{\beta_2}}{\sqrt{1+\frac{\beta_1^2}{\beta_2^2}}} + \beta_2 \frac{1}{\sqrt{1+\frac{\beta_1^2}{\beta_2^2}}} \\
&= \frac{\beta_1^2+\beta_2^2}{\operatorname{sign}(\beta_2)\sqrt{\beta_1^2+\beta_2^2}} \\
&= \operatorname{sign}(\beta_2)\sqrt{\beta_1^2+\beta_2^2}
\end{aligned}$$

und

$$\begin{aligned}
h_2(\beta,t_2) &= \beta_1 \cos\left(\operatorname{arccot}\frac{\beta_1}{\beta_2}+\pi\right) + \beta_2 \sin\left(\operatorname{arccot}\frac{\beta_1}{\beta_2}+\pi\right) \\
&= -\beta_1 \cos\left(\operatorname{arccot}\frac{\beta_1}{\beta_2}\right) - \beta_2 \sin\left(\operatorname{arccot}\frac{\beta_1}{\beta_2}\right) \\
&= -\operatorname{sign}(\beta_2)\sqrt{\beta_1^2+\beta_2^2}
\end{aligned}$$

folgen (3.29) und (3.30). □

Fassen wir zusammen: Wir wollen g_B auf $\Lambda(A)$ optimieren, indem wir g_B sukzessiv über alle Kurven mit den Richtungen $R_{kl}(1)$, $R_{kl}(i)$ und $S_m(i)$ optimieren. Dazu haben wir in Satz 3.6 gezeigt, daß sich g_B auf diesen Kurven wie die Funktion h_1 oder h_2 verhält. Mit Algorithmus 3.3 wurde ein Verfahren zur Maximierung von h_1 vorgestellt, aus dem sich durch geringfügige Modifikationen ein Verfahren zur Minimierung von h_1 ergibt. In Satz 3.8 haben wir die Extremalstellen von h_2 bestimmt. Dies alles zusammen liefert uns explizite Verfahren zur Maximierung und Minimierung von g_B auf $\Lambda(A)$. Ein Maximierungs-Sweep für g_B, also ein Zyklus, in dem g_B nacheinander über jede der Richtungen $R_{kl}(1)$, $R_{kl}(i)$ und $S_m(i)$ genau einmal maximiert wird, sieht demnach folgendermaßen aus:

Algorithmus 3.4 (Maximierungs-Sweep für g_B)
Eingabe: $X, B \in \mathbb{C}^{n,n}$

- Berechne $\beta_{1r} := \tau_{1r}(X, B)$, $r \in \{1, 2\}$, nach (3.19).
- Berechne $t_1 := t_+(\beta_{11}, \beta_{12})$ nach (3.27).
- Setze $X^{(12)} := V_1(e^{it_1})XV_1^*(e^{it_1})$.
- Für $l = 2, \dots, n$
  - für $k = 1, \dots, l-1$
    * berechne $\alpha_{klr}^{(1)} := \rho_{klr}\big(X^{(kl)}, B\big)$, $r \in \{1, \dots, 5\}$, nach (3.17),
    * berechne $c_{kl}^{(1)} := c_+\big(\alpha_{kl1}^{(1)}, \dots, \alpha_{kl5}^{(1)}\big)$ und $s_{kl}^{(1)} := s_+\big(\alpha_{kl1}^{(1)}, \dots, \alpha_{kl5}^{(1)}\big)$ mit Algorithmus 3.3,
    * setze $Y^{(kl)} := U_{kl}\big(c_{kl}^{(1)}, s_{kl}^{(1)}\big)X^{(kl)}U_{kl}^*\big(c_{kl}^{(1)}, s_{kl}^{(1)}\big)$,
    * berechne $\alpha_{klr}^{(2)} := \sigma_{klr}\big(Y^{(kl)}, B\big)$, $r \in \{1, \dots, 5\}$, nach (3.18),
    * berechne $c_{kl}^{(2)} := c_+\big(\alpha_{kl1}^{(2)}, \dots, \alpha_{kl5}^{(2)}\big)$ und $s_{kl}^{(2)} := s_+\big(\alpha_{kl1}^{(2)}, \dots, \alpha_{kl5}^{(2)}\big)$ mit Algorithmus 3.3,
    * setze $X^{(k+1,l)} := U_{kl}\big(c_{kl}^{(2)}, is_{kl}^{(2)}\big)Y^{(kl)}U_{kl}^*\big(c_{kl}^{(2)}, is_{kl}^{(2)}\big)$,
  - berechne $\beta_{lr} := \tau_{lr}(X^{(ll)}, B)$, $r \in \{1, 2\}$, nach (3.19),
  - berechne $t_l := t_+(\beta_{l1}, \beta_{l2})$ nach (3.27),
  - setze $X^{(1,l+1)} := V_l(e^{it_l})X^{(ll)}V_l^*(e^{it_l})$.

Ausgabe: $\mathrm{maxsweep}(X, B) := X^{(1,n+1)}$

Werden anstelle von $c_+(\alpha)$, $s_+(\alpha)$ und $t_+(\beta)$ die Werte $c_-(\alpha)$, $s_-(\alpha)$ und $t_-(\beta)$ berechnet, so erhält man einen Minimierungs-Sweep für g_B.

Da bei zwei aufeinanderfolgenden Iterierten im allgemeinen die Elemente von zwei ganzen Zeilen und Spalten differieren, die ρ_{klr} und σ_{klr} aber von allen Elementen der k-ten und l-ten Zeile und Spalte der Iterierten $X^{(kl)}$ bzw. $Y^{(kl)}$ abhängen, also insbesondere auch von Elementen, die sich nach der vorherigen Iteration geändert haben, sind der Maximierungs-Sweep Algorithmus 3.4 und der zugehörige Minimierungs-Sweep im Unterschied zu anderen jacobiartigen Verfahren nicht parallelisierbar.

Insgesamt erhalten wir das folgende jacobiartige Verfahren zur Berechnung einer Matrix aus $\Lambda(A)$, die die Funktion g_B maximiert, also nach (2.14) die Funktion f_B minimiert und damit das isospektrale Minimierungsproblem (2.6) löst. Als Abbruchkriterium wählen wir die Differenz der Funktionswerte von g_B zwischen den Iterierten am Ende zweier aufeinanderfolgender Sweeps, das heißt wir beenden das Verfahren nach dem ersten Sweep, in dem sich g_B um weniger als eine vorgegebene Abbruchschranke $\epsilon > 0$ vermehrt hat.

Algorithmus 3.5 (Jacobiartiges Verfahren zur Maximierung von g_B)
Eingabe: $A, B \in \mathbb{C}^{n,n}$, Abbruchschranke $\epsilon > 0$

- Setze $A^{(0)} := A$.
- Berechne $g^{(0)} := g_B\big(A^{(0)}\big)$.
- Für $q = 0, 1, \dots$
  - berechne $A^{(q+1)} := \mathrm{maxsweep}\big(A^{(q)}, B\big)$ mit Algorithmus 3.4,

- berechne $g^{(q+1)} := g_B\big(A^{(q+1)}\big)$,
- falls $g^{(q+1)} - g^{(q)} < \epsilon$, so beende die Schleife.

Ausgabe: $A_{\text{Jacobi}}(A,B) := A^{(q+1)}$

Da dieses Verfahren eine explizite Version von Algorithmus 3.1 ist, gilt der Konvergenzsatz 3.1.

Wird im obigen Verfahren anstelle eines Maximierungs-Sweeps ein Minimierungs-Sweep durchgeführt und das Abbruchkriterium

$$g^{(q+1)} - g^{(q)} < \epsilon$$

durch

$$g^{(q)} - g^{(q+1)} < \epsilon$$

ersetzt, so erhält man ein Verfahren zur Bestimmung einer zu A unitär ähnlichen Matrix, die die Funktion g_B minimiert, also die Funktion f_B nach (2.13) maximiert und so das isospektrale Maximierungsproblem (2.5) löst. Dieses Verfahren ist dann eine explizite Version von Algorithmus 3.2, weshalb sein Konvergenzverhalten durch Satz 3.1 beschrieben wird.

Es stellt sich die Frage, ob für das Maximierungsverfahren Algorithmus 3.5 und das zugehörige Minimierungsverfahren ein besseres Konvergenzresultat als Satz 3.1 erzielt werden kann. In Beispiel 3.12 im nächsten Abschnitt zeigen wir jedoch, daß die beiden Verfahren im allgemeinen keine globale Konvergenz aufweisen. Die Konvergenz von Algorithmus 3.5 wird in Abschnitt 5.1 anhand einiger Beispiele studiert.

Für diagonale B existieren einfachere, weniger aufwendige Verfahren zur Lösung des isospektralen Abstandsproblems, die wir im folgenden Abschnitt entwickeln werden.

3.1.2 Verfahren für diagonale B

Für $X \in \Lambda(A)$ und $1 \le k < l \le n$ sei

$$\mathcal{L}_{X,k,l} : \mathbb{R} \times \mathbb{R} \longrightarrow \Lambda(A),\ \mathcal{L}_{X,k,l}(\phi,\omega) = e^{R_{kl}(\phi e^{i\omega})} X e^{-R_{kl}(\phi e^{i\omega})}. \tag{3.31}$$

Dann ist $\mathcal{L}_{X,k,l}$ eine Fläche auf $\Lambda(A)$, die X enthält. Das folgende Lemma gibt Auskunft über die wichtigsten Eigenschaften dieser Flächen.

Lemma 3.9
Sei $X \in \mathbb{C}^{n,n}$ und $1 \le k < l \le n$. Dann ist für $\phi, \omega \in \mathbb{R}$

$$\mathcal{L}_{X,k,l}(\phi,\omega) = U_{kl}(\cos\phi, \sin\phi\, e^{i\omega}) X U_{kl}^*(\cos\phi, \sin\phi\, e^{i\omega}), \tag{3.32}$$

sowie

$$\{\mathcal{K}_{X,R_{kl}(1)}(t) : t \in \mathbb{R}\} \subset \{\mathcal{L}_{X,k,l}(\phi,\omega) : \phi,\omega \in \mathbb{R}\} \tag{3.33}$$

und

$$\{\mathcal{K}_{X,R_{kl}(i)}(t) : t \in \mathbb{R}\} \subset \{\mathcal{L}_{X,k,l}(\phi,\omega) : \phi,\omega \in \mathbb{R}\}. \tag{3.34}$$

Beweis:
Nach (3.6) ist

$$
\begin{aligned}
e^{R_{kl}(\phi e^{i\omega})} &= U_{kl}\left(\cos|\phi e^{i\omega}|, \frac{\phi e^{i\omega}}{|\phi e^{i\omega}|}\sin|\phi e^{i\omega}|\right) \\
&= U_{kl}\big(\cos\phi, \operatorname{sign}(\phi)e^{i\omega}\sin|\phi|\big) \\
&= U_{kl}(\cos\phi, \sin\phi\, e^{i\omega}),
\end{aligned}
$$

also

$$
\begin{aligned}
\mathcal{L}_{X,k,l}(\phi,\omega) &= e^{R_{kl}(\phi e^{i\omega})} X e^{-R_{kl}(\phi e^{i\omega})} \\
&= U_{kl}(\cos\phi, \sin\phi\, e^{i\omega}) X U_{kl}^*(\cos\phi, \sin\phi\, e^{i\omega}).
\end{aligned}
$$

Weiter ist nach (3.8)

$$
\begin{aligned}
\mathcal{K}_{X,R_{kl}(1)}(t) &= U_{kl}(\cos t, \sin t) X U_{kl}^*(\cos t, \sin t) \\
&= \mathcal{L}_{X,k,l}(t,0)
\end{aligned}
$$

und nach (3.9)

$$
\begin{aligned}
\mathcal{K}_{X,R_{kl}(i)}(t) &= U_{kl}(\cos t, i\sin t) X U_{kl}^*(\cos t, i\sin t) \\
&= U_{kl}(\cos t, \sin t\, e^{i\frac{\pi}{2}}) X U_{kl}^*(\cos t, \sin t\, e^{i\frac{\pi}{2}}) \\
&= \mathcal{L}_{X,k,l}\left(t, \frac{\pi}{2}\right).
\end{aligned}
$$

□

Im vorigen Abschnitt haben wir Verfahren zur Optimierung von g_B dadurch gewonnen, daß wir g_B nacheinander über die Kurven mit den Richtungen $R_{kl}(1)$, $R_{kl}(i)$ und $S_m(i)$ optimiert haben. Da die Kurven mit den Richtungen $R_{kl}(1)$ und $R_{kl}(i)$ auf den in (3.31) definierten Flächen liegen, können wir stattdessen auch g_B nacheinander auf diesen Flächen und auf den Kurven mit den Richtungen $S_m(i)$ optimieren. Dazu untersuchen wir, wie sich g_B für diagonale Matrizen B auf diesen Kurven und Flächen verhält.

Sei

$$
\begin{aligned}
& h : \mathbb{R}^3 \times \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}, \\
& h(\delta,\phi,\omega) = \delta_1 \cos^2\phi + \sin\phi\cos\phi(\delta_2\cos\omega + \delta_3\sin\omega).
\end{aligned}
\tag{3.35}
$$

Definiere für $1 \le k < l \le n$

$$
\kappa_{klr} : \mathbb{C}^{n,n} \times \mathbb{C}^{n,n} \longrightarrow \mathbb{R}, \quad r \in \{0,\dots,3\},
$$

durch

$$
\left|\begin{aligned}
\kappa_{kl0}(X,Y) &:= \operatorname{re}\left\{\sum_{\substack{j=1\\ j\notin\{k,l\}}}^{n} x_{jj}\overline{y_{jj}} + x_{ll}\overline{y_{kk}} + x_{kk}\overline{y_{ll}}\right\} \\
\kappa_{kl1}(X,Y) &:= \operatorname{re}\{(x_{kk}-x_{ll})(\overline{y_{kk}}-\overline{y_{ll}})\} \\
\kappa_{kl2}(X,Y) &:= \operatorname{re}\{(x_{kl}+x_{lk})(\overline{y_{kk}}-\overline{y_{ll}})\} \\
\kappa_{kl3}(X,Y) &:= \operatorname{im}\{(x_{kl}-x_{lk})(\overline{y_{kk}}-\overline{y_{ll}})\}
\end{aligned}\right|.
\tag{3.36}
$$

Dann gilt:

Satz 3.10
Seien $X, B \in \mathbb{C}^{n,n}$, B diagonal und $1 \le k < l \le n$, sowie $\phi, \omega \in \mathbb{R}$. Dann ist

$$g_B\big(\mathcal{L}_{X,k,l}(\phi,\omega)\big) = h\big(\kappa_{kl1}(X,B), \kappa_{kl2}(X,B), \kappa_{kl3}(X,B), \phi, \omega\big) + \kappa_{kl0}(X,B). \tag{3.37}$$

Weiter ist für $1 \le m \le n$ und $t \in \mathbb{R}$

$$g_B\big(\mathcal{K}_{X,S_m(i)}(t)\big) = g_B(X). \tag{3.38}$$

Beweis:
Für diagonales B ist nach (3.11)

$$\begin{aligned}
\eta_{kl0}(X,B) &= \operatorname{tr}(X_{kl11}B^*_{kl11}) + \operatorname{tr}(X_{kl33}B^*_{kl33}) + \operatorname{tr}(X_{kl55}B^*_{kl55}) \\
&= \sum_{\substack{j=1\\ j\notin\{k,l\}}}^{n} x_{jj}\overline{b_{jj}}, \\
\eta_{kl1}(X,B) &= \eta_{kl2}(X,B) = \eta_{kl3}(X,B) = 0, \\
\eta_{kl4}(X,B) &= x_{kk}\overline{b_{kk}} + x_{ll}\overline{b_{ll}}, \\
\eta_{kl5}(X,B) &= x_{lk}\overline{b_{kk}} - x_{lk}\overline{b_{ll}}, \\
\eta_{kl6}(X,B) &= x_{kl}\overline{b_{kk}} - x_{kl}\overline{b_{ll}}, \\
\eta_{kl7}(X,B) &= \eta_{kl8}(X,B) = 0
\end{aligned}$$

und

$$\eta_{kl9}(X,B) = x_{ll}\overline{b_{kk}} + x_{kk}\overline{b_{ll}},$$

also nach (3.32) und (3.13)

$$\begin{aligned}
&g_B\big(\mathcal{L}_{X,k,l}(\phi,\omega)\big) \\
&= g_B\big(U_{kl}(\cos\phi, \sin\phi\, e^{i\omega})XU^*_{kl}(\cos\phi, \sin\phi\, e^{i\omega})\big) \\
&= \operatorname{re}\big\{\operatorname{tr}\big(U_{kl}(\cos\phi, \sin\phi\, e^{i\omega})XU^*_{kl}(\cos\phi, \sin\phi\, e^{i\omega})B^*\big)\big\} \\
&= \operatorname{re}\{\eta_{kl0}(X,B) + \eta_{kl4}(X,B)\cos^2\phi + \eta_{kl5}(X,B)\cos\phi\sin\phi\, e^{i\omega} \\
&\quad + \eta_{kl6}(X,B)\cos\phi\sin\phi\, e^{-i\omega} + \eta_{kl9}(X,B)\sin^2\phi\} \\
&= \operatorname{re}\{\eta_{kl0}(X,B)\} + \operatorname{re}\{\eta_{kl4}(X,B)\}\cos^2\phi + \operatorname{re}\{\eta_{kl9}(X,B)\}(1-\cos^2\phi) \\
&\quad + \sin\phi\cos\phi\operatorname{re}\{\eta_{kl5}(X,B)(\cos\omega + i\sin\omega) + \eta_{kl6}(X,B)(\cos\omega - i\sin\omega)\} \\
&= \operatorname{re}\{\eta_{kl0}(X,B) + \eta_{kl9}(X,B)\} + \operatorname{re}\{\eta_{kl4}(X,B) - \eta_{kl9}(X,B)\}\cos^2\phi \\
&\quad + \sin\phi\cos\phi\big\{\operatorname{re}\big(\eta_{kl5}(X,B) + \eta_{kl6}(X,B)\big)\cos\omega \\
&\qquad\qquad + \operatorname{im}\big(\eta_{kl6}(X,B) - \eta_{kl5}(X,B)\big)\sin\omega\big\} \\
&= \kappa_{kl0}(X,B) + \kappa_{kl1}(X,B)\cos^2\phi + \sin\phi\cos\phi\{\kappa_{kl2}(X,B)\cos\omega + \kappa_{kl3}(X,B)\sin\omega\} \\
&= h\big(\kappa_{kl1}(X,B), \kappa_{kl2}(X,B), \kappa_{kl3}(X,B), \phi, \omega\big) + \kappa_{kl0}(X,B).
\end{aligned}$$

Weiter ist nach (3.12)

$$\theta_{m0}(X,B) = \operatorname{tr}(X_{m11}B^*_{m11}) + x_{mm}\overline{b_{mm}} + \operatorname{tr}(X_{m33}B^*_{m33}) = \sum_{j=1}^{n} x_{jj}\overline{b_{jj}} = \operatorname{tr}(XB^*)$$

und

$$\theta_{m1}(X,B) = \theta_{m2}(X,B) = 0,$$

also nach (3.10) und (3.14)

$$\begin{aligned} g_B\big(\mathcal{K}_{X,S_m(i)}(t)\big) &= g_B\big(V_m(e^{it})XV_m^*(e^{it})\big) = \mathrm{re}\{\theta_{m0}(X,B)\} = \mathrm{re}\{\mathrm{tr}(XB^*)\} \\ &= g_B(X). \end{aligned}$$

□

Da g_B für diagonale B auf den Kurven mit den Richtungen $S_m(i)$ konstant ist, brauchen wir g_B auch nicht über diese Kurven zu optimieren. Somit besteht ein jacobiartiges Verfahren zur Optimierung von g_B auf $\Lambda(A)$ darin, g_B nacheinander über die in (3.31) definierten Flächen zu optimieren. Dazu bestimmen wir die Extremalstellen der Funktion h. Wir bemerken an dieser Stelle, daß g_B für nichtdiagonale Matrizen B auf diesen Flächen eine wesentlich kompliziertere Funktion als h annimmt, deren Extremalstellen im allgemeinen analytisch nicht bestimmt werden können.

