Inaugural-Dissertation

zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich-Mathematischen
Gesamtfakultät
der
Ruprecht-Karls-Universität
Heidelberg

vorgelegt von
Diplom-Mathematiker Dmitry Logashenko
aus Kaliningrad (Russische Föderation)

Tag der mündlichen Prüfung: 8. Juni 2004

Verallgemeinerte filternde IBLU-Zerlegungen der Ordnung l

Gutachter: Prof. Dr. Gabriel Wittum
Prof. Dr. Peter Bastian

Danksagung

Die vorliegende Arbeit entstand in der Arbeitsgruppe von Prof. Dr. Gabriel Wittum an der Universität Heidelberg. In ihr habe ich die Ideen von Wittum und Buzdin zur Frequenzfilterung bei der Lösung großer schwachbesetzter Gleichungssysteme weiterentwickelt.

Für die Betreuung meiner Arbeit möchte ich mich vor allem bei Dr. Alexej Buzdin von der Universität Kaliningrad und bei Prof. Dr. Gabriel Wittum herzlich bedanken. Herrn Buzdin bin ich für die Problemstellung, die Einführung in die Theorie der frequenzfilternden Zerlegungen sowie viele wertvolle Diskussionen und sein Interesse an meiner Arbeit dankbar. Herrn Wittum danke ich für sein Interesse am Fortgang meiner Arbeit und die freundliche Aufnahme in seine Arbeitsgruppe, in der ich beste Voraussetzungen für mein wissenschaftliches Arbeiten fand und von deren Erfahrung im Bereich Wissenschaftliches Rechnen ich viel profitiert habe. Des Weiteren möchte ich mich bei PD. Dr. Klaus Neymeyr für seine hilfreichen Ratschläge zu Kapitel 5 dieser Arbeit bedanken.

Mein besonderer Dank gilt Dr. Torsten Fischer, ohne dessen sprachliche und inhaltliche Korrekturen sämtlicher Versionen diese Arbeit kaum möglich gewesen wäre.

PD. Dr. Christian Wagner danke ich für seine große Hilfe zu Beginn meiner Arbeit in Heidelberg, insbesondere bei wissenschaftlichen Problemen und bei der Benutzung des Software-Pakets UG.

Schließlich möchte ich mich noch herzlich bei allen meinen Kolleginnen und Kollegen aus den Jahren 1999 bis 2003 für die angenehme Zusammenarbeit und die Freundschaft bedanken.

Inhaltsverzeichnis

Zusammenfassung

Die vorliegende Arbeit befasst sich mit der Konstruktion frequenzfilternder IBLU-Zerlegungen für den Fall allgemeiner Gebiete. Die frequenzfilternden IBLU-Zerlegungen werden als Löser für die aus der Diskretisierung partieller Differentialgleichungen entstehenden großen schwachbesetzten linearen Gleichungssysteme benutzt. Die Vorgehensweise von Wittum (siehe [47]), Wagner ([41]) und Buzdin ([9], [11]) zur Konstruktion und theoretischen Untersuchung dieser Methoden wird in dieser Arbeit für die allgemeineren Fälle adaptiert und weiterentwickelt. Dabei führen wir eine neue Klasse solcher Verfahren ein, deren Grundidee die Approximation der Diagonalblöcke der vollständigen Blockzerlegung mit Kettenbrüchen ist.

Der Schwerpunkt dieser Arbeit liegt auf der analytischen Untersuchung der vorgestellten Klasse von Methoden. Wir betrachten im Detail die Existenz und die Konvergenzeigenschaften dieser Zerlegungen. Insbesondere bestimmen wir die Ordnung der Konvergenz für die optimale Wahl der Parameter. Wir bestätigen unsere theoretischen Ergebnisse experimentell. Als eine weitere Anwendung betrachten wir die Folgen dieser frequenzfilternden Zerlegungen als Vorkonditionierer für das Gradientenverfahren bei Eigenwertproblemen.

Die Arbeit ist wie folgt strukturiert. In Kapitel 1 beschreiben wir die allgemeine Problemstellung für Lösungsverfahren zu großen schwachbesetzten linearen Gleichungssystemen. Wir erläutern dabei die Diskretisierungsverfahren, bei denen solche Systeme entstehen und beschreiben kurz die wichtigsten Löser.

In Kapitel 2 geben wir eine Einführung in die frequenzfilternden IBLU-Zerlegungen. Wir definieren IBLU-Zerlegungen sowie den Begriff der Frequenzfilterung und beschreiben die Arbeiten von Wittum ([47]) und Wagner ([41]), welche wir weiter verallgemeinern.

In Kapitel 3 stellen wir die neue Klasse von frequenzfilternden Zerlegungen vor. Wir geben dabei eine allgemeine Definition und untersuchen die Existenz sowie die Konvergenzeigenschaften der Zerlegungen im Fall eines Modellproblems. Allgemeinere Probleme betrachten wir in Kapitel 4, in dem wir auch auf die bei der Implementierung unserer Verfahren entstehenden algorithmischen Probleme eingehen. Zudem stellen wir in den Kapiteln 3 und 4 auch unsere numerische Ergebnisse zu den eingeführten Zerlegungen, angewendet auf die Diskretisierungen von elliptischen partiellen Differentialgleichungen vor.

Kapitel 5 ist den Eigenwertproblemen und der Anwendung von Folgen von den in dieser Arbeit eingeführten Zerlegungen als Vorkonditionierer bei der Berechnung von Eigenvektoren gewidmet. Nach einer kurzen Einführung beweisen wir die Konvergenz des Block-Gradientenverfahren für die betrachtete Problemklasse. Die

quantitativen Konvergenzraten werden dann experimentell untersucht.

Für die vorgestellten numerischen Experimente haben wir die in dieser Arbeit eingeführten Zerlegungen mit Hilfe des Simulationssoftware-Pakets UG implementiert. Zu diesem geben wir in Anhang A eine kurze Beschreibung sowie eine Einleitung zur Benutzung der von uns geschriebenen neuen Module.

Abkürzungsverzeichnis

Spezielle Matrizen:

- diag $\{d_k\}$: Diagonalmatrix mit Diagonaleinträgen d_k.
- tridiag $\{a_k, d_k, b_k\}$: Tridiagonalmatrix,

$$\text{tridiag}\,\{a_k, d_k, b_k\} := \begin{pmatrix} d_1 & b_1 & & \\ a_2 & d_2 & b_2 & \\ & \ddots & \ddots & \ddots \end{pmatrix}.$$

Blockmatrizen:

- blockdiag $\{D_k\}$: eine blockdiagonale Matrix mit Diagonalblöcken D_k.
- blocktridiag $\{A_k, D_k, B_k\}$: eine blocktridiagonale Matrix,

$$\text{blocktridiag}\,\{A_k, D_k, B_k\} := \begin{pmatrix} D_1 & B_1 & & \\ A_2 & D_2 & B_2 & \\ & \ddots & \ddots & \ddots \end{pmatrix}.$$

Blockvektoren:

- blockvector $\{u_k\}$: ein Spaltenvektor aus den Blöcken u_k.

Lateinische Buchstaben:

$\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{C}$	Differentialoperatoren und Randbedingungen der kontinuierlichen Probleme, siehe (1.1.1), (5.1.1)
$\mathbf{A}$	Systemmatrix, Steifigkeitsmatrix
$\mathbf{B}$	Massenmatrix eines Eigenwertproblems, siehe Abschnitt 5.1
$\mathbb{C}$	Menge der komplexen Zahlen
D_k, L_k, U_k	Blöcke von $\mathbf{A}$, siehe (2.1.2)
D, L	Blöcke von $\mathbf{A}$ im Fall eines Modellproblems, siehe (2.1.16–2.1.17), (3.1.2–3.1.4)
$\mathbf{e}^{(i)}$, $e_k^{(i)}$	ein Testvektor und seine Blöcke: $\mathbf{e}^{(i)} = \text{blockvector}\,\{e_k^{(i)}\}$, siehe die Abschnitte 2.1 und 4.1
I	Einheitsmatrix
l	Ordnung der GIBLU(l)-Zerlegung, siehe Definition 3.1.2
N	Anzahl der Blockzeilen in der Systemmatrix $\mathbf{A}$
$\mathbb{N}$	Menge der natürlichen Zahlen (d.h. der ganzen Zahlen ≥ 1)
$\mathbb{N}_0$	$:= \mathbb{N} \cup \{0\}$

n, n_k	Anzahl der Knoten in einem (bei konstanter Blockgröße) bzw. dem k-ten Gitterblock
$\mathbb{R}$	Menge der reellen Zahlen
$\mathbb{R}^{m\times n}$	Raum der reellen Matrizen der Größe $m \times n$
$\mathbf{R}$, R_k	die Restmatrix einer IBLU-Zerlegung (siehe (2.1.7)) und ihre Blöcke: $\mathbf{R} = \text{blockdiag}\,\{R_k\}$
$\mathbf{R}^{(l)}$, $R_k^{(l)}$	die Restmatrix der GIBLU(l)-Zerlegung (siehe Definition 3.1.2) und ihre Blöcke: $\mathbf{R}^{(l)} = \text{blockdiag}\left\{R_k^{(l)}\right\}$
$\tilde{T}_k$, $\tilde{\mathbf{T}}$	die Diagonalblöcke einer IBLU-Zerlegung, siehe (2.1.7)
$\tilde{T}_k^{(l)}$, $\tilde{\mathbf{T}}^{(l)}$	die Diagonalblöcke der GIBLU(l)-Zerlegung, $\tilde{\mathbf{T}}^{(l)} = \text{blockdiag}\left\{\tilde{T}_k^{(l)}\right\}$, siehe Definition 3.1.2
$\mathfrak{u}$, $\mathfrak{f}$	die Unbekannte (oder die Eigenfunktion) und die rechte Seite der kontinuierlichen Probleme, siehe (1.1.1), (5.1.1)
$\mathbf{W}$, $\mathbf{W}_k$	ein Vorkonditionierer (siehe Abschnitt 1.2, Kapitel 5), typischerweise der einer IBLU-Zerlegung (siehe Kapitel 2)
$\mathbf{W}^{(l)}$	der GIBLU(l)-Vorkonditionierer, siehe Definition 3.1.2

Griechische Buchstaben:

$\kappa(M)$	Konditionszahl einer Matrix $M > 0$: $\kappa(M) = \dfrac{\lambda_{\max}}{\lambda_{\min}}$, wobei $\lambda_{\max}$ und $\lambda_{\min}$ der größte bzw. kleinste Eigenwert von M ist
$\lambda(u)$	Rayleigh-Quotient: $\lambda(u) = \dfrac{(\mathbf{A}u, u)}{(\mathbf{B}u, u)}$
$\hat{\mu}_i$, $\hat{\mu}_{\text{opt}}$	Parameter einer GIBLU(l)-Zerlegung, siehe Abschnitt 3.1
$\rho(M)$	Spektralradius einer Matrix M
$\sigma(M)$	Spektrum einer Matrix M
$\theta_k^{(l)}$	Koeffizienten einer GIBLU(l)-Zerlegung, siehe Definition 3.1.2
Ω	beschränktes Gebiet in $\mathbb{R}^2$
$\partial\Omega$	Rand von Ω
Ω_h	Gitter (ohne Dirichlet-Rand)
$\Omega_h^{(i)}$	Gitterblock mit Index i

Weitere Notationen:

$\{a_k\}_k$	Folge mit Gliedern a_k
$(a_{ij})_{ij}$	Matrix mit Einträgen a_{ij}
$(a_i)_i$	Spaltenvektor mit Einträgen a_i

Skalarprodukte und Normen: Überall in dieser Arbeit bezeichnen wir mit $(\cdot,\cdot)$ das euklidische Skalarprodukt. Die euklidische Norm wird mit $\|\cdot\|_2$ bezeichnet, $\|\cdot\|$ kann eine andere Norm bezeichnen. Das mit einer (positiv definiten) Matrix M assoziierte Energieskalarprodukt und die entsprechende Energienorm bezeichnen wir mit $(\cdot,\cdot)_M$ und $\|\cdot\|_M$, respektive. Für symmetrische Matrizen A und B bedeutet $A < B$ ($A \leq B$), dass die Matrix $B - A$ positiv (semi-) definit ist.

Kapitel 1

Einleitung

In diesem Kapitel fassen wir kurz die Problemstellung und den aktuellen Stand der Forschung im Bereich der Lösungsverfahren für große lineare schwachbesetzte Gleichungssysteme zusammen. Eine ausführliche Diskussion dieses Themas findet man z.B. in [18], [16], [24]. In dieser Arbeit verwenden wir die Terminologie und die Notation aus [18]. Dabei setzen wir die grundlegenden Begriffe und Aussagen der linearen Algebra als dem Leser bekannt voraus. Für einen Überblick über diese Grundlagen verweisen wir auf [18], [37] und [15].

1.1 Schwachbesetzte lineare Gleichungssysteme

Die in dieser Arbeit betrachteten Verfahren sind ursprünglich als Mittel zur numerischen Behandlung von Randwertproblemen

$$\begin{aligned} &\mathfrak{A}\mathfrak{u} = \mathfrak{f}, \\ &\mathfrak{C}(\mathfrak{u}|_{\partial\Omega}) = 0 \end{aligned} \tag{1.1.1}$$

auf einem beschränkten Gebiet $\Omega \subset \mathbb{R}^2$ entstanden. Hier ist $\mathfrak{u} : \Omega \to \mathbb{R}$ die unbekannte Funktion, $\mathfrak{f} : \Omega \to \mathbb{R}$ eine vorgegebene rechte Seite, $\mathfrak{C}$ eine Randbedingung und $\mathfrak{A}$ ein Differentialoperator der Form

$$\mathfrak{A}\mathfrak{u} = \nabla \cdot (-\mathbf{P}\nabla\mathfrak{u} + \mathbf{v}\mathfrak{u}) + c\mathfrak{u}. \tag{1.1.2}$$

Der symmetrische, positiv definite 2×2-Tensor $\mathbf{P}$, der Vektor $\mathbf{v}$ und das Skalar c können Konstanten oder Funktionen von $\mathbf{x} = (x, y) \in \Omega$ und $\mathfrak{u}$ sein. Die Probleme (1.1.1–1.1.2) treten bei vielen mathematischen Modellen auf, z.B. bei solchen von stationären Wärme- und Stofftransportprozessen, Strömungen in porösen Medien, Verteilung des elektrischen Potentials u.s.w. (siehe z.B. [34], [32], [29], [54], [51]). In dieser Arbeit betrachten wir hauptsächlich den elliptischen Fall, bei dem $\mathbf{v}$ gleich Null ist und $c \geq 0$.

Die analytische Lösung von (1.1.1–1.1.2) ist nur in einfachsten Fällen möglich (siehe z.B. [29], [54]). Für praktische Anwendungen spielt die numerische Behandlung die wichtigste Rolle. Sie kann in zwei Schritte aufgeteilt werden: Zunächst werden die Gleichungen (1.1.1) diskretisiert. Dadurch erhält man ein System von

algebraischen Gleichungen, welches man löst. Die Lösung beschreibt eine Näherung zur exakten Lösung von (1.1.1).

Wir stellen nun die gängigsten Diskretisierungsmethoden sowie die Problematik des Lösungsprozesses von den jeweils daraus entstehenden algebraischen Gleichungen vor. Eine detailliertere Betrachtung findet man in [16], [6] und [24].

Die einfachste Vorgehensweise zur Diskretisierung ist das Finite-Differenzen-Verfahren. Man betrachtet dabei eine endliche Menge $\overline{\Omega}_h \subset \Omega \cup \partial\Omega$ und ersetzt in ihren Punkten (außer denen auf dem Dirichlet-Rand) die partiellen Ableitungen durch die entsprechenden finiten Differenzen. Damit ergibt sich eine algebraische Gleichung in jedem Punkt $x \in \overline{\Omega}_h$, der nicht auf dem Dirichlet-Rand liegt. Diese Gleichungen bilden ein System, dessen Lösung die exakte Lösung von (1.1.1) auf $\overline{\Omega}_h$ approximiert. Die Menge $\overline{\Omega}_h$ nennt man ein *„Gitter“*. Bei Finite-Differenzen-Verfahren haben die Gitter typischerweise eine reguläre Struktur. Beispielsweise kann $\overline{\Omega}_h$ die Menge der Schnittpunkte von zwei zueinander senkrechten parallelen Geradenscharen (und dieser mit $\partial\Omega$) sein. Solche Gitter werden *strukturiert* genannt. Diese Diskretisierungsmethode ist aber nur mit erheblichem Mehraufwand auf komplizierte Gebiete übertragbar.

Eine andere Vorgehensweise ist das Finite-Elemente-Verfahren, bei dem das Gebiet Ω in elementare geometrischen Elemente (z.B. Dreiecke und Vierecke) zerlegt (*trianguliert*) wird. Diese Zerlegung wird ebenfalls als Gitter bezeichnet. Auf dem so dargestellten Gebiet betrachtet man den Raum von innerhalb jedes Elements linearen (im Fall eines Dreiecks) oder biliniaren (im Fall eines Vierecks) reellen Funktionen, die am Dirichlet-Rand den Randbedingungen genügen. In diesem Raum wird dann eine Funktion gewählt, für welche die Differenz zur exakten Lösung bezüglich der vom Operator (1.1.2) induzierten Energienorm minimal ist. Diese Bedingung führt auch zu einem algebraischen Gleichungssystem. Die Variablen sind hier die Knotenwerte der zu bestimmenden Funktion. Wir haben hier nur kurz einen einfachen Spezialfall von Finite-Elemente-Verfahren erläutert. Für eine ausführlichere Beschreibung verweisen wir z.B. auf [16], [24].

Eine weitere Klasse von Diskretisierungen bilden die Finite-Volumen-Verfahren oder Boxmethoden. Das Gebiet wird bei diesen auch in Elemente zerlegt, die aber eine kompliziertere Form als bei Finite-Elemente-Verfahren haben können. Typischerweise nimmt man das so genannte *„duale Gitter“* des für Finite-Elemente-Methode eingeführten Gitters (siehe [16], [6]). Gleichung (1.1.1) wird auf jedem einzelnen Element als Erhaltungsgesetz formuliert und mit der Hilfe des Gaußschen Satzes in eine Bedingung an die Werte von $\mathfrak{u}$ auf dessen Rand umgeschrieben. Physikalisch kann dies als die Bilanz von Ein- und Ausfluss für dieses Element interpretiert werden.

Finite-Elemente- und Finite-Volumen-Verfahren sind für die Anwendung auf komplizierte Gebietsgeometrien wesentlich besser geeignet als finite Differenzen. Insbesondere haben sie einen Vorteil bei Benutzung von adaptiven Methoden (siehe [38]). Daher sind sie bei heutigen Simulationen sehr verbreitet.

Der maximale Abstand zwischen zwei benachbarten Knoten bei Finite-Differenzen-Verfahren oder der maximale Durchmesser der geometrischen Elemente bei Finite-Elemente- und Finite-Volumen-Verfahren wird als *Gitterlänge* h bezeichnet. Diese Größe beschreibt die Genauigkeit der Diskretisierung: Wenn h gegen 0

geht, konvergiert die Lösung des algebraischen Systems gegen die Lösung des ursprünglichen Problems (1.1.1) (siehe [24], [16]). Je kleiner allerdings h ist, desto mehr Knoten enthält $\overline{\Omega}_h$ und desto mehrere Gleichungen hat das diskrete System. Damit sind wir zur ersten Eigenschaft von aus Diskretisierungen entstehenden Gleichungssysteme gekommen: *Sie sind sehr groß.* Und ihre Größe wächst mit der geforderten Genauigkeit der diskreten Lösung unbeschränkt. In praktischen Anwendungen enthalten diese Systeme nicht weniger als 1000 Gleichungen, typischerweise wesentlich mehr.

Jede Gleichung enthält aber nur eine beschränkte Anzahl von Variablen (ca. 5–10 bei zweidimensionalen Problemen). Diese spezielle Struktur der Matrix ist die zweite Eigenschaft solcher linearen Systeme: Jede Zeile der Matrix enthält nur eine *kleine und von h unabhängige* Zahl an von Null verschiedenen Einträgen. Solche Matrizen (und die entsprechenden linearen Gleichungssysteme) heißen *schwachbesetzt.* Wenn das diskrete Problem nicht-linear ist, wird es innerhalb der nicht-linearen Lösungsverfahren (z.B. der Newton-Iteration) linearisiert, und die daraus entstehenden Matrizen sind schwachbesetzt. Diese Eigenschaft ist in der lokalen Natur des Differentialoperators (1.1.2) begründet. In Gegensatz dazu sind die bei Diskretisierung von Integralgleichungen entstehenden Systeme vollbesetzt und müssen prinzipiell anders numerisch behandelt werden (siehe [19]).

Bei der Lösung großer schwachbesetzter linearer Systeme sind solche Algorithmen effizient, die die Matrizen als lineare Operatoren betrachten und nur die Matrix-Vektor-Multiplikation sowie die Vektor-Vektor-Operationen benutzen. Vor allem sind hier Iterationsverfahren geeignet. Deren Konstruktion stellen wir im nächsten Abschnitt vor. Bei diesen Verfahren müssen nur die von Null unterschiedliche Matrixeinträge gespeichert und bei Ausführung der Rechenoperationen berücksichtigt werten. Dadurch reduziert sich die Anzahl der arithmetischen Operationen und somit der Rechenaufwand. Der Aufwand solcher Algorithmen wie der Gaußschen Elimination oder des QR-Verfahrens, die mehrere neue von Null verschiedene Einträge produzieren, wächst oft sehr schneller mit der Anzahl der Gitterknoten (siehe z.B. [18], [33]). Daher ist die Effizienz dieser Methoden für feine Gitter gering.

Wir weisen aber darauf hin, dass die aus der Diskretisierung der Probleme (1.1.1–1.1.2) entstehenden Matrizen *schlecht konditioniert* sind: Ihre Konditionszahl (das Verhältnis des größten und des kleinsten Eigenwerts) wächst mit der Anzahl der Gitterknoten (siehe z.B. [18]). Diese Tatsache soll bei der Entwicklung der Iterationsverfahren berücksichtigt werden. Denn die Konditionszahl beeinflusst die Konvergenzrate.

Wir stellen nun ein einfaches Beispiel vor. Dieses findet man in vielen Lehrbüchern, insbesondere in [24], [16], [18]. Wir betrachten einen Spezialfall des Problems (1.1.1–1.1.2),

$$\begin{aligned} -\nabla \cdot (P \nabla \mathfrak{u}) &= \mathfrak{f}, \\ \mathfrak{u}|_{\partial\Omega} &= 1 \end{aligned} \tag{1.1.3}$$

mit

$$P = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}, \qquad a, b > 0, \tag{1.1.4}$$

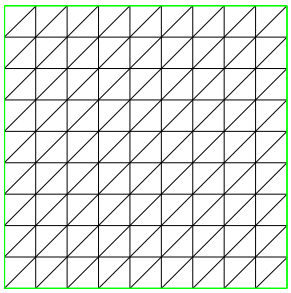

Abbildung 1.1: Die strukturierte Triangulierung des Einheitsquadrat. Die grünen Linien bezeichnen den Dirichlet-Rand.

auf dem Einheitsquadrat $\Omega = (0,1)^2$. Die Wahl

$$a = b = 1$$

führt zur Laplace-Gleichung, und für ungleiche Werte von a und b erhalten wir auf der linken Seite einen anisotropen elliptischen Operator. Zur Diskretisierung führen wir das strukturierte Gitter

$$\overline{\Omega}_h = \left\{ (x_i, y_j) : x_i = \frac{i}{N+1}, \quad y_j = \frac{j}{N+1}, \quad 0 \le i, j \le N+1 \right\} \tag{1.1.5}$$

ein und wenden das Finite-Differenzen-Verfahren an. Die unbekannte Gitterfunktion auf $\overline{\Omega}_h$ bezeichnen wir mit u und ihren Wert im Punkt (x_i, y_j) mit u_{ij}. Die Werte für die Indizes $i = 0$, $i = N+1$, $j = 0$ und $j = N+1$ sind durch die Dirichlet-Randbedingung festgelegt:

$$u_{0,j} = u_{N+1,j} = u_{i,0} = u_{i,N+1} = 1, \quad 0 \le i, j \le N+1. \tag{1.1.6}$$

Für die anderen Knoten ersetzen wir den Laplace-Operator Δ in (1.1.3) durch die finite Differenz und erhalten somit nach Multiplikation mit h^2, wobei $h = \frac{1}{N}$ die Gitterlänge ist, ein lineares System von N^2 Gleichungen:

$$-bu_{i,j-1} - au_{i-1,j} + 2(a+b) \cdot u_{ij} - au_{i+1,j} - bu_{i,j+1} = f_{ij}, \quad 1 \le i, j \le N, \tag{1.1.7}$$

wobei $f_{ij} = h^2 \mathfrak{f}(x_i, y_j)$. Der lineare Operator auf der linken Seite von (1.1.7) wird in [18] mit

$$\begin{bmatrix} & -b & \\ -a & 2(a+b) & -a \\ & -b & \end{bmatrix} \tag{1.1.8}$$

bezeichnet.

Bemerkung 1.1.1 Das Finite-Elemente-Verfahren mit linearen Basisfunktionen und die klassischen Finite-Volumen-Verfahren liefern für das in Abb. 1.1 dargestellte strukturierte Gitter den gleichen diskreten linearen Operator (1.1.8). Die rechten Seiten sind aber unterschiedlich. Sie konvergieren gegen den gleichen Grenzwert für $h \to 0$. (Siehe [16], [24]). □

Die Unbekannten in den Gleichungen (1.1.7) entsprechen den Punkten $\mathbf{x} \in \Omega_h \subset \overline{\Omega}_h$ mit

$$\Omega_h = \left\{(x_i, y_j) \in \overline{\Omega}_h : 1 \le i, j \le N\right\}. \tag{1.1.9}$$

Obwohl diese Gleichungen auch die u_{ij} mit $(x_i, y_j) \in \overline{\Omega}_h \backslash \Omega_h$ enthalten, gehören diese Werte nach (1.1.6) zur rechten Seite des linearen Gleichungssystems. Um das System (1.1.7) in Matrixform zu schreiben, müssen die Punkte in Ω_h geordnet werden. Das kann hier in zwei Schritten erfolgen. Zuerst führen wir die nummerierten Mengen

$$\Omega_h^{(j)} = \{(x_i, y_j) \in \Omega_h : y_j = jh, \quad 1 \le i \le N\}, \qquad 1 \le j \le N. \tag{1.1.10}$$

Dann werden innerhalb jedes *Blocks* $\Omega_h^{(j)}$ die Knoten mit dem Index i nummeriert. Dadurch lässt sich System (1.1.7) in der Form

$$\mathbf{A}u = f \tag{1.1.11}$$

schreiben, wobei

$$\mathbf{A} = \begin{pmatrix} D & -L & & \\ -L & D & \ddots & \\ & \ddots & \ddots & -L \\ & & -L & D \end{pmatrix} =: \text{blocktridiag}\{-L, D, -L\} \tag{1.1.12}$$

($\mathbf{A} \in \mathbb{R}^{N^2 \times N^2}$) mit

$$\begin{aligned} D &= \text{tridiag}\{-a, 2(a+b), -a\}, \\ L &= bI \end{aligned} \tag{1.1.13}$$

($D, L \in \mathbb{R}^{N \times N}$). Hier betrachten wir u als Vektor mit Einträgen u_{ij}. Der Vektor f bildet die rechte Seite des Systems (1.1.7) und enthält also nicht nur die f_{ij}, sondern auch die Randwerte (1.1.6). Nach (1.1.12) enthält $\mathbf{A}$ in jeder Zeile bis zu 5 von Null verschiedene Einträge. Diese befinden sich auf der Hauptdiagonalen und 4 Nebendiagonalen. Deswegen wird $\mathbf{A}$ als *Bandmatrix* bezeichnet (siehe [18], [16]). Die *Bandbreite* ist in diesem Fall $2N + 1$.

Das Spektrum der Matrix $\mathbf{A}$ kann analytisch untersucht werden:

$$\sigma(\mathbf{A}) = \left\{\lambda_{kl} = 4\left(a \cdot \sin^2 \tfrac{k\pi}{2(N+1)} + b \cdot \sin^2 \tfrac{l\pi}{2(N+1)}\right) : 1 \le k, l \le N\right\}$$

(siehe z.B. [18]). Wir erhalten also

$$\begin{aligned} \lambda_{\min} &= \min_{1 \le k,l \le N} \lambda_{kl} = 4(a+b)\sin^2 \tfrac{\pi}{2(N+1)} \\ \lambda_{\max} &= \max_{1 \le k,l \le N} \lambda_{kl} = 4(a+b)\sin^2 \tfrac{\pi N}{2(N+1)} \end{aligned}$$

und daher

$$\kappa(\mathbf{A}) = \frac{\lambda_{\max}}{\lambda_{\min}} = O(N^2).$$

Wie man sieht, wächst diese Konditionszahl sehr schnell, wenn N groß wird. Die Matrix $\mathbf{A}$ ist also bei großen N sehr schlecht konditioniert.

Wir bemerken hier aber ausdrücklich, dass die Matrix $\mathbf{A}$ eine *blocktridiagonale Struktur* hat. Lösungsverfahren, die diese Struktur ausnutzen, heißen *Blocklöser*. Die Menge Ω_h von Knoten, die nicht auf dem Dirichlet-Rand liegen (diese Menge nennen wir des Weiteren auch Gitter), hat hier die natürliche Zerlegung (1.1.10) in die Blöcke. Eine ähnliche Zerlegung kann auch für unstrukturierte Gitter erzeugt werden.

Dieses Beispiel ist zur theoretischen Untersuchung der Blocklöser für schwachbesetzte lineare Gleichungssysteme besonders gut geeignet. Einerseits ist es ein für praktische Anwendungen interessantes Problem. Andererseits kann man die Blöcke D und L analytisch behandeln: Da $D = 2b + \text{tridiag}\,\{-a, 2(a+b), -a\}$, erhalten wir so wie bei $\mathbf{A}$ das Spektrum

$$\sigma(D) = \left\{ 2b + 4a \cdot \sin^2 \tfrac{i\pi}{2(N+1)} : 1 \le i \le N \right\}. \tag{1.1.14}$$

Wir können also an diesem Beispiel unsere theoretischen Ergebnisse leicht überprüfen. Daher nennen wir es im Folgenden unser *Modellproblem*. In unseren weiteren Betrachtungen verweisen wir auf noch allgemeinere Modellprobleme, die aber stets den Fall (1.1.12–1.1.13) mit einschließen.

Bemerkung 1.1.2 Wenn wir statt des Dreiecksgitter ein strukturiertes Viereckgitter benutzen, liefert die Finite-Volumen-Diskretisierung des Laplace-Operators auf dem Einheitsquadrat eine Matrix mit dem 9-Punkt-Muster

$$\begin{bmatrix} -\frac{1}{4} & -\frac{1}{2} & -\frac{1}{4} \\ -\frac{1}{2} & 3 & -\frac{1}{2} \\ -\frac{1}{4} & -\frac{1}{2} & -\frac{1}{4} \end{bmatrix}. \tag{1.1.15}$$

Für die Blockstruktur (1.1.10) hat diese Matrix auch die Form (1.1.12), aber mit

$$D = \text{tridiag}\,\{-\tfrac{1}{2}, 3, -\tfrac{1}{2}\}, \qquad L = \text{tridiag}\,\{\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}\}. \tag{1.1.16}$$

Diese Blöcke können wie oben analytisch untersucht werden. Die im Folgenden betrachteten Modellprobleme schließen diesen Fall auch ein. □

1.2 Löser für große schwachbesetzte lineare Gleichungssysteme

Wir betrachten nun ein lineares Gleichungssystem

$$\mathbf{A}u = f \tag{1.2.1}$$

mit einer großen, schwachbesetzten und schlecht konditionierten Matrix $\mathbf{A}$. Prinzipiell sind Lösungsverfahren für solche Systeme aus der linearen Algebra bekannt. Aber

die Algorithmen, die sehr effizient für kleine Systeme sind, können für die jetzt speziell betrachtete Klasse von Problemen einen extrem hohen Rechenaufwand haben. Daher gibt es für die Lösung der großen schwachbesetzten linearen Gleichungssysteme eine eigene Theorie (siehe z.B. [18]), die wir hier kurz zusammenfassen.

Die Hauptidee zur Reduktion des Rechenaufwands ist die Benutzung von Algorithmen, welche Änderungen der Systemmatrixeinträge vermeiden und möglichst nur Vektor-Skalar-, Vektor-Vektor- und Vektor-Matrix-Operationen verwenden. Besonders wichtig sind die *Iterationsverfahren*. Diese konstruieren, ausgehend von einer beliebigen *Anfangsnäherung* $\tilde{u}_0$, rekursiv eine gegen Lösung u^* konvergierende Folge

$$\tilde{u}_0, \tilde{u}_1, \ldots, \tilde{u}_k, \ldots \tag{1.2.2}$$

Die einfachsten Methoden dieser Art sind lineare Iterationen der Form

$$\tilde{u}_{k+1} = \tilde{u}_k - \mathbf{W}^{-1}(f - \mathbf{A}\tilde{u}_k). \tag{1.2.3}$$

Hier ist $\mathbf{W}$ die so genannte *angenäherte Inverse* für $\mathbf{A}$, oder der *Vorkonditionierer*. Dabei ist vorausgesetzt, dass die Berechnung des Vektors $c = \mathbf{W}^{-1}r$ (z.B. durch das Lösen des Systems $\mathbf{W}c = r$) viel leichter als die Berechnung der Lösung von (1.2.1) ist. Dies gilt beispielsweise, wenn $\mathbf{W}$ eine diagonale oder Dreiecksmatrix ist, sowie ein Produkt von verschiedenen Dreiecksmatrizen. Zu dieser Klasse gehören z.B. die Gauß-Seidel-, SOR- und ILU-Verfahren. Die Eigenschaften solcher Iterationen werden von ihren Vorkonditionierern $\mathbf{W}$ bestimmt.

Die Effizienz der Verfahren (1.2.3) wird mit Hilfe der *Konvergenzrate*

$$\rho = \rho(I - \mathbf{W}^{-1}\mathbf{A})$$

gemessen. Ist $\rho < 1$, so konvergiert die Folge (1.2.2) gegen die Lösung von (1.2.1) und es gilt für fast alle Anfangsnäherungen

$$\lim_{k\to\infty} \frac{\|r_{k+1}\|}{\|r_k\|} = \rho,$$

wobei $r_k = f - \mathbf{A}\tilde{u}_k$ das *Residuum bzgl.* $\tilde{u}_k$ und $\|\cdot\|$ eine Vektornorm ist (siehe [18]). In diesem Fall heißt das Verfahren *konvergent*. (Jedes Verfahren kann durch geeignetes Dämpfen konvergent gemacht werden, siehe [18].) Je kleiner die Konvergenzrate ist, desto schneller dämpft die Iteration die Norm des Residuums. Typischerweise hängt ρ so von der Konditionszahl der Matrix $\mathbf{A}$ ab, dass $\rho \to 1$, wenn $\kappa(\mathbf{A})$ unbeschränkt wächst. Wie wir oben erwähnt haben, passiert dies bei Diskretisierungen, wenn die Gitterlänge h gegen Null geht und das diskretisierte System dementsprechend größer wird. Zum Beispiel hat das Gauß-Seidel-Verfahren für System (1.1.7) die Konvergenzrate $\rho_{\mathrm{GS}} = 1 - O(h^2)$, und für das SOR-Verfahren unter der optimalen Wahl des Parameters gilt $\rho_{\mathrm{SOR}} = 1 - O(h)$. Sogar diese Abschätzung zeigt, dass für hinreichend kleine h die SOR-Iteration besser als das Gauß-Seidel-Verfahren ist, da ρ_{GS} schneller als ρ_{SOR} gegen 1 konvergiert. Wenn für ein Verfahren die Konvergenzrate in der Form

$$\rho = 1 - O(h^p)$$

darstellbar ist, nennen wir die Zahl p die *Ordnung der Konvergenz* dieses Verfahrens.

Verfahren, deren Konvergenzrate für $h \to 0$ gegen 1 konvergiert, sind für sehr große Probleme ineffizient. Es gibt aber auch lineare Iterationen, deren Konvergenzrate von der Matrixgröße unabhängig sind. Die wichtigsten davon sind die geometrischen Mehrgitterverfahren (siehe [18], [35], [44]). Diese Algorithmen werden als die effizientesten Verfahren zur Lösung von großen schwachbesetzten schlecht konditionierten Systemen betrachtet. Für die Konstruktion dieser Methoden benötigt man aber nicht nur ein System (1.2.1), sondern eine Hierarchie solcher Gleichungen. Bei praktischen Anwendungen erhält man eine solche typischerweise durch die sukzessive Verfeinerung eines groben Anfangsgitters. Das Mehrgitterverfahren benötigt noch einen Löser für das System auf diesem gröbsten Gitter. In vielen Fällen ist dieses System relativ klein und kann mit der Gaußschen Elimination behandelt werden. Wenn aber die Geometrie des Gebietes kompliziert ist, wird das Gleichungssystem auf dem gröbsten Gitter groß und bedarf der Iterationsverfahren. Ein weiteres Problem ist die starke Abhängigkeit der Konvergenzrate des Standardmehrgitterverfahrens von Eigenschaften der linearen Systems: Starke Anisotropie (wenn z.B. $a \ll b$ in (1.1.3–1.1.4)) oder stark variierende Koeffizienten verschlechtern die Konvergenz dieser Iteration prinzipiell. Es wird deswegen gesagt, dass die Mehrgitterverfahren in ihrer Standardform *nicht robust* sind.

Diese Schwierigkeiten motivieren weitere Forschung auf dem Gebiet der linearen Löser. Als zwei Richtungen davon erwähnen wir hier die algebraischen Mehrgitterverfahren (siehe [35]), die selber die Gitterhierarchie automatisch konstruieren, und die unvollständigen Block-Zerlegungen (siehe Kapitel 2). Einer speziellen Art von diesen ist die vorliegende Arbeit gewidmet. Die beiden Methoden können sowohl als Grobgitterlöser (dies betrifft vor allem die Block-Zerlegungen) als auch als robuste selbständige Vorkonditionierer verwendet werden.

Obwohl wir hier nur lineare Iterationen betrachtet haben, können die Vorkonditionierer der oben beschriebenen Methoden (oder deren symmetrisierten Varianten, siehe [18]) auch in nicht-linearen Iterationsverfahren benutzt werden. Wir erwähnen hier z.B. das Verfahren der konjugierten Gradienten (*CG-Verfahren*), siehe [18]. Für dieses Verfahren ist es nicht nötig, dass die lineare Iteration mit dem gleichen Vorkonditionierer konvergent ist: Die von ihm generierte Folge (1.2.2) konvergiert immer gegen die Lösung des ursprünglichen Systems. Die Anwendung dieses nichtlinearen Verfahrens mit einem Vorkonditionierer lohnt sich immer, weil im Vergleich mit der linearen Iteration die Ordnung der Konvergenz halbiert wird (siehe [18]). Obwohl das CG-Verfahren nur für symmetrische Matrizen anwendbar ist, gibt es auch die Verallgemeinerungen, wie BiCGStab-Verfahren, für weitere Probleme (siehe [4]).

In allen Fällen werden die Eigenschaften des Verfahrens im Wesentlichen vom Vorkonditionierer bestimmt. Wir beschäftigen uns daher in dieser Arbeit mit der Entwicklung von effizienten Vorkonditionierern. Wir weisen darauf hin, dass diese Vorkonditionierer auch zur Lösung der algebraischen Eigenwertprobleme benutzt werden können. Darauf gehen wir in Kapitel 5 näher ein.

Bemerkung 1.2.1 Es existieren auch iterative Lösungsmethoden, die nicht zu den oben beschriebenen Klassen gehören. Als ein Beispiel nennen wir das Verfahren der alternierenden Richtungen (siehe [18], [39]), dessen Konvergenzrate für das

Modellproblem logarithmisch von der Gitterlänge abhängt. Des Weiteren erwähnen wir die linearen Iterationen der Form (1.2.3), bei denen der Vorkonditionierer $\mathbf{W}$ in jedem Schritt neu gewählt ist. Dazu zählen z.B. die Folgen der frequenzfilternden Zerlegungen (siehe [41], [9]), die eine gitterunabhängige Konvergenzrate haben.

Ein weiteres für praktische Anwendungen interessantes Thema ist die Konstruktion der Lösungsverfahren für indefinite lineare Gleichungssysteme und Systeme, die aus Diskretisierungen von Systemen partieller Differentialgleichungen, insbesondere aus Sattelpunktproblemen, entstehen (siehe z.B. [8]). Wir verweisen hier auf [46], [44], [48] und [12]. □

Kapitel 2

Filternde Unvollständige Blockzerlegungen

In diesem Kapitel beschreiben wir die Methode der unvollständigen Blockzerlegungen zur schnellen Lösung linearer Gleichungssysteme mit großen schwachbesetzten Matrizen. Nach der Definition dieser Klasse von Verfahren geben wir zunächst einen Überblick über bisher in der Literatur behandelte Beispiele und Möglichkeiten der theoretischen Untersuchungen von Konvergenzeigenschaften. Diese verwenden wir im nächstem Kapitel zur Analyse der in dieser Arbeit vorgestellten neuen Blockzerlegungen.

2.1 Idee, Definition und die Filterbedingung

Die Matrix $\mathbf{A}$ des linearen Systems

$$\mathbf{A}u = f \tag{2.1.1}$$

habe die blocktridiagonale Form:

$$\mathbf{A} = \begin{pmatrix} D_1 & -U_1 & & \\ -L_2 & D_2 & -U_2 & \\ & \ddots & \ddots & \ddots \\ & & -L_N & D_N \end{pmatrix} = \mathbf{L} + \mathbf{D} + \mathbf{U}. \tag{2.1.2}$$

Hierbei sind D_k, L_k und U_k Blöcke der Größe $n_k \times n_k$, $n_k \times n_{k-1}$ und $n_k \times n_{k+1}$, respektive, $\mathbf{D} = \text{blockdiag}\,\{D_k\}$, $\mathbf{L} = \text{blocktridiag}\,\{-L_k, 0, 0\}$, $\mathbf{U} = \text{blocktridiag}\,\{0, 0, -U_k\}$, $u = \text{blockvector}\,\{u_k\}$, $f = \text{blockvector}\,\{f_k\}$. Ein mögliches Verfahren zur Lösung von (2.1.1) ist die Blockvariante der Gaußschen Elimination (siehe [30]). Bei dieser Methode werden die Blockgleichungen

$$-L_k u_{k-1} + D_k u_k - U_k u_{k+1} = f_k$$

des Systems (2.1.1) sukzessive durch das Einsetzen von dem aus der $(k-1)$-en Gleichung berechneten u_{k-1} auf die Form

$$T_k u_k - U_k u_{k+1} = \hat{f}_k \tag{2.1.3}$$

transformiert. Die erste Blockgleichung hat bereits diese Form. Daraus ergeben sich die Rekursionsformeln für T_k und $\hat{f}_k$:

$$T_k = \begin{cases} D_1, & k = 1, \\ D_k - L_k T_{k-1}^{-1} U_{k-1}, & k \geq 2, \end{cases} \tag{2.1.4}$$

$$\hat{f}_k = \begin{cases} f_1, & k = 1, \\ f_k + L_k T_{k-1}^{-1} f_{k-1}, & k \geq 2. \end{cases} \tag{2.1.5}$$

Die Blöcke u_k der Lösung werden dann aus den Gleichungen (2.1.3) in absteigender Folge der Indizes berechnet.

Algorithmus 2.1.1 *(Blockvariante der Gaußschen Elimination)* Die Matrix $\mathbf{A}$ habe die Form (2.1.2). Der folgende Algorithmus findet die Lösung u des linearen Systems (2.1.1) für die beliebig gewählte rechte Seite f.

BEGIN
  $T_1 \leftarrow D_1$; $\hat{f}_1 \leftarrow f_1$;
  for all k **from** 2 **to** N **do**
    $T_k \leftarrow D_k - L_k T_{k-1}^{-1} U_{k-1}$;
    $\hat{f}_k \leftarrow f_k + L_k T_{k-1}^{-1} f_{k-1}$
  end for
  $u_N \leftarrow T_N^{-1} \hat{f}_N$;
  for all k **from** $N-1$ **to** 1 **do**
    $u_k \leftarrow T_k^{-1}(\hat{f}_k + U_k u_{k+1})$
  end for
END □

Diese Methode beruht auf einer *vollständigen Zerlegung* der Systemmatrix $\mathbf{A}$:

$$\mathbf{A} = (\mathbf{L} + \mathbf{T})\mathbf{T}^{-1}(\mathbf{T} + \mathbf{U}), \tag{2.1.6}$$

wobei $\mathbf{T} = \text{blockdiag}\,\{T_k\}$. Die Berechnung der Lösung u von (2.1.1) erfolgt also, indem die entsprechenden linearen Gleichungen für die in (2.1.6) angegeben Faktoren von $\mathbf{A}$ nacheinander gelöst werden.

Bemerkung 2.1.2 Im Allgemeinen existieren die Blöcke T_k mit der in (2.1.4) geforderten Eigenschaft nicht. Jedoch wurde die Existenz in wichtigen Fällen gezeigt, z.B. in [30]. Insbesondere existieren sie, wenn $\mathbf{A}$ positiv definit oder eine M-Matrix ist. □

Bei der praktischen Anwendung dieses Verfahrens auf schwachbesetzte Matrizen ergibt sich aber ein großes Problem: die Blöcke T_k sind wegen des Summanden $L_k T_{k-1}^{-1} U_{k-1}$ im allgemeinen vollbesetzt. Wenn die Größe der Blöcke mit der Anzahl der Gitterknoten wächst, benötigt die Behandlung solcher T_k, insbesondere die Berechnung von T_k^{-1}, sehr viel Rechenzeit. Deswegen ist diese Methode für große schwachbesetzte Matrizen nicht effizient.

Dieses Problem kann man umgehen, indem man in jedem Schritt der Rekursion (2.1.4) den Block T_k durch eine schwachbesetzte Approximierende $\tilde{T}_k$ ersetzt. Statt (2.1.6) erhalten wir dann für $\mathbf{A}$ die folgende Aufspaltung:

$$\mathbf{A} = \mathbf{W} - \mathbf{R}, \qquad \mathbf{W} = (\mathbf{L} + \tilde{\mathbf{T}})\tilde{\mathbf{T}}^{-1}(\tilde{\mathbf{T}} + \mathbf{U}) \tag{2.1.7}$$

mit $\tilde{\mathbf{T}} = \text{blockdiag}\left\{\tilde{T}_k\right\}$. Da jetzt $\mathbf{R} \neq 0$, wird im oben beschriebenen Prozess mit den approximierenden Blöcken $\tilde{T}_k$ nur $\mathbf{W}$ invertiert, also erhält man kein direktes Verfahren zur Lösung von (2.1.1). Die Matrix $\mathbf{W}$ kann aber als Vorkonditionierer in einem iterativen Löser verwendet werden.

Algorithmus 2.1.3 *(IBLU-Vorkonditionierer)* Die Matrix $\mathbf{A}$ habe die Form (2.1.2). Der folgende Algorithmus löst das lineare System $\mathbf{W}u = f$ mit $\mathbf{W}$ aus (2.1.7). Es wird angenommen, dass die Diagonalblöcke $\tilde{T}_k$ für alle $k \in \{1, \ldots, N\}$ schon berechnet wurden.

BEGIN
  $\hat{f}_1 \leftarrow f_1$;
  for all k **from** 2 **to** N **do**
    Berechne u_{k-1} aus $\tilde{T}_{k-1}u_{k-1} = f_{k-1}$; $\hat{f}_k \leftarrow f_k + L_k u_{k-1}$
  end for
  Berechne u_N aus $\tilde{T}_N u_N = \hat{f}_N$;
  for all k **from** $N-1$ **to** 1 **do**
    Berechne u_k aus $\tilde{T}_k u_k = \hat{f}_k + U_k u_{k+1}$
  end for
END □

Wir weisen darauf hin, dass die Blöcke $\tilde{T}_k$ nur einmal für den gesamten Lösungsprozess berechnet werden müssen, wobei Algorithmus 2.1.3 in jedem Schritt des Iterationsverfahren aufgerufen wird.

Definition 2.1.4 Die Aufspaltung (2.1.7) heißt *unvollständige Blockdreieckszerlegung (oder IBLU-Zerlegung)* der Matrix $\mathbf{A}$. Die Matrix $\mathbf{R}$ aus (2.1.7) heißt *Restmatrix* dieser IBLU-Zerlegung.

Man bemerkt sofort, dass einige bekannte Verfahren auch in dieser Form darstellbar sind. Zum Beispiel erhält man für $\tilde{T}_k = D_k$ das symmetrische Block-Gauß-Seidel-Verfahren und für $\tilde{T}_k = \omega D_k$ mit einer Konstanten ω das Block-SSOR-Verfahren (siehe [18]). Diese Fälle sind die einfachsten von IBLU-Zerlegungen. Durch eine andere Wahl der Diagonalblöcke $\tilde{T}_k$ kann man die Eigenschaften der Verfahren noch verbessern.

Jede IBLU-Zerlegung ist also durch die Regel zur Berechnung von $\tilde{T}_k$ gegeben. Üblicherweise geht man bei der Definition einer Approximierenden $\tilde{T}_k$ von der Rekursionsformel (2.1.4) aus. Das einfachste Beispiel dafür ist die so genannte Linien-ILU (ILLU) von Kettler (siehe [21]):

$$\tilde{T}_k = D_k - L_k \left(\tilde{T}_{k-1}^{-1}\right)^{(3)} U_{k-1}, \quad k \geq 2,$$

wobei für eine reelle Matrix $A = (a_{ij})_{ij}$

$$(A)^{(p)} := \begin{cases} a_{ij}, & \text{für } |i-j| \leq \frac{p-1}{2}, \\ 0, & \text{sonst} \end{cases}$$

ist. Die Berechnung des vollbesetzten $\tilde{T}_{k-1}^{-1}$ wäre sehr aufwendig. Allerdings ist es möglich, lediglich die Einträge auf der Haupt- und den beiden Nebendiagonalen

auszuwerten. In diesem Fall wächst die Anzahl der arithmetischen Operationen nur linear mit der Anzahl der Knoten in einem Block. Diese Zerlegung wurde als Vorkonditionierer für CG-Verfahren und als Glätter in Mehrgitterverfahren vorgeschlagen. Die linearen Systeme

$$\tilde{T}_k u_k = \tilde{f}_k,$$

die während des Lösungsprozesses bei solchen Zerlegungen entstehen, sind schwachbesetzt. Z.B. sind sie für die ILLU-Zerlegung im Fall der Diskretisierung (1.1.5–1.1.8) tridiagonal (siehe auch (1.1.11–1.1.13)) und lassen sich sehr effizient mit der Gaußschen Elimination lösen.

Die aktuelle Forschung interessiert sich für die IBLU-Zerlegungen vor allem, weil diese bei vielen Problemen mit stark variierenden Koeffizienten robust sind. Außerdem erlaubt die Freiheit in der Auswahl der Diagonalblöcke $\tilde{T}_k$ die Konstruktion von Verfahren mit vorgegebenen gewünschten Eigenschaften.

Eine für die Entwicklung von IBLU-Zerlegungen entscheidende Idee war, die Blöcke $\tilde{T}_k$ so zu konstruieren, dass die Matrizen $\mathbf{W}$ und $\mathbf{A}$ aus (2.1.7) $l+1$ vorgegebene Vektoren $\mathbf{e}^{(i)}$, $i \in \{0, \ldots, l\}$, gleich abbilden:

$$\mathbf{W}\mathbf{e}^{(i)} = \mathbf{A}\mathbf{e}^{(i)}, \tag{2.1.8}$$

also auf span $\left\{\mathbf{e}^{(0)}, \ldots, \mathbf{e}^{(l)}\right\}$ gleich sind. Solche Zerlegungen werden von vielen Autoren vorgeschlagen, siehe [47], [20], [41], [42], [1]. Die Vektoren $\mathbf{e}^{(i)}$ heißen *Testvektoren*. Da die Bedingung (2.1.8) zur Gleichung

$$\mathbf{R}\mathbf{e}^{(i)} = 0 \tag{2.1.9}$$

äquivalent ist, kann sie mit der Hilfe des folgenden Satzes zu einer Gleichung für die Diagonalblöcke $\tilde{T}_k$ umformuliert werden:

Satz 2.1.5 Gegeben sei eine IBLU-Zerlegung (2.1.7) einer blocktridiagonalen Matrix (2.1.2). Dann hat die Restmatrix $\mathbf{R}$ die blockdiagonale Form

$$\mathbf{R} = \text{blockdiag } \{R_k\}, \tag{2.1.10}$$

wobei die Blöcke R_k die Größe $n_k \times n_k$ haben. Für diese Blöcke gilt

$$R_k = \begin{cases} \tilde{T}_1 - D_1, & k = 1, \\ \tilde{T}_k - \left(D_k - L_k \tilde{T}_{k-1}^{-1} U_{k-1}\right), & k \geq 2. \end{cases} \tag{2.1.11}$$

Beweis: Nach (2.1.7) erhält man:

$$\begin{aligned} \mathbf{R} &= \mathbf{W} - \mathbf{A} = (\mathbf{L} + \tilde{\mathbf{T}})\tilde{\mathbf{T}}^{-1}(\tilde{\mathbf{T}} + \mathbf{U}) - \mathbf{D} - \mathbf{L} - \mathbf{U} \\ &= \tilde{\mathbf{T}} - (\mathbf{D} - \mathbf{L}\tilde{\mathbf{T}}^{-1}\mathbf{U}). \end{aligned} \tag{2.1.12}$$

Wegen der Struktur der Matrizen $\mathbf{L}$, $\mathbf{U}$ und $\tilde{\mathbf{T}}$ gilt

$$\mathbf{L}\tilde{\mathbf{T}}^{-1}\mathbf{U} = \text{blockdiag } \{K_k\}$$

mit

$$K_k = \begin{cases} 0, & k = 1, \\ L_k \tilde{T}_{k-1}^{-1} U_{k-1}, & k \geq 2, \end{cases}$$

also erfüllen die Blöcke von (2.1.12) die Gleichung (2.1.11). □

Dieser Satz zeigt, dass für die Testvektoren der Form

$$\mathbf{e}^{(i)} = \text{blockvector}\,\{e_k^{(i)}\}$$

Bedingung (2.1.8) genau dann erfüllt ist, wenn für jedes i

$$\begin{aligned} \tilde{T}_1 e_1^{(i)} &= D_1 e_1^{(i)}, \\ \tilde{T}_k e_k^{(i)} &= \left(D_k - L_k \tilde{T}_{k-1}^{-1} U_{k-1}\right) e_k^{(i)}, \quad k \geq 2, \end{aligned}$$

gilt.