Satz 3.11
Sei $\delta \in \mathbb{R}^3$, sowie

$$(3.39)\qquad \phi_+(\delta) := \begin{cases} \frac{1}{2}\operatorname{arccot}\frac{\delta_1}{\sqrt{\delta_2^2+\delta_3^2}} & , \quad \delta_3 > 0 \\ \frac{1}{2}\operatorname{arccot}\left(-\frac{\delta_1}{\sqrt{\delta_2^2+\delta_3^2}}\right) + \frac{\pi}{2} & , \quad \delta_3 < 0 \\ \frac{1}{2}\operatorname{arccot}\frac{\delta_1}{\delta_2} & , \quad \delta_2 > 0 = \delta_3 \\ \frac{1}{2}\operatorname{arccot}\frac{\delta_1}{\delta_2} + \frac{\pi}{2} & , \quad \delta_2 < 0 = \delta_3 \\ 0 & , \quad \delta_1 \geq 0 = \delta_2 = \delta_3 \\ \frac{\pi}{2} & , \quad \delta_1 < 0 = \delta_2 = \delta_3 \end{cases} ,$$

$$(3.40)\qquad \phi_-(\delta) := \begin{cases} \frac{1}{2}\operatorname{arccot}\frac{\delta_1}{\sqrt{\delta_2^2+\delta_3^2}} + \frac{\pi}{2} & , \quad \delta_3 > 0 \\ \frac{1}{2}\operatorname{arccot}\left(-\frac{\delta_1}{\sqrt{\delta_2^2+\delta_3^2}}\right) & , \quad \delta_3 < 0 \\ \frac{1}{2}\operatorname{arccot}\frac{\delta_1}{\delta_2} + \frac{\pi}{2} & , \quad \delta_2 > 0 = \delta_3 \\ \frac{1}{2}\operatorname{arccot}\frac{\delta_1}{\delta_2} & , \quad \delta_2 < 0 = \delta_3 \\ \frac{\pi}{2} & , \quad \delta_1 > 0 = \delta_2 = \delta_3 \\ 0 & , \quad \delta_1 \leq 0 = \delta_2 = \delta_3 \end{cases}$$

und

$$(3.41)\qquad \omega_0(\delta) := \begin{cases} \operatorname{arccot}\frac{\delta_2}{\delta_3} & , \quad \delta_3 \neq 0 \\ 0 & , \quad \delta_3 = 0 \end{cases} .$$

Dann gilt

$$(3.42)\qquad \max_{\phi,\omega\in\mathbb{R}} h(\delta,\phi,\omega) = h\big(\delta,\phi_+(\delta),\omega_0(\delta)\big) = \frac{1}{2}\left(\delta_1 + \sqrt{\delta_1^2+\delta_2^2+\delta_3^2}\right)$$

und

$$(3.43)\qquad \min_{\phi,\omega\in\mathbb{R}} h\big(\delta,\phi,\omega) = h(\delta,\phi_-(\delta),\omega_0(\delta)\big) = \frac{1}{2}\left(\delta_1 - \sqrt{\delta_1^2+\delta_2^2+\delta_3^2}\right) .$$

Beweis:
Zunächst einmal ist h periodisch in ϕ und ω, also

$$\{h(\delta,\phi,\omega) : \phi,\omega \in \mathbb{R}\} = \{h(\delta,\phi,\omega) : \phi,\omega \in [0,2\pi[\}.$$

Weiter ist für $\phi \in [0,\pi[$ und für $\omega \in [0,2\pi[$

$$\begin{aligned}
h(\delta,\pi+\phi,\omega) &= \delta_1 \cos^2(\pi+\phi) + \sin(\pi+\phi)\cos(\pi+\phi)\{\delta_2\cos\omega + \delta_3\sin\omega\}\\
&= \delta_1(-\cos\phi)^2 + (-\sin\phi)(-\cos\phi)\{\delta_2\cos\omega + \delta_3\sin\omega\}\\
&= \delta_1\cos^2\phi + \sin\phi\cos\phi\{\delta_2\cos\omega + \delta_3\sin\omega\}\\
&= h(\delta,\phi,\omega)
\end{aligned}$$

und für $\phi,\omega \in [0,\pi[$

$$\begin{aligned}
h(\delta,\phi,\pi+\omega) &= \delta_1\cos^2\phi + \sin\phi\cos\phi\{\delta_2\cos(\pi+\omega) + \delta_3\sin(\pi+\omega)\}\\
&= \delta_1\cos^2\phi - \sin\phi\cos\phi\{\delta_2\cos\omega + \delta_3\sin\omega\}\\
&= \delta_1(-\cos\phi)^2 + \sin\phi(-\cos\phi)\{\delta_2\cos\omega + \delta_3\sin\omega\}\\
&= \delta_1\cos^2(\pi-\phi) + \sin(\pi-\phi)\cos(\pi-\phi)\{\delta_2\cos\omega + \delta_3\sin\omega\}\\
&= h(\delta,\pi-\phi,\omega),
\end{aligned}$$

also

$$\{h(\delta,\phi,\omega) : \phi,\omega \in \mathbb{R}\} = \{h(\delta,\phi,\omega) : \phi,\omega \in [0,\pi[\}.$$

Als nächstes berechnen wir die partiellen Ableitungen von h. Es ist

$$\begin{aligned}
\frac{\partial h}{\partial\phi}(\delta,\phi,\omega) &= -2\delta_1\sin\phi\cos\phi + (\cos^2\phi - \sin^2\phi)\{\delta_2\cos\omega + \delta_3\sin\omega\}\\
&= -\delta_1\sin(2\phi) + \cos(2\phi)\{\delta_2\cos\omega + \delta_3\sin\omega\},\\
\frac{\partial h}{\partial\omega}(\delta,\phi,\omega) &= \sin\phi\cos\phi(\delta_3\cos\omega - \delta_2\sin\omega),\\
\frac{\partial^2 h}{\partial\phi^2}(\delta,\phi,\omega) &= -2\delta_1\cos(2\phi) - 2\sin(2\phi)\{\delta_2\cos\omega + \delta_3\sin\omega\},\\
\frac{\partial^2 h}{\partial\omega^2}(\delta,\phi,\omega) &= -\sin\phi\cos\phi(\delta_2\cos\omega + \delta_3\sin\omega)
\end{aligned}$$

und

$$\frac{\partial^2 h}{\partial\phi\partial\omega}(\delta,\phi,\omega) = \cos(2\phi)\{\delta_3\cos\omega - \delta_2\sin\omega\}.$$

Setze

$$d(\delta,\phi,\omega) := \frac{\partial^2 h}{\partial\phi^2}(\delta,\phi,\omega)\frac{\partial^2 h}{\partial\omega^2}(\delta,\phi,\omega) - \left(\frac{\partial^2 h}{\partial\phi\partial\omega}(\delta,\phi,\omega)\right)^2.$$

Dann gilt bekanntlich:

- Ist $\frac{\partial h}{\partial\phi}(\delta,\phi,\omega) = \frac{\partial h}{\partial\omega}(\delta,\phi,\omega) = 0$ und $d(\delta,\phi,\omega) > 0$, so besitzt h in (ϕ,ω) ein relatives Extremum.

- Ist $\frac{\partial h}{\partial\phi}(\delta,\phi,\omega) = \frac{\partial h}{\partial\omega}(\delta,\phi,\omega) = 0$ und $d(\delta,\phi,\omega) < 0$, so besitzt h in (ϕ,ω) kein relatives Extremum.

Im folgenden werden wir in Abhängigkeit von δ einige Fälle unterscheiden:

Fall 1: $\delta_2 = \delta_3 = 0$

Fall 1a: $\delta_1 = 0$
In diesem Fall ist $h(\delta, \phi, \omega) = 0$ für alle $\phi, \omega \in \mathbb{R}$. Also gelten (3.42) und (3.43) trivialerweise.

Fall 1b: $\delta_1 \neq 0$
In diesem Fall ist $h(\delta, \phi, \omega) = \delta_1 \cos^2 \phi$ unabhängig von ω. Weiter ist

$$\frac{\partial h}{\partial \phi}(\delta, \phi, \omega) = -\delta_1 \sin(2\phi) = 0$$

für $\phi_1 = 0$ und $\phi_2 = \frac{\pi}{2}$. Wegen $h(\delta, \phi_1, \omega) = \delta_1$ und $h(\delta, \phi_2, \omega) = 0$ ist für $\delta_1 > 0$

$$\max h(\delta, \phi, \omega) = \delta_1 = \frac{1}{2}(\delta_1 + |\delta_1|) = \frac{1}{2}\left(\delta_1 + \sqrt{\delta_1^2 + \delta_2^2 + \delta_3^2}\right) = h(\delta, \phi_1, 0)$$

und

$$\min h(\delta, \phi, \omega) = 0 = \frac{1}{2}(\delta_1 - |\delta_1|) = \frac{1}{2}\left(\delta_1 - \sqrt{\delta_1^2 + \delta_2^2 + \delta_3^2}\right) = h(\delta, \phi_2, 0),$$

woraus (3.42) und (3.43) folgen. Ebenso gelten (3.42) und (3.43) für $\delta_1 < 0$.

Fall 2: $(\delta_2, \delta_3) \neq (0, 0)$
In diesem Fall ist eine notwendige Bedingung für ein relatives Extremum, daß

$$0 = \frac{\partial h}{\partial \omega}(\delta, \phi, \omega) = \sin \phi \cos \phi (\delta_3 \cos \omega - \delta_2 \sin \omega)$$

ist, also entweder $\phi \in \{0, \frac{\pi}{2}\}$ oder $\delta_2 \sin \omega = \delta_3 \cos \omega$.

Fall 2a: $\delta_3 \neq 0$
Gäbe es ein $\omega \in [0, \pi[$, so daß $(0, \omega)$ oder $(\frac{\pi}{2}, \omega)$ Extremalstelle von h wäre, so müßte

$$0 = \frac{\partial h}{\partial \phi}(\delta, 0, \omega) = \delta_2 \cos \omega + \delta_3 \sin \omega$$

bzw.

$$0 = \frac{\partial h}{\partial \phi}\left(\delta, \frac{\pi}{2}, \omega\right) = -(\delta_2 \cos \omega + \delta_3 \sin \omega)$$

sein, also $\tan \omega = -\frac{\delta_2}{\delta_3}$ und

$$\omega = \arctan\left(-\frac{\delta_2}{\delta_3}\right) = -\arctan\frac{\delta_2}{\delta_3}.$$

Wegen

$$d(\delta, 0, \omega) = (-2\delta_1) \cdot 0 - (\delta_3 \cos \omega - \delta_2 \sin \omega)^2$$

und

$$d\left(\delta, \frac{\pi}{2}, \omega\right) = 2\delta_1 \cdot 0 - (\delta_2 \sin \omega - \delta_3 \cos \omega)^2$$

ist jedoch

$$
\begin{aligned}
d\left(\delta, 0, -\arctan\frac{\delta_2}{\delta_3}\right) &= d\left(\delta, \frac{\pi}{2}, -\arctan\frac{\delta_2}{\delta_3}\right) \\
&= -\left\{\delta_3 \cos\left(-\arctan\frac{\delta_2}{\delta_3}\right) - \delta_2 \sin\left(-\arctan\frac{\delta_2}{\delta_3}\right)\right\}^2 \\
&= -\left(\delta_3 \frac{1}{\sqrt{1+\frac{\delta_2^2}{\delta_3^2}}} + \delta_2 \frac{\frac{\delta_2}{\delta_3}}{\sqrt{1+\frac{\delta_2^2}{\delta_3^2}}}\right)^2 \\
&= -\left(\frac{\delta_2^2+\delta_3^2}{\operatorname{sign}(\delta_3)\sqrt{\delta_2^2+\delta_3^2}}\right)^2 \\
&< 0,
\end{aligned}
$$

also sind $(0, -\arctan\frac{\delta_2}{\delta_3})$ und $(\frac{\pi}{2}, -\arctan\frac{\delta_2}{\delta_3})$ keine Extremalstellen von h. Daher ist für das Vorliegen eines Extremums nötig, daß $\delta_2 \sin\omega = \delta_3 \cos\omega$ ist, also $\cot\omega = \frac{\delta_2}{\delta_3}$. Setze $\omega_0 := \operatorname{arccot}\frac{\delta_2}{\delta_3}$. Wegen

$$
\begin{aligned}
\delta_2 \cos\omega_0 + \delta_3 \sin\omega_0 &= \delta_2 \cos\left(\operatorname{arccot}\frac{\delta_2}{\delta_3}\right) + \delta_3 \sin\left(\operatorname{arccot}\frac{\delta_2}{\delta_3}\right) \\
&= \delta_2 \frac{\frac{\delta_2}{\delta_3}}{\sqrt{1+\frac{\delta_2^2}{\delta_3^2}}} + \delta_3 \frac{1}{\sqrt{1+\frac{\delta_2^2}{\delta_3^2}}} \\
&= \frac{\delta_2^2+\delta_3^2}{\operatorname{sign}(\delta_3)\sqrt{\delta_2^2+\delta_3^2}} \\
&= \operatorname{sign}(\delta_3)\sqrt{\delta_2^2+\delta_3^2}
\end{aligned}
$$

ist dann

$$
\begin{aligned}
0 &= \frac{\partial h}{\partial \phi}(\delta, \phi, \omega_0) \\
&= -\delta_1 \sin(2\phi) + \cos(2\phi)\{\delta_2 \cos\omega_0 + \delta_3 \sin\omega_0\} \\
&= -\delta_1 \sin(2\phi) + \cos(2\phi) \operatorname{sign}(\delta_3)\sqrt{\delta_2^2+\delta_3^2}
\end{aligned}
$$

genau dann, wenn

$$
\cot(2\phi) = \frac{\operatorname{sign}(\delta_3)\delta_1}{\sqrt{\delta_2^2+\delta_3^2}}
$$

ist, also für

$$
\phi_1 = \frac{1}{2}\operatorname{arccot}\frac{\operatorname{sign}(\delta_3)\delta_1}{\sqrt{\delta_2^2+\delta_3^2}} \in \left[0, \frac{\pi}{2}\right[
$$

und

$$
\phi_2 = \frac{1}{2}\operatorname{arccot}\frac{\operatorname{sign}(\delta_3)\delta_1}{\sqrt{\delta_2^2+\delta_3^2}} + \frac{\pi}{2} \in \left[\frac{\pi}{2}, \pi\right[.
$$

Aus

$$
\begin{aligned}
h(\delta,\phi_1,\omega_0) &= \delta_1\cos^2\phi_1 + \sin\phi_1\cos\phi_1(\delta_2\cos\omega_0+\delta_3\sin\omega_0)\\
&= \delta_1\frac{1+\cos(2\phi_1)}{2} + \frac{\sin(2\phi_1)}{2}\operatorname{sign}(\delta_3)\sqrt{\delta_2^2+\delta_3^2}\\
&= \frac{\delta_1}{2} + \frac{\delta_1}{2}\frac{\frac{\operatorname{sign}(\delta_3)\delta_1}{\sqrt{\delta_2^2+\delta_3^2}}}{\sqrt{1+\frac{\delta_1^2}{\delta_2^2+\delta_3^2}}} + \frac{1}{2}\frac{\operatorname{sign}(\delta_3)\sqrt{\delta_2^2+\delta_3^2}}{\sqrt{1+\frac{\delta_1^2}{\delta_2^2+\delta_3^2}}}\\
&= \frac{1}{2}\left(\delta_1 + \operatorname{sign}(\delta_3)\frac{\frac{\delta_1^2}{\sqrt{\delta_2^2+\delta_3^2}} + \sqrt{\delta_2^2+\delta_3^2}}{\sqrt{1+\frac{\delta_1^2}{\delta_2^2+\delta_3^2}}}\right)\\
&= \frac{1}{2}\left(\delta_1 + \operatorname{sign}(\delta_3)\frac{\delta_1^2+\delta_2^2+\delta_3^2}{\sqrt{\delta_1^2+\delta_2^2+\delta_3^2}}\right)\\
&= \frac{1}{2}\left(\delta_1 + \operatorname{sign}(\delta_3)\sqrt{\delta_1^2+\delta_2^2+\delta_3^2}\right)
\end{aligned}
$$

und

$$
\begin{aligned}
h(\delta,\phi_2,\omega_0) &= \delta_1\frac{1+\cos(2\phi_2)}{2} + \frac{\sin(2\phi_2)}{2}\operatorname{sign}(\delta_3)\sqrt{\delta_2^2+\delta_3^2}\\
&= \frac{\delta_1}{2} + \delta_1\frac{\cos(2\phi_1+\pi)}{2} + \frac{\sin(2\phi_1+\pi)}{2}\operatorname{sign}(\delta_3)\sqrt{\delta_2^2+\delta_3^2}\\
&= \frac{\delta_1}{2} - \delta_1\frac{\cos(2\phi_1)}{2} - \frac{\sin(2\phi_1)}{2}\operatorname{sign}(\delta_3)\sqrt{\delta_2^2+\delta_3^2}\\
&= \frac{1}{2}\left(\delta_1 - \operatorname{sign}(\delta_3)\sqrt{\delta_1^2+\delta_2^2+\delta_3^2}\right)
\end{aligned}
$$

folgen (3.42) und (3.43).

<u>Fall 2b: $\delta_3 = 0$</u>
Gäbe es in diesem Fall ein $\omega \in [0,\pi[$, so daß $(0,\omega)$ oder $(\frac{\pi}{2},\omega)$ Extremalstelle von h wäre, so müßte

$$0 = \frac{\partial h}{\partial\phi}(\delta,0,\omega) = \delta_2\cos\omega$$

bzw.

$$0 = \frac{\partial h}{\partial\phi}\left(\delta,\frac{\pi}{2},\omega\right) = -\delta_2\cos\omega$$

sein, also $\omega = \frac{\pi}{2}$. Es ist jedoch

$$\left\{\frac{\partial^2 h}{\partial\phi\partial\omega}\left(\delta,\phi,\frac{\pi}{2}\right)\right\}^2 = \{-\delta_2\cos(2\phi)\}^2 = \delta_2^2\cos^2(2\phi)$$

und

$$d\left(\delta,0,\frac{\pi}{2}\right) = d\left(\delta,\frac{\pi}{2},\frac{\pi}{2}\right) = -\left\{\frac{\partial^2 h}{\partial\phi\partial\omega}\left(\delta,0,\frac{\pi}{2}\right)\right\}^2 = -\delta_2^2 < 0,$$

also sind $(0,\frac{\pi}{2})$ und $(\frac{\pi}{2},\frac{\pi}{2})$ keine Extremalstellen von h. Daher muß für Extremalstellen $\delta_2\sin\omega = 0$ sein, was für $\omega_0 := 0$ erfüllt ist. Dann ist

$$0 = \frac{\partial h}{\partial\phi}(\delta,\phi,\omega_0) = -\delta_1\sin(2\phi) + \delta_2\cos(2\phi)$$

genau dann, wenn $\cot(2\phi) = \frac{\delta_1}{\delta_2}$ ist, also für

$$\phi_1 = \frac{1}{2}\operatorname{arccot}\frac{\delta_1}{\delta_2} \in \left[0, \frac{\pi}{2}\right[$$

und

$$\phi_2 = \frac{1}{2}\operatorname{arccot}\frac{\delta_1}{\delta_2} + \frac{\pi}{2} \in \left[\frac{\pi}{2}, \pi\right[.$$

Wegen

$$\begin{aligned}
h(\delta, \phi_1, \omega_0) &= \delta_1 \cos^2\phi_1 + \delta_2 \sin\phi_1 \cos\phi_1 \cos\omega_0 \\
&= \delta_1 \frac{1+\cos(2\phi_1)}{2} + \delta_2 \frac{\sin(2\phi_1)}{2} \\
&= \frac{\delta_1}{2} + \frac{\delta_1}{2} \frac{\frac{\delta_1}{\delta_2}}{\sqrt{1+\frac{\delta_1^2}{\delta_2^2}}} + \frac{\delta_2}{2} \frac{1}{\sqrt{1+\frac{\delta_1^2}{\delta_2^2}}} \\
&= \frac{1}{2}\left(\delta_1 + \operatorname{sign}(\delta_2)\sqrt{\delta_1^2+\delta_2^2+\delta_3^2}\right)
\end{aligned}$$

und

$$\begin{aligned}
h(\delta, \phi_2, \omega_0) &= \frac{\delta_1}{2} + \frac{\delta_1}{2}\cos(2\phi_2) + \frac{\delta_2}{2}\sin(2\phi_2) \\
&= \frac{\delta_1}{2} - \frac{\delta_1}{2}\cos(2\phi_1) - \frac{\delta_2}{2}\sin(2\phi_1) \\
&= \frac{1}{2}\left(\delta_1 - \operatorname{sign}(\delta_2)\sqrt{\delta_1^2+\delta_2^2+\delta_3^2}\right)
\end{aligned}$$

gelten (3.42) und (3.43). □

Wir haben bisher folgendes festgestellt: Die Kurven mit den Richtungen $R_{kl}(1)$ und $R_{kl}(i)$ liegen auf den in (3.31) definierten Flächen. Weiter nimmt g_B für diagonale Matrizen B auf diesen Flächen die Funktion h an und ist auf den Kurven mit den Richtungen $S_m(i)$ konstant. Für diagonale B besteht ein Optimierungs-Sweep für g_B also darin, daß g_B nacheinander über jede der in (3.31) definierten Flächen genau einmal optimiert wird. Dazu haben wir die Extremalstellen von g_B auf diesen Flächen, also die Extremalstellen der Funktion h im letzten Satz bestimmt. Ein Maximierungs-Sweep für g_B mit diagonaler Matrix B sieht also folgendermaßen aus:

Algorithmus 3.6 (Maximierungs-Sweep für g_B)
Eingabe: $X, B \in \mathbb{C}^{n,n}$ mit B diagonal

- Setze $X^{(11)} := X$.
- Für $l = 2, \ldots, n$
  - setze $X^{(1l)} := X^{(l-1,l-1)}$,
  - für $k = 1, \ldots, l-1$
    - berechne $\delta_{klr} := \kappa_{klr}\left(X^{(kl)}, B\right), r \in \{1,2,3\}$, nach (3.36),
    - berechne $\phi_+^{(kl)} := \phi_+(\delta_{kl1}, \delta_{kl2}, \delta_{kl3})$ nach (3.39),
    - berechne $\omega_0^{(kl)} := \omega_0(\delta_{kl1}, \delta_{kl2}, \delta_{kl3})$ nach (3.41),

* berechne $U^{(kl)} := U_{kl}\big(\cos\phi_+^{(kl)}, \sin\phi_+^{(kl)}\, e^{i\omega_0^{(kl)}}\big)$,
* setze $X^{(k+1,l)} := U^{(kl)} X^{(kl)} U^{(kl)*}$.

Ausgabe: $\operatorname{maxsweep}(X,B) := X^{(nn)}$

Wird in dem obigen Verfahren anstelle von $\phi_+(\delta)$ der Wert $\phi_-(\delta)$ berechnet, so erhalten wir einen Minimierungs-Sweep für g_B.

Beim Maximierungs-Sweep Algorithmus 3.6 und dem zugehörigen Minimierungs-Sweep unterscheiden sich zwei aufeinanderfolgende Iterierte in den Elementen zweier Zeilen und Spalten. Die κ_{klr} hängen jedoch nur von den vier Elementen der Iterierten $X^{(kl)}$ ab, die sich in den Positionen (k,k), (k,l), (l,k) und (l,l) befinden. Daher lassen sich die Paare $\{(k,l) : 1 \leq k < l \leq n\}$ so anordnen, daß die κ_{klr} unabhängig von den Elementen sind, die sich im Lauf der vorherigen Iteration geändert haben. Das hat zur Folge, daß Algorithmus 3.6 und der zugehörige Minimierungsalgorithmus im Gegensatz zu dem im vorigen Abschnitt verwendeten Maximierungs-Sweep Algorithmus 3.4 und dem dazu analogen Minimierungs-Sweep parallelisierbar sind.

Wir erhalten für diagonale B das folgende jacobiartige Verfahren zur Berechnung einer Matrix aus $\Lambda(A)$, die die Funktion g_B maximiert, also die Funktion f_B minimiert und damit das isospektrale Minimierungsproblem (2.6) löst.

Algorithmus 3.7 (Jacobiartiges Verfahren zur Maximierung von g_B)
Eingabe: $A, B \in \mathbb{C}^{n,n}$ mit B diagonal, Abbruchschranke $\epsilon > 0$

- Setze $A^{(0)} := A$.
- Berechne $g^{(0)} := g_B\big(A^{(0)}\big)$.
- Für $q = 0, 1, \ldots$
  - berechne $A^{(q+1)} := \operatorname{maxsweep}\big(A^{(q)}, B\big)$ mit Algorithmus 3.6,
  - berechne $g^{(q+1)} := g_B\big(A^{(q+1)}\big)$,
  - falls $g^{(q+1)} - g^{(q)} < \epsilon$, so beende die Schleife.

Ausgabe: $A_{\text{Jacobi}}(A,B) := A^{(q+1)}$

Wird im obigen Verfahren jeder Maximierungs- durch einen Minimierungs-Sweep ersetzt und das Abbruchkriterium so modifiziert, wie im letzten Abschnitt beschrieben, so erhalten wir ein jacobiartiges Verfahren zur Berechnung einer zur Matrix A unitär ähnlichen Matrix, die für diagonale Matrizen B die Funktion g_B auf $\Lambda(A)$ minimiert, also die Funktion f_B maximiert und damit das isospektrale Maximierungsproblem (2.5) löst.

Als nächstes stellt sich die Frage nach der Konvergenz dieser Verfahren. Da bei jedem Sweep der Verfahren g_B nacheinander über die in (3.31) definierten Flächen optimiert wird und diese Flächen die Kurven mit den Richtungen $R_{kl}(1)$ und $R_{kl}(i)$ enthalten, wird insbesondere g_B auch über jede dieser Kurven optimiert. Auf den Kurven mit den Richtungen $S_m(i)$ ist g_B konstant, also in jedem Punkt dieser Kurven optimal. Somit sind das Maximierungsverfahren Algorithmus 3.7 und das zugehörige Minimierungsverfahren Spezialfälle von Algorithmus 3.1 bzw. Algorithmus 3.2, was zur Folge hat, daß für sie Konvergenzsatz 3.1 gilt. Daß darüber hinaus keine besseren Konvergenzresultate erzielt werden können, also die Verfahren nicht global konvergent sind, zeigt das folgende Beispiel.