Diese Idee ist wie folgt weiterentwickelt worden. G. Wittum betrachtet in [47] eine Klasse von Blockzerlegungen, deren Diagonalblöcke die Form

$$\tilde{T}_k = \begin{cases} D_1, & k = 1, \\ D_k - \Theta_k, & k \geq 2, \end{cases}$$

haben, wobei die diagonalen Matrizen Θ_k so gewählt sind, dass Bedingung (2.1.8) gilt. Nach Satz 2.1.5 erhält man für die Einträge von Θ_k ein lineares System

$$\Theta_k e_k^{(0)} = L_k D_{k-1}^{-1} U_{k-1} e_k^{(0)}, \qquad 2 \leq j \leq N.$$

In [47] wurde vorgeschlagen, auf strukturierten Gittern $\Omega_h = \{(i,k) : 1 \leq i \leq n_k = n, 1 \leq j \leq N\}$ die Fourier-Moden

$$e_{ij}^{(0)} = \sin \tfrac{\pi\nu i}{n+1} \quad \text{oder} \quad e_{ij}^{(0)} = \cos \tfrac{\pi\nu i}{n+1} \tag{2.1.13}$$

als Testvektoren zu benutzen. Die entstehende Zerlegung dämpft dann den zum Parameter ν gehörigen Frequenzanteil des Fehlers. Wählt man eine hohe Frequenz, so erhält man einen Glätter für Mehrgitterverfahren. Wenn man hingegen eine niedrige Frequenz wählt, so ist diese Zerlegung ein guter Vorkonditionierer für das Verfahren der konjugierten Gradienten. Die beschriebene Methode wird *Frequenz-Filternde (FF) Zerlegung* genannt. Gleichung (2.1.8) heißt dann *Filterbedingung*.

Diese Methode wurde von C. Wagner weiterentwickelt. In [41] betrachtet er *Tangentiale Frequenzfilternde Zerlegungen* (TFF-Zerlegungen). Wählt man als Testvektoren (2.1.8) diejenigen aus (2.1.13), so werden auch die Sinus- und Kosinusvektoren mit Frequenz nahe bei ν stark gedämpft. Die Diagonalblöcke haben dann die Form

$$\tilde{T}_k = \begin{cases} D_1, & k = 1, \\ D_k + \Theta_{k,k-1} \tilde{T}_{k-1} \Theta_{k-1,k} - \Theta_{k,k-1} U_{k-1} - L_k \Theta_{k-1,k}, & k \geq 2, \end{cases} \tag{2.1.14}$$

Die *Transferoperatoren* Θ_{ij} werden aus Bedingung (2.1.9) berechnet, welche nach Satz 2.1.5

$$\left(\Theta_{k,k-1} \tilde{T}_{k-1} \Theta_{k-1,k} - \Theta_{k,k-1} U_{k-1} - L_k \Theta_{k-1,k}\right) e_k^{(0)} = -L_k D_{k-1}^{-1} U_{k-1} e_k^{(0)}$$

lautet und zur Gleichung

$$\left(\Theta_{k,k-1}\tilde{T}_{k-1} - L_{k-1}\right)\tilde{T}_{k-1}^{-1}\left(\tilde{T}_{k-1}\Theta_{k-1,k} - U_{k-1}\right)e_k^{(0)} = 0.$$

äquivalent ist. Hierfür ist

$$\left(\tilde{T}_{k-1}\Theta_{k-1,k} - U_{k-1}\right)e_k^{(0)} = 0 \tag{2.1.15}$$

eine hinreichende Bedingung. Für die strukturierten Gitter wählt man die diagonalen Θ_{ij} mit $\Theta_{ij} = \Theta_{ji}$.

Die ursprüngliche Idee dieser Klasse von Zerlegungen demonstrieren wir für ein einfaches Modellproblem, das bei der Diskretisierung von elliptischen partiellen Differentialgleichungen mit konstanten Koeffizienten entsteht.

Wir betrachten die Matrix $\mathbf{A}$ der Form

$$\mathbf{A} = \text{blocktridiag}\{-L, D, -L\} = \left.\begin{pmatrix} D & -L & & \\ -L & D & -L & \\ & \ddots & \ddots & \ddots \\ & & -L & D \end{pmatrix}\right\} N \text{ Blockzeilen}, \tag{2.1.16}$$

wobei die beiden Blöcke D und L von der Größe $n \times n$ sind. Wir nehmen weiter an, dass D und L symmetrisch und positiv definit sind und der Ungleichung

$$D > 2L \tag{2.1.17}$$

genügen. Diese Bedingungen sind sehr stark, aber z.B. für die Diskretisierung (1.1.5–1.1.8) (siehe (1.1.11–1.1.13)) erfüllt.

Da $\mathbf{A}$ einen Gitteroperator mit konstanten Koeffizienten beschreibt, nehmen wir in der Konstruktion von TFF-Zerlegungen (2.1.14) die Matrizen

$$\Theta_{k-1,k} = \Theta_{k,k-1} = \frac{1}{\lambda_{k-1}}I \tag{2.1.18}$$

mit $\lambda_k \in \mathbb{R}$. Nach (2.1.14) sind dann die Diagonalblöcke durch die Rekursionsformel

$$\tilde{T}_k = \begin{cases} D, & k = 1, \\ D + \dfrac{1}{\lambda_{k-1}^2}\tilde{T}_{k-1} - \dfrac{2}{\lambda_{k-1}}L, & k \geq 2, \end{cases} \tag{2.1.19}$$

gegeben. Die Koeffizienten λ_k sollen hier nach der Filterbedingung (2.1.9) durch die Wahl des Testvektors bestimmt werden. Als Testvektor nehmen wir hier $\mathbf{e}^{(0)} = \text{blockvector}\{\hat{e}\}$, wobei $\hat{e}$ ein Eigenvektor des Problems

$$D\hat{e} = \hat{\lambda}L\hat{e}$$

ist. Da die beiden Matrizen D und L positiv definit sind, besitzt dieses Problem ein System von n linear unabhängigen Eigenvektoren. Nach Bedingung (2.1.17) genügen ihre Eigenwerte der Ungleichung

$$\hat{\lambda} \geq 2.$$

Die Sinus-Gitterfunktionen (2.1.13) sind ein Beispiel für solche Testvektoren für die oben genannte Matrix (1.1.12–1.1.13).

Für die Koeffizienten λ_k fordern wir die Filterbedingung (2.1.9):

$$\mathbf{Re}^{(0)} = 0. \tag{2.1.20}$$

Des Weiteren möchten wir diese Koeffizienten nicht durch den Testvektor, sondern durch den entsprechenden Eigenwert $\hat{\lambda}$ beschreiben. Die folgenden Lemmata zeigen, dass Bedingung (2.1.20) zur Gleichung

$$\lambda_k = F_k(\hat{\lambda}) \tag{2.1.21}$$

äquivalent ist, wobei

$$F_k(\lambda) = \begin{cases} \lambda, & k = 1, \\ \lambda - \dfrac{1}{F_{k-1}(\lambda)}, & k \geq 2. \end{cases} \tag{2.1.22}$$

Lemma 2.1.6 Unabhängig von der Wahl der Parameter λ_k in (2.1.19) sind die Blöcke $\hat{T}_k := L^{-\frac{1}{2}} \tilde{T}_k L^{-\frac{1}{2}}$ lineare Funktionen von $\hat{D} := L^{-\frac{1}{2}} D L^{-\frac{1}{2}}$:

$$\hat{T}_k = \tilde{F}_k(\hat{D}), \tag{2.1.23}$$

wobei

$$\tilde{F}_k(\lambda) = \begin{cases} \lambda, & k = 1, \\ \lambda + \dfrac{\tilde{F}_{k-1}(\lambda)}{\lambda_{k-1}^2} - \dfrac{2}{\lambda_{k-1}}, & k \geq 2. \end{cases} \tag{2.1.24}$$

Für die Restmatrix gilt: $\mathbf{R} = \text{blockdiag}\,\{R_k\}$, wobei $R_k = L^{\frac{1}{2}} f_k(\hat{D}) L^{\frac{1}{2}}$ mit

$$f_k(\lambda) = \begin{cases} 0, & k = 1, \\ \dfrac{(\tilde{F}_{k-1}(\lambda) - \lambda_{k-1})^2}{\lambda_{k-1}^2 \tilde{F}_{k-1}(\lambda)}, & k \geq 2. \end{cases} \tag{2.1.25}$$

Beweis: Die Aussagen (2.1.23–2.1.24) erhält man direkt durch Multiplikation von (2.1.19) mit $L^{-\frac{1}{2}}$ von links und rechts. Die blockdiagonale Struktur von $\mathbf{R}$ und die Gleichheit $R_1 = 0$ folgen direkt aus Satz 2.1.5. Für $k \geq 2$ gilt

$$\begin{aligned} R_k &= \tilde{T}_k - D + L\tilde{T}_{k-1}^{-1} L = \frac{1}{\lambda_{k-1}^2}\tilde{T}_{k-1} - \frac{2}{\lambda_{k-1}} L + L\tilde{T}_{k-1}^{-1} L \\ &= \frac{1}{\lambda_{k-1}^2}(\tilde{T}_{k-1} - \lambda_{k-1} L)\tilde{T}_{k-1}^{-1}(\tilde{T}_{k-1} - \lambda_{k-1} L). \end{aligned}$$

Nach Multiplikation mit $L^{-\frac{1}{2}}$ von links und rechts erhält man:

$$L^{-\frac{1}{2}} R_k L^{-\frac{1}{2}} = \frac{1}{\lambda_{k-1}^2}(\tilde{F}_{k-1}(\hat{D}) - \lambda_{k-1} I)^2 \left(\tilde{F}_{k-1}(\hat{D})\right)^{-1} = f_k(\hat{D}).$$

□

Lemma 2.1.7 Für die IBLU-Zerlegung der Matrix (2.1.16) mit den Diagonalblöcken (2.1.19) sind die Filterbedingung (2.1.20) und Gleichung (2.1.21) einander äquivalent. Unter Voraussetzung (2.1.21) sind durch die Rekursion (2.1.24) die Funktionen

$$\tilde{F}_k(\lambda) = \begin{cases} \lambda, & k = 1, \\ F_k(\hat{\lambda}) + F_k'(\hat{\lambda})(\lambda - \hat{\lambda}), & k \geq 2 \end{cases} \tag{2.1.26}$$

eindeutig bestimmt. Und für die Funktionen f_k gilt

$$f_k(\lambda) = \begin{cases} 0, & k = 1, \\ \left(\dfrac{F_{k-1}'(\hat{\lambda})}{F_{k-1}(\hat{\lambda})}\right)^2 \dfrac{(\lambda - \hat{\lambda})^2}{\tilde{F}_{k-1}(\lambda)}, & k \geq 2. \end{cases} \tag{2.1.27}$$

Beweis: Da $L^{-\frac{1}{2}}\hat{e}$ ein Eigenvektor der Matrix $\hat{D}$ ist, lässt sich Bedingung (2.1.20) nach (2.1.25) in der Form

$$\lambda_{k-1} = \tilde{F}_{k-1}(\hat{\lambda}) \tag{2.1.28}$$

schreiben. Durch vollständige Induktion über k zeigt man unter Benutzung der Rekursionsformeln (2.1.22) und (2.1.24) die Gleichheit von (2.1.28) und (2.1.21). Formel (2.1.26) kann man durch Einsetzen von (2.1.26) in (2.1.24) direkt überprüfen. Dabei benutzt man

$$F_k'(\lambda) = \begin{cases} 1, & k = 1, \\ 1 + \dfrac{F_{k-1}'(\lambda)}{F_{k-1}^2(\lambda)}, & k \geq 2. \end{cases}$$

Gleichung (2.1.27) kann dann aus (2.1.25) mit der Hilfe von (2.1.26) hergeleitet werden. □

Wir betrachten nun die vollständige Zerlegung (2.1.6) der Matrix (2.1.16). Nach (2.1.4) sind die Diagonalblöcke T_k durch die Rekursionsformel

$$T_k = \begin{cases} D, & k = 1, \\ D - LT_{k-1}^{-1}L, & k \geq 2 \end{cases}$$

gegeben. Das bedeutet, dass

$$T_k = L^{\frac{1}{2}} F_k(\hat{D}) L^{\frac{1}{2}}$$

mit F_k aus (2.1.22). Lemma 2.1.7 zeigt, dass man die Diagonalblöcke $\tilde{T}_k$ der betrachteten TFF-Zerlegungen erhält, wenn man F_k durch $\tilde{F}_k$ in dieser Formel ersetzt. Nach (2.1.26) ist die Approximierende $\tilde{F}_k$ einfach die Tangente der Funktion F_k im Punkt $\hat{\lambda}$. Die Eigenwerte der Blöcke T_k werden in diesem Sinne also linear approximiert: Sowohl der Wert von F_k in $\hat{\lambda}$ als auch die Ableitung F_k' in diesem Punkt werden exakt rekonstruiert.

Das ist die ursprüngliche Idee bei diesen Zerlegungen (siehe [42]). Mit einer ähnlichen Motivation wurden in [47] filternde Zerlegungen von Wittum unter stärkeren Voraussetzungen als den von uns betrachteten eingeführt. Obwohl sich die Methoden zur theoretischen Behandlung nicht unmittelbar übertragen lassen, zeigen numerische Experimente für diese ähnlich gute Ergebnisse. So wurde in [47] gezeigt, dass bei

der optimalen Auswahl von Parametern der Spektralradius des Iterationsoperators von FF-Zerlegungen von der Größe $1 - O(n^{-1})$ ist, wie bei Block-SSOR-Verfahren, und von der Anzahl der Blöcke nicht abhängt. Diese Vorgehensweise besprechen wir im Detail in Abschnitt 2.2.

Die Filterbedingung (2.1.8) verlangt die exakte Gleichheit von $\mathbf{W}$ und $\mathbf{A}$ auf Testvektoren. Eine ähnliche Idee ist, diese Bedingung in einer schwachen Form zu stellen:

$$(\mathbf{We}^{(i)}, \mathbf{e}^{(i)}) = (\mathbf{Ae}^{(i)}, \mathbf{e}^{(i)}), \quad i \in \{0, \ldots, l\},$$

also

$$(\mathbf{Re}^{(i)}, \mathbf{e}^{(i)}) = 0, \quad i \in \{0, \ldots, l\}. \tag{2.1.29}$$

Das wurde von A. Buzdin gemacht. Er führt in [9] die *Tangentialen Zerlegungen* ein, wobei die Transferoperatoren $\Theta_{k,k-1}$ von TFF-Zerlegungen durch skalare Parameter ω_k ersetzt werden. Dann gilt

$$\tilde{T}_k = \begin{cases} D_1, & k = 1, \\ D_k + \omega_{k-1}^2 \tilde{T}_{k-1} - \omega_{k-1}(U_{k-1} + L_k), & k \geq 2, \end{cases} \tag{2.1.30}$$

Zur Berechnung von ω_k benutzt man die Formel

$$\omega_k = \frac{(U_k e_k^{(0)}, e_k^{(0)})}{(\tilde{T}_k e_k^{(0)}, e_k^{(0)})}. \tag{2.1.31}$$

Man gibt also nicht den Testvektor an, sondern nur die Testfrequenz. Für das erwähnte Modellproblem sind die TFF- und die tangentialen Zerlegungen gleich, und im Fall von variierenden Koeffizienten zeigen sie ähnliche Ergebnisse. Jedoch lässt sich die tangentiale Zerlegung einfacher berechnen.

Diese Konstruktion erlaubt eine Erweiterung auf zwei Testfrequenzen, die so genannte *Zweifrequenzzerlegung* (siehe [11]). Dabei ist $\tilde{F}_k$ keine Tangente, sondern eine lineare Funktion die mit F_k in zwei vorgegebenen Punkten $\hat{\lambda}^{(1)}$ und $\hat{\lambda}^{(2)}$ übereinstimmt. Die entsprechenden Eigenmoden der Systemmatrix werden dann im Fehler gedämpft. Als Grenzfall, wenn die beiden angegebenen Frequenzen gleich sind, erhält man die tangentiale Zerlegung.

In [11] beweisen Buzdin und Wittum, dass für das Modellproblem Zweifrequenzzerlegungen mit optimalen Parametern sogar die Konvergenzordnung $\frac{2}{3}$ haben, also asymptotisch besser sind als Block-SSOR-Verfahren.

2.2 Konvergenz von filternden IBLU-Zerlegungen

Es existiert keine allgemeine Methode, mit der man jede beliebige Zerlegung untersuchen kann. Es gibt vielmehr zu den verschiedenen Klassen von Zerlegungen jeweils eine eigene Theorie. Im Allgemeinen sind für jede IBLU-Zerlegung die drei folgenden Eigenschaften zu prüfen:

1. die Existenz der Zerlegung, die in diesem Fall bedeutet, dass alle Diagonalblöcke $\tilde{T}_k$ wohldefiniert und regulär sind,

2. die numerische Stabilität des Lösungsprozesses des linearen Systems $\mathbf{W}u = f$ und

3. die Konvergenzeigenschaften, wie der Spektralradius der Iterationsmatrix oder die Glättereigenschaft.

Existenzaussagen bestimmen die Anwendbarkeit der Zerlegungen auf verschiedene Probleme. Sie kann oft für große Klassen von Matrizen (wie M- oder positiv definite Matrizen) bewiesen werden. Die Existenztheorie der oben betrachteten filternden Zerlegungen wiederholen wir hier nicht und verweisen dafür auf [42], [9] und [11]. Die Untersuchung der Stabilität ist aber oft nur in Spezialfällen möglich. Trotzdem stellt dies keine Schwierigkeiten bei praktischen Anwendungen dar, wenn die numerischen Experimente keine Instabilität für Testbeispiele zeigen. Die Effizienz einer Zerlegung wird von den Konvergenzeigenschaften bestimmt, deren Untersuchung typischerweise ein besonders großes Problem darstellt.

Eine typische Vorgehensweise zur Konvergenzanalyse für viele Zerlegungen wie Linien-ILU von Kettler besteht in regulären Aufspaltungen (siehe [21]). Aber Abschätzungen der Konvergenzrate sind in allgemeinen Fällen entweder überhaupt nicht möglich oder nur sehr grob, sodass die Vorteile der Zerlegung daraus nicht ersichtlich sind. Um so wichtiger ist die Untersuchung von einfachen Modellproblemen, die in hohem Maße analysierbar sind und zugleich die wesentlichen Eigenschaften einer allgemeineren, für die Anwendung interessanten Problemklasse aufweisen. Ein Beispiel hierfür ist die Untersuchung der Zerlegungen aus dem letzten Abschnitt. Eine solche Vorgehensweise werden wir in dieser Arbeit weiterhin verwenden, und zwar zum Studium von weiter unten vorgestellten Zerlegungen. Zunächst erklären wir diese Strategie am Beispiel der TFF- und tangentialen Zerlegungen für das Modellproblem (2.1.16–2.1.17). Wie wir schon gesagt haben, sind diese Zerlegungen in diesem Fall gleich, und die Diagonalblöcke werden von der Rekursionsformel (2.1.19) definiert. Die weiter unten in diesem Abschnitt vorgestellte Theorie wurde im Wesentlichen von [11] übernommen, wo Zweifrequenzzerlegungen analysiert werden. (Siehe auch [13].)

Wie man aus Lemmata 2.1.6–2.1.7 sieht, können die Diagonalblöcke sowie auch die Restmatrix der Zerlegung mit der Hilfe von skalaren Funktionen beschrieben werden. Außerdem ist die Zerlegung eindeutig durch den skalaren Parameter $\hat{\lambda}$ angegeben. Diese Tatsache ermöglicht eine effiziente Untersuchung der Konvergenzeigenschaften: Unsere Aufgabe ist auf die Abschätzung von reellen Funktionen reduziert. Das ist der zentrale Punkt in dieser Vorgehensweise. Die dabei betrachteten Funktionen sind typischerweise gebrochen-rational, und solche können in vielen Fällen tief analytisch untersucht werden. Des Weiteren führt das zur Ersetzung der diskreten Mengen von Eigenwerten der Matrixblöcke durch kontinuierliche Intervalle und ermöglicht dadurch eine leichtere Untersuchung der Konvergenzordnung.

Als Beispiel schätzen wir den Spektralradius $\rho(\mathbf{W}^{-1}\mathbf{R})$ der Iterationsmatrix $\mathbf{W}^{-1}\mathbf{R} = I - \mathbf{W}^{-1}\mathbf{A}$ ab (siehe Abschnitt 1.2). Da $\mathbf{W}$ und $\mathbf{R}$ positiv semidefinit sind, reicht es dazu aus, ein $\omega \in [0, 1)$ zu finden, für das

$$\mathbf{R} \leq \omega \mathbf{W} \tag{2.2.1}$$

gilt. Dann erhalten wir

$$\rho(\mathbf{W}^{-1}\mathbf{R}) \leq \omega, \tag{2.2.2}$$

d.h. ω ist eine obere Schranke für die Konvergenzrate. Die Abschätzung (2.2.1) erfolgt in drei Schritten. Zuerst benutzen wir die Blockstruktur der Matrizen $\mathbf{R}$ und $\mathbf{W}$, um (2.2.1) in ein äquivalentes System von N Ungleichungen für Matrixblöcke umzuschreiben. Im zweiten Schritt werden die Blöcke durch skalare Funktionen dargestellt. Damit transformieren wir das System von Matrixungleichungen in ein System von Ungleichungen für skalare Funktionen auf kontinuierlichen Intervallen. Im dritten Schritt wird dieses durch das Auffinden einer oberen Schranke für eine einzige skalare Funktion ersetzt. Diese Schranke ist auch eine solche für ω aus (2.2.1).

Bemerkung 2.2.1 Aus den gleichen Gründen ist es notwendig, auch den Parameter $\hat{\lambda}$ als eine kontinuierliche Größe zu betrachten. Die Formeln (2.1.18), (2.1.21) und (2.1.22) definieren die beschriebenen Zerlegungen für alle reellen $\hat{\lambda}$. Aber die Existenz dieser Zerlegungen muss separat bewiesen werden. In [9] wurde gezeigt, dass solche Zerlegungen für ein beliebiges $\hat{\lambda} \in [2, +\infty)$ existieren. □

Die Reduktion von (2.2.1) auf Ungleichungen für die Matrixblöcke kann durch die Ersetzung der Matrix $\mathbf{W} = \mathbf{A} + \mathbf{R}$ durch eine blockdiagonale Matrix erreicht werden. Dazu wenden wir das folgende Lemma an:

Lemma 2.2.2 Für die Matrix $\mathbf{A}$ der Form (2.1.2) gilt die Abschätzung

$$\tilde{\mathbf{A}} \leq \mathbf{A}, \tag{2.2.3}$$

wobei $\tilde{\mathbf{A}} = \text{blockdiag}\,\{D - 2L\}$.

Beweis: $\mathbf{A} = \tilde{\mathbf{A}} + \text{blocktridiag}\,\{-L, 2L, -L\}$. Da $\text{blocktridiag}\,\{-L, 2L, -L\}$ positiv semidefinit ist, gilt (2.2.3). □

Mit Hilfe dieser Aussage erhalten wir eine hinreichende Bedingung für (2.2.1):

$$\mathbf{R} \leq \omega(\tilde{\mathbf{A}} + \mathbf{R}). \tag{2.2.4}$$

Da jetzt die Matrizen auf beiden Seiten tridiagonal sind, ist (2.2.4) zum System

$$R_k \leq \omega(D - 2L + R_k), \quad 1 \leq k \leq N \tag{2.2.5}$$

äquivalent.

Damit kommen wir zum zweiten Teil der Untersuchung. Multiplikation in (2.2.5) von links und rechts mit $L^{-\frac{1}{2}}$ und Anwendung von Lemma 2.1.6 ergeben

$$f_k(\hat{D}) \leq \omega(\hat{D} - 2I + f_k(\hat{D})), \quad 2 \leq k \leq N. \tag{2.2.6}$$

Eine dazu äquivalente Bedingung ist

$$f_k(\lambda) \leq \omega(\lambda - 2I + f_k(\lambda)), \quad 2 \leq k \leq N \tag{2.2.7}$$

für alle $\lambda \in \sigma(\hat{D})$. Wählt man $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ mit

$$\sigma(\hat{D}) \subseteq [\lambda_{\min}, \lambda_{\max}],$$

so sind (2.2.6) und somit (2.2.1) erfüllt.

Jetzt ersetzen wir alle f_k durch eine Funktion f_∞. Dazu benutzen wir das folgende Lemma.

Lemma 2.2.3 Für alle $k \geq 2$ und $\lambda \geq 2$ gilt

$$f_k(\lambda) \leq f_\infty(\lambda) := \left(\frac{F'_\infty(\hat{\lambda})}{F_\infty(\hat{\lambda})}\right)^2 \frac{(\lambda - \hat{\lambda})^2}{\tilde{F}_\infty(\lambda)}, \tag{2.2.8}$$

wobei

$$F_\infty(\lambda) = \frac{\lambda}{2} + \sqrt{\frac{\lambda^2}{4} - 1}, \qquad F'_\infty(\lambda) = \frac{F^2_\infty(\lambda)}{F^2_\infty(\lambda) - 1},$$

und

$$\tilde{F}_\infty(\lambda) = F_\infty(\hat{\lambda}) + F'_\infty(\hat{\lambda})(\lambda - \hat{\lambda}) = \lambda + \frac{\tilde{F}_\infty(\lambda)}{F^2_\infty(\hat{\lambda})} - \frac{2}{F_\infty(\hat{\lambda})}. \tag{2.2.9}$$

Beweis: Diese Aussagen wurden in einer ähnlichen Form in [11] bewiesen (siehe Lemmata 3.1–3.5 dort). Wir bemerken hier, dass f_∞ die Grenzfunktion der Folge $\{f_k\}_k$ bzgl. punktweiser Konvergenz für $k \to \infty$ ist. □

Mit der Hilfe dieses Lemmas reduzieren wir das System (2.2.7) auf eine Ungleichung:

Lemma 2.2.4 Für $\omega \in [0, 1)$ sei die Ungleichung

$$\forall \lambda \in [\lambda_{\min}, \lambda_{\max}] \quad g(\lambda) \leq \omega \tag{2.2.10}$$

erfüllt, wobei $g(\lambda) = \dfrac{f_\infty(\lambda)}{\lambda - 2 + f_\infty(\lambda)}$. Dann gilt (2.2.7).