Beispiel 3.12
Sei $B = \operatorname{diag}(-1, 1+i, 1-i)$,

$$H = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

und

$$A = B + BH - HB = \begin{pmatrix} -1 & -2-i & -2+i \\ 2+i & 1+i & 2i \\ 2-i & -2i & 1-i \end{pmatrix}.$$

Dann liefert der Maximierungs-Sweep Algorithmus 3.6 für A und B

$$\begin{aligned} X^{(11)} &= A \\ X^{(12)} &= X^{(11)} = A \end{aligned}$$

$$\begin{aligned} \delta_{121} = \kappa_{121}(A,B) &= \operatorname{re}\{(a_{11} - a_{22})(\overline{b_{11}} - \overline{b_{22}})\} = \operatorname{re}\{(-2-i)(-2+i)\} = 5 \\ \delta_{122} = \kappa_{122}(A,B) &= \operatorname{re}\{(a_{12} + a_{21})(\overline{b_{11}} - \overline{b_{22}})\} = \operatorname{re}\{0(-2+i)\} = 0 \\ \delta_{123} = \kappa_{123}(A,B) &= \operatorname{im}\{(a_{12} - a_{21})(\overline{b_{11}} - \overline{b_{22}})\} = \operatorname{im}\{2(-2-i)(-2+i)\} = 0 \end{aligned}$$

$$\begin{aligned} \phi_+^{(12)} &= \phi_+(\delta_{121}, \delta_{122}, \delta_{123}) = \phi_+(5,0,0) = 0 \\ \omega_0^{(12)} &= \omega_0(\delta_{121}, \delta_{122}, \delta_{123}) = \omega_0(5,0,0) = 0 \end{aligned}$$

$$\begin{aligned} U^{(12)} &= U_{12}\big(\cos\phi_+^{(12)}, \sin\phi_+^{(12)}\, e^{i\omega_0^{(12)}}\big) = U_{12}(1,0) = I_3 \\ X^{(22)} &= U^{(12)} X^{(12)} U^{(12)*} = I_3 A I_3 = A \\ X^{(13)} &= X^{(22)} = A \end{aligned}$$

$$\begin{aligned} \delta_{131} = \kappa_{131}(A,B) &= \operatorname{re}\{(a_{11} - a_{33})(\overline{b_{11}} - \overline{b_{33}})\} = \operatorname{re}\{(-2+i)(-2-i)\} = 5 \\ \delta_{132} = \kappa_{132}(A,B) &= \operatorname{re}\{(a_{13} + a_{31})(\overline{b_{11}} - \overline{b_{33}})\} = 0 \\ \delta_{133} = \kappa_{133}(A,B) &= \operatorname{im}\{(a_{13} - a_{31})(\overline{b_{11}} - \overline{b_{33}})\} = \operatorname{im}\{2(-2+i)(-2-i)\} = 0 \end{aligned}$$

$$\begin{aligned} \phi_+^{(13)} &= \phi_+(\delta_{131}, \delta_{132}, \delta_{133}) = \phi_+(5,0,0) = 0 \\ \omega_0^{(13)} &= \omega_0(\delta_{131}, \delta_{132}, \delta_{133}) = \omega_0(5,0,0) = 0 \end{aligned}$$

$$\begin{aligned} U^{(13)} &= U_{13}\big(\cos\phi_+^{(13)}, \sin\phi_+^{(13)}\, e^{i\omega_0^{(13)}}\big) = U_{13}(1,0) = I_3 \\ X^{(23)} &= U^{(13)} X^{(13)} U^{(13)*} = I_3 A I_3 = A \end{aligned}$$

$$\begin{aligned} \delta_{231} = \kappa_{231}(A,B) &= \operatorname{re}\{(a_{22} - a_{33})(\overline{b_{22}} - \overline{b_{33}})\} = \operatorname{re}\{(2i)(-2i)\} = 4 \\ \delta_{232} = \kappa_{232}(A,B) &= \operatorname{re}\{(a_{23} + a_{32})(\overline{b_{22}} - \overline{b_{33}})\} = 0 \\ \delta_{233} = \kappa_{233}(A,B) &= \operatorname{im}\{(a_{23} - a_{32})(\overline{b_{22}} - \overline{b_{33}})\} = \operatorname{im}\{2(2i)(-2i)\} = 0 \end{aligned}$$

$$\begin{aligned} \phi_+^{(23)} &= \phi_+(\delta_{231}, \delta_{232}, \delta_{233}) = \phi_+(4,0,0) = 0 \\ \omega_0^{(23)} &= \omega_0(\delta_{231}, \delta_{232}, \delta_{233}) = \omega_0(4,0,0) = 0 \end{aligned}$$

$$U^{(23)} = U_{23}\big(\cos\phi_+^{(23)}, \sin\phi_+^{(23)}\, e^{i\omega_0^{(23)}}\big) = U_{23}(1,0) = I_3$$
$$X^{(33)} = U^{(23)} X^{(23)} U^{(23)*} = I_3 A I_3 = A$$

$$\operatorname{maxsweep}(A,B) = X^{(33)} = A.$$

Algorithmus 3.6 stagniert also, was bedeutet, daß g_B auf den Flächen $\mathcal{L}_{A,R_{12}(1)}$, $\mathcal{L}_{A,R_{13}(1)}$ und $\mathcal{L}_{A,R_{23}(1)}$ maximal ist. Da die Kurven $\mathcal{K}_{A,R_{kl}(1)}$ und $\mathcal{K}_{A,R_{kl}(i)}$ mit $1 \le k < l \le 3$ auf diesen Flächen liegen, stagniert auch der Maximierungs-Sweep Algorithmus 3.4 aus dem vorigen Abschnitt. Dies hat zur Folge, daß sowohl Algorithmus 3.7 als auch Algorithmus 3.5 $A_{\text{Jacobi}}(A,B) = A$ ausgeben. Setze

$$Y = \begin{pmatrix} 0 & -1+4i & -1+4i \\ 1+4i & 0 & -1+4i \\ 1+4i & 1+4i & 0 \end{pmatrix} \in \mathcal{S}(3).$$

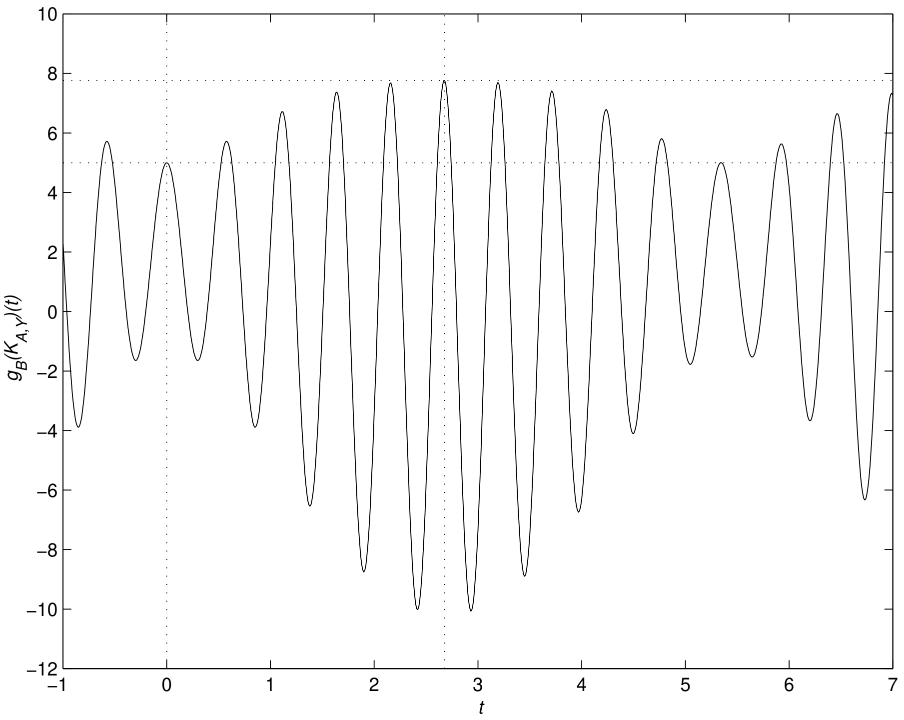

Abbildung 3.1: Verlauf von g_B auf der Kurve $\mathcal{K}_{A,Y}$ im Intervall $[-1,7]$

Dann ist $g_B\big(\mathcal{K}_{A,Y}(2.68)\big) = 7.7642 > 5 = g_B(A)$, wie Abbildung 3.1 zeigt, also ist $g_B(A) \neq G^+(A,B)$. Die beiden Algorithmen 3.5 und 3.7 konvergieren also nicht gegen ein globales Maximum von g_B auf $\Lambda(A)$. Werden die zugehörigen Minimierungsalgorithmen mit $-A$ gestartet, stagnieren auch sie, obwohl $g_B(-A) \neq G^-(-A,B)$ ist. Damit sind auch die beiden Minimierungsverfahren nicht global konvergent.

In Abschnitt 5.1 wird die Konvergenz von Algorithmus 3.7 in einigen Beispielen genauer untersucht. Zunächst werden wir im nächsten Abschnitt zeigen, wie mit den in diesem Abschnitt entwickelten Verfahren das isospektrale Abstandsproblem für normale Matrizen B gelöst werden kann.

3.1.3 Verfahren für normale B

Sei $A \in \mathbb{C}^{n,n}$ und $B \in \mathcal{N}(n)$. Dann existieren $W \in \mathcal{U}(n)$ und $D \in \mathbb{C}^{n,n}$ diagonal mit $B = WDW^*$. Nach (2.19) ist $G^+(A,B) = G^+(A,D)$ und nach (2.20) ist $G^-(A,B) = G^-(A,D)$. Seien $X_+, X_- \in \Lambda(A)$ mit $g_D(X_+) = G^+(A,D)$ und $g_D(X_-) = G^-(A,D)$. Dann ist für alle $X \in \Lambda(A)$

$$\begin{aligned} g_B(WXW^*) &= \mathrm{re}\{\mathrm{tr}(WXW^*B^*)\} = \mathrm{re}\{\mathrm{tr}(XW^*B^*W)\} = \mathrm{re}\{\mathrm{tr}(XD^*)\} \\ &= g_D(X), \end{aligned}$$

also

$$g_B(WX_+W^*) = g_D(X_+) = G^+(A,D) = G^+(A,B)$$

und

$$g_B(WX_-W^*) = g_D(X_-) = G^-(A,D) = G^-(A,B).$$

Damit liegt es nahe, zur Optimierung von g_B auf $\Lambda(A)$ zunächst einmal B zu diagonalisieren, dann den Algorithmen 3.7 bzw. den zugehörigen Minimierungsalgorithmus aus dem letzten Abschnitt anzuwenden und anschließend die Lösung wieder zurückzutransformieren. Ein Verfahren zur Maximierung von g_B lautet dann beispielsweise folgendermaßen:

- Bestimme $W \in \mathcal{U}(n)$ und $D \in \mathbb{C}^{n,n}$ diagonal mit $B = WDW^*$.
- Berechne $A_{\mathrm{Jacobi}}(A,D)$ mit Algorithmus 3.7.
- Setze $A_{\mathrm{Jacobi}}(A,B) := WA_{\mathrm{Jacobi}}(A,D)W^*$.

Bezüglich der Konvergenz gilt dasselbe, was im vorigen Abschnitt gesagt wurde.

3.2 Verfahren zur Berechnung einer nächstgelegenen normalen Matrix

In diesem Abschnitt wollen wir zu einer gegebenen Matrix $A \in \mathbb{C}^{n,n}$ eine in der Frobenius-Norm nächstgelegene normale Matrix finden. In Satz 2.15 hatten wir folgendes festgestellt: Ist VBV^* mit $V \in \mathcal{U}(n)$ und $B \in \mathbb{C}^{n,n}$ diagonal eine zu A nächstgelegene normale Matrix, so gibt es ein $H \in \mathcal{H}(n)$ mit

$$V^*AV = B + BH - HB.$$

Somit ist

$$B = \Delta(B + BH - HB) = \Delta(V^*AV) \tag{3.44}$$

und der Abstand von A zu einer nächstgelegenen normalen Matrix, also die Abweichung der Matrix A von der Normalität

$$\begin{aligned} \|A - VBV^*\|_F &= \|V^*AV - B\|_F = \|V^*AV - \Delta(V^*AV)\|_F \\ &= \sqrt{\mathrm{off}(V^*AV)}. \end{aligned} \tag{3.45}$$

Dies legt es nahe, ausgehend von A die Matrix V^*AV iterativ durch eine Folge von unitären Ähnlichkeitstransformationen zu bestimmen, wobei die Transformationen so gewählt werden, daß jede Iterierte eine kleinere Off-Norm besitzt als die vorherige. Ist V^*AV bekannt, so ist

$$V\Delta(V^*AV)V^* = VBV^* \tag{3.46}$$

eine zu A nächstgelegene normale Matrix.

Eine Möglichkeit die Off-Norm von A iterativ durch unitäre Ähnlichkeitstransformationen zu verringern ist es, in jedem Schritt den Abstand der Iterierten $X^{(k)}$ zu ihrer Diagonale zu vermindern, also die Funktion $g_{\Delta(X^{(k)})}$ zu vergrößern. Dann ist nämlich nach (2.10)

$$\begin{aligned}
\operatorname{off}\big(X^{(k)}\big) &= \big\|X^{(k)} - \Delta\big(X^{(k)}\big)\big\|_F^2 \\
&= \big\|X^{(k)}\big\|_F^2 + \big\|\Delta\big(X^{(k)}\big)\big\|_F^2 - 2g_{\Delta(X^{(k)})}\big(X^{(k)}\big) \\
&\geq \big\|X^{(k+1)}\big\|_F^2 + \big\|\Delta\big(X^{(k)}\big)\big\|_F^2 - 2g_{\Delta(X^{(k)})}\big(X^{(k+1)}\big) \\
&= \big\|X^{(k+1)} - \Delta\big(X^{(k)}\big)\big\|_F^2 \\
&= \big\|X^{(k+1)} - \Delta\big(X^{(k+1)}\big)\big\|_F^2 + \big\|\Delta\big(X^{(k+1)}\big) - \Delta\big(X^{(k)}\big)\big\|_F^2 \\
&\geq \big\|X^{(k+1)} - \Delta\big(X^{(k+1)}\big)\big\|_F^2 \\
&= \operatorname{off}\big(X^{(k+1)}\big).
\end{aligned}$$

Zur Durchführung modifizieren wir das in Abschnitt 3.1.2 entwickelte Verfahren zur Maximierung der Funktion g_B für diagonale Matrizen B auf $\Lambda(A)$ derart, daß wir in jedem Schritt anstelle der Funktion g_B mit fester Diagonalmatrix B die von der Iterierten $X^{(k)}$ abhängige Funktion $g_{\Delta(X^{(k)})}$ vergrößern. Aus Algorithmus 3.6, dem Maximierungs-Sweep für g_B, entsteht durch diese Modifikation der folgende Minimierungs-Sweep für die Off-Norm:

Algorithmus 3.8 (Minimierungs-Sweep für die Off-Norm)
Eingabe: $X \in \mathbb{C}^{n,n}$

- Setze $X^{(11)} := X$.
- Für $l = 2, \dots, n$
  - setze $X^{(1l)} := X^{(l-1,l-1)}$,
  - für $k = 1, \dots, l-1$
    - berechne $\delta_{klr} := \kappa_{klr}\big(X^{(kl)}, \Delta(X^{(kl)})\big), r \in \{1,2,3\}$, nach (3.36),
    - berechne $\phi^{(kl)} := \phi_+(\delta_{kl1}, \delta_{kl2}, \delta_{kl3})$ nach (3.39),
    - berechne $\omega^{(kl)} := \omega_0(\delta_{kl1}, \delta_{kl2}, \delta_{kl3})$ nach (3.41),
    - berechne $U^{(kl)} := U_{kl}\big(\cos\phi^{(kl)}, \sin\phi^{(kl)}\, e^{i\omega^{(kl)}}\big)$,
    - setze $X^{(k+1,l)} := U^{(kl)} X^{(kl)} U^{(kl)*}$.

Ausgabe: $\operatorname{sweep}(X) := X^{(nn)}$

Insgesamt erhalten wir so das folgende Verfahren zur Berechnung einer Matrix V^*AV, deren Off-Norm das Quadrat des Abstandes von A zu einer nächstgelegenen normalen Matrix ist:

Algorithmus 3.9 (1. Jacobiartiges Verfahren zur Min. der Off-Norm)
Eingabe: $A \in \mathbb{C}^{n,n}$, Abbruchschranke $\epsilon > 0$

- Setze $A^{(0)} := A$.
- Berechne $f^{(0)} := \big\|A^{(0)} - \Delta\big(A^{(0)}\big)\big\|_F$.
- Für $q = 0, 1, \dots$
  - berechne $A^{(q+1)} := \operatorname{sweep}\big(A^{(q)}\big)$ mit Algorithmus 3.8,
  - berechne $f^{(q+1)} := \big\|A^{(q+1)} - \Delta\big(A^{(q+1)}\big)\big\|_F$,
  - falls $f^{(q)} - f^{(q+1)} < \epsilon$, so beende die Schleife.

Ausgabe: $A_{\text{next}}(A) := A^{(q+1)}$

Das Verfahren kann enden ohne eine Matrix V^*AV mit der Eigenschaft erreicht zu haben, daß $V\Delta(V^*AV)V^*$ eine zu A nächstgelegene normale Matrix ist, wie das folgende Beispiel zeigt:

Beispiel 3.13
Seien A, B und Y wie in Beispiel 3.12. Dort haben wir gezeigt, daß Algorithmus 3.6 stagniert, jedoch A nicht das Maximum von g_B ist, da $g_B\big(\mathcal{K}_{A,Y}(2.68)\big) > g_B(A)$ ist. Weil $\Delta(A) = B$ ist, stagniert in diesem Fall auch Algorithmus 3.8. Es ist jedoch B keine nächstgelegene normale Matrix zu A, da

$$\begin{aligned}
\|B-A\|_F^2 &= f_B(A)^2 \\
&= \|A\|_F^2 + \|B\|_F^2 - 2g_B(A) \\
&> \big\|\mathcal{K}_{A,Y}(2.68)\big\|_F^2 + \|B\|_F^2 - 2g_B\big(\mathcal{K}_{A,Y}(2.68)\big) \\
&= f_B\big(\mathcal{K}_{A,Y}(2.68)\big)^2 \\
&= \big\|B - e^{2.68Y}Ae^{-2.68Y}\big\|_F^2 \\
&= \big\|e^{-2.68Y}Be^{2.68Y} - A\big\|_F^2
\end{aligned}$$

ist, also die normale Matrix $e^{-2.68Y}Be^{2.68Y}$ näher an A liegt als B.

Eine weitere Aussage von Satz 2.15 ist, daß für eine zu A nächstgelegene normale Matrix VBV^* mit $V \in \mathcal{U}(n)$ und $B \in \mathbb{C}^{n,n}$ diagonal die Matrix V^*AV eine ΔH-Matrix ist. Besitzt A jedoch mehrere gleiche Diagonaleinträge, so kann Algorithmus 3.9 abbrechen, ohne eine ΔH-Matrix erreicht zu haben:

Beispiel 3.14
Sei

$$A = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}.$$

Dann ist für alle $U \in \mathcal{U}(n)$

$$g_{\Delta(A)}(UAU^*) = \mathrm{re}\{\mathrm{tr}(UAU^*2I_2)\} = 2\,\mathrm{re}\{\mathrm{tr}(A)\} = 8$$

unabhängig von U, also $g_{\Delta(A)}$ auf $\Lambda(A)$ konstant. Somit liefert Algorithmus 3.9 $A_{\mathrm{next}}(A) = A$. Es ist jedoch weder A eine ΔH-Matrix, noch $\Delta(A)$ eine zu A nächstgelegene normale Matrix, da

$$\|A - \Delta(A)\|_F = 1 > \sqrt{\frac{1}{2}} = \left\| A - \begin{pmatrix} 2 & \frac{1}{2} \\ \frac{1}{2} & 2 \end{pmatrix} \right\|_F$$

ist, also die normale Matrix $\begin{pmatrix} 2 & \frac{1}{2} \\ \frac{1}{2} & 2 \end{pmatrix}$ näher an A liegt als $\Delta(A)$.

Eine andere Strategie zur Bestimmung von V^*AV, bei der dies vermieden wird, ist es, ausgehend von A eine Folge von unitären Ähnlichkeitstransformationen so zu wählen, daß sukzessiv alle 2×2 - Teilmatrizen in ΔH-Matrizen mit kleinerer Off-Norm transformiert werden. Dies ist die Idee des Verfahrens [21, Algorithm J]. Da die Herleitung dieses Verfahrens in [21] unvollständig ist, einer der im Verfahren angegebenen Winkel falsch ist und eine Abschätzung bezüglich der Verkleinerung der Off-Norm in jedem Iterationsschritt ebenfalls nicht stimmt, werden wir jetzt eine Methode zur Transformation von 2×2 - Matrizen in ΔH-Matrizen ausführlich darstellen.