Beweis: Da ω in $[0, 1)$ liegt, ist (2.2.10) zu $f_\infty(\lambda) \leq \dfrac{\omega}{1-\omega}$, $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ äquivalent. Nach Lemma 2.2.3 folgt daraus, dass für jedes $k \geq 2$ und $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ die Ungleichung $f_k(\lambda) \leq \dfrac{\omega}{1-\omega}$ gilt. Diese sind aber zu (2.2.7) äquivalent. □

Unsere Aufgabe ist also, die Funktion g auf dem Intervall $[\lambda_{\min}, \lambda_{\max}]$ von oben zu beschränken. Als ω können wir dann folgendes Maximum nehmen:

$$\omega(\hat{\lambda}, \lambda_{\min}, \lambda_{\max}) := \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} g(\lambda).$$

Da nach (2.2.8) und (2.2.9)

$$\begin{aligned} f_\infty(\lambda) &= \left(\frac{F'_\infty(\hat{\lambda})}{F_\infty(\hat{\lambda})}\right)^2 \frac{(\lambda - \hat{\lambda})^2}{\tilde{F}_\infty(\lambda)} = \frac{1}{F^2_\infty(\hat{\lambda})} \frac{(\tilde{F}_\infty(\lambda) - F_\infty(\hat{\lambda}))^2}{\tilde{F}_\infty(\lambda)} \\ &= \frac{\tilde{F}_\infty(\lambda)}{F^2_\infty(\hat{\lambda})} - \frac{2}{F_\infty(\hat{\lambda})} + \frac{1}{\tilde{F}_\infty(\lambda)} = \tilde{F}_\infty(\lambda) - \lambda + \frac{1}{\tilde{F}_\infty(\lambda)} \end{aligned}$$

gilt, erhalten wir eine einfachere Form für g:

$$\begin{aligned} g(\lambda) &= \frac{\tilde{F}_\infty(\lambda) f_\infty(\lambda)}{\tilde{F}_\infty^2(\lambda) - 2\tilde{F}_\infty(\lambda) + 1} \\ &= \left(\frac{F'_\infty(\hat{\lambda})}{F_\infty(\hat{\lambda})}\right)^2 \frac{(\lambda - \hat{\lambda})^2}{(\tilde{F}_\infty(\lambda) - 1)^2} \\ &= (\hat{g}(\lambda))^2, \end{aligned}$$

wobei

$$\hat{g}(\lambda) = \frac{F'_\infty(\hat{\lambda})}{F_\infty(\hat{\lambda})} \cdot \frac{\lambda - \hat{\lambda}}{\tilde{F}_\infty(\lambda) - 1} = \frac{F'_\infty(\hat{\lambda})}{F_\infty(\hat{\lambda})} \cdot \frac{\lambda - \hat{\lambda}}{F'_\infty(\hat{\lambda})(\lambda - \hat{\lambda}) + F_\infty(\hat{\lambda}) - 1}.$$

$\hat{g}$ ist eine gebrochen-lineare Funktion von λ und ist monoton steigend auf $(2, +\infty)$. Daraus folgt die Gleichheit

$$\omega(\hat{\lambda}, \lambda_{\min}, \lambda_{\max}) = \left(\max\{|\hat{g}(\lambda_{\min})|, |\hat{g}(\lambda_{\max})|\}\right)^2.$$

Insgesamt haben wir folgendes gezeigt:

Satz 2.2.5 Man betrachte die IBLU-Zerlegung mit den in (2.1.19), (2.1.21) und (2.1.22) definierten Diagonalblöcken für die Matrix (2.1.16–2.1.17). Für jedes $\hat{\lambda} \geq 2$ erfüllt diese Zerlegung die Abschätzung (2.2.2) mit

$$\omega = \omega(\hat{\lambda}, \lambda_{\min}, \lambda_{\max}) = \left(\frac{F'_\infty(\hat{\lambda})}{F_\infty(\hat{\lambda})}\right)^2 \max\left\{\frac{(\lambda_{\min} - \hat{\lambda})^2}{(\tilde{F}_\infty(\lambda_{\min}) - 1)^2}, \frac{(\lambda_{\max} - \hat{\lambda})^2}{(\tilde{F}_\infty(\lambda_{\max}) - 1)^2}\right\}. \tag{2.2.11}$$

Bemerkung 2.2.6 Die Abschätzung aus Satz (2.2.5) hängt von der Anzahl N der Blöcke nicht ab. Das ist eine typische Eigenschaft der IBLU-Zerlegungen. Sie folgt hier aus der Bedingung $\lambda_{\min} > 2$, die unter Voraussetzung (2.1.17) gilt. □

Bisher haben wir den Parameter $\hat{\lambda}$ beliebig aus dem Intervall $[2, +\infty)$ gewählt. Wir weisen aber darauf hin, dass bei der festen Wahl dieses Parameters die Asymptotik

$$\left(\frac{F'_\infty(\hat{\lambda})}{F_\infty(\hat{\lambda})}\right)^2 \frac{(\lambda_{\min} - \hat{\lambda})^2}{(\tilde{F}_\infty(\lambda_{\min}) - 1)^2} = 1 - O(\lambda_{\min} - 2)$$

gilt, aus der die Abschätzung

$$\omega(\hat{\lambda}, \lambda_{\min}, \lambda_{\max}) = 1 - O(\lambda_{\min} - 2)$$

folgt. Für die Diskretisierung (1.1.5–1.1.8) des Randwertproblems (1.1.3) erhalten wir damit die Konvergenzrate $1 - O(\frac{1}{n^2})$, die die gleiche Ordnung bzgl. der Blockgröße n hat, wie z.B. das Gauß-Seidel-Verfahren (obwohl die absolute Konvergenzrate der tangentialen Zerlegung für dieses Modellproblem für fast jedes $\hat{\lambda}$ besser ist). Um eine bessere Konvergenzordnung zu erhalten, muss man den Parameter $\hat{\lambda}$ abhängig

von n, also von $\lambda_{\min}$ und $\lambda_{\max}$, wählen. Das kann mit Hilfe von Satz 2.2.5 erfolgen: Wir wählen den Parameter $\hat{\lambda}$ für die vorgegebenen $\lambda_{\min}$ und $\lambda_{\max}$ so, dass der Wert $\omega(\hat{\lambda}, \lambda_{\min}, \lambda_{\max})$ aus (2.2.11) eine bessere Ordnung bzgl. $\lambda_{\min}$ hat.

Dazu ist es bequemer, die Variablen λ und $\hat{\lambda}$ durch

$$\nu = \lambda - 2 \quad \text{und} \quad x = 1 - \frac{1}{F_\infty(\hat{\lambda})} = \frac{F_\infty(\hat{\lambda}) - 1}{F_\infty(\hat{\lambda})} \tag{2.2.12}$$

zu ersetzen. Diese Transformation bildet $(\lambda, \hat{\lambda}) \in [2, +\infty) \times [2, +\infty)$ bijektiv auf $(\nu, x) \in [0, +\infty) \times [0, 1)$ ab. Dabei sind $\nu(\lambda)$ und $x(\hat{\lambda})$ monoton steigend. Damit erhalten wir:

$$\begin{aligned} \lambda - \hat{\lambda} &= \lambda - \frac{F_\infty^2(\hat{\lambda}) + 1}{F_\infty(\hat{\lambda})} = \nu - \frac{(F_\infty(\hat{\lambda}) - 1)^2}{F_\infty(\hat{\lambda})} = F_\infty(\hat{\lambda}) \left(\frac{\nu}{F_\infty(\hat{\lambda})} - \frac{(F_\infty(\hat{\lambda}) - 1)^2}{F_\infty^2(\hat{\lambda})} \right) \\ &= F_\infty(\hat{\lambda})((1 - x)\nu - x^2) \end{aligned}$$

und analog

$$\begin{aligned} \tilde{F}_\infty(\lambda) - 1 &= \frac{F_\infty^2(\hat{\lambda})}{F_\infty^2(\hat{\lambda}) - 1} \nu - \frac{F_\infty^2(\hat{\lambda})}{F_\infty^2(\hat{\lambda}) - 1} + F_\infty(\lambda) - 1 \\ &= \frac{F_\infty^2(\hat{\lambda})}{F_\infty^2(\hat{\lambda}) - 1} \nu + \frac{(F_\infty(\hat{\lambda}) - 1)^2}{F_\infty^2(\hat{\lambda}) - 1} \\ &= F'_\infty(\hat{\lambda})(\nu + x^2). \end{aligned}$$

Bzgl. der neuen Variablen schreibt sich $\hat{g}$ wie folgt:

$$\hat{g}(\nu) = \frac{(1 - x)\nu - x^2}{\nu + x^2}.$$

Das Intervall $[\lambda_{\min}, \lambda_{\max}]$ wird auf $[\nu_{\min}, \nu_{\max}]$ abgebildet mit $\nu_{\min} = \lambda_{\min} - 2$, $\nu_{\max} = \lambda_{\max} - 2$. Die Funktion $\omega(\hat{\lambda}, \lambda_{\min}, \lambda_{\max})$ schreibt sich nun als

$$\omega(x, \nu_{\min}, \nu_{\max}) = \max \left\{ \left(\frac{(1 - x)\nu_{\min} - x^2}{\nu_{\min} + x^2} \right)^2, \left(\frac{(1 - x)\nu_{\max} - x^2}{\nu_{\max} + x^2} \right)^2 \right\}. \tag{2.2.13}$$

Jetzt nehmen wir zwei Vereinfachungen vor. Zuerst bemerken wir, dass die Funktion $\hat{g}(\nu)$ steigt und

$$\lim_{\nu \to +\infty} \hat{g}(\nu) = 1 - x < 1.$$

Daher können wir $\nu_{\max}$ aus (2.2.13) eliminieren, wenn wir die zweite Schranke durch den Grenzwert $(1 - x)^2$ ersetzen. Zweitens betrachten wir nur $\hat{\lambda} \geq \lambda_{\min}$. Dadurch ist $\hat{g}(\nu_{\min})$ negativ. Damit erhalten wir eine neue Schranke $\overline{\omega}(x, \nu_{\min})$ für ω in (2.2.10):

$$\overline{\omega}(x, \nu_{\min}) := (\max\{-\hat{g}(\nu_{\min}), 1 - x\})^2 \geq \omega(x, \nu_{\min}, \nu_{\max}). \tag{2.2.14}$$

Für diese neue Abschätzung suchen wir den Parameter $x = x_{\text{opt}}(\nu_{\min})$, bei dem die beiden Schranken gleich sind:

$$-\hat{g}(\nu_{\min}) = 1 - x_{\text{opt}}(\nu_{\min}). \tag{2.2.15}$$

Die entsprechende Abschätzung wird im Folgenden mit $\omega_{\text{opt}}(\nu_{\min}) := \overline{\omega}(x_{\text{opt}}(\nu_{\min}), \nu_{\min})$ bezeichnet. Nach (2.2.14) erhält man

$$\omega_{\text{opt}}(\nu_{\min}) = (1 - x_{\text{opt}}(\nu_{\min}))^2. \tag{2.2.16}$$

Gleichung (2.2.15) lässt sich auf folgende algebraische Form reduzieren:

$$x_{\text{opt}}^3 + 2\nu_{\min} x_{\text{opt}} - 2\nu_{\min} = 0. \tag{2.2.17}$$

Anwendung der Cardanoschen Formel zeigt, dass (2.2.17) nur eine reelle Lösung hat und zwar

$$x_{\text{opt}}(\nu_{\min}) = \sqrt[3]{\nu_{\min}} \left[\sqrt[3]{1 + \sqrt{1 + \tfrac{8}{27}\nu_{\min}}} + \sqrt[3]{1 - \sqrt{1 + \tfrac{8}{27}\nu_{\min}}} \right]. \tag{2.2.18}$$

Da der Term in eckigen Klammern eine stetige, positive, monoton fallende Funktion in $\nu_{\min} \in [0, +\infty)$ ist, die für $\nu_{\min} = 0$ gleich $\sqrt[3]{2}$ ist, hat $x_{\text{opt}}(\nu_{\min})$ die folgende asymptotische Darstellung:

$$x_{\text{opt}}(\nu_{\min}) = \sqrt[3]{2}\nu_{\min}^{\frac{1}{3}} + o(\nu_{\min}^{\frac{1}{3}}),$$

was nach (2.2.16) zu

$$\omega_{\text{opt}}(\nu_{\min}) = 1 - 2\sqrt[3]{2}\nu_{\min}^{\frac{1}{3}} + o(\nu_{\min}^{\frac{1}{3}})$$

führt. Damit erhalten wir die folgende Aussage:

Satz 2.2.7 Bei der Wahl des Parameters

$$\hat{\lambda} = \hat{\lambda}_{\text{opt}}(\lambda_{\min}) = 1 - x_{\text{opt}}(\lambda_{\min} - 2) + \frac{1}{1 - x_{\text{opt}}(\lambda_{\min} - 2)}$$

mit $x_{\text{opt}}(\nu)$ aus (2.2.18) genügt die IBLU-Zerlegung mit den in (2.1.19), (2.1.21) und (2.1.22) definierten Diagonalblöcken für die Matrix (2.1.16–2.1.17) Abschätzung (2.2.2) mit

$$\omega = \omega_{\text{opt}}(\lambda_{\min} - 2) = 1 - 2\sqrt[3]{2}(\lambda_{\min} - 2)^{\frac{1}{3}} + o\left((\lambda_{\min} - 2)^{\frac{1}{3}}\right). \tag{2.2.19}$$

Für die Diskretisierung (1.1.5–1.1.8) (mit $a = b = 1$) des Laplace-Operators auf dem Einheitsquadrat erhalten wir damit Abschätzung $\omega = 1 - O\left(n^{-\frac{2}{3}}\right)$ für die Konvergenzrate. (Siehe Abbildung 2.1.)

Beweis: Die Behauptung folgt aus den vorherigen Überlegungen, wenn wir die zu (2.2.12) inverse Variablentransformation anwenden:

$$\lambda = \nu + 2, \qquad \hat{\lambda} = 1 - x + \frac{1}{1 - x}.$$

Für die Diskretisierung (1.1.5–1.1.8) des Laplace-Operators, d.h. für die Matrix (1.1.12–1.1.13) mit $a = b = 1$, ist $\nu_{\min} = 4\sin^2 \frac{\pi}{2(n+1)} = O(n^{-2})$ (siehe (1.1.14), woraus man die letzte Aussage des Satzes erhält. □

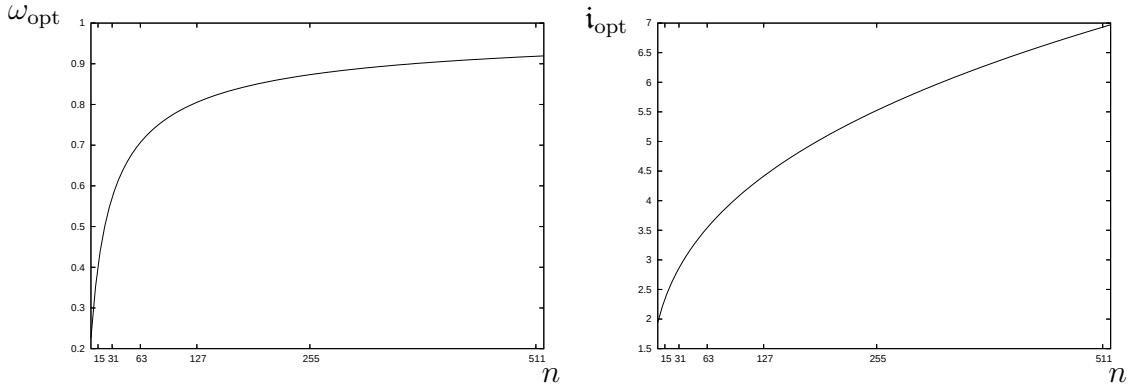

Abbildung 2.1: Konvergenzeigenschaften der tangentialen Zerlegung für die 5-Punkt-Stern-Diskretisierung des Laplace-Operators als Funktionen der Blockgröße n. Links: Die Abschätzung (2.2.16), (2.2.18) des Spektralradius der Iterationsmatrix. Rechts: Der Eigenwertindex $\mathrm{i}_{\text{opt}} = \frac{2(n+1)}{\pi}\sqrt{\arcsin\frac{\hat{\lambda}_{\text{opt}}-2}{4}}$ des optimalen Parameters $\hat{\lambda}_{\text{opt}}$.

Bemerkung 2.2.8 Interessant ist die Abhängigkeit der in Satz 2.2.7 vorgestellten Abschätzung von der Anisotropie des Problems. Wir betrachten wieder das mit dem 5-Punkt-Stern (1.1.8) diskretisiertes Problem (1.1.3–1.1.4) auf dem Einheitsquadrat. Wie es in Abschnitt 1.1 erwähnt wurde, schreibt sich die Matrix der dadurch entstehenden Gleichungssystem in Form (2.1.16–2.1.17) mit Blöcken (1.1.13). Nach (1.1.14) hat die Matrix $\hat{D} = L^{-\frac{1}{2}}DL^{-\frac{1}{2}} = \frac{1}{b}D$ das Spektrum

$$\sigma(\hat{D}) = \left\{\lambda_i = 2 + 4\tfrac{a}{b}\sin^2\tfrac{i\pi}{2(n+1)} : 1 \le i \le n\right\}.$$

(In diesem Fall ist $n = N$.) Also gilt $\lambda_{\min} = \lambda_1 = 2 + 4\frac{a}{b}\sin^2\frac{\pi}{2(n+1)} = 2 + \frac{a}{b}\cdot\frac{\pi^2}{(n+1)^2} + O((n+1)^{-4})$. Die Abschätzung (2.2.19) schreibt sich dann wie folgt:

$$\omega = 1 - 2\sqrt[3]{2\tfrac{a}{b}}\left(\frac{\pi}{n+1}\right)^{\frac{2}{3}} + O\left(\frac{1}{(n+1)^{\frac{4}{3}}}\right). \tag{2.2.20}$$

Für $a > b$ erhalten wir eine bessere Abschätzung der Konvergenzrate als für das isotrope Problem. Allerdings liegt für $\frac{a}{b} \ll 1$ Schranke (2.2.20) nahe bei 1. Das bedeutet, dass diese Abschätzung ungenau wird: Im Grenzfall, wenn $\frac{a}{b} = 0$ ist, enthält $\sigma(\hat{D})$ nur den Eigenwert $\lambda = 2$, und für die tangentiale Zerlegung mit $\hat{\lambda} = 2$ erhalten wir $\mathbf{W} = \mathbf{A}$, also ist die Konvergenzrate gleich 0. Für eine genauere Abschätzung der Konvergenzrate im Fall $\frac{a}{b} \ll 1$ reicht unsere hier vorgenommene Betrachtung einer Obermenge des Spektrums von $\hat{D}$ nicht aus. Vielmehr müsste man seine diskrete Struktur im Detail untersuchen, was wir hier aber nicht tun. Die Ordnung $\frac{1}{3}$ der Abschätzung (2.2.20) gilt aber unabhängig von dem Wert $\frac{a}{b}$. Diese Vereinfachung haben wir vorgenommen, um eine bessere Abschätzung der Konvergenzordnung zu erhalten. □

Diese Ergebnisse gelten nur im Fall des Modellproblems (2.1.16–2.1.17). Was kann man in einem allgemeineren Fall erwarten? Zuerst existieren die TFF- und

tangentialen Zerlegungen für eine viel größere Klasse von blocktridiagonalen Matrizen. Wenn eine solche Matrix positiv definit ist, sind sowohl die Vorkonditionierer $\mathbf{W}$ dieser Zerlegungen als auch die Restmatrizen $\mathbf{R}$ positiv semidefinit. Also ist die mit $\mathbf{W}$ vorkonditionierte lineare Iteration konvergent, und die Zerlegung kann auch im Verfahren der konjugierten Gradienten angewandt werden. Abschätzungen der Konvergenzraten sind für solche Fälle aber bislang unbekannt. Die hier beschriebe Theorie beruht im Wesentlichen auf der Möglichkeit, die Untersuchung der Konvergenzrate auf ein skalares Problem zu reduzieren und lässt sich daher nicht auf andere Probleme übertragen.

Die numerischen Experimente zeigen aber, dass die TFF- und tangentialen Zerlegungen auch in allgemeineren Fällen robust sind und die Konvergenzraten von dem Problem nicht sehr stark abhängen (siehe [9], [11], [41], [47]). Die theoretischen Aussagen über das Modellproblem könnten also einen Hinweis zur Wahl der Parameter geben.

Die in diesem Abschnitt von uns durchgeführte Untersuchung betrifft die Anwendung der TFF- und tangentialen Zerlegungen mit dem gleichen Parameter für alle Schritte des Lösungsverfahrens. Eine wesentliche Verbesserung der Konvergenzeigenschaften erhält man auch durch die schrittabhängige Wahl der Parameter. In der Literatur findet man zwei solche Vorgehensweisen. In der ersten wendet man eine bestimmte finite Folge von den Zerlegungen mit im voraus berechneten Parametern an, sodass alle Fourier-Moden in dem Fehler stark gedämpft sind. Dadurch erzielt man eine gitterunabhängige Konvergenzrate, wenn die Anzahl der Zerlegungen logarithmisch mit der Blockgröße n wächst (siehe [41], [9]). Im zweiten Verfahren, das als adaptive Iteration bekannt ist, wählt man den Testvektor auf jedem Schritt abhängig von den Resultaten der vorherigen (siehe [43]).

2.3 Zerlegungen auf unstrukturierten Gittern

Für die Anwendung der IBLU-Zerlegungen soll die Steifigkeitsmatrix $\mathbf{A}$ des linearen Systems (2.1.1) die blocktridiagonale Form (2.1.2) haben. Diese Struktur wird aber nicht direkt von den Diskretisierungsverfahren erzeugt. Im Allgemeinen erhält man ein System, in dem jedem Knoten eine lineare Gleichung zugeordnet ist, welche die Variablen für diesen und die benachbarten Knoten enthält. Für die Matrix $\mathbf{A}$ ist also nur der Graph bekannt, aus dem man das Muster der schwachbesetzten Matrix ablesen kann. Die geeignete Zerlegung des Gitters in Blöcke ist nicht unmittelbar aus dem Graphen erkennbar. Dazu sind wiederum spezielle Algorithmen notwendig.

Diese Situation kann wie folgt formalisiert werden. Man betrachte ein Gitter mit der Knotenmenge Ω_h. Es sei $V(\Omega_h) = \{u : \Omega_h \to \mathbb{R}\}$ der lineare Raum aller reellwertigen Gitterfunktionen auf Ω_h. Für diesen Raum führen wir die kanonische Basis $\{e_\alpha\}_{\alpha\in\Omega_h}$ ein, definiert durch

$$e_\alpha(\beta) = \begin{cases} 1, & \beta = \alpha, \\ 0, & \beta \neq \alpha. \end{cases}$$

Ist $(\cdot,\cdot)$ das euklidische Skalarprodukt in $V(\Omega_h)$, $(u,v) = \sum_{\alpha\in\Omega_h} u(\alpha)\cdot v(\alpha)$, $u, v \in$

$V(\Omega_h)$, lassen sich die Einträge des Gitteroperators $\mathbf{A}$ in dieser Basis durch

$$\mathbf{A} = (a_{\alpha,\beta})_{\alpha,\beta\in\Omega_h}, \qquad a_{\alpha,\beta} = (\mathbf{A}e_\alpha, e_\beta)$$

beschreiben. Für jede Gitterfunktion $u \in \Omega_h$ haben wir dann:

$$\mathbf{A}u = \sum_{\alpha\in\Omega_h} \left(\sum_{\beta\in\Omega_h} a_{\alpha,\beta} u(\beta) \right) e_\alpha. \tag{2.3.1}$$

Wir betrachten jetzt eine Zerlegung von Ω_h in die *Blöcke* $\Omega_h^{(1)}, \ldots, \Omega_h^{(N)}$, sodass für jedes $i \in \{1, \ldots, N\}$ die Menge $\Omega_h^{(i)}$ nicht leer ist und

$$\Omega_h^{(1)} \cup \cdots \cup \Omega_h^{(N)} = \Omega_h \quad \text{und} \quad \Omega_h^{(1)} \cap \cdots \cap \Omega_h^{(N)} = \emptyset$$

gilt. Die Matrix $\mathbf{A}$ lässt sich in der Form

$$\mathbf{A} = \begin{pmatrix} A_{11} & \ldots & A_{1N} \\ \vdots & \ddots & \vdots \\ A_{N1} & \ldots & A_{NN} \end{pmatrix} \tag{2.3.2}$$

und jede Gitterfunktion $u = \sum_{\alpha\in\Omega_h} u_\alpha e_\alpha$ in der Form

$$u = \text{blockvector}\,\{u_1, \ldots, u_N\}$$

darstellen, wobei $A_{ij} = (a_{\alpha,\beta})_{\alpha\in\Omega_h^{(i)},\beta\in\Omega_h^{(j)}}$, $u_i = (u(\alpha))_{\alpha\in\Omega_h^{(i)}}$. Mit der Definition

$$A_{ij}u_j = \sum_{\alpha\in\Omega_h^{(i)}} \left(\sum_{\beta\in\Omega_h^{(j)}} a_{\alpha,\beta} u(\beta) \right) e_\alpha$$

stimmt dann die formale Multiplikation in dieser Blockdarstellung mit Formel (2.3.1) überein. Die Zerlegung des Gitters definiert also eine Blockstruktur der Matrix und Gitterfunktionen. Wählt man $\Omega_h^{(i)}$ so, dass $A_{ij} = 0$ für $|i-j| \geq 2$, wird (2.3.2) gerade die blocktridiagonale Form (2.1.2).

In Fällen von strukturierten Gittern ist eine solche Zerlegung offensichtlich. Ein Beispiel dazu wurde von uns in Abschnitt 1.1 betrachtet. Bei unstrukturierten Gittern ist sie oft auch möglich. Es entstehen hier aber die zwei folgenden Probleme.

Erstens ist es nicht zu erwarten, dass alle $\Omega_h^{(i)}$ die gleiche Größe haben. Zwar haben wir für das Problem (2.1.1–2.1.2) verschiedene Größen n_k für die Diagonalblöcke erlaubt, aber die in Abschnitt 2.1 eingeführten tangentialen und TFF-Zerlegungen lassen sich nur im Fall konstanter Blockgrößen definieren. Zum Beispiel ist Formel (2.1.30) nur dann sinnvoll, wenn die Matrizen U_{k-1} und L_k quadratisch sind, d.h. wenn alle n_k gleich sind. Das gleiche gilt auch für die TFF-Zerlegung. Die Transferoperatoren Θ_{ij} müssen im Allgemeinen zwar nicht quadratisch sein, jedoch scheinen sich bislang alle Untersuchungen auf diesen Fall zu beschränken.

Zweitens sollen die Diagonalblöcke $\tilde{T}_k$ schwachbesetzt und z.B. durch die Gaußsche Elimination „leicht invertierbar“ sein, was aber bei gegebenen Rekursionsvorschritten (2.1.30) und (2.1.14) nicht für jede Gitterzerlegung gewährleistet ist.

In [40] setzt Wagner die FF-Zerlegungen von Wittum auf unstrukturierten Gittern erfolgreich als Glätter in Mehrgitterverfahren ein. Für die Aufspaltung des Gitters auf die Blöcke wurde hier der folgende Algorithmus benutzt:

Algorithmus 2.3.1 *(C. Wagner, [40])* Seien zu dem Gitter Ω_h und der Matrix $\mathbf{A} = (a_{\alpha,\beta})_{\alpha,\beta\in\Omega_h}$ Blöcke $\Omega_h^{(1)}, \ldots, \Omega_h^{(k)}$ bereits gefunden. Der folgende Algorithmus konstruiert den nächsten Block $\Omega_h^{(k+1)}$. Dabei wird angenommen, dass die Knoten in allen Blöcken $\Omega_h^{(i)}$, $1 \le i \le k$, geordnet sind, und der Algorithmus legt die Ordnung der Knoten im Block $\Omega_h^{(k+1)}$ fest.

BEGIN
$\Omega_h^{(k+1)} \leftarrow \emptyset$;
for all $\alpha \in \Omega_h^{(k)}$ **do**
$\Omega' \leftarrow \left\{\beta \in \Omega_h : a_{\alpha,\beta} \neq 0,\ \beta \notin \Omega_h^{(1)} \cup \cdots \cup \Omega_h^{(k)}\right\}$;
Füge alle Knoten aus Ω' in $\Omega_h^{(k+1)}$ ein, und zwar in der folgenden Reihenfolge:
1. Zuerst die Knoten, die mit den zuletzt in $\Omega_h^{(k+1)}$ eingefügten Knoten verbunden sind.
2. Für die restlichen Elemente von Ω': Knoten β vor Knoten γ, wenn β mit mehr Knoten in $\Omega_h^{(k+1)}$ verbunden ist, als γ. Falls die Ordnung daraus nicht bestimmt werden kann, dann β vor γ, wenn β mit wenigeren Knoten in $\Omega_h^{(k)}$ verbunden ist, als γ.
end for
END □

Die wiederholte Anwendung dieses Algorithmus konstruiert die Folge $\Omega_h^{(1)}, \ldots, \Omega_h^{(N)}$, bei der die Matrix $\mathbf{A}$ blocktridiagonal ist. Wir bemerken aber, dass dieses Thema noch bei weitem nicht abgeschlossen ist. C. Wagner nennt in [41] die folgenden Nachteile solcher Algorithmen, die jeden Block aus den Nachbarknoten des vorhergehenden Blocks bilden. Erstens wird die Blockstruktur durch diese Prozedur nur bis auf die Angabe des ersten Blocks definiert. Außerdem können bei Gittern, die mehrere stark verfeinerte Bereiche beinhalten, Blöcke entstehen, bei denen unabhängig von der Anordnung der Knoten die Gaußsche Elimination einen hohen Aufwand erfordert.

Für die in dieser Arbeit vorgestellten Verfahren werden etwas andere Bedingungen an die Struktur der Blöcke gestellt. Um die Blockstruktur der Matrix anzulegen, betrachten wir Methoden, die auch als Variationen von Algorithmus 2.3.1 betrachtet werden können.

Bemerkung 2.3.2 Es gibt noch eine ganz andere Vorgehensweise zur Anwendung der TFF-Zerlegungen auf unstrukturierten Gittern, nämlich die so genannten *divide-and-conquer frequenzfilternden Zerlegungen* (siehe [41]). Wir betrachten diese

Methode hier nicht, da sie nicht auf die in dieser Arbeit vorgestellten Zerlegungen anwendbar ist. □

Kapitel 3

Approximation der Diagonalblöcke mit Kettenbrüchen

In diesem Kapitel führen wir eine neue Klasse von filternden IBLU-Zerlegungen ein, die auf der Approximation der Diagonalblöcke von der vollständigen Zerlegung mit Kettenbrüchen beruht. Wir definieren zunächst allgemein diese Zerlegungen. Dann betrachten wir hauptsächlich den Fall strukturierter Gitter und erläutern für diesen insbesondere die Behandlung der Diagonalblöcke. Die Konvergenzeigenschaften dieser Zerlegungen untersuchen wir für ein Modellproblem nach der in Kapitel 2 beschriebenen Vorgehensweise. Dieses ist aber allgemeiner als Modellproblem (2.1.16–2.1.17). Die Anwendung dieser Zerlegungen auf unstrukturierte Gitter wird dann in Kapitel 4 untersucht.

3.1 Filternde Zerlegungen der Ordnung l

Wir beginnen mit der Beschreibung des Modellproblems. Dazu betrachten wir das lineare System

$$\mathbf{A}u = f \tag{3.1.1}$$

mit der blocktridiagonalen Matrix

$$\mathbf{A} = \text{blocktridiag}\left\{-L, D, -L^T\right\} = \left.\begin{pmatrix} D & -L^T & & \\ -L & D & -L^T & \\ & \ddots & \ddots & \ddots \\ & & -L & D \end{pmatrix}\right\} N \text{ Blockzeilen}, \tag{3.1.2}$$

wobei D und L Blöcke der Größe $n \times n$ sind und folgende Bedingungen erfüllen:

$$D = D^T > 0, \tag{3.1.3}$$

$$LD^{-1}L^T = L^T D^{-1} L < \tfrac{1}{4} D. \tag{3.1.4}$$

Wir weisen darauf hin, dass die Modellprobleme aus Abschnitt 1.1 dieser Klasse gehören. Das Problem (2.1.16–2.1.17) ist ein Spezialfall von (3.1.2–3.1.4) mit symmetrischem, positiv definitem L.

Wie wir in Lemma 3.3.2 (siehe Seite 52) zeigen, ist die Matrix $\mathbf{A}$ wegen der Bedingungen (3.1.3–3.1.4) positiv definit. Sie besitzt also die vollständige Blockzerlegung

$$\mathbf{A} = (\mathbf{L} + \mathbf{T})\mathbf{T}^{-1}(\mathbf{T} + \mathbf{L}^T) \tag{3.1.5}$$

mit $\mathbf{L} = \text{blocktridiag}\{-L, 0, 0\}$ und $\mathbf{T} = \text{blockdiag}\{T_k\}$, wobei die Blöcke $T_k \in \mathbb{R}^{n\times n}$ $(1 \le k \le N)$ durch die Rekursion

$$T_k = \begin{cases} D, & k = 1, \\ D - LT_{k-1}^{-1}L^T, & k \ge 2 \end{cases} \tag{3.1.6}$$

gegeben sind.

Nach (3.1.3) existiert die Matrix $D^{-\frac{1}{2}}$. Wir definieren $C_k := D^{-\frac{1}{2}}T_kD^{-\frac{1}{2}}$. Nach (3.1.6) erhalten wir:

$$\begin{aligned} C_1 &= I \quad \text{und} \\ C_k &= I - D^{-\frac{1}{2}}LD^{-\frac{1}{2}}C_{k-1}^{-1}D^{-\frac{1}{2}}L^TD^{-\frac{1}{2}} = I - BC_{k-1}^{-1}B^T, \quad k \ge 2, \end{aligned} \tag{3.1.7}$$

mit

$$B = D^{-\frac{1}{2}}LD^{-\frac{1}{2}}. \tag{3.1.8}$$

Multiplikation der Gleichung in (3.1.4) mit $D^{-\frac{1}{2}}$ von links und rechts liefert

$$BB^T = B^TB,$$

d.h. B ist eine normale Matrix. Folglich besitzen B und B^T ein gemeinsames System von Eigenvektoren. Aus (3.1.7) folgt, dass B und B^T mit allen C_k kommutieren. Es gilt also

$$C_k = \begin{cases} I, & k = 1, \\ I - BB^TC_{k-1}^{-1}, & k \ge 2. \end{cases}$$

Daraus erhalten wir, dass jeder Block C_k eine gebrochen-rationale Funktion von BB^T ist:

$$C_k = \tau_k(BB^T),$$

wobei

$$\tau_k(\mu) = \begin{cases} 1, & k = 1, \\ 1 - \dfrac{\mu}{\tau_{k-1}(\mu)}, & k \ge 2. \end{cases} \tag{3.1.9}$$

Wir haben also gezeigt, dass

$$T_k = D^{\frac{1}{2}}\tau_k(BB^T)D^{\frac{1}{2}}. \tag{3.1.10}$$

Nach (3.1.4) gilt $BB^T < \frac{1}{4}$, und somit liegt das Spektrum von BB^T in $[0, \frac{1}{4})$. Wir interessieren uns daher für die Einschränkungen der Funktionen τ_k auf dieses Intervall. Der folgende Satz beschreibt die Struktur von τ_k.

Satz 3.1.1 Für jedes $k \in \{1, \ldots, N\}$ gilt:

$$\tau_k(\mu) = \frac{Q_{k+1}(\mu)}{Q_k(\mu)}, \tag{3.1.11}$$

wobei die Polynome Q_k der Rekursionsformel

$$Q_{k+2}(\mu) = Q_{k+1}(\mu) - \mu Q_k(\mu), \tag{3.1.12}$$
$$Q_1(\mu) = Q_2(\mu) = 1 \tag{3.1.13}$$

genügen. Der Grad des Polynoms Q_k ist

$$\deg Q_k = \left\lfloor \tfrac{k-1}{2} \right\rfloor. \tag{3.1.14}$$

Bei ungeraden k sind also sowohl Zähler als auch Nenner von τ_k vom gleichen Grad $\frac{k-1}{2}$. Bei geraden k hat der Zähler den Grad $\frac{k}{2}$ und der Nenner den Grad $\frac{k}{2}-1$.

Beweis: Wir betrachten die Funktion τ_k aus (3.1.11–3.1.13). Nach (3.1.13) gilt $\tau_1(\mu) = 1$. Aus (3.1.12) erhalten wir:

$$\tau_{k+1}(\mu) = \frac{Q_{k+1}(\mu) - \mu Q_k(\mu)}{Q_{k+1}(\mu)} = 1 - \frac{\mu}{\tau_k(\mu)}.$$

Also genügen die in (3.1.11) definierten Funktionen τ_k der Rekursionsformel (3.1.9). Gleichung (3.1.14) folgt induktiv aus (3.1.12–3.1.13). □

Analog zur Vorgehensweise aus Abschnitt 2.1 erhält man eine unvollständige Blockzerlegung, wenn die Funktionen τ_k durch Approximierende $\tilde{\tau}_k$ ersetzt werden. Wir approximieren die gebrochen-rationale Funktion τ_k durch eine solche mit beschränktem, von k unabhängigem Zähler- und Nennergrad. Wir wählen dazu eine ganze Zahl $l \geq 0$ und suchen eine Approximierende

$$\tilde{\tau}_k^{(l)}(\mu) = \frac{\tilde{P}_k^{(l)}(\mu)}{\tilde{Q}_k^{(l)}(\mu)} \tag{3.1.15}$$

mit der Eigenschaft

$$\deg \tilde{P}_k^{(l)} \leq \left\lceil \frac{l}{2} \right\rceil, \qquad \deg \tilde{Q}_k^{(l)} \leq \left\lfloor \frac{l}{2} \right\rfloor \tag{3.1.16}$$

für jedes $k \in \{1, \ldots, N\}$.

Nach Satz 3.1.1 erfüllen die Funktionen $\tau_1, \ldots, \tau_{l+1}$ schon Bedingung (3.1.16). Wir nehmen daher

$$\tilde{\tau}_k^{(l)} = \tau_k, \quad 1 \leq k \leq l+1. \tag{3.1.17}$$

Bei größerem k sind Zähler und Nenner von τ_k Polynome von höherem Grad, als in (3.1.16) erlaubt. Wir versuchen, τ_k durch $\tilde{\tau}_k^{(l)}$ mit

$$\tilde{P}_k^{(l)}(\mu) = \sum_{i=0}^{\lceil l/2 \rceil} \tilde{p}_{k,i}^{(l)} \mu^i, \quad \tilde{Q}_k^{(l)}(\mu) = \sum_{i=0}^{\lfloor l/2 \rfloor} \tilde{q}_{k,i}^{(l)} \mu^i \tag{3.1.18}$$

möglichst gut zu approximieren. Da die $l+2$ Koeffizienten $\tilde{p}_{k,i}^{(l)}$ und $\tilde{q}_{k,i}^{(l)}$ bis auf einem skalaren Faktor bestimmt werden sollen, brauchen wir hier $l+1$ Bedingungen, um $\tilde{\tau}_k^{(l)}$ festzulegen. Dazu wählen wir $l+1$ Punkte

$$\hat{\mu}_0 \leq \hat{\mu}_1 \leq \cdots \leq \hat{\mu}_l, \qquad \hat{\mu}_i \in [0, \tfrac{1}{4}).$$

Diese liegen also in der Nähe der Eigenwerte der Matrix BB^T und stellen für jedes k ein Interpolationsproblem

$$\tilde{\tau}_k^{(l)}(\hat{\mu}_i) = \tau_k(\hat{\mu}_i), \quad i \in \{0, \ldots, l\}. \tag{3.1.19}$$

Für den Fall, dass genau $s+1$ Punkte $\hat{\mu}_i, \ldots, \hat{\mu}_{i+s}$ gleich sind, fordern wir, dass τ_k und $\tilde{\tau}_k^{(l)}$ an dieser Stelle auch bis zur s-ten Ableitung übereinstimmen:

$$\frac{d^r}{d\mu^r}\tilde{\tau}_k^{(l)}(\hat{\mu}_i) = \frac{d^r}{d\mu^r}\tau_k(\hat{\mu}_i), \quad 0 \le r \le s. \tag{3.1.20}$$

Die Funktion $\tilde{\tau}_k^{(l)}$ heißt *Padé Approximation* für τ_k. Die Theorie solcher Approximation wird detailliert in [3] besprochen. Es wird dort z.B. gezeigt, dass diese Approximation besser als die polynomiale ist. Eine kurze Beschreibung der Existenz und der Eigenschaften der Lösung rationaler Interpolationsprobleme findet man auch in [33]. Wir betrachten einige theoretischen Fragen zu dieser Approximation weiter unten und beschäftigen uns jetzt mit ihrer Anwendung auf die Konstruktion von unvollständigen Zerlegungen.

Für jedes fest gewählte l betrachte man eine Blockdreieckszerlegung

$$\mathbf{A} = \left(\mathbf{L} + \tilde{\mathbf{T}}^{(l)}\right)\left(\tilde{\mathbf{T}}^{(l)}\right)^{-1}\left(\tilde{\mathbf{T}}^{(l)} + \mathbf{U}\right) - \mathbf{R}^{(l)}, \tag{3.1.21}$$

mit $\tilde{\mathbf{T}}^{(l)} = \text{blockdiag}\left\{\tilde{T}_k^{(l)}\right\}$, wobei

$$\tilde{T}_k^{(l)} = D^{\frac{1}{2}}\tilde{\tau}_k^{(l)}(BB^T)D^{\frac{1}{2}}. \tag{3.1.22}$$

Die Behandlung dieser Blöcke $\tilde{T}_k^{(l)}$ bedarf einer speziellen Erklärung. Im einfachsten Fall $l = 0$ sind die Funktionen $\tilde{\tau}_k^{(0)}$ konstant:

$$\tilde{\tau}_k^{(0)}(\mu) = \theta_k^{(0)},$$

wobei nach Bedingung (3.1.19) gilt: $\theta_k^{(0)} = \tau_k(\hat{\mu}_0)$. Daraus ergibt sich die Formel für die Diagonalblöcke der Zerlegung im Falle des Modellproblems:

$$\tilde{T}_k^{(0)} = \theta_k^{(0)} D. \tag{3.1.23}$$

Wenn D „leicht invertierbar“ ist, so gilt dies auch für die Blöcke (3.1.23). Für $l \ge 1$ enthalten die $\tilde{T}_k^{(l)}$ Faktoren D^{-1} oder Inverse von noch komplizierteren Matrizen, sind also nicht mehr schwachbesetzt. Die Auswertung dieser Blöcke wäre so aufwändig wie bei der vollständigen Zerlegung, und die betrachteten Verfahren wären ineffizient. Es gibt aber eine andere Methode, die auf der folgenden Darstellung der Funktion $\tilde{\tau}_k^{(l)}$ als Kettenbruch beruht: Für $k \ge l+2$ gilt

$$\tilde{\tau}_k^{(l)} = \hat{\tau}_k^{(l)}, \quad \text{wobei} \quad \hat{\tau}_k^{(i)}(\mu) = \begin{cases} \theta_k^{(0)}, & i = 0, \\ \theta_k^{(i)} - \dfrac{\mu}{\hat{\tau}_{k-1}^{(i-1)}(\mu)}, & i \ge 1 \end{cases} \tag{3.1.24}$$

mit $\theta_j^{(i)} \in \mathbb{R}$. Wir weisen darauf hin, dass gemäß (3.1.9) und (3.1.17) für $k \in \{1, \ldots, l+1\}$ die Funktionen $\tilde{\tau}_k^{(l)}$ diese Form bereits haben. Für $k \geq l+2$ kann man dazu eine ähnliche Methode benutzen wie bei der Berechnung des rationalen Interpolierenden mit Kettenbrüchen (siehe [33]). Wir zerlegen $\hat{\tau}_k^{(l)}(\mu) = \dfrac{\hat{P}_k^{(l)}(\mu)}{\hat{Q}_k^{(l)}(\mu)}$ in zwei Summanden

$$\hat{\tau}_k^{(l)}(\mu) = \hat{\tau}_k^{(l)}(0) - \left(\frac{\hat{P}_k^{(l)}(0)}{\hat{Q}_k^{(l)}(0)} - \frac{\hat{P}_k^{(l)}(\mu)}{\hat{Q}_k^{(l)}(\mu)} \right) = \theta_k^{(l)} - \frac{\frac{\hat{P}_k^{(l)}(0)}{\hat{Q}_k^{(l)}(0)} \hat{Q}_k^{(l)}(\mu) - \hat{P}_k^{(l)}(\mu)}{\hat{Q}_k^{(l)}(\mu)}$$

mit $\theta_k^{(l)} = \hat{\tau}_k^{(l)}(0)$. Da der Zähler des letzten Bruchs die Nullstelle $\mu = 0$ hat, können wir ihn als Produkt $\mu P(\mu)$ mit einem Polynom P vom Grad $\lfloor \frac{l-1}{2} \rfloor$ darstellen:

$$\hat{\tau}_k^{(l)}(\mu) = \theta_k^{(l)} - \frac{\mu P(\mu)}{\hat{Q}_k^{(l)}(\mu)} = \theta_k^{(l)} - \frac{\mu}{\hat{\tau}_{k-1}^{(l-1)}(\mu)},$$

wobei $\hat{\tau}_{k-1}^{(l-1)}(\mu) = \dfrac{\hat{Q}_k^{(l)}(\mu)}{P(\mu)}$. Die rekursive Anwendung dieser Prozedur führt schließlich zu (3.1.24), falls alle daraus entstehenden Funktionen $\hat{\tau}_j^{(i)}(\mu)$ in $\mu = 0$ wohldefiniert sind.

Die zentrale Idee des Verfahrens ist, mit Hilfe der Darstellung (3.1.24) die Blöcke $\tilde{T}_k^{(l)}$ aus (3.1.22) für $k \geq l+2$ rekursiv zu definieren:

$$\begin{aligned} \tilde{T}_k^{(l)} &= D^{\frac{1}{2}} \hat{\tau}_k^{(l)}(BB^T) D^{\frac{1}{2}} = \theta_k^{(l)} D - D^{\frac{1}{2}} BB^T \left(\hat{\tau}_{k-1}^{(l-1)}(BB^T) \right)^{-1} D^{\frac{1}{2}} \\ &= \theta_k^{(l)} D - D^{\frac{1}{2}} B \left(\hat{\tau}_{k-1}^{(l-1)}(BB^T) \right)^{-1} B^T D^{\frac{1}{2}} \\ &= \theta_k^{(l)} D - L \left(\tilde{T}_{k-1}^{(l-1)} \right)^{-1} L^T, \end{aligned}$$

also

$$\tilde{T}_k^{(l)} = \begin{cases} \theta_k^{(0)} D, & l = 0, \\ \theta_k^{(l)} D - L \left(\tilde{T}_{k-1}^{(l-1)} \right)^{-1} L^T, & l \geq 1, \end{cases} \tag{3.1.25}$$

wobei $\theta_j^{(i)} \in \mathbb{R}$. Diese Blöcke sind vollständig durch die Koeffizienten $\theta_k^{(0)}, \ldots, \theta_k^{(l)}$ bestimmt. Da für $k \in \{1, \ldots, l+1\}$ die Funktionen $\tilde{\tau}_k^{(l)}$ und τ_k gleich sind, erhalten wir

$$\tilde{T}_k^{(l)} = T_k, \quad 1 \leq k \leq l+1,$$

und diese Blöcke werden durch Rekursionsformel (3.1.6) definiert. Für $k = l+1$ sind (3.1.25) und (3.1.6) allerdings identisch, da die Funktion $\tilde{\tau}_{l+1}^{(l)} = \tau_{l+1}$ sich auch in Form (3.1.24) mit Koeffizienten

$$\theta_1^{(0)} = \ldots = \theta_{l+1}^{(l)} = 1 \tag{3.1.26}$$

schreibt. Gleichung (3.1.26) stimmt mit (3.1.19–3.1.20) für jede Wahl von $\hat{\mu}_i$ überein.

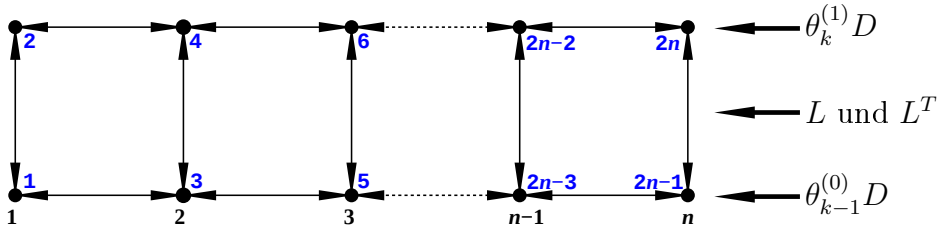

Abbildung 3.1: Der Graph des Systems (3.1.30). Die schwarzen Zahlen in der unteren Reihe geben die Nummerierung in jedem Gitterblock an. Die blauen Zahlen zeigen die neue Anordnung der Knoten, unter der die Systemmatrix die geforderte Struktur hat.

Wir verwenden (3.1.25) für die wichtigste Operation in Algorithmus 2.1.3, das Lösen von Systemen der Form

$$\tilde{T}_k^{(l)} u_k = \tilde{f}_k. \tag{3.1.27}$$

Und zwar erhalten wir u_k für $k \geq l+2$ aus der Lösung des folgenden Gleichungssystems:

$$\begin{pmatrix} \theta_{k-l}^{(0)}D & -L^T & & \\ -L & \theta_{k-l+1}^{(1)}D & \ddots & \\ & \ddots & \ddots & -L^T \\ & & -L & \theta_k^{(l)}D \end{pmatrix} \begin{pmatrix} \hat{u}_{k-l} \\ \vdots \\ \hat{u}_{k-1} \\ u_k \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ f_k^{(l)} \end{pmatrix}. \tag{3.1.28}$$

Nach sukzessiver Elimination der Vektorblöcke $\hat{u}_{k-l}, \ldots, \hat{u}_{k-1}$ erhalten wir das ursprüngliche System (3.1.27). Die Matrix des Systems (3.1.28) ist schwachbesetzt und besteht nur aus den skalierten Blöcken D und L. Die entsprechenden Systeme für $k \in \{1, \ldots, l+1\}$ haben die Form

$$\left.\begin{pmatrix} D & -L^T & & \\ -L & D & \ddots & \\ & \ddots & \ddots & -L^T \\ & & -L & D \end{pmatrix} \begin{pmatrix} \hat{u}_0 \\ \vdots \\ \hat{u}_{k-1} \\ u_k \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ f_k^{(l)} \end{pmatrix}\right\} k \text{ Blockzeilen.} \tag{3.1.29}$$

Die Systeme (3.1.28–3.1.29) haben für schwachbesetzte Matrizen $\mathbf{A}$ eine einfache Struktur und lassen sich relativ effizient lösen. Wir betrachten nun konkret den Fall $l = 1$, in dem (3.1.28) die Form

$$\begin{pmatrix} \theta_{k-1}^{(0)}D & -L^T \\ -L & \theta_k^{(1)}D \end{pmatrix} \begin{pmatrix} \hat{u}_k \\ u_k \end{pmatrix} = \begin{pmatrix} 0 \\ \tilde{f}_k \end{pmatrix} \tag{3.1.30}$$

hat. Falls D tridiagonal und L diagonal ist, wie bei dem 5-Punkt-Stern (1.1.8), hat die Matrix dieses Systems den in Abbildung 3.1 gezeigten Graph. Unter der speziellen, in Abbildung 3.1 gezeigten Anordnung der Knoten besitzt diese Matrix die blocktridiagonale Struktur mit Blöcken der Größe 2×2. Dieses System kann

also mit der Blockvariante der Gaußschen Elimination (Algorithmus 2.1.1) gelöst werden, die in diesem Fall sehr effizient ist. Für ein beliebiges l würden wir dann ein blocktridiagonales System aus den Blöcken $(l+1)\times(l+1)$ erhalten. Für praktische Anwendungen sind diese Zerlegungen nur sinnvoll, wenn l nicht sehr groß ist, also 1 oder 2. Wir zeigen, dass bereits diese Werte von l zu einer wesentlichen Verbesserung der Konvergenzeigenschaften führen.

Wir haben bisher nur das Modellproblem betrachtet. Die Rekursionsformel (3.1.25) lässt sich aber leicht auf eine beliebige blocktridiagonale Matrix übertragen. Dies ermöglicht, die hier für das Modellproblem eingeführten Zerlegungen im allgemeinen Fall zu betrachten:

Definition 3.1.2 Es sei

$$\mathbf{A} = \text{blocktridiag}\,\{-L_k, D_k, -U_k\} = \mathbf{L} + \mathbf{D} + \mathbf{U}$$

mit

$$\begin{aligned} \mathbf{L} &= \text{blocktridiag}\,\{-L_k, 0, 0\}, \\ \mathbf{U} &= \text{blocktridiag}\,\{0, 0, -U_k\}, \\ \mathbf{D} &= \text{blockdiag}\,\{D_k\}. \end{aligned}$$

Sei $l \geq 0$ beliebig, aber fest gewählt. Sei $\tilde{\mathbf{T}}^{(l)} = \text{blockdiag}\left\{\tilde{T}_k^{(l)}\right\}$ mit rekursiv über k definierten Blöcken

$$\tilde{T}_k^{(l)} = \begin{cases} D_1, & k = 1, \\ D_k - L_k \left(\tilde{T}_{k-1}^{(l)}\right)^{-1} U_{k-1}, & k \geq 2 \end{cases} \tag{3.1.31}$$

für $k \in \{1, \ldots, l+1\}$ und

$$\tilde{T}_k^{(l)} = \begin{cases} \theta_k^{(0)} D_k, & l = 0, \\ \theta_k^{(l)} D_k - L_k \left(\tilde{T}_{k-1}^{(l-1)}\right)^{-1} U_{k-1}, & l \geq 1 \end{cases} \tag{3.1.32}$$

für $k \geq l+2$. Dann nennen wir die IBLU-Zerlegung

$$\mathbf{A} = \mathbf{W}^{(l)} - \mathbf{R}^{(l)} \quad \text{mit} \quad \mathbf{W}^{(l)} = \left(\mathbf{L} + \tilde{\mathbf{T}}^{(l)}\right)\left(\tilde{\mathbf{T}}^{(l)}\right)^{-1}\left(\tilde{\mathbf{T}}^{(l)} + \mathbf{U}\right) \tag{3.1.33}$$

verallgemeinerte IBLU-Zerlegung der Ordnung l zu den Koeffizienten $\theta_k^{(i)}$. Des Weiteren bezeichnen wir diese Zerlegungen auch mit GIBLU(l) (Abkürzung für *„Generalized IBLU of order l“*), falls die Wahl der Parameter klar ist.

Wir weisen darauf hin, dass wir in dieser Definition keine weiteren Bedingungen an die Struktur der Blöcke von $\mathbf{A}$ stellen. Die Diagonalblöcke D_k können auch unterschiedliche Größen haben. Die Anwendung von diesen Zerlegungen im Fall unstrukturierter Gitter ist also auch möglich.

Für die spezielle Wahl $\theta_k^{(i)} = 1$ für alle $i \geq 1$ ist die GIBLU(l)-Zerlegung gerade die IBLU(l)-Zerlegung aus [10]. Im Allgemeinen führt eine andere Wahl dieser Parameter zu besseren Konvergenzeigenschaften.

Wir betrachten jetzt die wichtigsten Fälle von GIBLU(l)-Zerlegungen, nämlich solche der Ordnung 0, 1 und 2. Die Zerlegungen für $l = 0$ haben die sehr einfache Struktur:

$$\tilde{T}_k^{(0)} = \theta_k^{(0)} D_k$$

und brauchen daher keine spezielle Behandlung. Diese Zerlegung existiert, wenn $\theta_k^{(0)} \neq 0$ für alle k ist. Wählt man alle Koeffizienten $\theta_k^{(0)}$ gleich, so ist die Matrix $\mathbf{W}$ dieser Zerlegung der Vorkonditionierer des Block-SSOR-Verfahrens (siehe [18]). Wie wir oben schon erwähnt haben, ist im Falle des Modellproblems $\tilde{T}_k^{(0)} = \tau_k(\hat{\mu}_0) D$, wobei der Punkt $\hat{\mu}_0 \in [0, \frac{1}{4})$ als Parameter gewählt werden soll.

Die GIBLU(1)-Zerlegung hat für $k \geq 2$ die Diagonalblöcke

$$\tilde{T}_k^{(1)} = \theta_k^{(1)} D_k - \frac{1}{\theta_{k-1}^{(0)}} L_k D_{k-1}^{-1} U_{k-1} \tag{3.1.34}$$

(mit $\theta_1^{(0)} = \theta_2^{(1)} = 1$), die durch die Koeffizienten $\theta_k^{(1)}$ und $\theta_k^{(0)}$ bestimmt sind. Die Idee zur Invertierung solcher Matrizen ist im wesentlichen die gleiche wie die zur Behandlung des Systems (3.1.30).