Wie kann eine beliebige 2×2 - Matrix M durch eine unitäre Ähnlichkeitstransformation in eine ΔH-Matrix mit kleinerer Off-Norm verwandelt werden? Ist M ein Vielfaches der Einheitsmatrix, so ist M insbesondere eine ΔH-Matrix und $\operatorname{off}(M) = 0$ ist minimal, das heißt es ist nichts mehr zu tun. Sei im folgenden M kein Vielfaches der Einheitsmatrix. Dann gehen wir folgendermaßen vor: Zunächst bestimmen wir $m_0 \in \mathbb{C}$ und $\psi_0 \in]-\pi, \pi]$ so, daß

$$M(m,\psi) := e^{i\psi}(M - mI_2) \tag{3.47}$$

für (m_0, ψ_0) Spur Null und reelle, nichtpositive Determinante hat. Anschließend wählen wir $W \in \mathcal{U}(2)$ so, daß

$$WM(m_0,\psi_0)W^* = \begin{pmatrix} q_1 & q_2 \\ -\overline{q_2} & -q_1 \end{pmatrix} =: Q \tag{3.48}$$

ist mit $q_1 \in \mathbb{R} \setminus \{0\}$. Dann ist

$$\begin{aligned}
WMW^* &= W\big(e^{-i\psi_0}M(m_0,\psi_0) + m_0I_2\big)W^* \\
&= e^{-i\psi_0}WM(m_0,\psi_0)W^* + m_0I_2 \\
&= e^{-i\psi_0}Q + m_0I_2 \\
&= \begin{pmatrix} e^{-i\psi_0}q_1 + m_0 & e^{-i\psi_0}q_2 \\ -e^{-i\psi_0}\overline{q_2} & -e^{-i\psi_0}q_1 + m_0 \end{pmatrix} \\
&= \begin{pmatrix} e^{-i\psi_0}q_1 + m_0 & 0 \\ 0 & -e^{-i\psi_0}q_1 + m_0 \end{pmatrix} + \begin{pmatrix} 0 & 2e^{-i\psi_0}q_1\frac{q_2}{2q_1} \\ -2e^{-i\psi_0}q_1\frac{\overline{q_2}}{2q_1} & 0 \end{pmatrix} \\
&= \begin{pmatrix} e^{-i\psi_0}q_1 + m_0 & 0 \\ 0 & -e^{-i\psi_0}q_1 + m_0 \end{pmatrix} \\
&\quad + \begin{pmatrix} e^{-i\psi_0}q_1 + m_0 & 0 \\ 0 & -e^{-i\psi_0}q_1 + m_0 \end{pmatrix}\begin{pmatrix} 0 & \frac{q_2}{2q_1} \\ \frac{\overline{q_2}}{2q_1} & 0 \end{pmatrix} \\
&\quad - \begin{pmatrix} 0 & \frac{q_2}{2q_1} \\ \frac{\overline{q_2}}{2q_1} & 0 \end{pmatrix}\begin{pmatrix} e^{-i\psi_0}q_1 + m_0 & 0 \\ 0 & -e^{-i\psi_0}q_1 + m_0 \end{pmatrix}
\end{aligned}$$

eine ΔH-Matrix. Wir zeigen im folgenden, daß solche m_0, ψ_0 und W existieren. Als erstes bestimmen wir m_0 und ψ_0. Es ist

$$\operatorname{tr}\big(M(m,\psi)\big) = e^{i\psi}(m_{11} - m) + e^{i\psi}(m_{22} - m) = e^{i\psi}(m_{11} + m_{22} - 2m) = 0$$

genau dann, wenn $m = \frac{m_{11}+m_{22}}{2} =: m_0$ ist. Sei $\delta := \det(M - m_0I_2)$. Dann ist

$$\begin{aligned}
\det M(m_0,\psi) &= \det\{e^{i\psi}(M - m_0I_2)\} \\
&= (e^{i\psi})^2 \det(M - m_0I_2) \\
&= e^{2i\psi}|\delta|e^{i\arg\delta} \\
&= e^{i(2\psi + \arg\delta)}|\delta| \\
&\le 0
\end{aligned}$$

nur, wenn $2\psi + \arg\delta = \pi$ ist, also für $\psi_0 := -\frac{1}{2}\arg\delta + \frac{\pi}{2}$. Somit sind m_0 und ψ_0 eindeutig bestimmt. Zur Vereinfachung der Notation setzen wir $N := M(m_0,\psi_0)$. Dann ist

$$\begin{aligned}
0 &\ge \det N \\
&= \det\begin{pmatrix} n_{11} & n_{12} \\ n_{21} & -n_{11} \end{pmatrix} \\
&= -n_{11}^2 - n_{12}n_{21} \\
&= -\operatorname{re}(n_{11})^2 + \operatorname{im}(n_{11})^2 - \operatorname{re}(n_{12}n_{21}) - \{2\operatorname{re}(n_{11})\operatorname{im}(n_{11}) + \operatorname{im}(n_{12}n_{21})\}i,
\end{aligned}$$

also

$$0 \geq -\operatorname{re}(n_{11})^2 + \operatorname{im}(n_{11})^2 - \operatorname{re}(n_{12}n_{21}) \tag{3.49}$$

und

$$2\operatorname{re}(n_{11})\operatorname{im}(n_{11}) + \operatorname{im}(n_{12}n_{21}) = 0. \tag{3.50}$$

Für $\operatorname{im}(n_{11}) \neq 0$ folgt daraus

$$\begin{aligned}
0 &\geq 4\operatorname{im}(n_{11})^2\{\operatorname{im}(n_{11})^2 - \operatorname{re}(n_{11})^2 - \operatorname{re}(n_{12}n_{21})\} \\
&= 4\operatorname{im}(n_{11})^2 \left\{\operatorname{im}(n_{11})^2 - \left(\frac{\operatorname{im}(n_{12}n_{21})}{2\operatorname{im}(n_{11})}\right)^2 - \operatorname{re}(n_{12}n_{21})\right\} \\
&= 4\operatorname{im}(n_{11})^4 - 4\operatorname{re}(n_{12}n_{21})\operatorname{im}(n_{11})^2 - \operatorname{im}(n_{12}n_{21})^2 \\
&= \big(2\operatorname{im}(n_{11})^2 - \operatorname{re}(n_{12}n_{21}) - |n_{12}n_{21}|\big)\big(2\operatorname{im}(n_{11})^2 - \operatorname{re}(n_{12}n_{21}) + |n_{12}n_{21}|\big)
\end{aligned}$$

und somit

$$2\operatorname{im}(n_{11})^2 - \operatorname{re}(n_{12}n_{21}) - |n_{12}n_{21}| \leq 0. \tag{3.51}$$

Diese Abschätzung gilt auch für $\operatorname{im}(n_{11}) = 0$, da in diesem Fall aus (3.50) folgt, daß $\operatorname{im}(n_{12}n_{21}) = 0$ ist, also $|n_{12}n_{21}| = |\operatorname{re}(n_{12}n_{21})|$. Wie wir später sehen werden, garantiert sie uns, daß $\operatorname{off}(WMW^*) \leq \operatorname{off}(M)$ ist.

Es bleibt noch ein $W \in \mathcal{U}(2)$ zu bestimmen, das (3.48) erfüllt. Dazu vergewissern wir uns, daß

$$W := \begin{pmatrix} \cos\phi & \sin\phi\, e^{i\omega} \\ -\sin\phi\, e^{-i\omega} & \cos\phi \end{pmatrix}$$

mit $\omega := \arg(n_{12} + \overline{n_{21}})$ und

$$\phi := \begin{cases} \frac{1}{2}\arctan\frac{|n_{12}+\overline{n_{21}}|}{2\operatorname{re}(n_{11})} & , \quad \operatorname{re}(n_{11}) \neq 0 \\ \frac{\pi}{4} & , \quad \operatorname{re}(n_{11}) = 0 \end{cases}$$

dies leistet. Zunächst betrachten wir den Fall, daß $\operatorname{re}(n_{11}) \neq 0$ ist. In diesem Fall ist

$$\begin{aligned}
\cos(2\phi) &= \cos\left(\arctan\frac{|n_{12}+\overline{n_{21}}|}{2\operatorname{re}(n_{11})}\right) \\
&= \frac{1}{\sqrt{1 + \frac{|n_{12}+\overline{n_{21}}|^2}{\left(2\operatorname{re}(n_{11})\right)^2}}} \\
&= \frac{2\operatorname{sign}\big(\operatorname{re}(n_{11})\big)\operatorname{re}(n_{11})}{\sqrt{4\operatorname{re}(n_{11})^2 + |n_{12}+\overline{n_{21}}|^2}}
\end{aligned}$$

und

$$\begin{aligned}
\sin(2\phi) &= \sin\left(\arctan\frac{|n_{12}+\overline{n_{21}}|}{2\operatorname{re}(n_{11})}\right) \\
&= \frac{\frac{|n_{12}+\overline{n_{21}}|}{2\operatorname{re}(n_{11})}}{\sqrt{1 + \frac{|n_{12}+\overline{n_{21}}|^2}{\left(2\operatorname{re}(n_{11})\right)^2}}} \\
&= \frac{\operatorname{sign}\big(\operatorname{re}(n_{11})\big)|n_{12}+\overline{n_{21}}|}{\sqrt{4\operatorname{re}(n_{11})^2 + |n_{12}+\overline{n_{21}}|^2}}.
\end{aligned}$$

Daraus folgt

$$
\begin{aligned}
[WH(N)W^*]_{11} &= \operatorname{re}(n_{11})\cos^2\phi + \frac{n_{21}+\overline{n_{12}}}{2}\sin\phi\cos\phi\, e^{i\omega} \\
&\quad + \frac{n_{12}+\overline{n_{21}}}{2}\sin\phi\cos\phi\, e^{-i\omega} - \operatorname{re}(n_{11})\sin^2\phi \\
&= \operatorname{re}(n_{11})(\cos^2\phi - \sin^2\phi) + \frac{1}{2}\sin\phi\cos\phi\Big\{(n_{21}+\overline{n_{12}})e^{i\omega} \\
&\quad + (n_{12}+\overline{n_{21}})e^{-i\omega}\Big\} \\
&= \operatorname{re}(n_{11})\cos(2\phi) + \frac{1}{4}\sin(2\phi)2|n_{12}+\overline{n_{21}}| \\
&= \operatorname{re}(n_{11})\frac{2\operatorname{sign}\big(\operatorname{re}(n_{11})\big)\operatorname{re}(n_{11})}{\sqrt{4\operatorname{re}(n_{11})^2+|n_{12}+\overline{n_{21}}|^2}} \\
&\quad + \frac{1}{2}|n_{12}+\overline{n_{21}}|\frac{\operatorname{sign}\big(\operatorname{re}(n_{11})\big)|n_{12}+\overline{n_{21}}|}{\sqrt{4\operatorname{re}(n_{11})^2+|n_{12}+\overline{n_{21}}|^2}} \\
&= \frac{1}{2}\operatorname{sign}\big(\operatorname{re}(n_{11})\big)\sqrt{4\operatorname{re}(n_{11})^2+|n_{12}+\overline{n_{21}}|^2} \\
&=: q_1, \\
[WH(N)W^*]_{12} &= -\operatorname{re}(n_{11})\sin\phi\cos\phi\, e^{i\omega} - \frac{n_{21}+\overline{n_{12}}}{2}\sin^2\phi\, e^{2i\omega} \\
&\quad + \frac{n_{12}+\overline{n_{21}}}{2}\cos^2\phi - \operatorname{re}(n_{11})\sin\phi\cos\phi\, e^{i\omega} \\
&= \left\{-2\operatorname{re}(n_{11})\sin\phi\cos\phi + \frac{|n_{12}+\overline{n_{21}}|}{2}(\cos^2\phi-\sin^2\phi)\right\}e^{i\omega} \\
&= \left\{-\operatorname{re}(n_{11})\sin(2\phi) + \frac{|n_{12}+\overline{n_{21}}|}{2}\cos(2\phi)\right\}e^{i\omega} \\
&= \left\{-\operatorname{re}(n_{11})\frac{\operatorname{sign}\big(\operatorname{re}(n_{11})\big)|n_{12}+\overline{n_{21}}|}{\sqrt{4\operatorname{re}(n_{11})^2+|n_{12}+\overline{n_{21}}|^2}}\right. \\
&\quad \left.+\frac{|n_{12}+\overline{n_{21}}|}{2}\frac{2\operatorname{sign}\big(\operatorname{re}(n_{11})\big)\operatorname{re}(n_{11})}{\sqrt{4\operatorname{re}(n_{11})^2+|n_{12}+\overline{n_{21}}|^2}}\right\}e^{i\omega} \\
&= 0, \\
[WH(N)W^*]_{21} &= \overline{[WH(N)W^*]_{12}} \\
&= 0, \\
[WH(N)W^*]_{22} &= \operatorname{tr}\big(WH(N)W^*\big) - [WH(N)W^*]_{11} \\
&= \operatorname{re}\{\operatorname{tr}(N)\} - [WH(N)W^*]_{11} \\
&= -q_1, \\
[WS(N)W^*]_{11} &= \operatorname{im}(n_{11})i\cos^2\phi + \frac{n_{21}-\overline{n_{12}}}{2}\sin\phi\cos\phi\, e^{i\omega} \\
&\quad + \frac{n_{12}-\overline{n_{21}}}{2}\sin\phi\cos\phi\, e^{-i\omega} - \operatorname{im}(n_{11})i\sin^2\phi \\
&= \operatorname{im}(n_{11})i\cos(2\phi) + \frac{1}{4}\sin(2\phi)\Big\{(n_{21}-\overline{n_{12}})e^{i\omega} \\
&\quad + (n_{12}-\overline{n_{21}})e^{-i\omega}\Big\}
\end{aligned}
$$

$$
\begin{aligned}
&= \operatorname{im}(n_{11}) i \frac{2\,\operatorname{sign}\bigl(\operatorname{re}(n_{11})\bigr)\operatorname{re}(n_{11})}{\sqrt{4\operatorname{re}(n_{11})^2 + |n_{12}+\overline{n_{21}}|^2}} \\
&\quad + \frac{1}{4}\frac{\operatorname{sign}\bigl(\operatorname{re}(n_{11})\bigr)|n_{12}+\overline{n_{21}}|}{\sqrt{4\operatorname{re}(n_{11})^2 + |n_{12}+\overline{n_{21}}|^2}}\Bigl\{(n_{21}-\overline{n_{12}})e^{i\omega} \\
&\quad + (n_{12}-\overline{n_{21}})e^{-i\omega}\Bigr\} \\
&= \frac{\operatorname{sign}\bigl(\operatorname{re}(n_{11})\bigr)}{\sqrt{4\operatorname{re}(n_{11})^2 + |n_{12}+\overline{n_{21}}|^2}}\Bigl\{2\operatorname{re}(n_{11})\operatorname{im}(n_{11})i \\
&\quad + \frac{1}{4}\bigl\{(n_{21}-\overline{n_{12}})(n_{12}+\overline{n_{21}}) + (\overline{n_{12}}+n_{21})(n_{12}-\overline{n_{21}})\bigr\}\Bigr\} \\
&= \frac{\operatorname{sign}\bigl(\operatorname{re}(n_{11})\bigr)}{\sqrt{4\operatorname{re}(n_{11})^2 + |n_{12}+\overline{n_{21}}|^2}}\Bigl\{\frac{1}{2}\bigl(\operatorname{re}(n_{11}) + \operatorname{im}(n_{11})i\bigr)^2 \\
&\quad - \frac{1}{2}\bigl(\operatorname{re}(n_{11}) - \operatorname{im}(n_{11})i\bigr)^2 + \frac{1}{4}\bigl\{n_{12}n_{21} + |n_{21}|^2 - |n_{12}|^2 - \overline{n_{12}n_{21}} \\
&\quad + |n_{12}|^2 - \overline{n_{12}n_{21}} + n_{12}n_{21} - |n_{21}|^2\bigr\}\Bigr\} \\
&= \frac{\operatorname{sign}\bigl(\operatorname{re}(n_{11})\bigr)}{2\sqrt{4\operatorname{re}(n_{11})^2 + |n_{12}+\overline{n_{21}}|^2}}\Bigl\{n_{11}^2 - \overline{n_{11}}^2 + n_{12}n_{21} - \overline{n_{12}n_{21}}\Bigr\} \\
&= \frac{\operatorname{sign}\bigl(\operatorname{re}(n_{11})\bigr)}{2\sqrt{4\operatorname{re}(n_{11})^2 + |n_{12}+\overline{n_{21}}|^2}}\Bigl\{-\det N + \overline{\det N}\Bigr\} \\
&= 0, \\
[WS(N)W^*]_{12} &= -\overline{[WS(N)W^*]_{21}} \\
&=: q_2
\end{aligned}
$$

und

$$
\begin{aligned}
[WS(N)W^*]_{22} &= \operatorname{tr}\bigl(WS(N)W^*\bigr) - [WS(N)W^*]_{11} \\
&= \operatorname{im}\{\operatorname{tr}(N)\} - 0 \\
&= 0,
\end{aligned}
$$

also

$$
\begin{aligned}
WNW^* &= WH(N)W^* + WS(N)W^* \\
&= \begin{pmatrix} q_1 & 0 \\ 0 & -q_1 \end{pmatrix} + \begin{pmatrix} 0 & q_2 \\ -\overline{q_2} & 0 \end{pmatrix} \\
&= \begin{pmatrix} q_1 & q_2 \\ -\overline{q_2} & -q_1 \end{pmatrix}
\end{aligned}
$$

mit

$$
q_1 = \operatorname{sign}\bigl(\operatorname{re}(n_{11})\bigr)\sqrt{\operatorname{re}(n_{11})^2 + \frac{1}{4}|n_{12}+\overline{n_{21}}|^2} \in \mathbb{R}\setminus\{0\}. \tag{3.52}
$$

Somit ist (3.48) erfüllt.

Als nächstes betrachten wir den Fall, daß $\mathrm{re}(n_{11}) = 0$ ist. In diesem Fall folgt aus (3.50), daß $n_{12}n_{21} \in \mathbb{R}$ ist. Damit vereinfacht sich (3.49) zu

$$0 \geq \mathrm{im}(n_{11})^2 - n_{12}n_{21}. \tag{3.53}$$

Angenommen es wäre $n_{12} + \overline{n_{21}} = 0$. Dann folgte $n_{12} = -\overline{n_{21}}$, also nach (3.53) $0 \geq \mathrm{im}(n_{11})^2 + |n_{12}|^2$ und damit $\mathrm{im}(n_{11}) = n_{12} = n_{21} = 0$. Dies hieße aber, daß $N = 0$ wäre und nach (3.47)

$$M = e^{-i\psi_0} M(m_0, \psi_0) + m_0 I_2 = e^{-i\psi_0} N + m_0 I_2 = m_0 I_2$$

ein Vielfaches der Einheitsmatrix, was wir zu Beginn ausgeschlossen haben. Daher ist $n_{12} + \overline{n_{21}} \neq 0$ und

$$\begin{aligned}
[WH(N)W^*]_{11} &= \frac{n_{21} + \overline{n_{12}}}{2} \sin\frac{\pi}{4} \cos\frac{\pi}{4}\, e^{i\omega} + \frac{n_{12} + \overline{n_{21}}}{2} \sin\frac{\pi}{4} \cos\frac{\pi}{4}\, e^{-i\omega} \\
&= \frac{1}{2}|n_{12} + \overline{n_{21}}| \\
&=: q_1, \\
[WH(N)W^*]_{12} &= -\frac{n_{21} + \overline{n_{12}}}{2} \sin^2\frac{\pi}{4}\, e^{2i\omega} + \frac{n_{12} + \overline{n_{21}}}{2} \cos^2\frac{\pi}{4} \\
&= \frac{1}{4}\Big\{-(n_{12} + \overline{n_{21}}) + n_{12} + \overline{n_{21}}\Big\} \\
&= 0, \\
[WH(N)W^*]_{21} &= 0, \\
[WH(N)W^*]_{22} &= -q_1, \\
[WS(N)W^*]_{11} &= \mathrm{im}(n_{11}) i \cos\frac{\pi}{2} + \frac{1}{4}\sin\frac{\pi}{2}\{(n_{21} - \overline{n_{12}})e^{i\omega} + (n_{12} - \overline{n_{21}})e^{-i\omega}\} \\
&= \frac{1}{4}\,\frac{(n_{21} - \overline{n_{12}})(n_{12} + \overline{n_{21}}) + (n_{12} - \overline{n_{21}})(n_{21} + \overline{n_{12}})}{|n_{12} + \overline{n_{21}}|} \\
&= \frac{2(n_{12}n_{21} - \overline{n_{12}n_{21}})}{4|n_{12} + \overline{n_{21}}|} \\
&= 0, \\
[WS(N)W^*]_{12} &= -\overline{[WS(N)W^*]_{21}} \\
&=: q_2,
\end{aligned}$$

sowie

$$[WS(N)W^*]_{22} = 0,$$

also

$$WNW^* = \begin{pmatrix} q_1 & q_2 \\ -\overline{q_2} & -q_1 \end{pmatrix}$$

mit

$$q_1 = \frac{1}{2}|n_{12} + \overline{n_{21}}| > 0. \tag{3.54}$$

Damit ist auch in diesem Fall (3.48) erfüllt.

Wir zeigen jetzt, daß die Nichtpositivität der Determinante von $N = M(m_0, \psi_0)$ zur Folge hat, daß die Off-Norm von WMW^* kleiner ist als die Off-Norm von M. Nach (3.52) und (3.54) ist $q_1^2 = \mathrm{re}(n_{11})^2 + \frac{1}{4}|n_{12} + \overline{n_{21}}|^2$, also folgt aus (3.51)

$$
\begin{aligned}
\mathrm{off}(M) - \mathrm{off}(WMW^*) &= \mathrm{off}(e^{-i\psi_0}N + m_0 I_2) - \mathrm{off}\big(W(e^{-i\psi_0}N + m_0 I_2)W^*\big) \\
&= \mathrm{off}(N) - \mathrm{off}(Q) \\
&= \|N\|_F^2 - 2|n_{11}|^2 - \big(\|Q\|_F^2 - 2|q_1|^2\big) \\
&= 2\big(q_1^2 - |n_{11}|^2\big) \\
&= 2\left\{\mathrm{re}(n_{11})^2 + \frac{1}{4}|n_{12} + \overline{n_{21}}|^2 - \mathrm{re}(n_{11})^2 - \mathrm{im}(n_{11})^2\right\} \\
&= \frac{1}{2}|n_{12} + \overline{n_{21}}|^2 - 2\,\mathrm{im}(n_{11})^2 \\
&= \frac{1}{2}\big\{|n_{12}|^2 + |n_{21}|^2 + 2\,\mathrm{re}(n_{12}n_{21})\big\} - 2\,\mathrm{im}(n_{11})^2 \\
&= \frac{1}{2}\big(|n_{12}| - |n_{21}|\big)^2 + |n_{12}n_{21}| + \mathrm{re}(n_{12}n_{21}) - 2\,\mathrm{im}(n_{11})^2 \\
&\geq \frac{1}{2}\big(|n_{12}| - |n_{21}|\big)^2 \\
&\geq 0.
\end{aligned}
$$

Wir haben also eine Methode gefunden, eine beliebige 2×2 - Matrix durch eine unitäre Ähnlichkeitstransformation in eine ΔH-Matrix mit kleinerer Off-Norm zu verwandeln. Damit sieht ein Minimierungs-Sweep für die Off-Norm, also ein Zyklus, in dem zu einer gegebenen Matrix alle 2×2 - Teilmatrizen nacheinander in eine ΔH-Matrix mit kleinerer Off-Norm transformiert werden, folgendermaßen aus:

Algorithmus 3.10 (Minimierungs-Sweep für die Off-Norm)
Eingabe: $X \in \mathbb{C}^{n,n}$

- Setze $X^{(11)} := X$.
- Für $l = 2, \ldots, n$
  - setze $X^{(1l)} := X^{(l-1,l-1)}$,
  - für $k = 1, \ldots, l-1$
    - setze $Y^{(kl)} := \left\{ \begin{pmatrix} x_{kk}^{(kl)} & x_{kl}^{(kl)} \\ x_{lk}^{(kl)} & x_{ll}^{(kl)} \end{pmatrix} - \frac{x_{kk}^{(kl)} + x_{ll}^{(kl)}}{2} I_2 \right\}$,
    - berechne $\psi_{kl} := -\frac{1}{2}\arg\big(\det Y^{(kl)}\big) + \frac{\pi}{2}$,
    - setze $N^{(kl)} := e^{i\psi_{kl}} Y^{(kl)}$,
    - falls $N^{(kl)} = 0$, so setze $\phi_{kl} := 0$ und $\omega_{kl} := 0$,
    - sonst setze $\omega_{kl} := \arg\big(n_{12}^{(kl)} + \overline{n_{21}}^{(kl)}\big)$ und
      - falls $\mathrm{re}\big(n_{11}^{(kl)}\big) \neq 0$, so setze $\phi_{kl} := \frac{1}{2}\arctan \frac{\left|n_{12}^{(kl)} + \overline{n_{21}}^{(kl)}\right|}{2\,\mathrm{re}\left(n_{11}^{(kl)}\right)}$,
      - sonst setze $\phi_{kl} := \frac{\pi}{4}$,
    - berechne $W^{(kl)} := U_{kl}\big(\cos\phi^{(kl)}, \sin\phi^{(kl)}\, e^{i\omega^{(kl)}}\big)$,
    - setze $X^{(k+1,l)} := W^{(kl)} X^{(kl)} W^{(kl)*}$.

Ausgabe: $\mathrm{sweep}(X) := X^{(nn)}$

Um eine Matrix V^*AV mit der Eigenschaft, daß $V\Delta(V^*AV)V^*$ eine zu A nächstgelegene normale Matrix ist, zu bestimmen, erhalten wir das folgende Verfahren:

Algorithmus 3.11 (2. Jacobiartiges Verfahren zur Min. der Off-Norm)
Eingabe: $A \in \mathbb{C}^{n,n}$, Abbruchschranke $\epsilon > 0$

- Setze $A^{(0)} := A$.
- Berechne $f^{(0)} := \left\| A^{(0)} - \Delta\big(A^{(0)}\big) \right\|_F$.
- Für $q = 0, 1, \ldots$
    - berechne $A^{(q+1)} := \operatorname{sweep}\big(A^{(q)}\big)$ mit Algorithmus 3.10,
    - berechne $f^{(q+1)} := \left\| A^{(q+1)} - \Delta\big(A^{(q+1)}\big) \right\|_F$,
    - falls $f^{(q)} - f^{(q+1)} < \epsilon$, so beende die Schleife.