Die GIBLU(2)-Zerlegung erfolgt auf ähnliche Weise. Die Diagonalblöcke haben für $k \geq 3$ die Form

$$\tilde{T}_k^{(2)} = \theta_k^{(2)} D_k - L_k \left(\theta_{k-1}^{(1)} D_{k-1} - \frac{1}{\theta_{k-1}^{(0)}} L_{k-1} D_{k-2}^{-1} U_{k-2} \right)^{-1} U_{k-1}.$$

Zur Lösung der Gleichungen $\tilde{T}_k^{(2)} x_k = \hat{f}_k$ werden also die Hilfssysteme

$$\begin{pmatrix} \theta_{k-2}^{(0)} D_{k-2} & U_{k-2} & \\ L_{k-1} & \theta_{k-1}^{(1)} D_{k-1} & U_{k-1} \\ & L_k & \theta_k^{(2)} D_k \end{pmatrix} \begin{pmatrix} \hat{u}_k^{(2)} \\ \hat{u}_k^{(1)} \\ u_k \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \hat{f}_k \end{pmatrix}$$

betrachtet. Die Struktur dieser Systeme ist in vielen Fällen recht einfach, z.B. für die 5-Punkt-Diskretisierung (1.1.8). In diesem Fall ist die Systemmatrix bei entsprechender Anordnung der Knoten blocktridiagonal mit Blöcken der Größe 3×3.

3.2 Existenz der GIBLU(l)-Zerlegungen im Fall des Modellproblems

Im Folgenden zeigen wir die Existenz der GIBLU(1)- und GIBLU(2)-Zerlegungen für das Modellproblem (3.1.1–3.1.4). Wir benötigen den folgenden Satz (siehe auch [10]):

Satz 3.2.1 1. Die durch Rekursionsformel (3.1.9) definierten Funktionen τ_k sind für alle $\mu \in [0, \frac{1}{4})$ wohldefiniert, und für sie gilt

$$\tau_k(\mu) = \tau_\infty(\mu) \cdot \frac{1 - \left(\frac{1 - \tau_\infty(\mu)}{\tau_\infty(\mu)} \right)^{k+1}}{1 - \left(\frac{1 - \tau_\infty(\mu)}{\tau_\infty(\mu)} \right)^{k}}, \tag{3.2.1}$$

wobei $\tau_\infty(\mu) = \frac{1}{2} + \sqrt{\frac{1}{4} - \mu}$. Für jedes $k \geq 2$ ist die Funktion τ_k positiv, monoton fallend und konkav auf $[0, \frac{1}{4})$.

2. Die Ableitung $\tau_k'(\mu)$ existiert für alle $\mu \in [0, \frac{1}{4})$ und ist differenzierbar. Sie genügt der Rekursion

$$\tau_k'(\mu) = \begin{cases} 0, & k = 1, \\ \dfrac{\mu\tau_{k-1}'(\mu) - \tau_{k-1}(\mu)}{(\tau_{k-1}(\mu))^2}, & k \geq 2. \end{cases} \tag{3.2.2}$$

Für $k \geq 2$ ist $\tau_k'(\mu)$ negativ und nicht-steigend, für $k \geq 3$ sogar monoton fallend.

3. Für jedes $\mu \in [0, \frac{1}{4})$ ist die Folge $\{\tau_k(\mu)\}_k$ nicht-steigend und konvergiert gegen $\tau_\infty(\mu)$. Die Folgen $\{\tau_k'(\mu)\}_k$ und $\{\tau_k''(\mu)\}_k$ konvergieren für jedes $\mu \in [0, \frac{1}{4})$ gegen $\tau_\infty'(\mu)$ und $\tau_\infty''(\mu)$, respektive.

Beweis: Die Rekursionsgleichung (3.1.12) hat die allgemeine Lösung $Q_k = \alpha q_1^k + \beta q_2^k$, wobei $q_{1,2} = \frac{1}{2} \pm \sqrt{\frac{1}{4} - \mu}$. Die Koeffizienten α und β erhält man aus den Anfangsbedingungen (3.1.13):

$$\begin{aligned} Q_k(\mu) &= \tfrac{1}{2\sqrt{\frac{1}{4}-\mu}} \left[\left(\tfrac{1}{2} + \sqrt{\tfrac{1}{4} - \mu} \right)^k - \left(\tfrac{1}{2} - \sqrt{\tfrac{1}{4} - \mu} \right)^k \right] \\ &= \tfrac{1}{2\sqrt{\frac{1}{4}-\mu}} \left[(\tau_\infty(\mu))^k - (1 - \tau_\infty(\mu))^k \right]. \end{aligned}$$

Für $\mu \in [0, \frac{1}{4})$ liegt $\tau_\infty(\mu)$ in $(\frac{1}{2}, 1]$. Also sind $Q_k(\mu)$ für alle $k \geq 1$ positiv. Folglich sind wegen (3.1.11) die durch Rekursionsformel (3.1.9) definierten Funktionen τ_k auf $[0, \frac{1}{4})$ wohldefiniert und positiv, und für sie gilt

$$\tau_k(\mu) = \frac{(\tau_\infty(\mu))^{k+1} - (1 - \tau_\infty(\mu))^{k+1}}{(\tau_\infty(\mu))^k - (1 - \tau_\infty(\mu))^k} = \tau_\infty(\mu) \cdot \frac{1 - \left(\frac{1-\tau_\infty(\mu)}{\tau_\infty(\mu)} \right)^{k+1}}{1 - \left(\frac{1-\tau_\infty(\mu)}{\tau_\infty(\mu)} \right)^k}.$$

Differentiation von (3.1.9) liefert (3.2.2) und

$$\tau_k''(\mu) = \begin{cases} 0, & k = 1, \\ \dfrac{2\tau_{k-1}'(\mu) + \mu\tau_{k-1}''(\mu)}{(\tau_{k-1}(\mu))^2} - \dfrac{2\mu \left(\tau_{k-1}'(\mu)\right)^2}{(\tau_{k-1}(\mu))^3}, & k \geq 2. \end{cases} \tag{3.2.3}$$

Da sowohl (3.2.2) als auch (3.2.3) linear in $\tau_k'(\mu)$ und $\tau_k''(\mu)$ sind, definieren sie diese Ableitungen wohl auf dem ganzen Halbintervall $[0, \frac{1}{4})$. Wir untersuchen nun die Vorzeichen von $\tau_k'(\mu)$ und $\tau_k''(\mu)$ in jedem Punkt $\mu \in [0, \frac{1}{4})$ und für $k \geq 2$. Für $k = 2$ gilt

$$\tau_2(\mu) = 1 - \mu, \quad \tau_2'(\mu) = -1, \quad \tau_2''(\mu) = 0.$$

Wir nehmen nun an, dass für ein $k \geq 2$ die Aussagen $\tau_k'(\mu) < 0$ und $\tau_k''(\mu) \leq 0$ gelten. Nach (3.2.2) und (3.2.3) erhalten wir dann: $\tau_{k+1}'(\mu) < 0$, $\tau_{k+1}''(\mu) < 0$. Also ist die erste Ableitung von τ_k auf $[0, \frac{1}{4})$ negativ für alle $k \geq 2$, und die zweite Ableitung

ist nicht-positiv. Für $k \geq 3$ ist τ_k'' sogar negativ. Das beweist, dass τ_k für $k \geq 2$ monoton fallend und konkav ist, wobei τ_k' strikt monoton fallend für $k \geq 3$ ist.

Der Wert $\sigma := \frac{1-\tau_\infty(\mu)}{\tau_\infty(\mu)}$ liegt für $\mu \in [0, \frac{1}{4})$ in $[0, 1)$. Da die Funktion $x \mapsto \frac{1-\sigma x}{1-x}$ auf $[0, 1)$ monoton steigend ist und für $x \to 0$ gegen 1 geht, konvergiert die Folge $\left\{\tau_k(\mu) = \tau_\infty(\mu) \cdot \frac{1-\sigma\cdot\sigma^k}{1-\sigma^k}\right\}_k$ für jedes $\mu \in [0, \frac{1}{4})$ nicht-steigend gegen $\tau_\infty(\mu)$. Die Konvergenz der Folgen $\{\tau_k'(\mu)\}_k$ und $\{\tau_k''(\mu)\}_k$ erhält man durch ähnliche Überlegungen. □

Im Fall des Modellproblems basieren die GIBLU(1)-Zerlegungen auf der linearen Approximation der Funktion $\tau_k(\mu)$: Nach den Bedingungen (3.1.16) gelten $\deg \tilde{P}_k^{(1)}(\mu) \leq 1$ und $\deg \tilde{Q}_k^{(1)}(\mu) = 0$, also

$$\tilde{\tau}_k^{(1)}(\mu) = \begin{cases} 1, & k = 1, \\ \theta_k^{(1)} - \dfrac{\mu}{\theta_{k-1}^{(0)}}, & k \geq 2. \end{cases} \tag{3.2.4}$$

Die Koeffizienten $\theta_k^{(1)}$ und $\theta_k^{(0)}$ sind für $k \geq 2$ durch die Bedingungen (3.1.19) oder (3.1.20) bestimmt. Wir können entweder zwei verschiedene Punkte $\hat{\mu}_0$ und $\hat{\mu}_1$ wählen und die Gleichheit von τ_k und $\tilde{\tau}_k^{(1)}$ an diesen Stellen fordern oder die Gleichheit von τ_k und $\tilde{\tau}_k^{(1)}$ in ihren Funktionswerten und in ihren ersten Ableitungen in nur einem gewählten Punkt $\hat{\mu}$. Die erste Variante liefert dann

$$\theta_k^{(1)} = \frac{\hat{\mu}_1\tau_k(\hat{\mu}_0) - \hat{\mu}_0\tau_k(\hat{\mu}_1)}{\hat{\mu}_1 - \hat{\mu}_0}, \quad \theta_{k-1}^{(0)} = \frac{\hat{\mu}_1 - \hat{\mu}_0}{\tau_k(\hat{\mu}_0) - \tau_k(\hat{\mu}_1)} \tag{3.2.5}$$

und die zweite

$$\theta_k^{(1)} = \tau_k(\hat{\mu}) - \hat{\mu}\tau_k'(\hat{\mu}), \quad \theta_{k-1}^{(0)} = -\frac{1}{\tau_k'(\hat{\mu})}. \tag{3.2.6}$$

Solche Zerlegungen existieren immer, wie folgender Satz zeigt.

Satz 3.2.2 Die Parameter (3.2.5) und (3.2.6) sind wohldefiniert für jede Wahl der Punkte $\hat{\mu}_0, \hat{\mu}_1 \in [0, \frac{1}{4})$, $\hat{\mu}_0 \neq \hat{\mu}_1$, bzw. $\hat{\mu} \in [0, \frac{1}{4})$. Die Diagonalblöcke der GIBLU(1)-Zerlegung sind in beiden Fällen positiv definit:

$$\tilde{T}_k^{(1)} > 0. \tag{3.2.7}$$

Beweis: Die Parameter (3.2.5) und (3.2.6) sind wohldefiniert, da nach Satz 3.2.1 die Ungleichungen $\tau_k(\hat{\mu}_0) \neq \tau_k(\hat{\mu}_1)$ und $\tau_k'(\hat{\mu}) \neq 0$ gelten. Da in diesem Fall $\tilde{T}_k^{(1)} = D^{\frac{1}{2}}\tilde{\tau}_k^{(1)}(BB^T)D^{\frac{1}{2}}$, folgt (3.2.7) aus der Aussage, dass auf $\sigma(BB^T) \subseteq [0, \frac{1}{4})$ die in (3.2.4) definierte Funktion $\tilde{\tau}_k^{(1)}$ positiv ist. Für $k \geq 2$ erhalten wir dies aus der Konkavität der Funktion τ_k (Satz 3.2.1). Im Fall von Parametern (3.2.6) ist $\tilde{\tau}_k^{(1)}(\mu)$ die Tangente an $\tau_k(\mu)$ im Punkt $\hat{\mu}$, und deswegen $\tilde{\tau}_k^{(1)}(\mu) \geq \tau_k(\mu) > 0$ für jedes $\mu \in [0, \frac{1}{4})$. Im Fall (3.2.5) können wir annehmen, dass $\hat{\mu}_0 < \hat{\mu}_1$ ist. Da dann $\tau_k(\hat{\mu}_0) > \tau_k(\hat{\mu}_1)$ gilt, ist $\tilde{\tau}_k^{(1)}$ monoton fallend. Daraus erhalten wir: Für $\mu \in [0, \hat{\mu}_1]$ ist $\tilde{\tau}_k^{(1)}(\mu) \geq \tau_k(\hat{\mu}_1) > 0$, und für $\mu \in (\hat{\mu}_1, \frac{1}{4})$ gilt wegen der Konkavität $\tilde{\tau}_k^{(1)}(\mu) \geq \tau_k(\mu) > 0$. □

Bemerkung 3.2.3 Die GIBLU(1)-Zerlegungen sind in gewissem Sinne analog zu den zwei-Frequenz- und filternden Zerlegungen aus [11] und [9] (siehe Abschnitt 2.1). Sie sind aber schon für das Modellproblem nicht identisch: Die Rekursionsformeln (3.1.9) und (2.1.22) definieren miteinander verwandte Funktionen, die durch folgende Transformation ineinander überführt werden:

$$F_k(\lambda) = \lambda \tau_k \left(\frac{1}{\lambda^2} \right).$$

Das Intervall $(2, +\infty)$ für λ entspricht dabei $(0, \frac{1}{4})$ für μ. Der Fall $\mu = 0$, d.h. $\lambda = \infty$, wurde in Kapitel 2 ausgeschlossen, wird aber nun auch betrachtet. Da diese Transformation nicht-linear ist, bildet sie lineare Approximierende zu F_k nicht auf solche zu τ_k ab. □

Die GIBLU(2)-Zerlegungen beruhen auf der gebrochen-rationalen Approximation von $\tau_k(\mu)$ für $k \geq 4$ durch Kettenbrüche

$$\tilde{\tau}_k^{(2)}(\mu) = \theta_k^{(2)} - \cfrac{\mu}{\theta_{k-1}^{(1)} - \cfrac{\mu}{\theta_{k-2}^{(0)}}}. \tag{3.2.8}$$

Die Koeffizienten $\theta_k^{(2)}$, $\theta_{k-1}^{(1)}$ und $\theta_{k-2}^{(0)}$ werden von den Bedingungen (3.1.19–3.1.20) durch die Wahl der drei Punkte $\hat{\mu}_0$, $\hat{\mu}_1$ und $\hat{\mu}_2$ ($\hat{\mu}_0 \leq \hat{\mu}_1 \leq \hat{\mu}_2$) bestimmt. Die Existenz dieser Zerlegungen wird in folgendem Satz untersucht.

Satz 3.2.4 Wir betrachten die Matrix (3.1.2–3.1.4). Deren GIBLU(2)-Zerlegung zu den Koeffizienten $\theta_k^{(2)}$, $\theta_k^{(1)}$ und $\theta_k^{(0)}$, die durch die Bedingungen (3.1.19–3.1.20) bestimmt werden, existiert für jede Wahl der Punkte $\hat{\mu}_0$, $\hat{\mu}_1$ und $\hat{\mu}_2$, $0 \leq \hat{\mu}_0 \leq \hat{\mu}_1 \leq \hat{\mu}_2 < \frac{1}{4}$. Dabei gilt $\tilde{T}_k^{(2)} > 0$ und die Koeffizienten $\theta_k^{(2)}$, $\theta_k^{(1)}$, $\theta_k^{(0)}$ sind positiv. Die Funktionen $\tilde{\tau}_k^{(2)}$ sind auf $[0, \frac{1}{4})$ für $k \geq 3$ monoton fallend.

Beweis: Wir beweisen für jedes $k \geq 3$ die drei folgenden Behauptungen:

(a) Die Bedingungen (3.1.19–3.1.20) definieren die Koeffizienten $\theta_k^{(2)}$, $\theta_{k-1}^{(1)}$ und $\theta_{k-2}^{(0)}$, d.h. diese existieren und sind eindeutig.

(b) Die Funktion $\tilde{\tau}_k^{(2)}$ ist auf dem ganzen Intervall $[0, \frac{1}{4})$ wohldefiniert, positiv und monoton fallend. Insbesondere folgt daraus $\tilde{T}_k^{(2)} > 0$ (siehe (3.1.22)).

(c) $\theta_k^{(2)}, \theta_{k-1}^{(1)}, \theta_{k-2}^{(0)} > 0$.

Zunächst bemerken wir, dass sich für jedes $k \geq 3$ die Funktion $\tilde{\tau}_k^{(2)}$ in der Form

$$\tilde{\tau}_k^{(2)}(\mu) = \frac{a_k \mu + b_k}{\mu + d_k} \tag{3.2.9}$$

darstellen lässt. Nach der Umformung von (3.2.8) erhalten wir

$$\tilde{\tau}_k^{(2)}(\mu) = \frac{(\theta_{k-2}^{(0)} + \theta_k^{(2)})\mu - \theta_k^{(2)} \theta_{k-1}^{(1)} \theta_{k-2}^{(0)}}{\mu - \theta_{k-1}^{(1)} \theta_{k-2}^{(0)}}.$$

Wir setzen

$$a_k = \theta_{k-2}^{(0)} + \theta_k^{(2)}, \quad b_k = -\theta_k^{(2)}\theta_{k-1}^{(1)}\theta_{k-2}^{(0)}, \quad d_k = -\theta_{k-1}^{(1)}\theta_{k-2}^{(0)}. \tag{3.2.10}$$

Weiter unten zeigen wir, dass $d_k \neq 0$. Dann gelten $\theta_{k-1}^{(1)} \neq 0$ und $\theta_{k-2}^{(0)} \neq 0$, und die Substitution (3.2.10) lässt sich invertieren:

$$\theta_k^{(2)} = \frac{b_k}{d_k}, \quad \theta_{k-2}^{(0)} = a_k - \theta_k^{(2)}, \quad \theta_{k-1}^{(1)} = -\frac{d_k}{\theta_{k-2}^{(0)}}. \tag{3.2.11}$$

Folglich sind für $d_k \neq 0$ die Existenz bzw. die Eindeutigkeit der Systeme $(\theta_k^{(2)}, \theta_{k-1}^{(1)}, \theta_{k-2}^{(0)})$ und (a_k, b_k, d_k) einander äquivalent.

Für Koeffizienten a_k, b_k, d_k betrachten wir zunächst den Fall

$$0 \leq \hat{\mu}_0 < \hat{\mu}_1 < \hat{\mu}_2 < \frac{1}{4}. \tag{3.2.12}$$

Die Bedingungen (3.1.19) lauten dann

$$\tilde{\tau}_k^{(2)}(\hat{\mu}_i) = \tau_k(\hat{\mu}_i), \quad 0 \leq i \leq 2,$$

und liefern nach (3.2.9) das lineare System

$$\hat{\mu}_i a_k - \tau_k(\hat{\mu}_i) \cdot d_k + b_k = \hat{\mu}_i \tau_k(\hat{\mu}_i), \quad 0 \leq i \leq 2, \tag{3.2.13}$$

bezüglich a_k, b_k und d_k. Die Determinante dieses Systems ist

$$\begin{aligned}
\Delta_k &= \begin{vmatrix} \hat{\mu}_0 & -\tau_k(\hat{\mu}_0) & 1 \\ \hat{\mu}_1 & -\tau_k(\hat{\mu}_1) & 1 \\ \hat{\mu}_2 & -\tau_k(\hat{\mu}_2) & 1 \end{vmatrix} = - \begin{vmatrix} \hat{\mu}_1 - \hat{\mu}_0 & \tau_k(\hat{\mu}_1) - \tau_k(\hat{\mu}_0) \\ \hat{\mu}_2 - \hat{\mu}_1 & \tau_k(\hat{\mu}_2) - \tau_k(\hat{\mu}_1) \end{vmatrix} \\
&= -(\hat{\mu}_1 - \hat{\mu}_0)(\hat{\mu}_2 - \hat{\mu}_1) \left[\frac{\tau_k(\hat{\mu}_2) - \tau_k(\hat{\mu}_1)}{\hat{\mu}_2 - \hat{\mu}_1} - \frac{\tau_k(\hat{\mu}_1) - \tau_k(\hat{\mu}_0)}{\hat{\mu}_1 - \hat{\mu}_0} \right] \\
&= -(\hat{\mu}_1 - \hat{\mu}_0)(\hat{\mu}_2 - \hat{\mu}_1) \left[\tau_k'(\xi_1) - \tau_k'(\xi_0) \right]
\end{aligned}$$

mit $\xi_0 \in (\hat{\mu}_0, \hat{\mu}_1)$, $\xi_1 \in (\hat{\mu}_1, \hat{\mu}_2)$. Da für $k \geq 3$ die Funktion τ_k' monoton fallend ist, erhalten wir: $\tau_k'(\xi_0) > \tau_k'(\xi_1)$. Daraus folgt, dass $\Delta_k > 0$ und das System (3.2.13) eindeutig lösbar ist. Dabei gilt

$$d_k = \frac{\Delta_k^{(d)}}{\Delta_k} \tag{3.2.14}$$

mit

$$\begin{aligned}
\Delta_k^{(d)} &= \begin{vmatrix} \hat{\mu}_0 & \hat{\mu}_0\tau_k(\hat{\mu}_0) & 1 \\ \hat{\mu}_1 & \hat{\mu}_1\tau_k(\hat{\mu}_1) & 1 \\ \hat{\mu}_2 & \hat{\mu}_2\tau_k(\hat{\mu}_2) & 1 \end{vmatrix} = \begin{vmatrix} \hat{\mu}_1 - \hat{\mu}_0 & \hat{\mu}_1\tau_k(\hat{\mu}_1) - \hat{\mu}_0\tau_k(\hat{\mu}_0) \\ \hat{\mu}_2 - \hat{\mu}_1 & \hat{\mu}_2\tau_k(\hat{\mu}_2) - \hat{\mu}_1\tau_k(\hat{\mu}_1) \end{vmatrix} \\
&= (\hat{\mu}_1 - \hat{\mu}_0)(\hat{\mu}_2 - \hat{\mu}_1) \left[\frac{\hat{\mu}_2\tau_k(\hat{\mu}_2) - \hat{\mu}_1\tau_k(\hat{\mu}_1)}{\hat{\mu}_2 - \hat{\mu}_1} - \frac{\hat{\mu}_1\tau_k(\hat{\mu}_1) - \hat{\mu}_0\tau_k(\hat{\mu}_0)}{\hat{\mu}_1 - \hat{\mu}_0} \right] \\
&= (\hat{\mu}_1 - \hat{\mu}_0)(\hat{\mu}_2 - \hat{\mu}_1) \left[(\mu\tau_k(\mu))'\big|_{\mu=\zeta_1} - (\mu\tau_k(\mu))'\big|_{\mu=\zeta_0} \right],
\end{aligned}$$

wobei $\zeta_0 \in (\hat{\mu}_0, \hat{\mu}_1)$, $\zeta_1 \in (\hat{\mu}_1, \hat{\mu}_2)$. Auf $[0, \frac{1}{4})$ ist die Funktion $(\mu\tau_k(\mu))'$ monoton fallend: Da

$$(\mu\tau_k(\mu))' = \tau_k(\mu) + \mu\tau_k'(\mu), \tag{3.2.15}$$

gilt $(\mu\tau_k(\mu))'' = 2\tau_k'(\mu) + \mu\tau_k''(\mu) < 0$. Analog zur Positivität von Δ_k erhalten wir daher $\Delta_k^{(d)} < 0$. Folglich gilt $d_k \neq 0$. Wie wir oben schon gezeigt haben, folgt daraus unter Annahme (3.2.12) die Wohldefinierheit der Koeffizienten $\theta_k^{(2)}$, $\theta_{k-1}^{(1)}$ und $\theta_{k-2}^{(0)}$ von Bedingung (3.1.19).

Die Koeffizienten a_k und b_k lassen sich auch leicht in der Determinantenform darstellen:

$$a_k = \frac{\Delta_k^{(a)}}{\Delta_k}, \quad b_k = \frac{\Delta_k^{(b)}}{\Delta_k}, \tag{3.2.16}$$

wobei

$$\begin{aligned}
\Delta_k^{(a)} &= \begin{vmatrix} \hat{\mu}_0\tau_k(\hat{\mu}_0) & -\tau_k(\hat{\mu}_0) & 1 \\ \hat{\mu}_1\tau_k(\hat{\mu}_1) & -\tau_k(\hat{\mu}_1) & 1 \\ \hat{\mu}_2\tau_k(\hat{\mu}_2) & -\tau_k(\hat{\mu}_2) & 1 \end{vmatrix} \\
&= (\hat{\mu}_1 - \hat{\mu}_0)(\hat{\mu}_2 - \hat{\mu}_1)\frac{\tau_k(\hat{\mu}_1) - \tau_k(\hat{\mu}_0)}{\hat{\mu}_1 - \hat{\mu}_0} \cdot \frac{\tau_k(\hat{\mu}_2) - \tau_k(\hat{\mu}_1)}{\hat{\mu}_2 - \hat{\mu}_1} \\
&\quad \times \left[\frac{\hat{\mu}_2\tau_k(\hat{\mu}_2) - \hat{\mu}_1\tau_k(\hat{\mu}_1)}{\tau_k(\hat{\mu}_2) - \tau_k(\hat{\mu}_1)} - \frac{\hat{\mu}_1\tau_k(\hat{\mu}_1) - \hat{\mu}_0\tau_k(\hat{\mu}_0)}{\tau_k(\hat{\mu}_1) - \tau_k(\hat{\mu}_0)} \right],
\end{aligned}$$

$$\begin{aligned}
\Delta_k^{(b)} &= \begin{vmatrix} \hat{\mu}_0 & -\tau_k(\hat{\mu}_0) & \hat{\mu}_0\tau_k(\hat{\mu}_0) \\ \hat{\mu}_1 & -\tau_k(\hat{\mu}_1) & \hat{\mu}_1\tau_k(\hat{\mu}_1) \\ \hat{\mu}_2 & -\tau_k(\hat{\mu}_2) & \hat{\mu}_2\tau_k(\hat{\mu}_2) \end{vmatrix} = \tau_k(\hat{\mu}_0)\tau_k(\hat{\mu}_1)\tau_k(\hat{\mu}_2) \begin{vmatrix} \frac{\hat{\mu}_0}{\tau_k(\hat{\mu}_0)} & -1 & \hat{\mu}_0 \\ \frac{\hat{\mu}_1}{\tau_k(\hat{\mu}_1)} & -1 & \hat{\mu}_1 \\ \frac{\hat{\mu}_2}{\tau_k(\hat{\mu}_2)} & -1 & \hat{\mu}_2 \end{vmatrix} \\
&= \tau_k(\hat{\mu}_0)\tau_k(\hat{\mu}_1)\tau_k(\hat{\mu}_2)(\hat{\mu}_1 - \hat{\mu}_0)(\hat{\mu}_2 - \hat{\mu}_1) \\
&\quad \times \left[\frac{\frac{\hat{\mu}_1}{\tau_k(\hat{\mu}_1)} - \frac{\hat{\mu}_0}{\tau_k(\hat{\mu}_0)}}{\hat{\mu}_1 - \hat{\mu}_0} - \frac{\frac{\hat{\mu}_2}{\tau_k(\hat{\mu}_2)} - \frac{\hat{\mu}_1}{\tau_k(\hat{\mu}_1)}}{\hat{\mu}_2 - \hat{\mu}_1} \right].
\end{aligned}$$

Da (3.2.14) und (3.2.16) nur die Differenzquotienten von $\tau_k(\mu)$, $\mu\tau_k(\mu)$ und $\mu/\tau_k(\mu)$ enthalten, und diese Funktionen auf $[0, \frac{1}{4})$ zweimal stetig differenzierbar sind, lassen sich diese Formeln auch für die Grenzfälle $\hat{\mu}_0 = \hat{\mu}_1$ und $\hat{\mu}_1 = \hat{\mu}_2$ übertragen. Falls dabei $\hat{\mu}_0 \neq \hat{\mu}_2$ gilt, zeigt man $d_k \neq 0$ wie oben.

Im Fall $\hat{\mu}_0 = \hat{\mu}_1 = \hat{\mu}_2 =: \hat{\mu}$ dividieren wir Zähler und Nenner in (3.2.14) und (3.2.16) durch $\hat{\mu}_2 - \hat{\mu}_0$. Dann konvergieren die erhaltenen Differenzquotienten in den Formeln für b_k und d_k gegen die zweiten Ableitungen der Funktionen $\tau_k(\mu)$, $\mu\tau_k(\mu)$ und $\mu/\tau_k(\mu)$. In der Formel für a_k kann $\Delta_k^{(a)}$ in der folgenden Form dargestellt

werden:

$$\begin{aligned}\Delta_k^{(a)} &= (\hat\mu_2-\hat\mu_0)(\hat\mu_1-\hat\mu_0)(\hat\mu_2-\hat\mu_1)\\ &\times\frac{\tau_k(\hat\mu_1)-\tau_k(\hat\mu_0)}{\hat\mu_1-\hat\mu_0}\cdot\frac{\tau_k(\hat\mu_2)-\tau_k(\hat\mu_1)}{\hat\mu_2-\hat\mu_1}\cdot\frac{\tau_k(\hat\mu_2)-\tau_k(\hat\mu_0)}{\hat\mu_2-\hat\mu_0}\\ &\times\frac{\dfrac{\hat\mu_2\tau_k(\hat\mu_2)-\hat\mu_1\tau_k(\hat\mu_1)}{\tau_k(\hat\mu_2)-\tau_k(\hat\mu_1)}-\dfrac{\hat\mu_1\tau_k(\hat\mu_1)-\hat\mu_0\tau_k(\hat\mu_0)}{\tau_k(\hat\mu_1)-\tau_k(\hat\mu_0)}}{\tau_k(\hat\mu_2)-\tau_k(\hat\mu_0)}.\end{aligned}\tag{3.2.17}$$

Der letzte Bruch konvergiert hier gegen die zweite Ableitung der Funktion $\mu(\tau)\cdot\tau$ bzgl. τ, wobei $\mu(\tau)$ die Inverse zur $\tau_k(\mu)$ ist. Diese Inverse ist wohldefiniert, weil τ_k streng monoton ist. Für diese Ableitung erhalten wir:

$$(\mu\tau_k(\mu))''_{\tau_k(\mu),\tau_k(\mu)}=\frac{2\tau_k'^2(\mu)-\tau_k''(\mu)}{\tau_k'^3(\mu)}.\tag{3.2.18}$$

Da weder τ_k'' noch $(\mu\tau_k(\mu))''$ auf $[0,\frac14)$ ungleich 0 sind, ist in diesem Fall auch $d_k\neq0$. Also sind für alle $\hat\mu_0,\hat\mu_1,\hat\mu_2\in[0,\frac14)$, $\hat\mu_0\le\hat\mu_1\le\hat\mu_2$, die Koeffizienten $\theta_k^{(2)}$, $\theta_{k-1}^{(1)}$ und $\theta_{k-2}^{(0)}$ eindeutig definiert. Damit ist die Behauptung (a) bewiesen.

Da die in (3.2.9) definierte Funktion $\tilde\tau_k^{(2)}$ gebrochen-linear ist, ist sie auf ganz $\mathbb{R}$ außer in einem Punkt definiert. Links und rechts von diesem Punkt hat sie das gleiche Monotonieverhalten: Je nach den Parametern a_k, b_k und d_k ist sie entweder steigend oder fallend. Des Weiteren ist die Funktion auf dem linken Intervall kleiner als auf dem rechten, wenn sie fallend ist, und sonst umgekehrt. In unserem Fall gilt $\hat\mu_0\le\hat\mu_1\le\hat\mu_2$ und dabei $\tau_k(\hat\mu_0)\ge\tau_k(\hat\mu_1)\ge\tau_k(\hat\mu_2)$. Daher kann $\tilde\tau_k^{(2)}$ nur fallend sein. Da τ_k konkav ist, liegen die Interpolationspunkte $(\hat\mu_i,\tau_k(\hat\mu_i))$ so zueinander, dass $\tilde\tau_k^{(2)}$ auch auf $[0,\frac14)$ nur konkav sein kann. Das bedeutet, dass alle drei Interpolationspunkte auf dem linken Teil des Graphen der fallenden gebrochen-linearen Funktion $\tilde\tau_k^{(2)}$ liegen. (Der rechte Teil des Graphen solcher Funktionen ist konvex.) Der Unstetigkeitspunkt $\mathring\mu$ dieser Funktion ist also rechts von $\hat\mu_2$ und $\lim\limits_{\mu\to\mathring\mu-0}\tilde\tau_k^{(2)}(\mu)=-\infty$. Wir zeigen nun, dass $\tilde\tau_k^{(2)}$ in keinem Punkt $\mu\in[\hat\mu_2,\frac14)$ kleiner als τ_k ist. Damit ist dann bewiesen, dass $\mathring\mu\ge\frac14$ und $\tilde\tau_k^{(2)}$ auf $[0,\frac14)$ positiv ist.

Wir beweisen folgende Aussage durch Induktion über k: Für $k\ge3$ ist $\tilde\tau_k^{(2)}$ auf $[0,\frac14)$ wohldefiniert und genügt der Ungleichung

$$\tilde\tau_k^{(2)}(\mu)\ge\tau_k(\mu),\quad\mu\in[\hat\mu_2,\tfrac14).\tag{3.2.19}$$

Für $k=3$ ist diese Aussage erfüllt, da $\tilde\tau_3^{(2)}=\tau_3$. Wir nehmen nun an, dass sie für $\tilde\tau_{k-1}^{(2)}$ gilt.

Wir betrachten die Funktion

$$\hat\tau_k(\mu):=1-\frac{\mu}{\tilde\tau_{k-1}^{(2)}(\mu)}.\tag{3.2.20}$$

Sie ist nach der Induktionsannahme auf dem ganzen Intervall $[0, \frac{1}{4})$ wohldefiniert. Aus (3.2.19) folgt dabei $\hat{\tau}_k(\mu) \geq \tau_k(\mu)$ für $\mu \in [\hat{\mu}_2, \frac{1}{4})$. Nach (3.2.9) erhalten wir:

$$\hat{\tau}_k(\mu) = \frac{-\mu^2 + (a_{k-1} - d_{k-1})\mu + b_{k-1}}{a_{k-1}\mu + b_{k-1}}.$$

Die Funktionen $\hat{\tau}_k$ und $\tilde{\tau}_k^{(2)}$ sind in $\hat{\mu}_0$, $\hat{\mu}_1$ und $\hat{\mu}_2$ wegen (3.1.19–3.1.20) gleich:

$$\hat{\tau}_k(\hat{\mu}_i) = \tilde{\tau}_k^{(2)}(\hat{\mu}_i), \quad i \in \{0, 1, 2\}. \tag{3.2.21}$$

(Wenn die Punkte $\hat{\mu}_i$ einander gleich sind, betrachten wir die Gleichheit der entsprechenden Ableitungen.) Der Zähler von $\hat{\tau}_k$ hat den Grad 2. Also können diese Funktionen nicht überall gleich sein. Damit ist ausgeschlossen, dass $\hat{\tau}_k$ und $\tilde{\tau}_k^{(2)}$ in einem vierten Punkt gleich sind. Sie wären sonst überall identisch (siehe [33]). Daher sind $\hat{\mu}_0$, $\hat{\mu}_1$ und $\hat{\mu}_2$ die einzigen Punkte, in denen $\hat{\tau}_k$ und $\tilde{\tau}_k^{(2)}$ (jeweils mit Vielfachheit 1) gleich sind.

Der Wert der Funktion $\tilde{\tau}_k^{(2)}$ in $\mu = 0$ ist

$$\tilde{\tau}_k^{(2)}(0) = \frac{b_k}{d_k}.$$

Wie wir oben gezeigt haben, hängen b_k und d_k, also auch $\tilde{\tau}_k^{(2)}(0)$, für jedes $\epsilon > 0$ stetig von $(\hat{\mu}_0, \hat{\mu}_1, \hat{\mu}_2)$ aus

$$M = \{(\mu_0, \mu_1, \mu_2) : \epsilon \leq \mu_0 \leq \mu_1 \leq \mu_2 \leq \tfrac{1}{4} - \epsilon\}$$

ab. Daraus folgt, dass der Wert $\tilde{\tau}_k^{(2)}(0)$ für alle Parameter $(\hat{\mu}_0, \hat{\mu}_1, \hat{\mu}_2) \in M$ entweder nur kleiner oder nur größer als 1 ist. In der Tat, wenn $\tilde{\tau}_k^{(2)}(0)$ je nach Wahl der Parameter größer und kleiner 1 sein könnte, so gäbe es $(\hat{\mu}_0, \hat{\mu}_1, \hat{\mu}_2) \in M$ mit $\tilde{\tau}_k^{(2)}(0) = 1$. Dann wären aber für diese Parameter $\tilde{\tau}_k^{(2)}$ und $\hat{\tau}_k$ in den vier Punkten 0, $\hat{\mu}_0$, $\hat{\mu}_1$ und $\hat{\mu}_2$ gleich, was unmöglich ist. Wir müssen folglich nur für eine Wahl von Parametern testen, ob $\tilde{\tau}_k^{(2)}(0)$ größer oder kleiner als 1 ist.

Dazu untersuchen wir den Fall $\hat{\mu}_0 = \hat{\mu}_1 = \hat{\mu}_2 =: \hat{\mu}$. Aus den Formeln (3.2.14), (3.2.16) erhalten wir:

$$\frac{b_k}{d_k} = -\tau_k^3(\hat{\mu})\, \frac{(\mu/\tau_k(\mu))''}{(\mu\tau_k(\mu))''}\bigg|_{\mu=\hat{\mu}} = \tau_k(\hat{\mu}) - \frac{2\hat{\mu}(\tau_k'(\hat{\mu}))^2}{2\tau_k'(\hat{\mu}) + \hat{\mu}\tau_k''(\hat{\mu})}.$$

Diese Funktion von $\hat{\mu}$ ist monoton fallend auf $[0, \frac{1}{4})$. Da für $\hat{\mu} = 0$ die Gleichung $\tilde{\tau}_k^{(2)}(0) = 1$ gilt, gibt es einen Wert $\hat{\mu} \in (0, \frac{1}{4})$, für den in diesem Fall $\tilde{\tau}_k^{(2)}(0)$ kleiner 1 ist.

Damit haben wir bewiesen, dass für alle $\hat{\mu}_0$, $\hat{\mu}_1$ und $\hat{\mu}_2$, $0 < \hat{\mu}_0 \leq \hat{\mu}_1 \leq \hat{\mu}_2 < \frac{1}{4}$,

$$\tilde{\tau}_k^{(2)}(0) < 1 = \hat{\tau}_k(0).$$

Da außerdem $\tilde{\tau}_k^{(2)}(\mu) - \hat{\tau}_k(\mu)$ nur die Nullstellen $\hat{\mu}_0$, $\hat{\mu}_1$ und $\hat{\mu}_2$ hat und diese jeweils einfach sind, müssen die folgenden Ungleichungen gelten:

$$\begin{array}{rcll} \tilde{\tau}_k^{(2)}(\mu) & < & \hat{\tau}_k(\mu) & \text{für } 0 \le \mu < \hat{\mu}_0, \\ \tilde{\tau}_k^{(2)}(\mu) & > & \hat{\tau}_k(\mu) & \text{für } \hat{\mu}_0 < \mu < \hat{\mu}_1, \\ \tilde{\tau}_k^{(2)}(\mu) & < & \hat{\tau}_k(\mu) & \text{für } \hat{\mu}_1 < \mu < \hat{\mu}_2, \\ \tilde{\tau}_k^{(2)}(\mu) & > & \hat{\tau}_k(\mu) \ge \tau_k(\mu) & \text{für } \hat{\mu}_2 < \mu < \frac{1}{4}. \end{array} \tag{3.2.22}$$

Es kann also für diese $\hat{\mu}_0$, $\hat{\mu}_1$ und $\hat{\mu}_2$ kein $\mu \in [\hat{\mu}_2, \frac{1}{4})$ mit $\tilde{\tau}_k^{(2)}(\mu) \le \tau_k(\mu)$ existieren. Da a_k, b_k und d_k stetig von $\hat{\mu}_i$ abhängen, erhalten wir aus (3.2.22)

$$\tilde{\tau}_k^{(2)}(\mu) \ge \tau_k(\mu), \quad \mu \in [\hat{\mu}_2, \tfrac{1}{4}),$$

für alle $\hat{\mu}_0, \hat{\mu}_1, \hat{\mu}_2 \in [0, \frac{1}{4})$. Das beweist die Induktionsbehauptung und somit die Aussage (b).

Wenn ein von Koeffizienten $\theta_k^{(2)}$, $\theta_{k-1}^{(1)}$, $\theta_{k-2}^{(0)}$ gleich 0 wäre, wäre $\tilde{\tau}_k^{(2)}$ in $\mu = 0$ entweder nicht definiert oder kleiner als $\tau_k(0)$. Da diese Koeffizienten stetig von $\hat{\mu}_i$ abhängen, können sie also für alle $\hat{\mu}_0, \hat{\mu}_1, \hat{\mu}_2 \in [0, \frac{1}{4})$ entweder nur positiv oder nur negativ sein. Für $\hat{\mu}_0 = \hat{\mu}_1 = \hat{\mu}_2 = 0$ erhalten wir aus (3.2.11), (3.2.14) und (3.2.16) $\theta_k^{(2)} = \theta_{k-1}^{(1)} = \theta_{k-2}^{(0)} = 1$. Daraus folgt die Behauptung (c). □

Aus diesem Beweis erhalten wir auch die Formeln zur praktischen Berechnung von $\theta_k^{(2)}$, $\theta_{k-1}^{(1)}$ und $\theta_{k-2}^{(0)}$ im Fall des Modellproblems (3.1.2–3.1.4) (siehe (3.2.11), (3.2.14), (3.2.16) und (3.1.9)):

$$\theta_k^{(2)} = \frac{b_k}{d_k}, \quad \theta_{k-2}^{(0)} = a_k - \theta_k^{(2)}, \quad \theta_{k-1}^{(1)} = -\frac{d_k}{\theta_{k-2}^{(0)}} \tag{3.2.23}$$

mit

$$\begin{aligned} a_k &= -\frac{\hat{\delta}_{2,1,k}\delta_{1,0,k} - \hat{\delta}_{1,0,k}\delta_{2,1,k}}{\delta_{2,1,k} - \delta_{1,0,k}}, \\ b_k &= -\tau_k(\hat{\mu}_0)\tau_k(\hat{\mu}_1)\tau_k(\hat{\mu}_2)\frac{\delta_{2,1,k+1} - \delta_{1,0,k+1}}{\delta_{2,1,k} - \delta_{1,0,k}}, \\ d_k &= -\frac{\hat{\delta}_{2,1,k} - \hat{\delta}_{1,0,k}}{\delta_{2,1,k} - \delta_{1,0,k}}, \end{aligned} \tag{3.2.24}$$

wobei

$$\delta_{ijk} = \begin{cases} \dfrac{\tau_k(\hat{\mu}_i) - \tau_k(\hat{\mu}_j)}{\hat{\mu}_i - \hat{\mu}_j}, & \hat{\mu}_i \ne \hat{\mu}_j, \\ \tau_k'(\hat{\mu}_i), & \hat{\mu}_i = \hat{\mu}_j \end{cases} \tag{3.2.25}$$

und

$$\hat{\delta}_{ijk} = \begin{cases} \dfrac{\hat{\mu}_i\tau_k(\hat{\mu}_i) - \hat{\mu}_j\tau_k(\hat{\mu}_j)}{\hat{\mu}_i - \hat{\mu}_j}, & \hat{\mu}_i \ne \hat{\mu}_j, \\ (\mu\tau_k(\mu))'|_{\mu=\hat{\mu}_i}, & \hat{\mu}_i = \hat{\mu}_j. \end{cases} \tag{3.2.26}$$

Die Werte $\tau_k'(\hat{\mu}_i)$ und $(\mu\tau_k(\mu))'|_{\mu=\hat{\mu}_i}$ lassen sich dann nach Formeln (3.2.2) bzw. (3.2.15) berechnen. Hierbei haben wir vorausgesetzt, dass mindestens eine der Ungleichungen $\hat{\mu}_0 \ne \hat{\mu}_1$ oder $\hat{\mu}_1 \ne \hat{\mu}_2$ gilt. Für $\hat{\mu}_0 = \hat{\mu}_1 = \hat{\mu}_2 =: \hat{\mu}$ gelten wegen (3.2.17)

und (3.2.18) die folgenden Formeln für a_k, b_k und d_k ($k \geq 3$):

$$\begin{aligned} a_k &= \frac{\tau_k''(\hat{\mu}) - 2\tau_k'(\hat{\mu})}{\tau_k''(\hat{\mu})}, \\ b_k &= -\tau_k^3(\hat{\mu})\frac{\tau_{k+1}''(\hat{\mu})}{\tau_k''(\hat{\mu})}, \\ d_k &= -\frac{2\tau_k'(\hat{\mu}) + \hat{\mu}\tau_k''(\hat{\mu})}{\tau_k''(\hat{\mu})}. \end{aligned} \tag{3.2.27}$$

Die Werte $\tau_k''(\hat{\mu})$ können nach Rekursionsformel

$$\tau_k''(\mu) = \begin{cases} 0, & k = 1, \\ \dfrac{2\tau_{k-1}'(\mu) + \mu\tau_{k-1}''(\mu)}{\tau_{k-1}^2(\mu)} - \dfrac{2\mu\tau_{k-1}'^2(\mu)}{\tau_{k-1}^3(\mu)}, & k \geq 2 \end{cases}$$

berechnet werden (siehe (3.2.2)).

3.3 Konvergenz der GIBLU(l)-Zerlegungen

In diesem Abschnitt untersuchen wir die Konvergenzeigenschaften der oben eingeführten GIBLU(1)- und GIBLU(2)-Zerlegungen für das Modellproblem (3.1.1–3.1.4). Dabei gehen wir wie in Abschnitt 2.2 beschrieben vor. Wir wollen Werte $\omega_1^{(l)} \leq 0$ und $\omega_2^{(l)} \in [0, 1)$ finden, für die folgende Ungleichung gilt:

$$\omega_1^{(l)}(\mathbf{A} + \mathbf{R}^{(l)}) \leq \mathbf{R}^{(l)} \leq \omega_2^{(l)}(\mathbf{A} + \mathbf{R}^{(l)}). \tag{3.3.1}$$

Diese Werte schätzen den Spektralradius der Iterationsmatrix $(\mathbf{W}^{(l)})^{-1}\mathbf{R}^{(l)}$ ab und dadurch die Konvergenzraten der Iterationsverfahren:

Satz 3.3.1 Unter der Voraussetzung (3.3.1) gilt die Abschätzung

$$\gamma_1^{(l)}\mathbf{W}^{(l)} \leq \mathbf{A} \leq \gamma_2^{(l)}\mathbf{W}^{(l)} \tag{3.3.2}$$

mit

$$\gamma_1^{(l)} = 1 - \omega_2^{(l)}, \quad \gamma_2^{(l)} = 1 - \omega_1^{(l)}. \tag{3.3.3}$$

Die Konditionszahl der Matrix $(\mathbf{W}^{(l)})^{-1}\mathbf{A}$ besitzt die folgende Abschätzung:

$$\kappa((\mathbf{W}^{(l)})^{-1}\mathbf{A}) \leq \frac{\gamma_2^{(l)}}{\gamma_1^{(l)}} = \frac{1 - \omega_1^{(l)}}{1 - \omega_2^{(l)}} =: \kappa_l. \tag{3.3.4}$$

Die Konvergenzrate der mit $\mathbf{W}^{(l)}$ vorkonditionierten linearen Iteration wird durch $\rho(I - (\mathbf{W}^{(l)})^{-1}\mathbf{A}) \leq \max\{-\omega_1^{(l)}, \omega_2^{(l)}\} =: \rho_l$ abgeschätzt, und für das CG-Verfahren gilt

$$\|e^m\|_{\mathbf{A}} \leq \frac{2c^m}{1 + c^{2m}} \cdot \|e^0\|_{\mathbf{A}},$$

wobei $c = \dfrac{\sqrt{\kappa_l} - 1}{\sqrt{\kappa_l} + 1}$ und e^m der Fehler in der m-ten Iteration ist.

Beweis: (3.3.2–3.3.3) erhält man aus (3.3.1) durch Einsetzen der Gleichung $\mathbf{W}^{(l)} = \mathbf{A} + \mathbf{R}^{(l)}$ und Umordnen von Summanden. Die anderen Aussagen folgen aus (3.3.2–3.3.3) (siehe [18]). □

Nach unserer in Abschnitt 2.2 erläuterten Vorgehensweise reduzieren wir jetzt Ungleichung (3.3.1) im Fall des Modellproblems auf ein System von Ungleichungen für die Matrixblöcke und letztlich auf eine Ungleichung für eine skalare Funktion. Dazu wenden wir den Satz 2.1.5 auf die GIBLU(l)-Zerlegung (3.1.33) an und erhalten

$$\mathbf{R}^{(l)} = \text{blockdiag}\left\{R_k^{(l)}\right\}, \tag{3.3.5}$$

wobei

$$R_k^{(l)} = \begin{cases} 0, & k = 1, \\ \tilde{T}_k^{(l)} - \left(D - L\left(\tilde{T}_{k-1}^{(l)}\right)^{-1} L^T\right), & k \geq 2. \end{cases}$$

Wegen (3.1.31) gilt $R_k^{(l)} = 0$ für alle $k \in \{1, \ldots, l+1\}$. Für $k \geq l+2$ erhalten wir nach (3.1.22):

$$D^{-\frac{1}{2}} R_k^{(l)} D^{-\frac{1}{2}} = \tilde{\tau}_k^{(l)}(BB^T) - 1 + BB^T \left(\tilde{\tau}_{k-1}^{(l)}(BB^T)\right)^{-1}.$$

Mit der Bezeichnung

$$f_k^{(l)}(\mu) := \tilde{\tau}_k^{(l)}(\mu) - 1 + \frac{\mu}{\tilde{\tau}_{k-1}^{(l)}(\mu)}, \tag{3.3.6}$$

schreiben sich also die Blöcke $R_k^{(l)}$ für $k \geq l+2$ in der Form

$$R_k^{(l)} = D^{\frac{1}{2}} f_k^{(l)}(BB^T) D^{\frac{1}{2}}.$$

Die zweite für unsere Reduktion zentrale Aussage ist das Analogon zu Lemma 2.2.2:

Lemma 3.3.2 *(Siehe [10])* Es sei $\hat{\mathbf{A}} = \text{blockdiag}\left\{I - 2(BB^T)^{\frac{1}{2}}\right\}$ mit B aus (3.1.8). Dann ist $\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} \geq \hat{\mathbf{A}} > 0$.

Beweis: Nach der Definition von $\mathbf{A}$ und B erhalten wir:

$$\begin{aligned} \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} &= \text{blockdiag}\left\{I - 2(BB^T)^{\frac{1}{2}}\right\} + \text{blocktridiag}\left\{-B, 2(BB^T)^{\frac{1}{2}}, -B^T\right\} \\ &=: \hat{\mathbf{A}} + \mathbf{K}. \end{aligned}$$

Wir zeigen nun, dass die Matrix $\mathbf{K}$ positiv semidefinit ist.

Da die Matrix B normal ist, besitzt sie ein System von N orthogonalen Eigenvektoren $u^{(1)}, \ldots, u^{(n)} \in \mathbb{R}^n$ mit Eigenwerten $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$, respektive. Für jedes $i \in \{1, \ldots, n\}$ betrachten wir dann die hermitesche Matrix $K_i = \text{tridiag}\left\{-\lambda_i, 2|\lambda_i|, -\overline{\lambda_i}\right\}$ der Größe $N \times N$. Jede solche Matrix hat ein System von N orthogonalen Eigenvektoren $v^{(1,i)}, \ldots, v^{(N,i)} \in \mathbb{C}^N$ zu den Eigenwerten $\nu_1^{(i)}, \ldots, \nu_N^{(i)} \in \mathbb{R}$, respektive. Dann gilt für alle $i \in \{1, \ldots, n\}$ und $j \in \{1, \ldots, N\}$

$$\begin{aligned} -Bv_{k-1}^{(j,i)}u^{(i)} + 2(BB^T)^{\frac{1}{2}}v_k^{(j,i)}u^{(i)} - Bv_{k+1}^{(j,i)}u^{(i)} &= \left(-\lambda_i v_{k-1}^{(j,i)} + 2|\lambda_i| v_k^{(j,i)} - \overline{\lambda}_i v_{k+1}^{(j,i)}\right) u^{(i)} \\ &= \nu_j^{(i)} u^{(i)}, \end{aligned}$$

also ist $\mathbf{u}^{(i,j)} = \text{blockvector}\{v_1^{(j,i)}u^{(i)}, \ldots, v_n^{(j,i)}u^{(i)}\}$ Eigenvektor von $\mathbf{K}$ zum Eigenwert $\nu_j^{(i)}$. Da für $(i,j) \neq (i',j')$ die Vektoren $\mathbf{u}^{(i,j)}$ und $\mathbf{u}^{(i',j')}$ offensichtlich orthogonal zueinander sind, bilden sie ein vollständiges System von Nn Eigenvektoren der Matrix $\mathbf{K}$, d.h. $\{\nu_j^{(i)} : 1 \leq i \leq n, 1 \leq j \leq N\}$ ist das Spektrum von $\mathbf{K}$. Die Anwendung vom Gershgorinschen Satz auf K_i zeigt, dass alle $\nu_j^{(i)}$ nicht-negativ sind. Das liefert $\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} - \hat{\mathbf{A}} = \mathbf{K} \geq 0$.

Die Ungleichung $\hat{\mathbf{A}} > 0$ folgt aus $BB^T \leq \frac{1}{4}$ (siehe (3.1.4)). □

Nach diesem Lemma formulieren wir eine hinreichende Bedingung für (3.3.1):

$$\omega_1^{(l)}(\hat{\mathbf{A}} + \mathbf{D}^{-\frac{1}{2}}\mathbf{R}^{(l)}\mathbf{D}^{-\frac{1}{2}}) \leq \mathbf{D}^{-\frac{1}{2}}\mathbf{R}^{(l)}\mathbf{D}^{-\frac{1}{2}} \leq \omega_2^{(l)}(\hat{\mathbf{A}} + \mathbf{D}^{-\frac{1}{2}}\mathbf{R}^{(l)}\mathbf{D}^{-\frac{1}{2}}).$$

Da hier alle Matrizen blockdiagonal sind, lässt sich diese Ungleichung in Block-Form schreiben:

$$\omega_1^{(l)}(I - 2(BB^T)^{\frac{1}{2}} + f_k^{(l)}(BB^T)) \leq f_k^{(l)}(BB^T) \leq \omega_2^{(l)}(I - 2(BB^T)^{\frac{1}{2}} + f_k^{(l)}(BB^T)), \tag{3.3.7}$$

$l+2 \leq k \leq N$. Für $\sigma(BB^T) \subseteq [\mu_{\min}, \mu_{\max}] \subseteq [0, \frac{1}{4})$ können wir (3.3.7) durch folgendes System von skalaren Ungleichungen ersetzen:

$$\omega_1^{(l)}(1 - 2\sqrt{\mu} + f_k^{(l)}(\mu)) \leq f_k^{(l)}(\mu) \leq \omega_2^{(l)}(1 - 2\sqrt{\mu} + f_k^{(l)}(\mu)), \tag{3.3.8}$$

$l+2 \leq k \leq N$, $\mu_{\min} \leq \mu \leq \mu_{\max}$. Die weitere Untersuchung dieses Systems beruht auf dessen Reduktion auf eine Ungleichung, die wir mit der Hilfe folgendes Lemmas erreichen.