Ausgabe: $A_{\text{next}}(A) := A^{(q+1)}$

Das letzte Verfahren konvergiert gegen eine ΔH-Matrix, wobei genauere analytische Konvergenzresultate bislang nicht vorliegen. In Abschnitt 5.2 werden wir die Konvergenz der Algorithmen 3.9 und 3.11 anhand einiger Beispiele testen. Dabei zeigt sich in Beispiel 5.9, daß das letzte Verfahren zumindest numerisch nicht global konvergent ist.

Kapitel 4

Gradientenverfahren

4.1 Verfahren zur Lösung des isospektralen Abstandsproblems

Im vorigen Kapitel haben wir in Abschnitt 3.1.1 jacobiartige Verfahren zur Optimierung von g_B auf $\Lambda(A)$, und damit zur Lösung des isospektralen Abstandsproblems entwickelt, die darauf beruhen, daß g_B nacheinander über Kurven auf $\Lambda(A)$ mit vorgegebenen Richtungen optimiert wird. Genauer haben wir eine Basis von $\mathcal{S}(n)$ gewählt und nacheinander die Extrema von g_B auf den Kurven bestimmt, deren Richtungen durch die Basiselemente von $\mathcal{S}(n)$ gegeben werden. Eine andere Strategie ist es, keine festen Richtungen vorzugeben, sondern stets in die Richtung zu gehen, in der g_B am stärksten ansteigt bzw. abfällt. Da die Kurven, die durch diese Richtungen gegeben werden, im allgemeinen nicht geschlossen sind, können wir die Extremalstellen von g_B auf diesen Kurven jedoch in den meisten Fällen nicht exakt bestimmen. Die sich aus dieser Strategie ergebenden Gradientenverfahren sind in einem gewissen Sinne das Gegenstück zu den jacobiartigen Verfahren aus dem letzten Kapitel. Während wir bei den jacobiartigen Verfahren nichtoptimale Richtungen wählen und g_B auf den durch diese Richtungen bestimmten Kurven exakt optimieren, wählen wir bei den Gradientenverfahren zwar optimale Richtungen, können aber die Extrema von g_B auf den dadurch gegebenen Kurven nicht genau lokalisieren. Bereits in der Arbeit [10] wurde ein solches Gradientenverfahren zur Maximierung der zu g_B ähnlichen Funktion

$$\Lambda(A) \longrightarrow \mathbb{R},\ X \longmapsto |\operatorname{tr}(XB^*)|,$$

die bei der Untersuchung der Signalstärke in spektroskopischen Experimenten eine wichtige Rolle spielt, vorgestellt. Zur Bestimmung der Richtung, in der g_B den größten Anstieg verzeichnet, führen wir auf $\mathcal{S}(n)$ das folgende Skalarprodukt ein, bei dem es sich um eine Restriktion des Standard-Skalarproduktes auf $\mathbb{C}^{n,n}$ handelt.

Sei

$$\langle \cdot, \cdot \rangle : \mathcal{S}(n) \times \mathcal{S}(n) \longrightarrow \mathbb{R},\ \langle R, S \rangle = -\operatorname{tr}(RS). \tag{4.1}$$

Dann gilt:

Lemma 4.1
$\langle \cdot, \cdot \rangle$ ist ein Skalarprodukt auf dem reellen Vektorraum $\mathcal{S}(n)$. Die induzierte Norm $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$ ist die Frobenius-Norm.

Beweis:
Seien $R, S \in \mathcal{S}(n)$, $U \in \mathcal{U}(n)$ und D diagonal mit $R = UDU^*$. Dann ist

$$\langle R, S\rangle = -\operatorname{tr}(RS) = -\operatorname{tr}(UDU^*S) = -\operatorname{tr}(DU^*SU) = -\sum_{j=1}^{n} d_{jj}[U^*SU]_{jj}$$
$$= \sum_{j=1}^{n} \operatorname{im}\{d_{jj}\} \operatorname{im}\{[U^*SU]_{jj}\} \in \mathbb{R}.$$

Weiter ist $\langle \cdot, \cdot \rangle$ linear und symmetrisch, da die Spur linear ist und $\operatorname{tr}(RS) = \operatorname{tr}(SR)$, und

$$\langle R, R\rangle = -\operatorname{tr}(UDU^*UDU^*) = -\operatorname{tr}(DD) = -\sum_{j=1}^{n} d_{jj}^2 = \sum_{j=1}^{n} \operatorname{im}(d_{jj})^2 > 0$$

für $R \neq 0$, also $\langle \cdot, \cdot \rangle$ positiv definit. Aus

$$\langle R, R\rangle = -\operatorname{tr}(RR) = \operatorname{tr}(RR^*) = \|R\|_F^2$$

folgt, daß die durch das Skalarprodukt induzierte Norm die Frobenius-Norm ist. □

Da nach (2.24) $[H(X), H(B)] - [S(X), S(B)] \in \mathcal{S}(n)$ ist, gilt nach (2.25) für alle $X \in \Lambda(A)$ und $Y \in \mathcal{S}(n)$

$$\begin{aligned}\frac{d}{dt} g_B(\mathcal{K}_{X,Y})(0) &= \operatorname{tr}\big\{\big([H(X), H(B)] - [S(X), S(B)]\big)Y\big\}\\ &= \big\langle -\big([H(X), H(B)] - [S(X), S(B)]\big), Y\big\rangle\end{aligned}$$

und aus der Cauchy-Schwarzschen Ungleichung folgt

$$\begin{aligned}&\max\left\{\frac{d}{dt} g_B(\mathcal{K}_{X,Y})(0) : Y \in \mathcal{S}(n), \|Y\|_F = \big\|[H(X), H(B)] - [S(X), S(B)]\big\|_F\right\}\\ &= \max\Big\{\big\langle -\big([H(X), H(B)] - [S(X), S(B)]\big), Y\big\rangle :\\ &\qquad Y \in \mathcal{S}(n), \|Y\|_F = \big\|[H(X), H(B)] - [S(X), S(B)]\big\|_F\Big\}\\ &= \big\|[H(X), H(B)] - [S(X), S(B)]\big\|_F^2\\ &= \big\langle -\big([H(X), H(B)] - [S(X), S(B)]\big), -\big([H(X), H(B)] - [S(X), S(B)]\big)\big\rangle\\ &= \frac{d}{dt} g_B(\mathcal{K}_{X,\Gamma_B(X)})(0)\end{aligned}$$

mit

$$\Gamma_B : \Lambda(A) \longrightarrow \mathcal{S}(n),\ \Gamma_B(X) = -\big([H(X), H(B)] - [S(X), S(B)]\big). \tag{4.2}$$

Also steigt g_B am steilsten in der Gradientenrichtung Γ_B an. Als nächstes müssen wir eine Schrittweite wählen, das heißt wir müssen entscheiden, wie weit wir in Gradientenrichtung gehen. Ausführliche Informationen zu Schrittweitenstrategien sind in [24], Kapitel 6, zu finden. Wir beschränken uns hier auf eine kurze Herleitung der Armijo-Schrittweite, die wir im folgenden benutzen werden.

Sei $y : \mathbb{R} \longrightarrow \mathbb{R}$ differenzierbar und beschränkt und $x \in \mathbb{R}$ mit $y'(x) > 0$. Dann existieren für $\delta \in]0, 1[$ Zahlen $\zeta(x, \delta) \geq \epsilon(x, \delta) > 0$ so, daß

$$\frac{y(x + t) - y(x)}{t} \geq \delta y'(x) \quad \text{für} \quad t \in]0, \epsilon(x, \delta)] \tag{4.3}$$

und

$$\frac{y(x+t)-y(x)}{t} < \delta y'(x) \quad \text{für} \quad t > \zeta(x,\delta) \tag{4.4}$$

ist. (4.3) gilt wegen der Differenzierbarkeit von y und (4.4) folgt aus der Beschränktheit von y. Wir suchen ausgehend von einem Startwert t_0 eine möglichst große Schrittweite t mit

$$y(x+t) \geq y(x) + \delta t y'(x). \tag{4.5}$$

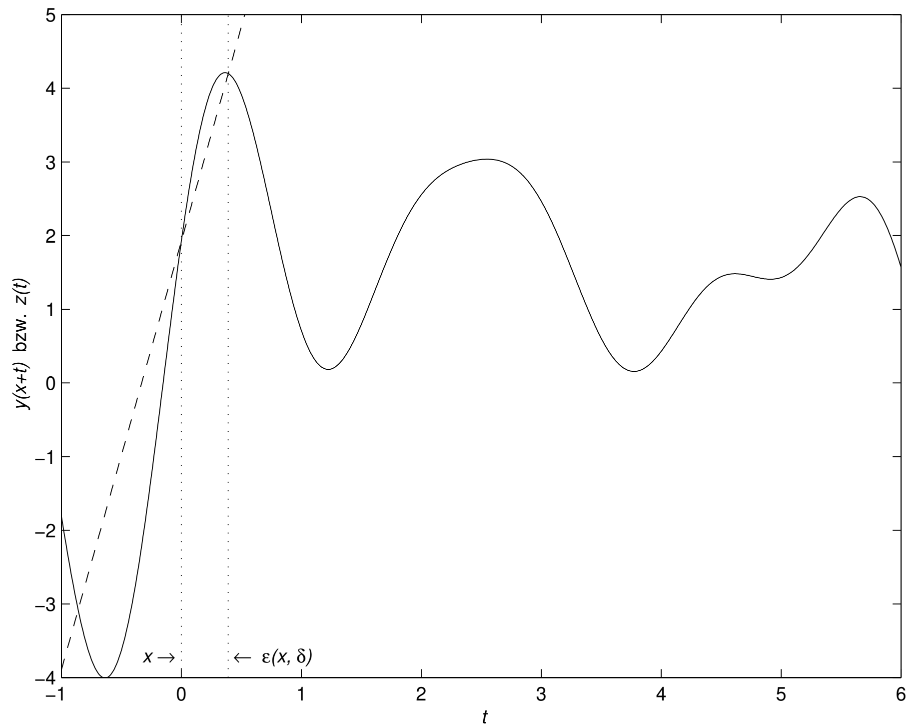

Abbildung 4.1: Verlauf von $y(x+t)$ (durchgezogene Linie) und $z(t) := y(x)+\delta t y'(x)$ (gestrichelte Linie) für $y = g_B(\mathcal{K}_{X,\Gamma_B(X)})$ mit zufälligen $X, B \in \mathbb{C}^{6,6}$, $x = 0$ und $\delta = \frac{1}{2}$.

Dabei gehen wir folgendermaßen vor: Ist (4.5) für t_0 nicht erfüllt, so verkleinern wir t_0 solange um den Faktor $\gamma \in]0,1[$, bis (4.5) erfüllt ist. Gilt (4.5) für t_0, so vergrößern wir t_0 solange um den Faktor $\frac{1}{\gamma}$, wie (4.5) richtig bleibt. Genauer sieht das Verfahren so aus:

Gegeben: $x \in \mathbb{R}$ mit $y'(x) > 0$, $\gamma, \delta \in]0,1[$

- Wähle $t_0 > 0$.
- Falls $y(x+t_0) \geq y(x) + \delta t_0 y'(x)$, so
  - setze $t_1^- := t_0$,
  - für $k = 1, 2, \ldots$
    - setze $t_k^+ := \frac{t_k^-}{\gamma}$,

* $*$ falls $y(x+t_k^+) < y(x) + \delta t_k^+ y'(x)$, so setze $t := t_k^-$ und beende die Schleife,
    * $*$ setze $t_{k+1}^- := t_k^+$,

- sonst
  - setze $t_1^+ := t_0$,
  - für $k = 1, 2, \dots$
    * $*$ setze $t_k^- := \gamma t_k^+$,
    * $*$ falls $y(x+t_k^-) \geq y(x) + \delta t_k^- y'(x)$, so setze $t := t_k^-$ und beende die Schleife,
    * $*$ setze $t_{k+1}^+ := t_k^-$.

Wegen (4.3) und (4.4) endet das Verfahren nach endlich vielen Schritten. Angewandt mit $y = g_B(\mathcal{K}_{X,\Gamma_B(X)})$, $x = 0$, $\gamma = \delta = \frac{1}{2}$ und $t_0 = 1$ ergibt sich das folgende Verfahren:

Algorithmus 4.1 (Gradientenschritt zur Maximierung von g_B)
Eingabe: $X, B \in \mathbb{C}^{n,n}$

- Berechne $Y := \Gamma_B(X)$.
- Falls $Y = 0$, so setze $\operatorname{maxstep}(X, B) := X$ und beende das Verfahren.
- Berechne $y := g_B(X)$.
- Berechne $y' := \|Y\|_F^2$.
- Setze $t_0 := 1$.
- Falls $g_B(\mathcal{K}_{X,Y})(t_0) \geq y + \frac{1}{2} t_0 y'$, so
  - setze $t_1^- := t_0$,
  - für $k = 1, 2, \dots$
    * $*$ setze $t_k^+ := 2t_k^-$,
    * $*$ falls $g_B(\mathcal{K}_{X,Y})(t_k^+) < y + \frac{1}{2} t_k^+ y'$, so setze $t := t_k^-$ und beende die Schleife,
    * $*$ setze $t_{k+1}^- := t_k^+$,
- sonst
  - setze $t_1^+ := t_0$,
  - für $k = 1, 2, \dots$
    * $*$ setze $t_k^- := \frac{1}{2} t_k^+$,
    * $*$ falls $g_B(\mathcal{K}_{X,Y})(t_k^-) \geq y + \frac{1}{2} t_k^- y'$, so setze $t := t_k^-$ und beende die Schleife,
    * $*$ setze $t_{k+1}^+ := t_k^-$.

Ausgabe: $\operatorname{maxstep}(X, B) := \mathcal{K}_{X,Y}(t)$

Der Rechenaufwand eines Gradientenschrittes ist im wesentlichen proportional zu der Anzahl der Vergrößerungen bzw. Verkleinerungen des Startwertes t_0 und daher kaum vorhersagbar. Zur Minimierung von g_B gehen wir in die negative Gradientenrichtung. Die oben angegebene Schrittweitenstrategie läßt sich auf diesen Fall durch geringfügige Modifikationen übertragen.

Insgesamt erhalten wir das folgende Verfahren zur Berechnung einer Matrix aus $\Lambda(A)$, die g_B maximiert, also nach (2.14) f_B minimiert und damit das isospektrale Minimierungsproblem (2.6) löst:

Algorithmus 4.2 (Gradientenverfahren zur Maximierung von g_B)
Eingabe: $A, B \in \mathbb{C}^{n,n}$, Abbruchschranke $\epsilon > 0$

- Setze $A^{(0)} := A$.
- Berechne $g^{(0)} := g_B\big(A^{(0)}\big)$.
- Für $q = 0, 1, \ldots$
  - berechne $A^{(q+1)} := \operatorname{maxstep}\big(A^{(q)}, B\big)$ mit Algorithmus 4.1,
  - berechne $g^{(q+1)} := g_B\big(A^{(q+1)}\big)$,
  - falls $g^{(q+1)} - g^{(q)} < \epsilon$, so beende die Schleife.

Ausgabe: $A_{\text{Gradient}}(A, B) := A^{(q+1)}$

Ein Verfahren zur Berechnung einer zur Matrix A unitär ähnlichen Matrix, die g_B minimiert, also f_B nach (2.13) maximiert und so das isospektrale Maximierungsproblem (2.5) löst, läßt sich analog gewinnen.

Im Gegensatz zu den jacobiartigen Verfahren im vorigen Kapitel wissen wir bei den Gradientenverfahren nur, daß sie gegen irgendeinen stationären Punkt von g_B konvergieren. Anhand einiger Beispiele studieren wir die Konvergenz von Algorithmus 4.2 in Abschnitt 5.1.

4.2 Ein Verfahren zur Berechnung einer nächstgelegenen normalen Matrix

In Abschnitt 3.2 haben wir für $A \in \mathbb{C}^{n,n}$ ein Verfahren zur Berechnung einer zu A unitär ähnlichen Matrix V^*AV mit der Eigenschaft, daß $V\Delta(V^*AV)V^*$ eine zu A nächstgelegene normalen Matrix ist, hergeleitet, das darauf beruht, daß die Off-Norm von A durch eine Folge von unitären Ähnlichkeitstransformationen verringert wird. Zu diesem Zweck haben wir das jacobiartige Verfahren 3.7 zur Maximierung von g_B auf $\Lambda(A)$ so modifiziert, daß anstelle der Funktion g_B die von der Iterierten $X^{(k)}$ abhängige Funktion $g_{\Delta(X^{(k)})}$ vergrößert wird. Modifizieren wir Algorithmus 4.2 in derselben Weise, so erhalten wir ein Gradientenverfahren zur Berechnung von V^*AV:

Algorithmus 4.3 (Gradientenverfahren zur Minimierung der Off-Norm)
Eingabe: $A \in \mathbb{C}^{n,n}$, Abbruchschranke $\epsilon > 0$

- Setze $A^{(0)} := A$.
- Berechne $f^{(0)} := \big\|A^{(0)} - \Delta\big(A^{(0)}\big)\big\|_F$.
- Für $q = 0, 1, \ldots$
  - berechne $A^{(q+1)} := \operatorname{maxstep}\big(A^{(q)}, \Delta(A^{(q)})\big)$ mit Algorithmus 4.1,
  - berechne $f^{(q+1)} := \big\|A^{(q+1)} - \Delta\big(A^{(q+1)}\big)\big\|_F$,
  - falls $f^{(q)} - f^{(q+1)} < \epsilon$, so beende die Schleife.

Ausgabe: $A_{\text{next}}(A) := A^{(q+1)}$

Das Verfahren ist nicht global konvergent, da sich die Beispiele 3.13 und 3.14 aus dem vorigen Kapitel übertragen lassen. In Abschnitt 5.2 testen wir die Konvergenz dieses Verfahrens in einigen Beispielen.

Kapitel 5

Numerische Beispiele

In diesem Kapitel wollen wir das Konvergenzverhalten der in den Kapiteln 3 und 4 entwickelten Verfahren zur Lösung des isospektralen Minimierungsproblems (2.6) und zur Berechnung einer nächstgelegenen normalen Matrix anhand einiger Beispiele untersuchen. Da die Verfahren zur Lösung des isospektralen Maximierungsproblems (2.5) sich bezüglich ihrer Konvergenz vollkommen analog zu den Verfahren zur Lösung des isospektralen Minimierungsproblems verhalten, verzichten wir auf numerische Tests zu diesen Verfahren. Sämtliche Berechnungen wurden mit dem Programmpaket MATLAB durchgeführt, als Abbruchschranke wurde bei allen Verfahren $\epsilon = 10^{-12}$ gewählt. Das Konvergenzverhalten der Verfahren in den hier untersuchten Beispielen ist insofern typisch, als daß in mehreren hundert Simulationen mit Zufallsmatrizen ähnliche Ergebnisse erzielt wurden. Zur Erzeugung der Beispielmatrizen benutzen wir die beiden schiefhermiteschen Matrizen

$$Y_1 := \begin{pmatrix} 2i & 4+6i & 4+5i & -i & 3-6i & 9-5i & -5+9i \\ -4+6i & -8i & -9+7i & -8+4i & 8+7i & -1-4i & -7+2i \\ -4+5i & 9+7i & -i & 3i & 7+i & 7i & 2-9i \\ -i & 8+4i & 3i & -5i & -7-i & -1-i & -9+5i \\ -3-6i & -8+7i & -7+i & 7-i & 6i & 6-3i & 4-3i \\ -9-5i & 1-4i & 7i & 1-i & -6-3i & 0 & -2+3i \\ 5+9i & 7+2i & -2-9i & 9+5i & -4-3i & 2+3i & -9i \end{pmatrix}$$

und

$$Y_2 := \begin{pmatrix} 5i & 3 & 4+4i & 8-3i & 2+9i & -9-4i & -1+6i \\ -3 & -6i & 8-i & -1 & -7+3i & 3+i & -9-6i \\ -4+4i & -8-i & -9i & 3-2i & -3+i & 8-4i & 4+4i \\ -8-3i & 1 & -3-2i & 0 & 6i & 9-5i & -1+2i \\ -2+9i & 7+3i & 3+i & 6i & -8i & 7+4i & 5+9i \\ 9-4i & -3+i & -8-4i & -9-5i & -7+4i & -6i & 4-4i \\ 1+6i & 9-6i & -4+4i & 1+2i & -5+9i & -4-4i & 8i \end{pmatrix}.$$

Zunächst studieren wir das Konvergenzverhalten der Verfahren zur Lösung des isospektralen Minimierungsproblems.

5.1 Beispiele zur Lösung des isospektralen Minimierungsproblems

In diesem Abschnitt untersuchen wir die Konvergenz der in den Abschnitten 3.1 und 4.1 entwickelten Verfahren zur Lösung des isospektralen Minimierungsproblems

(2.6). Dies sind die jacobiartigen Verfahren Algorithmus 3.5 und Algorithmus 3.7, wobei letzterer nur für diagonale Matrizen B durchführbar ist, sowie das Gradientenverfahren Algorithmus 4.2. Das exakte Minimum $F^-(A,B)$ der Funktion f_B auf $\Lambda(A)$ wurde in den Beispielen mit Hilfe von Satz 2.9 ermittelt. Wir bemerken schon an dieser Stelle, daß in allen untersuchten Beispielen mit diagonalem B die Konvergenzkurven der jacobiartigen Algorithmen 3.5 und 3.7 in den Grafiken so dicht beieinander liegen, daß kein Unterschied zu erkennen ist. Zunächst untersuchen wir die Konvergenz der Verfahren für reelle Diagonalmatrizen B.

Beispiel 5.1
Sei

$$A = e^{Y_1}\operatorname{diag}(-1,2,-7,4,5,-3,9)e^{-Y_1} + e^{Y_2}\operatorname{diag}(i,3i,2i,4i,-5i,8i,-7i)e^{-Y_2},$$

also eine unstrukturierte komplexe 7×7 - Matrix, und

$$B = \operatorname{diag}(-9,-4,-1,0,1,4,9)$$

eine reelle Diagonalmatrix mit paarweise verschiedenen Eigenwerten. Dann konvergieren die jacobiartigen Algorithmen 3.5 und 3.7 quadratisch gegen eine Lösung von (2.6), wie Abbildung 5.1 zeigt. Das Gradientenverfahren Algorithmus 4.2 konvergiert ebenfalls gegen eine Lösung von (2.6), besitzt aber keine quadratische Konvergenz und benötigt erheblich mehr Schritte als die jacobiartigen Verfahren, wie in Abbildung 5.2 zu sehen ist.

Beispiel 5.2
Sei

$$A = e^{Y_1}\operatorname{diag}(-1,2,-7,4,5,-3,9)e^{-Y_1}$$

eine hermitesche Matrix und

$$B = \operatorname{diag}(1,4,9,0,1,4,9)$$

eine reelle Diagonalmatrix mit mehrfachen Eigenwerten. Die jacobiartigen Verfahren konvergieren hier langsamer gegen eine Lösung von (2.6) als im vorigen Beispiel, da sie, bedingt durch die Vielfachheit der Eigenwerte von B, nicht mehr quadratisch konvergent sind. Das Gradientenverfahren konvergiert dagegen schneller als im vorigen Beispiel. Siehe dazu die Abbildungen 5.3 und 5.4.

Insgesamt ist zu beobachten, daß die jacobiartigen Verfahren quadratisch konvergent sind, wenn B eine reelle Diagonalmatrix mit paarweise verschiedenen Eigenwerten ist. Bezüglich der Konvergenz ist zwischen den beiden jacobiartigen Verfahren kein Unterschied zu erkennen; da ein Sweep von Algorithmus 3.5 jedoch erheblich aufwendiger ist als ein Sweep von Algorithmus 3.7, benötigt Algorithmus 3.7 wesentlich weniger Rechenzeit und ist daher schneller. Das Gradientenverfahren ist in diesem Fall nicht konkurrenzfähig, da es die Diagonalität von B nicht ausnutzt. Besitzt B mehrfache Eigenwerte, so verschlechtert sich die Konvergenz der jacobiartigen Verfahren, insbesondere geht die quadratische Konvergenz verloren. Die Konvergenz des Gradientenverfahrens wird dagegen durch mehrfache Eigenwerte beschleunigt. Ist B keine reelle, sondern eine rein imaginäre Diagonalmatrix, so verhalten sich die Verfahren genauso.

Als nächstes untersuchen wir, wie die Verfahren konvergieren, wenn B zwar hermitesch, aber keine Diagonalmatrix ist. In diesem Fall kann Algorithmus 3.7 nicht verwendet werden.

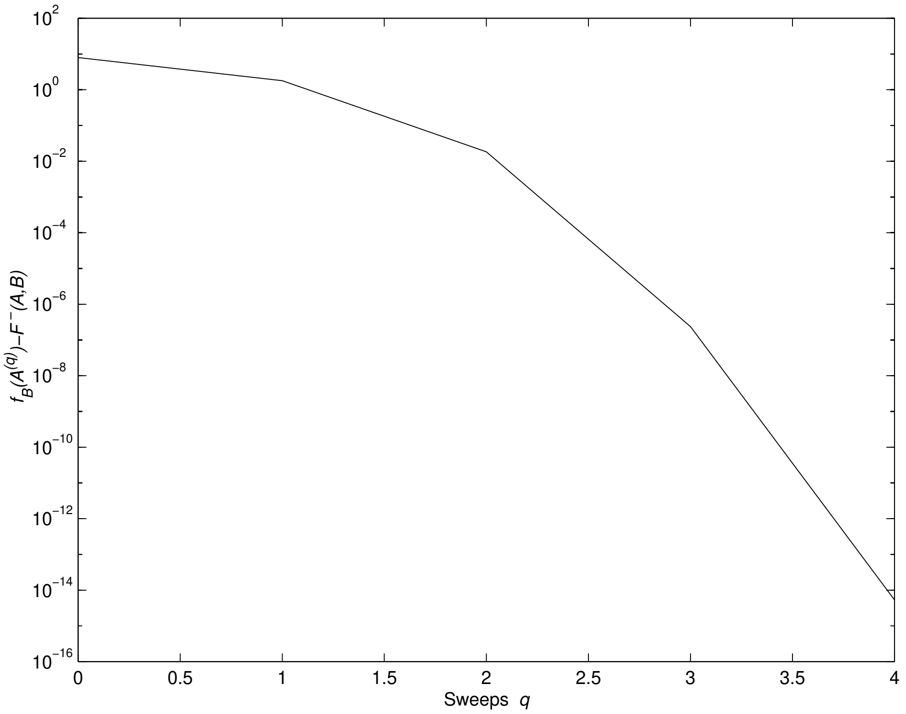

Abbildung 5.1: Konvergenz der jacobiartigen Verfahren Algorithmus 3.5 und Algorithmus 3.7 in Beispiel 5.1

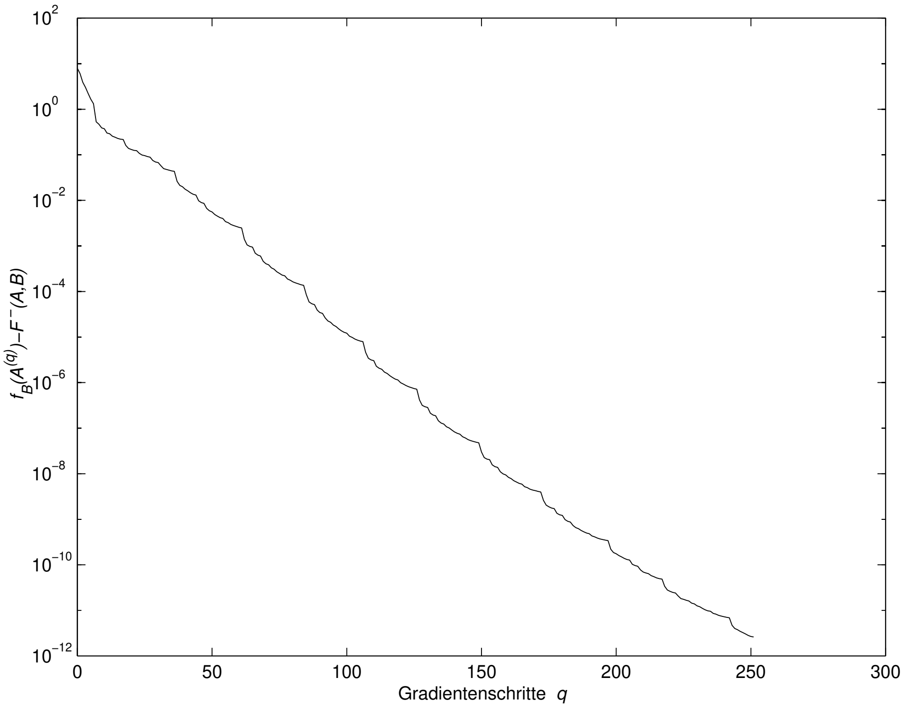

Abbildung 5.2: Konvergenz des Gradientenverfahrens Algorithmus 4.2 in Beispiel 5.1

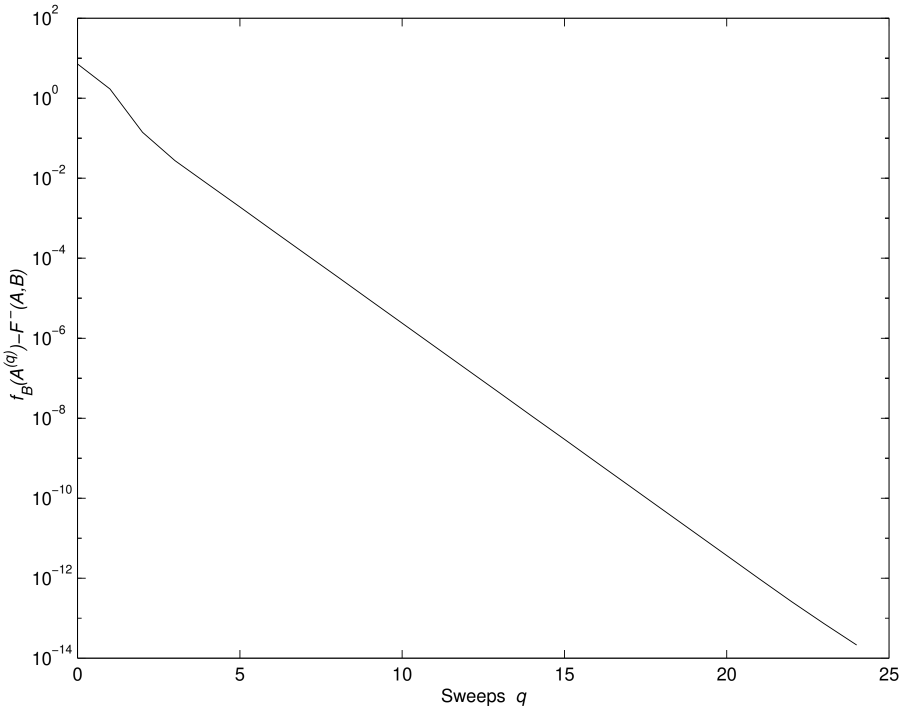

Abbildung 5.3: Konvergenz der jacobiartigen Verfahren Algorithmus 3.5 und Algorithmus 3.7 in Beispiel 5.2

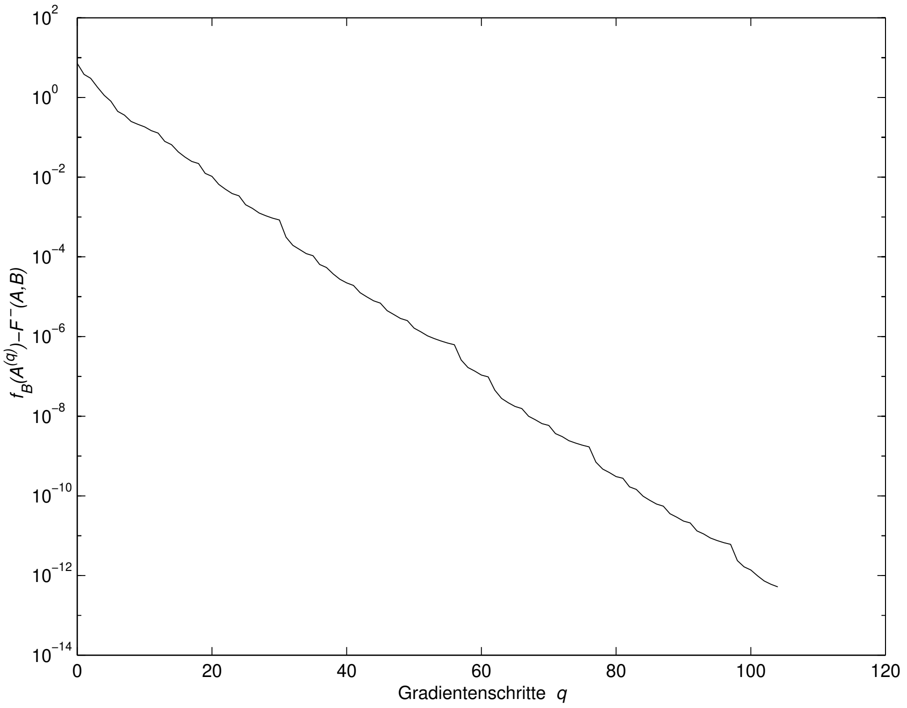

Abbildung 5.4: Konvergenz des Gradientenverfahrens Algorithmus 4.2 in Beispiel 5.2

Beispiel 5.3
Sei

$$A = e^{Y_1} \operatorname{diag}(-1, 2, -7, 4, 5, -3, 9) e^{-Y_1}$$

und

$$B = e^{Y_2} \operatorname{diag}(-9, -4, -1, 0, 1, 4, 9) e^{-Y_2}.$$

Dann konvergieren der jacobiartige Algorithmus 3.5 und das Gradientenverfahren Algorithmus 4.2 gegen eine Lösung von (2.6). Obwohl alle Eigenwerte von B einfach sind, besitzt das jacobiartige Verfahren keine quadratische Konvergenz, zudem ist die Konvergenz noch erheblich langsamer als in Beispiel 5.2. Das Gradientenverfahren konvergiert etwa gleich schnell wie in Beispiel 5.1. Siehe dazu Abbildung 5.5.

Beispiel 5.4
Sei

$$A = e^{Y_1} \operatorname{diag}(-1, 2, -7, 4, 5, -3, 9) e^{-Y_1}$$

und

$$B = e^{Y_2} \operatorname{diag}(1, 4, 9, 0, 1, 4, 9) e^{-Y_2}.$$

In diesem Fall besitzt B mehrfache Eigenwerte, was zur Folge hat, daß sowohl das jacobiartige, als auch das Gradientenverfahren schneller konvergieren als im vorigen Beispiel, wie Abbildung 5.6 zeigt.

Für hermitesche B verschlechtert sich die Konvergenz des jacobiartigen Algorithmus 3.5 also gravierend, wenn B nicht mehr diagonal ist. Das Gradientenverfahren benötigt zwar in den meisten dieser Fälle immer noch mehr Schritte als das jacobiartige Verfahren, ein Gradientenschritt ist aber weniger aufwendig als ein Sweep von Algorithmus 3.5. Daher konvergiert das Gradientenverfahren insgesamt schneller als das jacobiartige Verfahren. Für schiefhermitesche Matrizen B verhalten sich die Verfahren ebenso.

Damit bietet sich zur Berechnung einer Lösung des Minimierungsproblems (2.6) mit nichtdiagonaler Matrix B folgende Strategie an: Ist B normal, so sollte, wie in Abschnitt 3.1.3 beschrieben, B zunächst einmal diagonalisiert werden und dann das jacobiartige Verfahren Algorithmus 3.7 angewendet werden. Ist B nicht normal, so sollte das Gradientenverfahren verwendet werden.

Als letztes untersuchen wir in diesem Abschnitt die Konvergenz der Verfahren für normale Matrizen A und B.

Beispiel 5.5
Sei

$$B = \operatorname{diag}(8i, 7+8i, -7-i, -1-7i, -5+3i, -2, -1+7i)$$

und

$$a = (6-2i, 7-8i, 5+5i, -4+4i, 1-4i, 1+3i, -8).$$

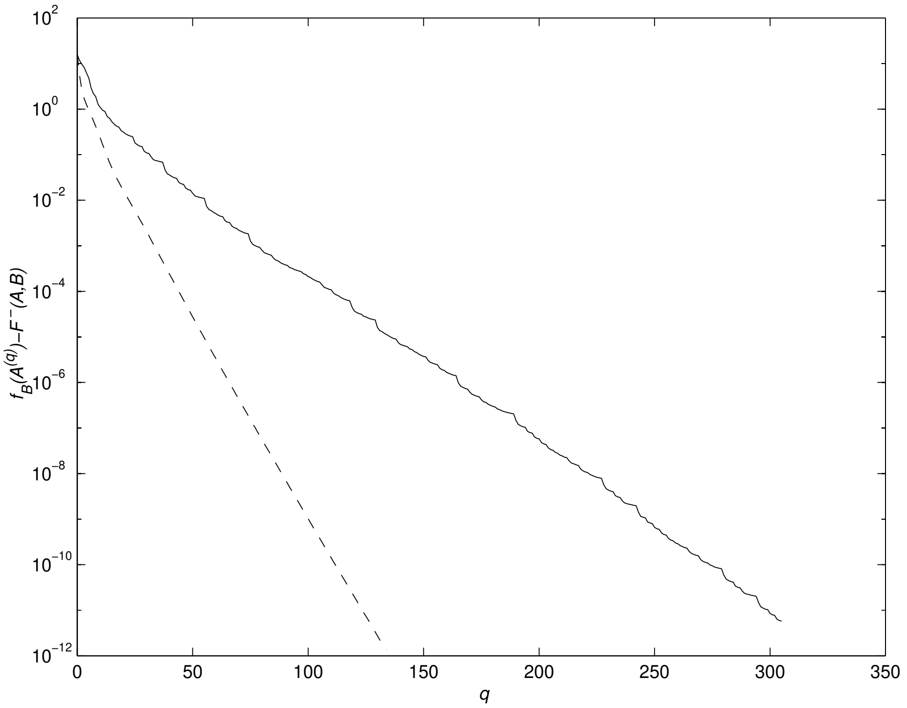

Abbildung 5.5: Konvergenz des jacobiartigen Verfahrens Algorithmus 3.5 (gestrichelte Linie) und des Gradientenverfahrens Algorithmus 4.2 (durchgezogene Linie) in Beispiel 5.3

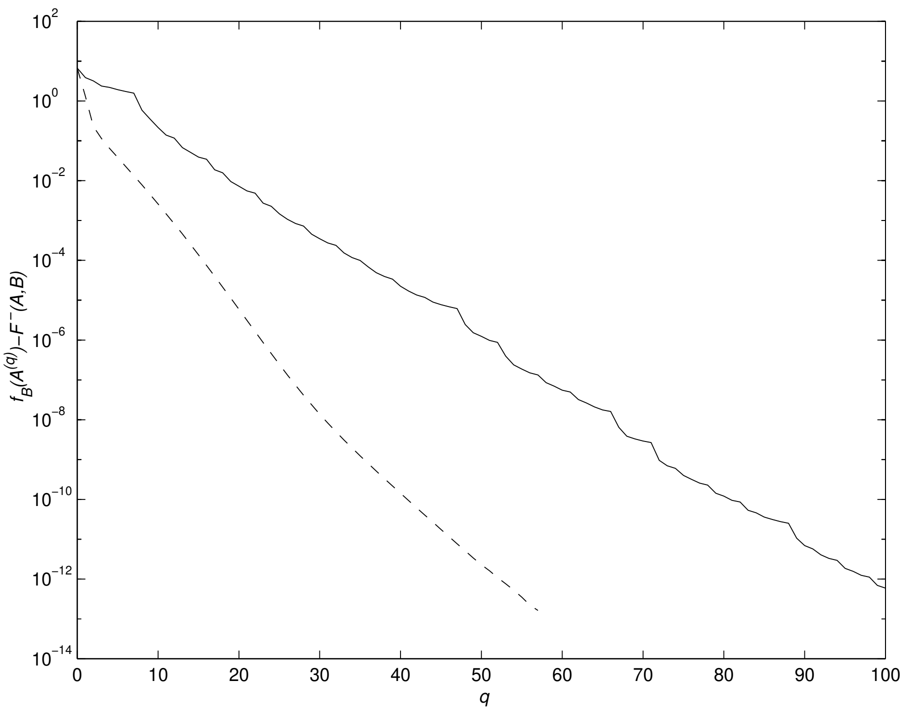

Abbildung 5.6: Konvergenz des jacobiartigen Verfahrens Algorithmus 3.5 (gestrichelte Linie) und des Gradientenverfahrens Algorithmus 4.2 (durchgezogene Linie) in Beispiel 5.4

Fall 1: $A = e^{Y_1} \operatorname{diag}(a_1, a_2, a_3, a_4, a_5, a_6, a_7)e^{-Y_1}$
In diesem Fall konvergieren die jacobiartigen Algorithmen 3.5 und 3.7 quadratisch gegen eine Lösung des isospektralen Minimierungsproblems (2.6). Das Gradientenverfahren Algorithmus 4.2 konvergiert ebenfalls gegen eine Lösung von (2.6), ist aber erheblich langsamer. Siehe dazu die Abbildungen 5.7 und 5.8.

Fall 2: $A = e^{Y_1} \operatorname{diag}(a_2, a_3, a_1, a_4, a_5, a_6, a_7)e^{-Y_1}$
In diesem Fall konvergieren weder die jacobiartigen Verfahren noch das Gradientenverfahren gegen eine Lösung von (2.6), wie die Abbildungen 5.9 und 5.10 zeigen.

Fall 3: $A = e^{Y_1} \operatorname{diag}(a_4, a_6, a_5, a_1, a_7, a_2, a_3)e^{-Y_1}$
In diesem Fall konvergieren die jacobiartigen Verfahren gegen eine Lösung von (2.6), das Gradientenverfahren jedoch nicht.

Fall 4: $A = e^{Y_1} \operatorname{diag}(a_3, a_2, a_1, a_4, a_5, a_6, a_7)e^{-Y_1}$
In diesem Fall konvergieren die jacobiartigen Verfahren nicht gegen eine Lösung von (2.6), dafür jedoch das Gradientenverfahren.

In den Fällen 3 und 4 verzichten wir auf grafische Darstellungen des Konvergenzverhaltens.

Beispiel 5.6
Sei

$$A = e^{Y_1} \operatorname{diag}(1 + i, -3 - 2i, 4 + 3i, -5 + 2i, 6 + 5i, 8 + 7i, 9 + 9i)e^{-Y_1}$$

und

$$B = \operatorname{diag}(1 - 2i, -3 + 4i, -6 - 4i, 7 + 5i, 5 + 7i, -8 - 5i, 8 + 8i).$$

Dann sind zwar die Real- und Imaginärteile der Eigenwerte von A und B paarweise verschieden, aber wegen

$$\begin{aligned}
&\operatorname{re}\big\{\big((1+i) - (4+3i)\big)\big(\overline{(1-2i)} - \overline{(-3+4i)}\big)\big\} \\
&= \operatorname{re}\big\{\big((-3-2i) - (-5+2i)\big)\big(\overline{(-6-4i)} - \overline{(-8-5i)}\big)\big\} \\
&= \operatorname{re}\big\{\big((6+5i) - (8+7i)\big)\big(\overline{(7+5i)} - \overline{(5+7i)}\big)\big\} \\
&= 0
\end{aligned}$$

ist die Lösung des isospektralen Minimierungsproblems (2.6) nicht eindeutig (vergleiche Beispiel 2.11). Dies führt dazu, daß die jacobiartigen Verfahren zwar gegen eine Lösung von (2.6) konvergieren, aber die Konvergenz nicht mehr quadratisch ist. Dennoch konvergieren sie immer noch erheblich schneller als das Gradientenverfahren. Siehe dazu die Abbildungen 5.11 und 5.12.

Zusammenfassend stellen wir folgendes fest: Selbst wenn A und B normal sind, so kann es passieren, daß sowohl die jacobiartigen Algorithmen 3.5 und 3.7, als auch das Gradientenverfahren Algorithmus 4.2 nicht gegen eine Lösung des isospektralen Minimierungsproblems (2.6) konvergieren, wobei die Konvergenz der jacobiartigen Verfahren und die des Gradientenverfahrens unabhängig voneinander sind. Sind die Verfahren konvergent und ist die Lösung von (2.6) eindeutig, so konvergieren die jacobiartigen Verfahren quadratisch, existieren mehrere Lösungen, so geht die quadratische Konvergenz zwar verloren, aber die jacobiartigen Verfahren sind immer noch schneller als das Gradientenverfahren.

Im nächsten Abschnitt beschäftigen wir uns mit der Konvergenz der Verfahren zur Berechnung einer nächstgelegenen normalen Matrix.

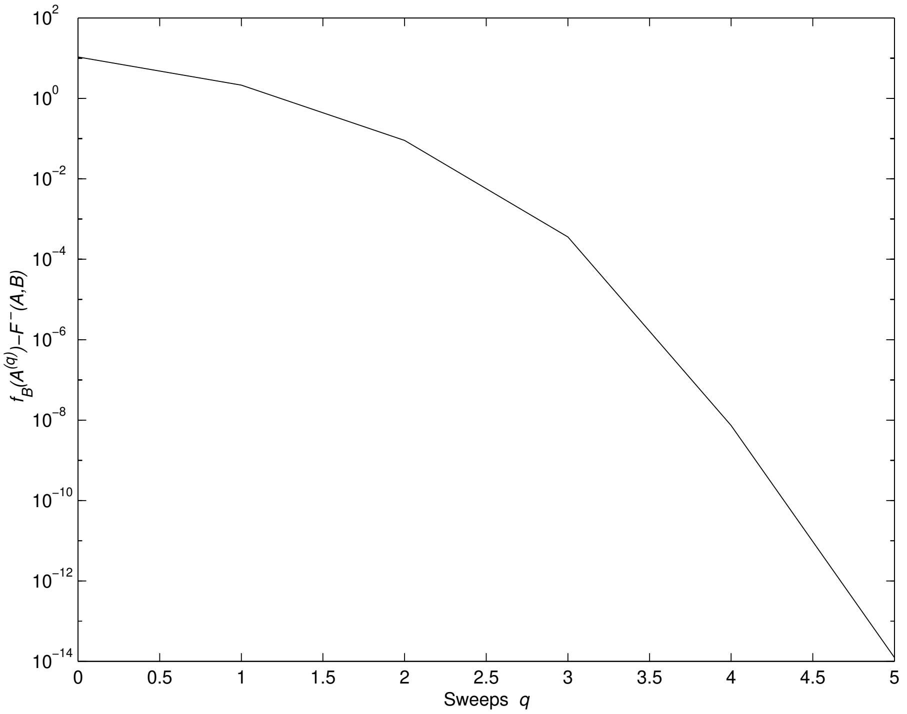

Abbildung 5.7: Konvergenz der jacobiartigen Verfahren Algorithmus 3.5 und Algorithmus 3.7 in Beispiel 5.5, Fall 1

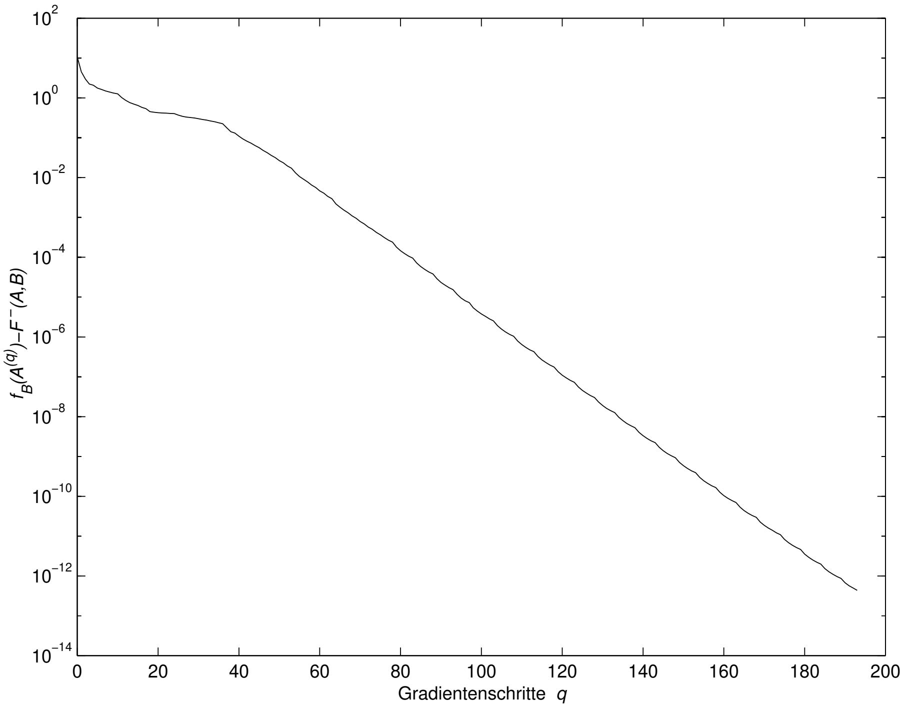

Abbildung 5.8: Konvergenz des Gradientenverfahrens Algorithmus 4.2 in Beispiel 5.5, Fall 1

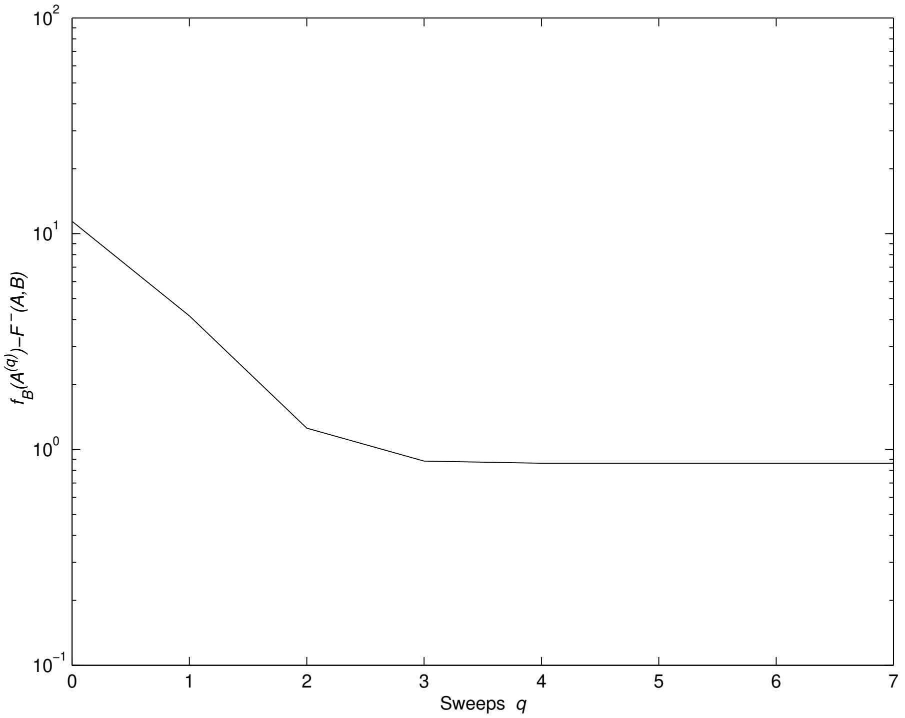

Abbildung 5.9: Konvergenz der jacobiartigen Verfahren Algorithmus 3.5 und Algorithmus 3.7 in Beispiel 5.5, Fall 2

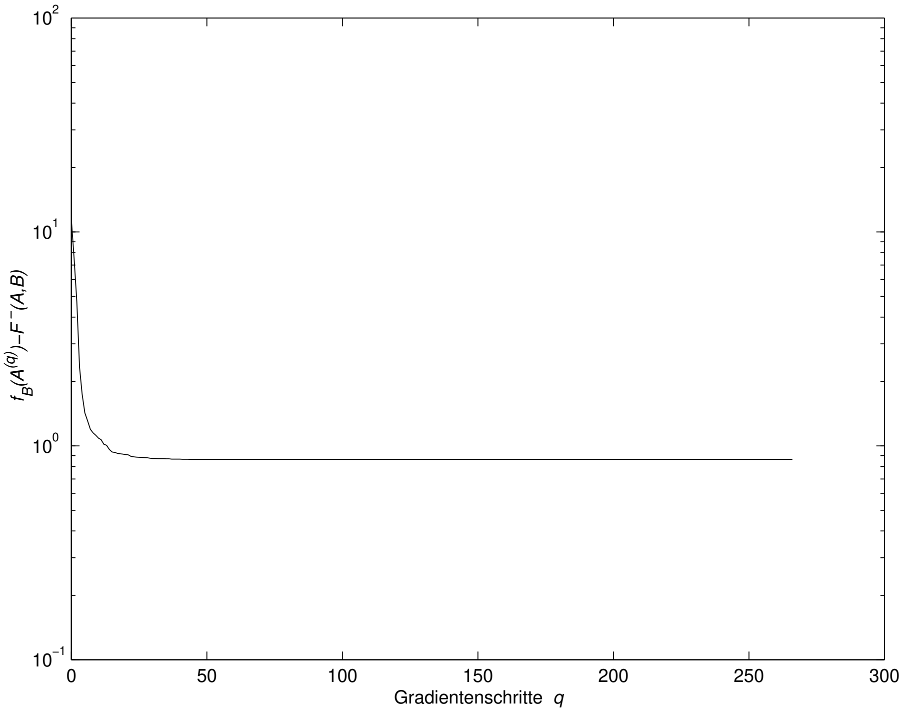

Abbildung 5.10: Konvergenz des Gradientenverfahrens Algorithmus 4.2 in Beispiel 5.5, Fall 2

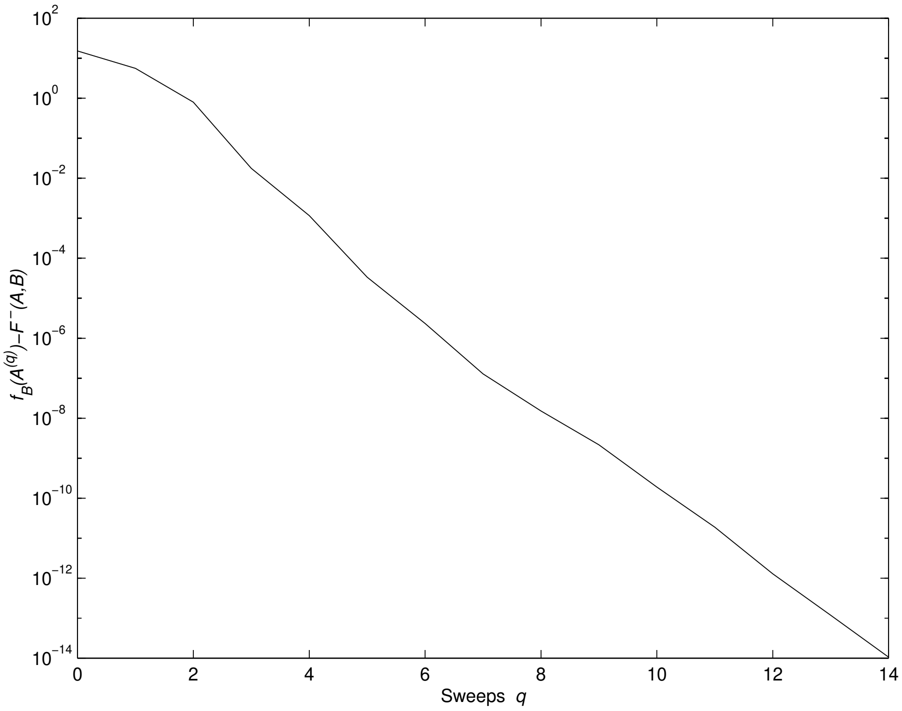

Abbildung 5.11: Konvergenz der jacobiartigen Verfahren Algorithmus 3.5 und Algorithmus 3.7 in Beispiel 5.6

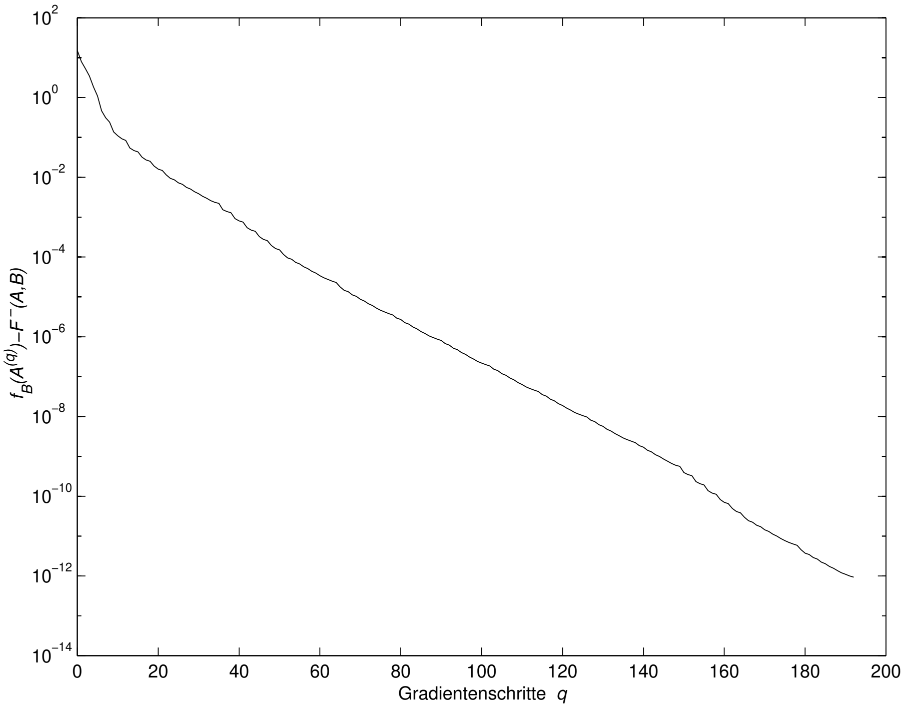

Abbildung 5.12: Konvergenz des Gradientenverfahrens Algorithmus 4.2 in Beispiel 5.6

5.2 Beispiele zur Berechnung einer nächstgelegenen normalen Matrix

Sei A eine komplexe Matrix. Wir haben in den Abschnitten 3.2 und 4.2 die jacobiartigen Verfahren Algorithmus 3.9 und Algorithmus 3.11 und das Gradientenverfahren Algorithmus 4.3 zur Berechnung einer zu A unitär ähnlichen ΔH-Matrix V^*AV derart, daß die Matrix $V\Delta(V^*AV)V^*$ eine zu A nächstgelegene normale Matrix ist, vorgestellt. Zunächst untersuchen wir, wie die Verfahren für normale A konvergieren. Da der Abstand der Matrix V^*AV zu ihrer Diagonale gleich dem Abstand von A zu einer nächstgelegenen normalen Matrix ist, ist in diesem Fall V^*AV eine Diagonalmatrix, deren Einträge die Eigenwerte von A sind, und wir erwarten, daß die Verfahren A diagonalisieren.

Beispiel 5.7
Sei

$$A = e^{Y_1} \operatorname{diag}(-6+5i, -4-4i, 9, 7-2i, 8+3i, 9i, -7)e^{-Y_1}$$

eine normale Matrix mit paarweise verschiedenen Eigenwerten. Dann konvergieren die jacobiartigen Algorithmen 3.9 und 3.11 quadratisch gegen eine Diagonalmatrix. Das Gradientenverfahren Algorithmus 4.3 konvergiert ebenfalls gegen eine Diagonalmatrix, benötigt aber zum einen wesentlich mehr Schritte als die jacobiartigen Verfahren und ist zum anderen ziemlich ungenau, da die Matrix, bei der das Verfahren endet, von einer Diagonalmatrix um einige Größenordnungen weiter entfernt ist als die Abbruchschranke $\epsilon = 10^{-12}$. Die Abbildungen 5.13 und 5.14 zeigen die Konvergenzkurven der Verfahren.

Beispiel 5.8
Sei

$$A = e^{Y_1} \operatorname{diag}(-6+5i, -4-4i, 9, 7-2i, -6+5i, -4-4i, 9)e^{-Y_1}$$

eine normale Matrix mit mehrfachen Eigenwerten. Im Vergleich zum vorigen Beispiel benötigen die jacobiartigen Verfahren zwar kaum mehr Schritte, sind aber nicht mehr quadratisch konvergent. Das Gradientenverfahren konvergiert ungefähr genauso schnell wie im vorigen Beispiel und ist fast genauso ungenau. Siehe dazu die Abbildungen 5.15 und 5.16.

Wir stellen fest: Für normale Matrizen A konvergiert der jacobiartige Algorithmus 3.11 gegen eine zu A ähnliche Diagonalmatrix; sind die Eigenwerte von A einfach, so ist die Konvergenz quadratisch. Der jacobiartige Algorithmus 3.9 und das Gradientenverfahren Algorithmus 4.3 konvergieren dagegen nicht unbedingt gegen eine Diagonalmatrix, für die hermitesche Matrix

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

etwa stagnieren sie, da für alle $U \in \mathcal{U}(2)$

$$g_{\Delta(A)}(UAU^*) = \operatorname{re}\{\operatorname{tr}(UAU^* 2I_2)\} = 2\operatorname{re}\{\operatorname{tr}(A)\} = 8$$

ist, also $g_{\Delta(A)}$ auf $\Lambda(A)$ konstant ist (vergleiche Beispiel 3.14). Konvergiert Algorithmus 3.9 gegen eine Diagonalmatrix, so benötigt er ungefähr genausoviele Schritte wie Algorithmus 3.11, insbesondere ist er quadratisch konvergent, wenn A nur einfache Eigenwerte besitzt. Das Gradientenverfahren konvergiert dagegen erheblich langsamer als die jacobiartigen Verfahren und ist zudem noch wesentlich ungenauer.

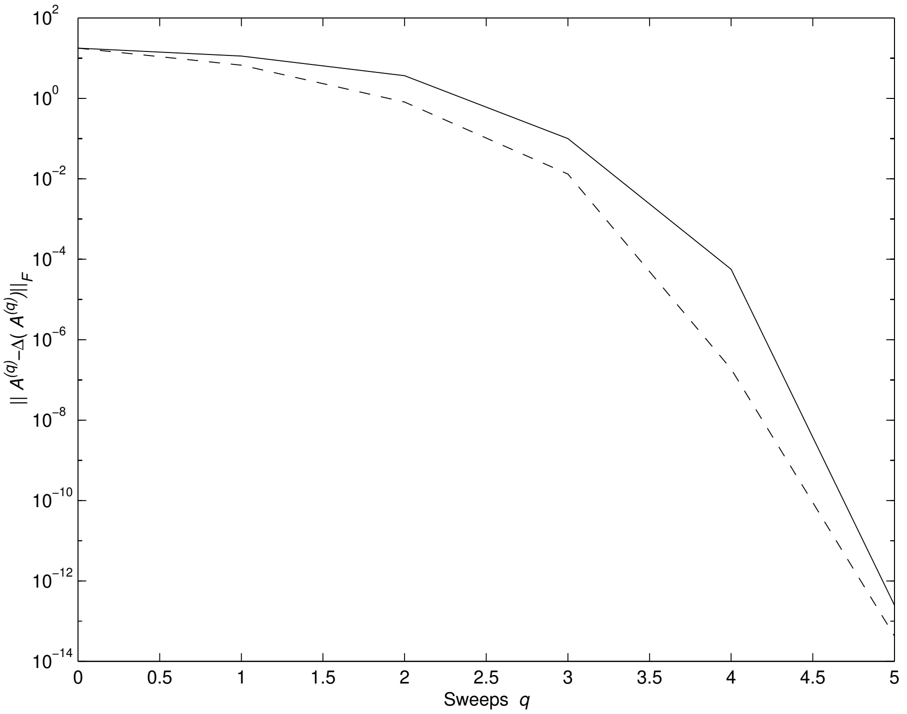

Abbildung 5.13: Konvergenz der jacobiartigen Verfahren Algorithmus 3.9 (durchgezogene Linie) und Algorithmus 3.11 (gestrichelte Linie) in Beispiel 5.7

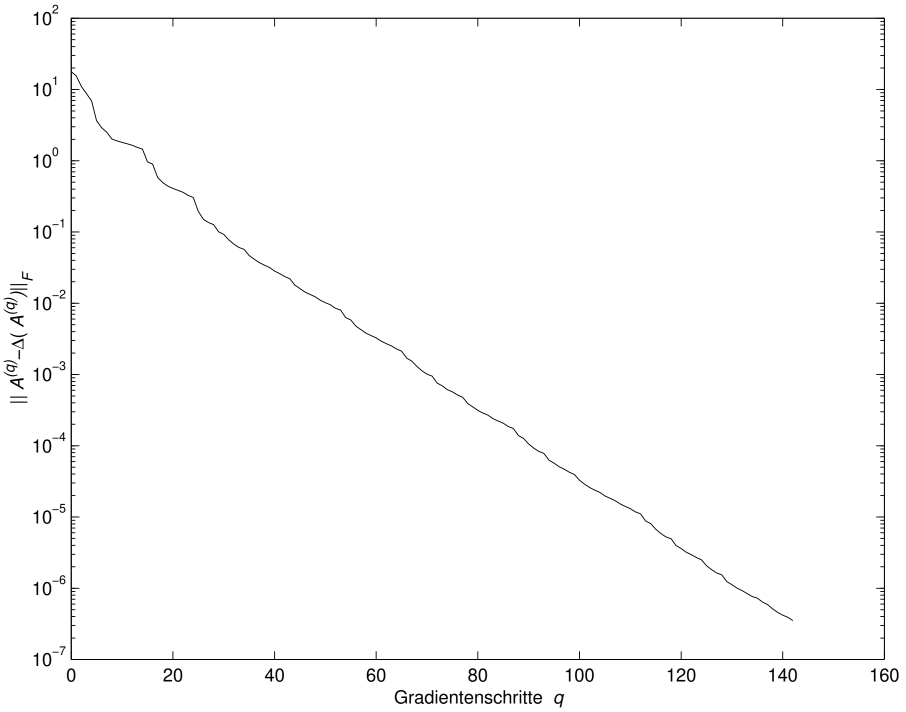

Abbildung 5.14: Konvergenz des Gradientenverfahrens Algorithmus 4.3 in Beispiel 5.7

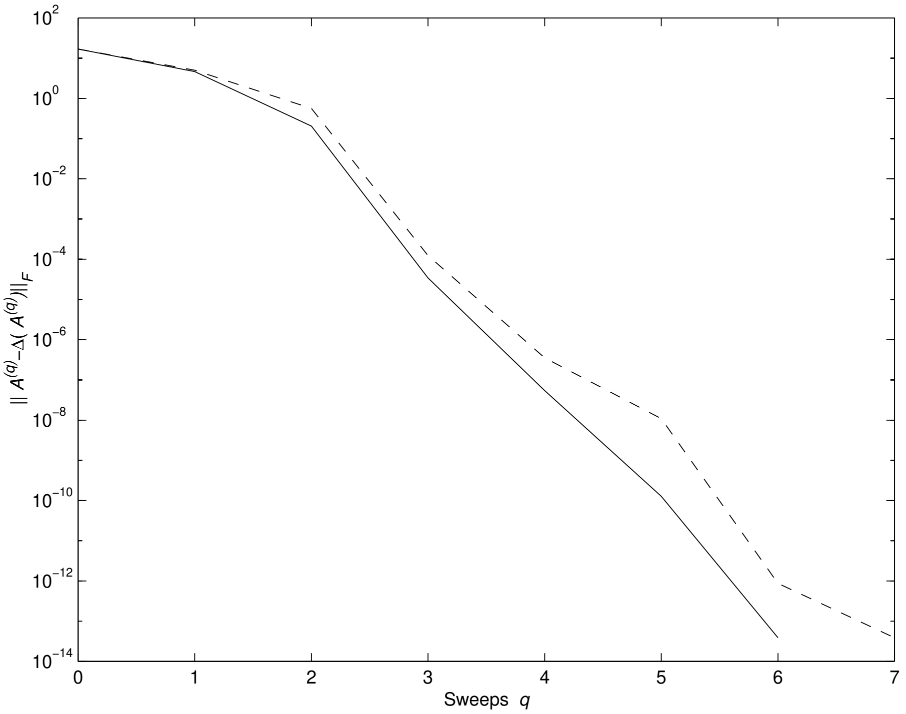

Abbildung 5.15: Konvergenz der jacobiartigen Verfahren Algorithmus 3.9 (durchgezogene Linie) und Algorithmus 3.11 (gestrichelte Linie) in Beispiel 5.8

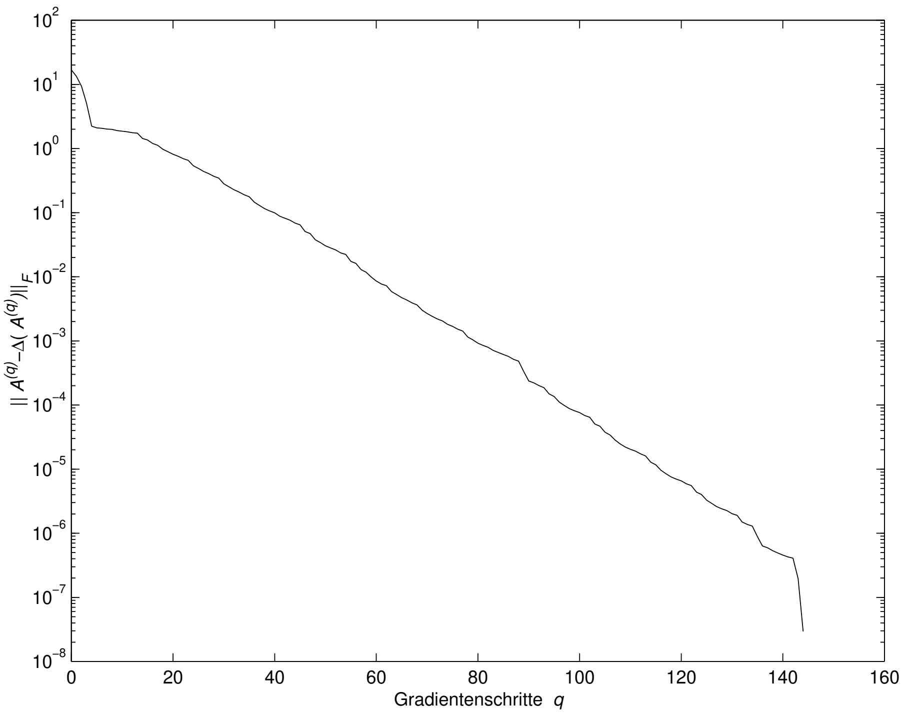

Abbildung 5.16: Konvergenz des Gradientenverfahrens Algorithmus 4.3 in Beispiel 5.8

Wie sieht es mit der Konvergenz der Verfahren aus, wenn A nichtnormal ist? In diesem Fall können wir den Abstand von A zu einer nächstgelegenen normalen Matrix im allgemeinen nicht analytisch bestimmen. Daher messen wir zur Konvergenzanalyse der Verfahren den Abstand der Norm $\|A^{(q)} - \Delta\big(A^{(q)}\big)\|_F$ der Nichtdiagonalteile der Iterierten $A^{(q)}$ zu

$$\xi := \min\big\{\big\|A_{\text{next}}^{(j)} - \Delta\big(A_{\text{next}}^{(j)}\big)\big\|_F : j \in \{1,2,3\}\big\} - \epsilon,$$

wobei $A_{\text{next}}^{(j)}$ die jeweils letzten Iterierten der drei Verfahren sind. Dem liegt die Annahme zugrunde, daß wenigstens eines der Verfahren gegen eine ΔH-Matrix, deren Abstand zu ihrer Diagonale gleich dem Abstand von A zu einer nächstgelegenen normalen Matrix ist, konvergiert, und die letzte Iterierte dieses Verfahrens noch einen Fehler von ϵ aufweist.

Beispiel 5.9
Sei

$$A = e^{Y_1}\,\mathrm{diag}(x,8,1,-7,5,3,4)e^{-Y_1} + e^{Y_2}\,\mathrm{diag}(i,3i,9i,-5i,-7i,-9i,2i)e^{-Y_2}.$$

Fall 1: $x = -2$
In diesem Fall konvergieren alle Verfahren gegen ΔH-Matrizen, deren Diagonaleinträge bis auf Permutationen übereinstimmen. Algorithmus 3.11 konvergiert von den Verfahren am schnellsten, ist allerdings erheblich langsamer als in den vorigen Beispielen mit normalen Matrizen. Algorithmus 3.9 benötigt ein paar Schritte mehr als Algorithmus 3.11, ist aber immer noch um einiges schneller als das Gradientenverfahren Algorithmus 4.3, wie die Abbildungen 5.17 und 5.18 zeigen.

Fall 2: $x = -1.655$
Auch in diesem Fall konvergieren alle Verfahren gegen ΔH-Matrizen mit (bis auf Permutationen) gleichen Diagonaleinträgen. Während Algorithmus 3.9 und Algorithmus 4.3 ungefähr gleich schnell konvergieren wie im vorigen Fall, durchläuft Algorithmus 3.11 eine lange Stagnationsphase und ist deshalb langsamer als Algorithmus 3.9, aber immer noch schneller als das Gradientenverfahren. Die Abbildungen 5.19 und 5.20 zeigen die Konvergenzkurven der Verfahren.

Fall 3: $x = -1$
In diesem Fall konvergieren die Algorithmen 3.9 und 4.3 wie in den Fällen 1 und 2 gegen ΔH-Matrizen, deren Diagonaleinträge sich nur durch Permutationen unterscheiden. Algorithmus 3.11 hingegen konvergiert gegen eine ΔH-Matrix mit anderen Diagonaleinträgen. Diese hat einen größeren Abstand zu ihrer Diagonale als die ΔH-Matrizen, gegen die die anderen Verfahren konvergieren. Also ist insbesondere der Abstand dieser ΔH-Matrix zu ihrer Diagonale größer als der Abstand von A zu einer nächstgelegenen normalen Matrix. Siehe dazu die Abbildungen 5.21 und 5.22.

Wie wir in Beispiel 3.14 gesehen haben, kann der jacobiartige Algorithmus 3.9 enden, ohne eine ΔH-Matrix erreicht zu haben. Auch das Gradientenverfahren konvergiert in diesem Beispiel nicht gegen eine ΔH-Matrix. Selbst wenn diese beiden Verfahren gegen eine zu A ähnliche ΔH-Matrix U^*AU konvergieren, so kann es sein, daß deren rücktransformierter Diagonalteil $U\Delta(U^*AU)U^*$ keine nächstgelegene normale Matrix zu A ist, so etwa in Beispiel 3.13. Im Gegensatz dazu konvergiert Algorithmus 3.11 stets gegen eine ΔH-Matrix. Wie Fall 3 im letzten Beispiel zeigt, kann es aber auch hier passieren, daß deren rücktransformierter Diagonalteil keine nächstgelegene normale Matrix zu A ist.

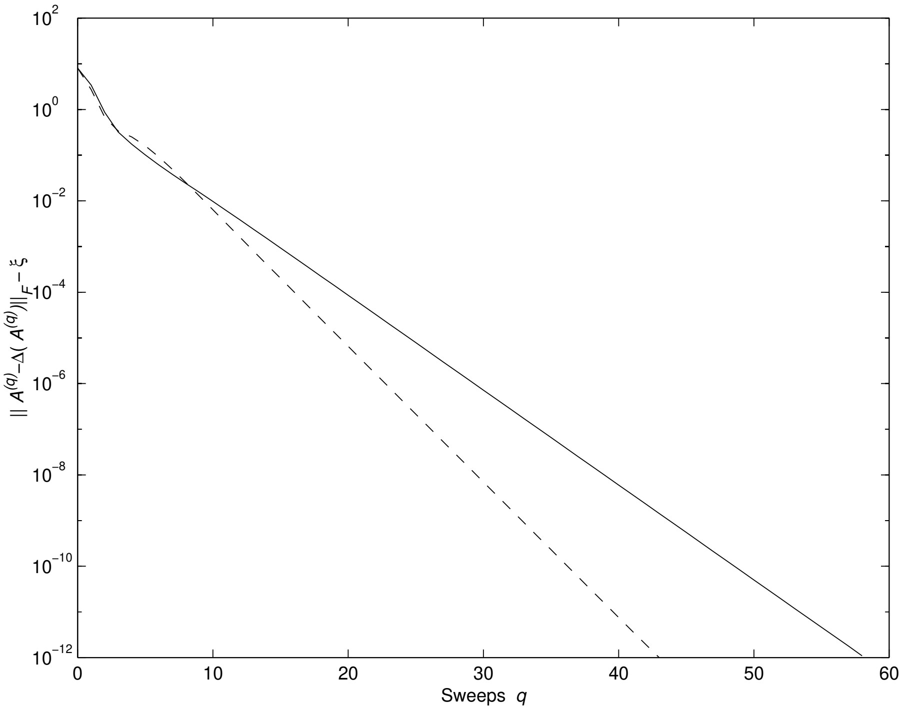

Abbildung 5.17: Konvergenz der jacobiartigen Verfahren Algorithmus 3.9 (durchgezogene Linie) und Algorithmus 3.11 (gestrichelte Linie) in Beispiel 5.9, Fall 1

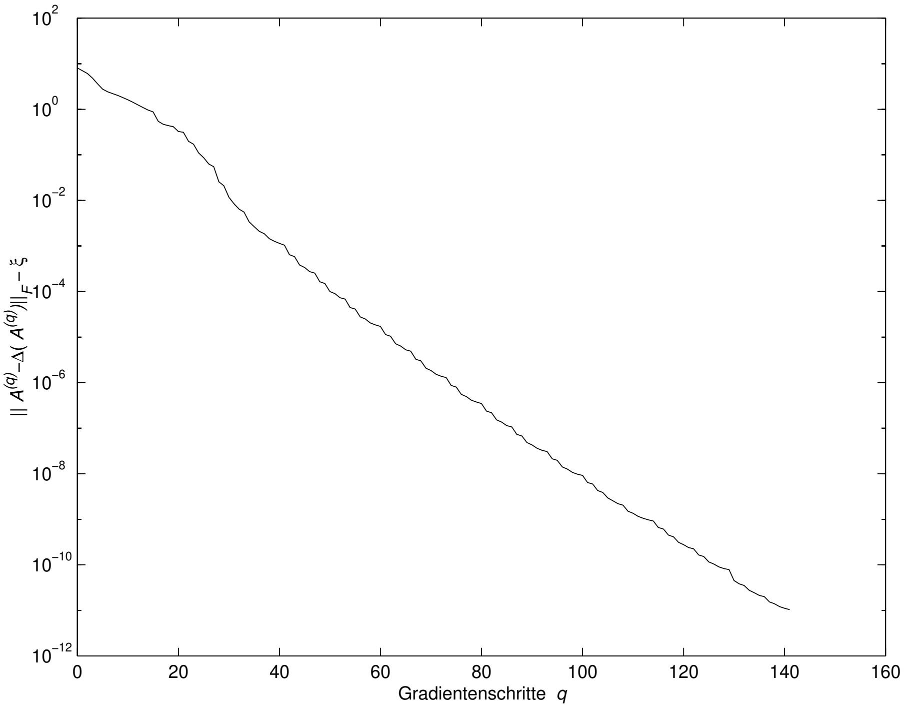

Abbildung 5.18: Konvergenz des Gradientenverfahrens Algorithmus 4.3 in Beispiel 5.9, Fall 1

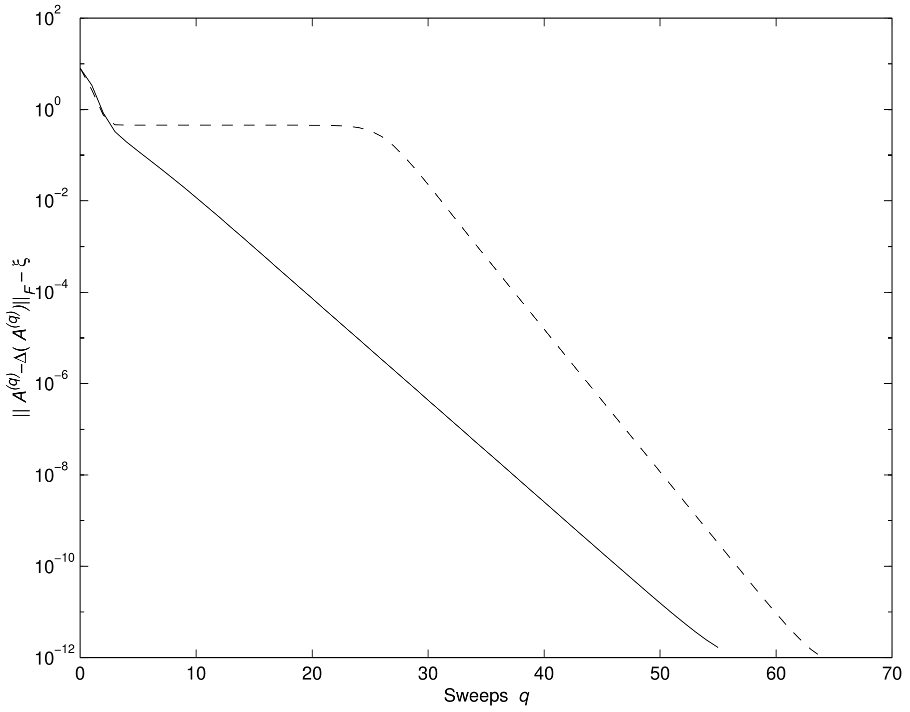

Abbildung 5.19: Konvergenz der jacobiartigen Verfahren Algorithmus 3.9 (durchgezogene Linie) und Algorithmus 3.11 (gestrichelte Linie) in Beispiel 5.9, Fall 2

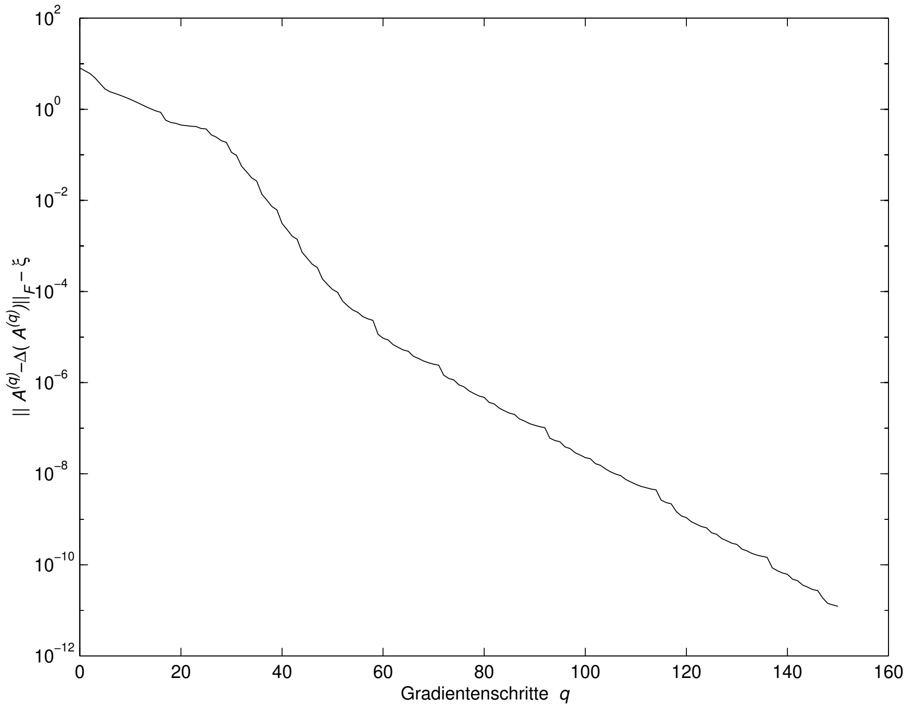

Abbildung 5.20: Konvergenz des Gradientenverfahrens Algorithmus 4.3 in Beispiel 5.9, Fall 2

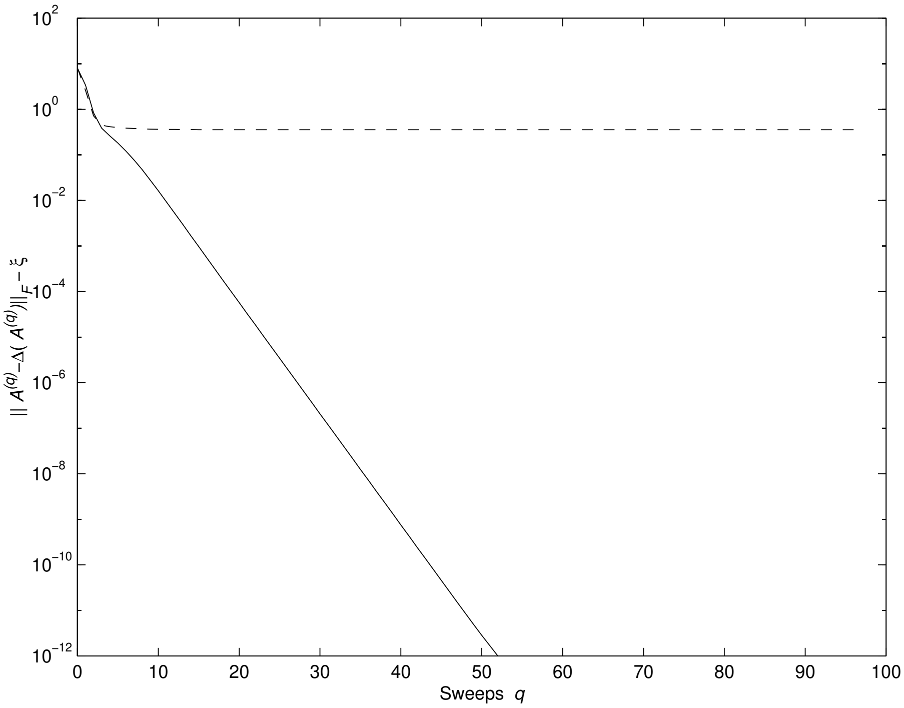

Abbildung 5.21: Konvergenz der jacobiartigen Verfahren Algorithmus 3.9 (durchgezogene Linie) und Algorithmus 3.11 (gestrichelte Linie) in Beispiel 5.9, Fall 3

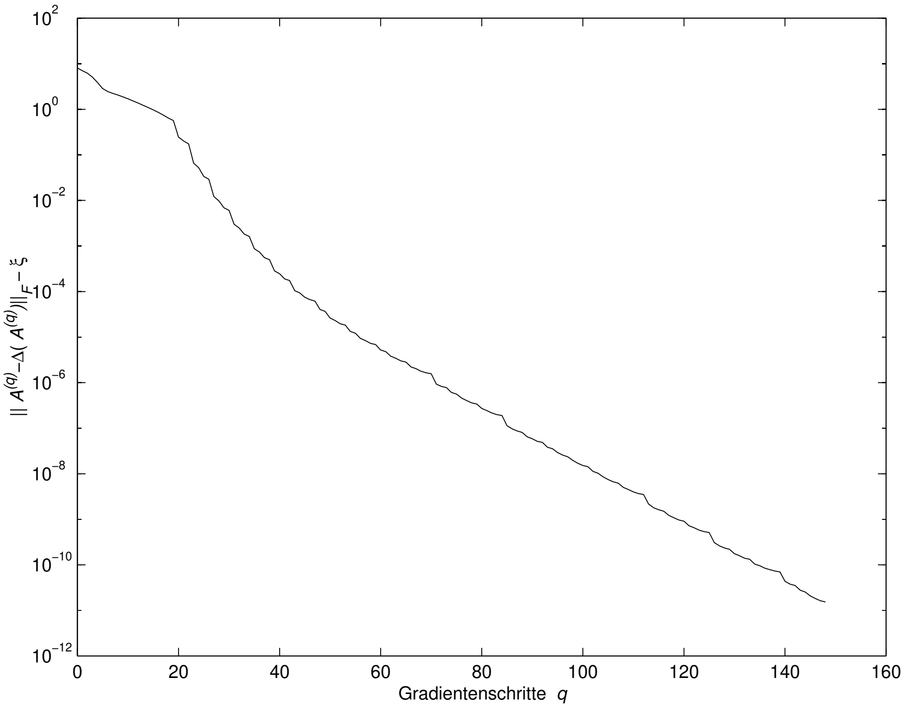

Abbildung 5.22: Konvergenz des Gradientenverfahrens Algorithmus 4.3 in Beispiel 5.9, Fall 3

In der überwiegenden Zahl der Fälle konvergieren allerdings alle drei Verfahren für nichtnormale Matrizen gegen ΔH-Matrizen, deren Diagonaleinträge bis auf Permutationen übereinstimmen. Der jacobiartige Algorithmus 3.11 ist dabei fast immer etwas schneller als der jacobiartige Algorithmus 3.9, während das Gradientenverfahren Algorithmus 4.3 um einiges langsamer als die jacobiartigen Verfahren ist und sich daher nicht zur Berechnung einer nächstgelegenen normalen Matrix anbietet.

Literaturverzeichnis

[1] Rajendra Bhatia
persönliche Mitteilung

[2] Richard P. Brent, Franklin T. Luk
The solution of singular-value and symmetric eigenvalue problems on multiprocessor arrays
SIAM Journal on Scientific and Statistical Computing 6 (1985), no. 1, 69–84

[3] Robert L. Causey
On closest normal matrices
Technical Report CS-10, Department of Computer Science, Stanford, 1964

[4] James Demmel, Krešimir Veselić
Jacobi's method is more accurate than QR
SIAM Journal on Matrix Analysis and Applications 13 (1992), no. 4, 1204–1245

[5] Patricia J. Eberlein
A Jacobi-like method for the automatic computation of eigenvalues and eigenvectors of an arbitrary matrix
Journal of the Society for Industrial and Applied Mathematics 10 (1962), 74–88

[6] Ludwig Elsner
A note on the Hoffman-Wielandt theorem
Linear Algebra and its Applications 182 (1993), 235–237

[7] Ludwig Elsner, M. H. C. Paardekooper
On measures of nonnormality of matrices
Linear Algebra and its Applications 92 (1987), 107–123

[8] Richard Gabriel
Zur besten normalen Approximation komplexer Matrizen in der Euklidischen Norm
Mathematische Zeitschrift 200 (1989), no. 4, 591–600

[9] Herman H. Goldstine, L. P. Horwitz
A procedure for the diagonalization of normal matrices
Journal of the Association for Computing Machinery 6 (1959), 176–195

[10] S. J. Glaser, T. Schulte-Herbrüggen, M. Sieveking, O. Schedletzky, N. C. Nielsen, O. W. Sørensen, C. Griesinger
Unitary Control in Quantum Ensembles: Maximizing Signal Intensity in Coherent Spectroscopy
Science 280 (1998), 421–424

[11] Gene H. Golub, Charles F. Van Loan
Matrix computations
Johns Hopkins University Press, Baltimore, Maryland, 1983

[12] Jürgen Götze
On the parallel implementation of Jacobi and Kogbetliantz algorithms
SIAM Journal on Scientific Computing 15 (1994), no. 6, 1331–1348

[13] Robert Grone, Charles R. Johnson, Eduardo M. Sa, Henry Wolkowicz
Normal matrices
Linear Algebra and its Applications 87 (1987), 213–225

[14] Vjeran Hari
On sharp quadratic convergence bounds for the serial Jacobi methods
Numerische Mathematik 60 (1991), no. 3, 375–406

[15] Uwe Helmke, Knut Hüper
Structure and convergence of Jacobi-type methods
Technical Report TUM-LNS-TR-95-2, Technische Universität München, 1995

[16] Uwe Helmke, Knut Hüper, John B. Moore
Structure and convergence of conventional Jacobi-type methods minimizing the off-norm function
Proceedings of the 35th IEEE International Conference on Decision and Control, Kobe, Japan, 1996, 3688–3689

[17] Peter Henrici
On the speed of convergence of cyclic and quasicyclic Jacobi methods for computing eigenvalues of Hermitian matrices
Journal of the Society for Industrial and Applied Mathematics 6 (1958), 144–162

[18] Roger A. Horn, Charles R. Johnson
Matrix analysis
Cambridge University Press, Cambridge-New York, 1985

[19] Carl Gustav Jacob Jacobi
Über ein leichtes Verfahren die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen
Journal für die reine und angewandte Mathematik 30 (1846), 51–94

[20] Roy Mathias
Accurate eigensystem computations by Jacobi methods
SIAM Journal on Matrix Analysis and Applications 16 (1995), no. 3, 977–1003

[21] Axel Ruhe
Closest normal matrix finally found!
BIT 27 (1987), no. 4, 585–598

[22] Arnold Schönhage
Zur Konvergenz des Jacobi-Verfahrens
Numerische Mathematik 3 (1961), 374–380

[23] Arnold Schönhage
Zur quadratischen Konvergenz des Jacobi-Verfahrens
Numerische Mathematik 6 (1964), 410–412

[24] Hubert Schwetlick
Numerische Lösung nichtlinearer Gleichungen
VEB Deutscher Verlag der Wissenschaften, Berlin, 1979

[25] G. W. Stewart
A Jacobi-like algorithm for computing the Schur decomposition of a non-Hermitian matrix
SIAM Journal on Scientific and Statistical Computing 6 (1985), no. 4, 853–864

[26] H. P. M. van Kempen
On the quadratic convergence of the special cyclic Jacobi method
Numerische Mathematik 9 (1966), 19–22

[27] James H. Wilkinson
Note on the quadratic convergence of the cyclic Jacobi process
Numerische Mathematik 4 (1962), 296–300