Lemma 3.3.3 Sei $f^{(l)}$ eine auf $\mu_{\min} \leq \mu \leq \mu_{\max}$ wohldefinierte Funktion, die den folgenden Bedingungen genügt:

$$1 - 2\sqrt{\mu} + f^{(l)}(\mu) > 0, \quad \mu_{\min} \leq \mu \leq \mu_{\max} \tag{3.3.9}$$

und für jedes $k \in \{l+2, \ldots, N\}$ gilt

$$|f_k^{(l)}(\mu)| \leq |f^{(l)}(\mu)|, \quad \mu_{\min} \leq \mu \leq \mu_{\max}. \tag{3.3.10}$$

Für die Zahlen $\omega_1^{(l)} \leq 0$ und $\omega_2^{(l)} \in [0, 1)$, welche die Ungleichung

$$\omega_1^{(l)} \leq \frac{f^{(l)}(\mu)}{1 - 2\sqrt{\mu} + f^{(l)}(\mu)} \leq \omega_2^{(l)}, \quad \mu_{\min} \leq \mu \leq \mu_{\max} \tag{3.3.11}$$

erfüllen, gelten die Ungleichungen (3.3.8).

Beweis: Für jedes $k \in \{l+2, \ldots, N\}$ lässt sich die Ungleichung (3.3.8) in folgende äquivalente Form umschreiben:

$$\frac{\omega_1^{(l)}}{1 - \omega_1^{(l)}} \leq \frac{f_k^{(l)}(\mu)}{1 - 2\sqrt{\mu}} \leq \frac{\omega_2^{(l)}}{1 - \omega_2^{(l)}}. \tag{3.3.12}$$

Unter der Voraussetzung (3.3.10) folgt (3.3.12) aus

$$\frac{\omega_1^{(l)}}{1-\omega_1^{(l)}} \leq \frac{f^{(l)}(\mu)}{1-2\sqrt{\mu}} \leq \frac{\omega_2^{(l)}}{1-\omega_2^{(l)}},$$

was zur Ungleichung

$$\omega_1^{(l)}(1-2\sqrt{\mu}+f^{(l)}(\mu)) \leq f^{(l)}(\mu) \leq \omega_2^{(l)}(1-2\sqrt{\mu}+f^{(l)}(\mu))$$

und wegen (3.3.9) auch zu Ungleichung (3.3.11) äquivalent ist. □

Die Bestimmung von $\omega_1^{(l)}$ und $\omega_2^{(l)}$ aus (3.3.11) erfolgt durch die analytische Untersuchung der skalaren Funktion $\dfrac{f^{(l)}(\mu)}{1-2\sqrt{\mu}+f^{(l)}(\mu)}$. Wir wollen eine Funktion $f^{(l)}(\mu)$ finden, die den Bedingungen (3.3.9–3.3.10) genügt und zusätzlich eine gute Abschätzung der Konvergenzrate liefert, die in numerischen Experimenten beobachtet wird. Wir beschreiben zuerst die allgemeine Vorgehensweise zur analytischen Untersuchung. In den Abschnitten 3.3.1 und 3.3.2 wenden wir diese Überlegungen separat auf die Fälle $l=1$ und $l=2$ an.

Wie in Abschnitt 2.2 nehmen wir $f^{(l)} = f_\infty^{(l)}$, die Grenzfunktion der Folge $\left\{f_k^{(l)}\right\}_k$ bzgl. punktweiser Konvergenz. Wir bezeichnen mit

$$g_l(\mu) := \frac{f_\infty^{(l)}(\mu)}{1-2\sqrt{\mu}+f_\infty^{(l)}(\mu)}$$

die in (3.3.11) abzuschätzende Funktion und untersuchen sie für $\mu \in [0, \frac{1}{4})$. Zur Vereinfachung der Darstellung von g_l bemerken wir, dass für $k \geq l+2$

$$f_k^{(l)}(\mu) = \frac{\tilde{P}_k^{(l)}}{\tilde{Q}_k^{(l)}} - 1 + \frac{\mu\tilde{Q}_{k-1}^{(l)}}{\tilde{P}_{k-1}^{(l)}} = \frac{\tilde{P}_k^{(l)}\tilde{P}_{k-1}^{(l)} - \tilde{Q}_k^{(l)}\tilde{P}_{k-1}^{(l)} + \mu\tilde{Q}_{k-1}^{(l)}\tilde{Q}_k^{(l)}}{\tilde{Q}_k^{(l)}\tilde{P}_{k-1}^{(l)}} \tag{3.3.13}$$

gilt (siehe (3.3.6) und (3.1.15)). Falls die Polynome $\tilde{P}_i^{(l)}$ und $\tilde{Q}_i^{(l)}$ den maximalen in (3.1.18) möglichen Grad haben, gilt $\deg \tilde{Q}_k^{(l)}\tilde{P}_{k-1}^{(l)} = l$ und

$$\begin{aligned}
\deg \tilde{P}_k^{(l)}\tilde{P}_{k-1}^{(l)} &= \begin{cases} l, & \text{für gerades } l, \\ l+1, & \text{für ungerades } l, \end{cases} \\
\deg \mu\tilde{Q}_{k-1}^{(l)}\tilde{Q}_k^{(l)} &= \begin{cases} l+1, & \text{für gerades } l, \\ l, & \text{für ungerades } l. \end{cases}
\end{aligned}$$

Folglich hat der Nenner von $f_k^{(l)}$ den Grad $l+1$. Da $f_k^{(l)}$ nach (3.1.19–3.1.20) $l+1$ Nullstellen $\hat{\mu}_0, \ldots, \hat{\mu}_l$ hat, lässt sich (3.3.13) in der Form

$$f_k^{(l)}(\mu) = \frac{M_k^{(l)}(\mu-\hat{\mu}_0)\cdots(\mu-\hat{\mu}_l)}{\tilde{Q}_k^{(l)}(\mu)\tilde{P}_{k-1}^{(l)}(\mu)} \tag{3.3.14}$$

darstellen, wobei $M_k^{(l)}$ der führende Koeffizient von $\tilde{Q}_{k-1}^{(l)}\tilde{Q}_k^{(l)}$ (für gerades l) oder von $\tilde{P}_k^{(l)}\tilde{P}_{k-1}^{(l)}$ (für ungerades l) ist. Wenn $M_k^{(l)}$, $\tilde{P}_k^{(l)}$ und $\tilde{Q}_k^{(l)}$ gegen $M_\infty^{(l)}$, $\tilde{P}_\infty^{(l)}$ und $\tilde{Q}_\infty^{(l)}$, respektive, konvergieren, folgt

$$f_\infty^{(l)}(\mu) = \frac{M_\infty^{(l)}(\mu-\hat{\mu}_0)\cdots(\mu-\hat{\mu}_l)}{\tilde{Q}_\infty^{(l)}(\mu)\tilde{P}_\infty^{(l)}(\mu)}. \tag{3.3.15}$$

Außerdem gilt nach (3.3.6)

$$f_\infty^{(l)}(\mu) = \tilde{\tau}_\infty^{(l)}(\mu) - 1 + \frac{\mu}{\tilde{\tau}_\infty^{(l)}(\mu)}, \tag{3.3.16}$$

wobei

$$\tilde{\tau}_\infty^{(l)}(\mu) = \frac{\tilde{P}_\infty^{(l)}(\mu)}{\tilde{Q}_\infty^{(l)}(\mu)}.$$

Nach (3.3.15) und (3.3.16) erhalten wir:

$$\begin{aligned} g_l(\mu) &= \frac{f_\infty^{(l)}(\mu)}{1-2\sqrt{\mu}+f_\infty^{(l)}(\mu)} = \frac{f_\infty^{(l)}(\mu)}{1-2\sqrt{\mu}+\tilde{\tau}_\infty^{(l)}(\mu)-1+\frac{\mu}{\tilde{\tau}_\infty^{(l)}(\mu)}} \\ &= \frac{\tilde{\tau}_\infty^{(l)}(\mu)f_\infty^{(l)}(\mu)}{\left(\tilde{\tau}_\infty^{(l)}(\mu)-\sqrt{\mu}\right)^2} = \frac{M_\infty^{(l)}(\mu-\hat{\mu}_0)\cdots(\mu-\hat{\mu}_l)}{\left(\tilde{Q}_\infty^{(l)}(\mu)(\tilde{\tau}_\infty^{(l)}(\mu)-\sqrt{\mu})\right)^2} \\ &= \frac{M_\infty^{(l)}(\mu-\hat{\mu}_0)\cdots(\mu-\hat{\mu}_l)}{\left(\tilde{P}_\infty^{(l)}(\mu)-\sqrt{\mu}\tilde{Q}_\infty^{(l)}(\mu)\right)^2} \end{aligned}$$

Wir wählen für $\omega_1^{(l)}$ und $\omega_2^{(l)}$ die untere, bzw. obere Schranke der Funktion

$$g_l(\mu) = \frac{M_\infty^{(l)}(\mu-\hat{\mu}_0)\cdots(\mu-\hat{\mu}_l)}{\left(\tilde{P}_\infty^{(l)}(\mu)-\sqrt{\mu}\tilde{Q}_\infty^{(l)}(\mu)\right)^2}. \tag{3.3.17}$$

Dabei erwähnen wir, dass $M_\infty^{(l)}$ für gerades l das Quadrat des führenden Koeffizienten von $\tilde{Q}_\infty^{(l)}(\mu)$ ist und für ungerades l das Quadrat des führenden Koeffizienten von $\tilde{P}_\infty^{(l)}(\mu)$. Schließlich können wir folgenden Satz beweisen.

Satz 3.3.4 Es gelte $l \geq 0$, und die folgenden Bedingungen seien erfüllt:

1. Die Folge $\left\{\tilde{\tau}_k^{(l)}\right\}_k$ konvergiert punktweise gegen eine Grenzfunktion $\tilde{\tau}_\infty^{(l)}$. Dabei gilt:
$$\forall \mu \in [\mu_{\min}, \mu_{\max}] \quad \tilde{\tau}_\infty^{(l)}(\mu) > \sqrt{\mu}. \tag{3.3.18}$$
2. Für die Grenzfunktion $f_\infty^{(l)}$ der Folge $\left\{f_k^{(l)}\right\}_k$ gilt:
$$\left|f_k^{(l)}(\mu)\right| \leq \left|f_\infty^{(l)}(\mu)\right|, \quad \mu_{\min} \leq \mu \leq \mu_{\max}, \quad k \geq l+2. \tag{3.3.19}$$

Genügt dann die in (3.3.17) definierte Funktion g_l der Ungleichung

$$\forall \mu \in [\mu_{\min}, \mu_{\max}] \quad \omega_1^{(l)} \leq g_l(\mu) \leq \omega_2^{(l)} \tag{3.3.20}$$

mit $\omega_1^{(l)} \leq 0$ und $\omega_2^{(l)} \in [0,1)$, so gilt Abschätzung (3.3.1) mit den gleichen Konstanten $\omega_1^{(l)}$ und $\omega_2^{(l)}$.

Beweis: Nach (3.3.6) und den Bedingungen des Satzes ist die Funktion $f_\infty^{(l)}$ als punktweiser Grenzwert der Folge $\left\{ f_k^{(l)} \right\}_k$ auf $[\mu_{\min}, \mu_{\max}]$ wohldefiniert. Diese Funktion genügt (3.3.9), da nach (3.3.16) und (3.3.18)

$$1 - 2\sqrt{\mu} + f_\infty^{(l)}(\mu) = \tilde{\tau}_\infty^{(l)}(\mu) - 2\sqrt{\mu} + \frac{\mu}{\tilde{\tau}_\infty^{(l)}(\mu)} = \frac{\left(\tilde{\tau}_\infty^{(l)}(\mu) - \sqrt{\mu}\right)^2}{\tilde{\tau}_\infty^{(l)}(\mu)} > 0$$

gilt. Nach (3.3.19) und Lemmata 3.3.2–3.3.3 folgt (3.3.1) aus (3.3.20). □

3.3.1 Konvergenz der GIBLU(1)-Zerlegung

Obwohl die Funktionen $\tilde{\tau}_k^{(1)}$ linear sind, ist die Konvergenzanalyse im Fall der GIBLU(1)-Zerlegung etwas komplizierter als die für die tangentialen und Zwei-Frequenz-Zerlegungen aus Kapitel 2 (siehe Bemerkung 3.2.3). Wie wir aber zeigen, sind die Eigenschaften dieser Verfahren für das Modellproblem (2.1.16–2.1.17) asymptotisch gleich.

Nach (3.1.19–3.1.20) und (3.2.4–3.2.6) gilt:

$$\tilde{\tau}_k^{(1)}(\mu) = \tau_k(\hat{\mu}_0) - A_k(\mu - \hat{\mu}_0), \quad k \geq 1,$$

mit $A_k = \dfrac{\tau_k(\hat{\mu}_0) - \tau_k(\hat{\mu}_1)}{\hat{\mu}_1 - \hat{\mu}_0}$ für $\hat{\mu}_0 \neq \hat{\mu}_1$ und $A_k = -\tau_k'(\hat{\mu}_0)$ für $\hat{\mu}_0 = \hat{\mu}_1$. Da $A_k > 0$ für alle $\hat{\mu}_0, \hat{\mu}_1 \in [0, \frac{1}{4})$ und $k \geq 2$ ist (siehe Satz 3.2.1), erhalten wir nach (3.3.14):

$$f_k^{(1)}(\mu) = \frac{A_k A_{k-1}(\mu - \hat{\mu}_0)(\mu - \hat{\mu}_1)}{\tau_{k-1}(\hat{\mu}_0) - A_{k-1}(\mu - \hat{\mu}_0)}. \tag{3.3.21}$$

Dies führt zur folgenden Aussage.

Lemma 3.3.5 Für alle $\hat{\mu}_0, \hat{\mu}_1 \in [0, \frac{1}{4})$ und $k \geq 3$ gilt:

$$|f_k^{(1)}(\mu)| \leq |f_\infty^{(1)}(\mu)|, \qquad \mu \in [0, \tfrac{1}{4}),$$

wobei

$$f_\infty^{(1)}(\mu) = \frac{A_\infty^2(\mu - \hat{\mu}_0)(\mu - \hat{\mu}_1)}{\tau_\infty(\hat{\mu}_0) - A_\infty(\mu - \hat{\mu}_0)} = \tilde{\tau}_\infty^{(1)}(\mu) - 1 + \frac{\mu}{\tilde{\tau}_\infty^{(1)}(\mu)} \tag{3.3.22}$$

mit

$$\tilde{\tau}_\infty^{(1)}(\mu) = \tau_\infty(\hat{\mu}_0) - A_\infty(\mu - \hat{\mu}_0) \tag{3.3.23}$$

und

$$A_\infty = \begin{cases} \dfrac{\tau_\infty(\hat\mu_0) - \tau_\infty(\hat\mu_1)}{\hat\mu_1 - \hat\mu_0}, & \hat\mu_0 \neq \hat\mu_1, \\ -\tau'_\infty(\hat\mu_0), & \hat\mu_0 = \hat\mu_1. \end{cases}$$

Beweis: Wir führen diesen Beweis nur für $\hat\mu_0 \neq \hat\mu_1$, da sich alle Beweisschritte leicht auf den Fall $\hat\mu_0 = \hat\mu_1$ übertragen lassen. Wir zeigen durch Induktion über k, dass $A_k \leq A_{k+1}$. Diese Aussage gilt für $k = 1$, da $A_1 = 0 \leq 1 = A_2$. Für $k \geq 2$ erhalten wir:

$$\begin{aligned} \tau_k(\hat\mu_0) - \tau_k(\hat\mu_1) &= \frac{\hat\mu_1}{\tau_{k-1}(\hat\mu_1)} - \frac{\hat\mu_0}{\tau_{k-1}(\hat\mu_0)} = \frac{\hat\mu_1\tau_{k-1}(\hat\mu_0) - \hat\mu_0\tau_{k-1}(\hat\mu_1)}{\tau_{k-1}(\hat\mu_0)\tau_{k-1}(\hat\mu_1)} \\ &= \frac{\hat\mu_1 - \hat\mu_0}{\tau_{k-1}(\hat\mu_1)} + \frac{\hat\mu_0}{\tau_{k-1}(\hat\mu_0)\tau_{k-1}(\hat\mu_1)}(\tau_{k-1}(\hat\mu_0) - \tau_{k-1}(\hat\mu_1)). \end{aligned}$$

Nach Induktionsannahme und Satz 3.2.1 folgt daraus

$$A_k = \frac{1}{\tau_{k-1}(\hat\mu_1)} + \frac{\hat\mu_0}{\tau_{k-1}(\hat\mu_0)\tau_{k-1}(\hat\mu_1)}A_{k-1} \leq \frac{1}{\tau_k(\hat\mu_1)} + \frac{\hat\mu_0}{\tau_k(\hat\mu_0)\tau_k(\hat\mu_1)}A_k = A_{k+1}.$$

Die Folge $\{A_k\}_k$ ist also nicht-fallend und von oben beschränkt. Ihr Grenzwert ist A_∞. Daher erhalten wir für jedes $\mu \in [0, \frac{1}{4})$ und $k \geq 3$

$$|f_k^{(1)}(\mu)| = \frac{A_k|(\mu - \hat\mu_0)(\mu - \hat\mu_1)|}{\dfrac{\tau_{k-1}(\hat\mu_0)}{A_{k-1}} - (\mu - \hat\mu_0)} \leq \frac{A_\infty|(\mu - \hat\mu_0)(\mu - \hat\mu_1)|}{\dfrac{\tau_\infty(\hat\mu_0)}{A_\infty} - (\mu - \hat\mu_0)} = |f_\infty^{(1)}(\mu)|,$$

da $\tau_k(\mu) \geq \tau_\infty(\mu) \geq 0$. Die zweite Gleichung in (3.3.22) folgt aus (3.3.16). □

Die Darstellung von $\tilde\tau_\infty^{(1)}$ lässt sich weiter vereinfachen. Nach (3.3.23) gilt

$$\tilde\tau_\infty^{(1)}(\mu) = B_\infty - A_\infty\mu, \tag{3.3.24}$$

mit

$$B_\infty = \begin{cases} \dfrac{\hat\mu_1\tau_\infty(\hat\mu_0) - \hat\mu_0\tau_\infty(\hat\mu_1)}{\hat\mu_1 - \hat\mu_0}, & \hat\mu_0 \neq \hat\mu_1, \\ \tau_\infty(\hat\mu_0) - \tau'_\infty(\hat\mu_0)\hat\mu_0, & \hat\mu_0 = \hat\mu_1. \end{cases}$$

Da $\tau_\infty^2(\mu) - \tau_\infty(\mu) + \mu = 0$ für $\mu \in [0, \frac{1}{4})$ gilt, erhalten wir

$$A_\infty = \frac{1}{\tau_\infty(\hat\mu_0) + \tau_\infty(\hat\mu_1) - 1}, \qquad \frac{B_\infty}{A_\infty} = \tau_\infty(\hat\mu_0)\tau_\infty(\hat\mu_1). \tag{3.3.25}$$

Damit können wir zeigen, dass die erste Voraussetzung in Satz 3.3.4 für $\tilde\tau_\infty^{(1)}$ erfüllt ist:

Lemma 3.3.6 Für alle $\hat\mu_0, \hat\mu_1 \in [0, \frac{1}{4})$ gilt:

$$\tilde\tau_\infty^{(1)}(\mu) > \tfrac{1}{4} + \mu \geq \sqrt{\mu}, \quad \mu \in [0, \tfrac{1}{4}]. \tag{3.3.26}$$

Beweis: Nach (3.3.24) und (3.3.25) erhalten wir für $\mu \in [0, \frac{1}{4}]$:

$$\begin{aligned}\tilde{\tau}_\infty^{(1)}(\mu) - (\tfrac{1}{4} + \mu) &= A_\infty \left(\frac{B_\infty}{A_\infty} - \left(\frac{1}{A_\infty} - 1 \right) \mu - \frac{1}{4A_\infty} \right) \\ &= A_\infty \left(\tau_\infty(\hat{\mu}_0)\tau_\infty(\hat{\mu}_1) - (\tfrac{1}{4} + \mu)(\tau_\infty(\hat{\mu}_0) + \tau_\infty(\hat{\mu}_1)) + \tfrac{1}{4} \right) \\ &\geq A_\infty \left(\tau_\infty(\hat{\mu}_0)\tau_\infty(\hat{\mu}_1) - \tfrac{1}{2}(\tau_\infty(\hat{\mu}_0) + \tau_\infty(\hat{\mu}_1)) + \tfrac{1}{4} \right) \\ &= A_\infty \left(\tau_\infty(\hat{\mu}_0) - \tfrac{1}{2} \right) \left(\tau_\infty(\hat{\mu}_1) - \tfrac{1}{2} \right) > 0.\end{aligned}$$

Die zweite Ungleichung in (3.3.26) folgt aus $\frac{1}{4} - \sqrt{\mu} + \mu = (\frac{1}{2} - \sqrt{\mu})^2 \geq 0$. □

Nach Lemmata 3.3.5–3.3.6 und Satz 3.3.4 reduziert sich unsere Aufgabe auf die Untersuchung der Funktion g_1 aus (3.3.17). Da $\tilde{P}_\infty^{(1)}(\mu) = B_\infty - A_\infty \mu$ und $\tilde{Q}_\infty^{(1)}(\mu) = 1$, erhalten wir

$$g_1(\mu) = \frac{A_\infty^2(\mu - \hat{\mu}_0)(\mu - \hat{\mu}_1)}{(B_\infty - A_\infty\mu - \sqrt{\mu})^2} = \frac{(\mu - \hat{\mu}_0)(\mu - \hat{\mu}_1)}{\left(\frac{B_\infty}{A_\infty} - \frac{\sqrt{\mu}}{A_\infty} - \mu \right)^2}.$$

Wir suchen nun $\omega_1^{(1)} < 0$ und $\omega_2^{(1)} \in [0, 1)$, für die (3.3.20) erfüllt ist.

Wir betrachten nun den Spezialfall $\hat{\mu}_0 = \hat{\mu}_1 =: \hat{\mu} \in [0, \frac{1}{4})$. Für diesen gilt

$$g_1(\mu) = \frac{(\mu - \hat{\mu})^2}{\left(\frac{B_\infty}{A_\infty} - \frac{\sqrt{\mu}}{A_\infty} - \mu \right)^2} = G_1^2(\mu, \hat{\mu}),$$

wobei

$$G_1(\mu, \hat{\mu}) := \frac{\mu - \hat{\mu}}{\frac{B_\infty}{A_\infty} - \frac{\sqrt{\mu}}{A_\infty} - \mu} = \frac{\mu - \hat{\mu}}{\tau_\infty^2(\hat{\mu}) - (2\tau_\infty(\hat{\mu}) - 1)\sqrt{\mu} - \mu}. \tag{3.3.27}$$

Die Funktion g_1 ist in diesem Fall nicht-negativ, und $g_1(\hat{\mu}) = 0$. Deswegen gilt (3.3.20) mit

$$\omega_1^{(1)} = 0 \tag{3.3.28}$$

und einer Konstanten $\omega_2^{(1)} \geq 0$, die der Ungleichung

$$\forall \mu \in [\mu_{\min}, \mu_{\max}] \quad |G_1(\mu, \hat{\mu})| \leq \sqrt{\omega_2^{(1)}} \tag{3.3.29}$$

genügt. Die Untersuchung der Funktion G_1 ist aber wegen des Wurzelterms $\sqrt{\mu}$ im Nenner recht kompliziert. Daher ersetzen wir sie mit einer gebrochen-linearen Funktion gemäß dem folgenden Lemma:

Lemma 3.3.7 Für jedes $\hat{\mu} \in [0, \frac{1}{4})$ gilt

$$|G_1(\mu, \hat{\mu})| \leq |\tilde{G}_1(\mu, \hat{\mu})|, \quad \mu \in [0, \tfrac{1}{4}) \tag{3.3.30}$$

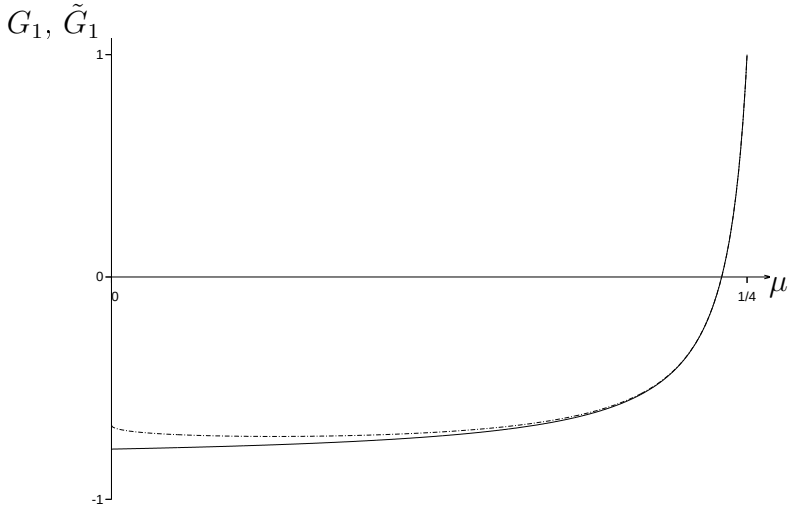

Abbildung 3.2: Die Funktionen $\tilde{G}_1(\mu, 0.24)$ (durchgezeichnete Linie) und $G_1(\mu, 0.24)$ (gestrichelt) auf $[0, \frac{1}{4})$

mit

$$\tilde{G}_1(\mu, \hat{\mu}) = \frac{\mu - \hat{\mu}}{\tau_\infty^2(\hat{\mu}) - \frac{1}{2}\tau_\infty(\hat{\mu}) + \frac{1}{4} - 2\tau_\infty(\hat{\mu})\mu}.$$

Die Funktion $\tilde{G}_1(\cdot, \hat{\mu})$ ist auf $[0, \frac{1}{4}]$ wohldefiniert und steigend. Ihr Nenner ist auf diesem Intervall positiv.

Beweis: Der Nenner von $\tilde{G}_1$ ist gleich $\left(\tilde{\tau}_\infty^{(1)}(\mu) - (\frac{1}{4} + \mu)\right)/A_\infty$ (siehe (3.3.24) und (3.3.25)). Er ist nach Lemma 3.3.6 für jedes $\mu \in [0, \frac{1}{4}]$ positiv. Somit ist $\tilde{G}_1$ auf diesem Intervall wohldefiniert. Des Weiteren ist dieser Nenner für $\mu \geq 0$ kleiner als jener von G_1 (siehe (3.3.27)):

$$\frac{\tilde{\tau}_\infty^{(1)}(\mu) - (\frac{1}{4} + \mu)}{A_\infty} \leq \frac{\tilde{\tau}_\infty^{(1)}(\mu) - \sqrt{\mu}}{A_\infty} = \frac{B_\infty}{A_\infty} - \frac{\sqrt{\mu}}{A_\infty} - \mu.$$

Damit ist (3.3.30) bewiesen. □

Bemerkung 3.3.8 Die Funktionen $G_1(\cdot, \hat{\mu})$ und $\tilde{G}_1(\cdot, \hat{\mu})$ sind für $\hat{\mu} = 0.24$ auf Abbildung 3.2 dargestellt. Man erhält $\tilde{G}_1$ aus G_1 durch Ersetzen von $\sqrt{\mu}$ mit $\frac{1}{4} + \mu$. Da $\sqrt{\mu} = \frac{1}{4} + \mu + O\left((\frac{1}{4} - \mu)^2\right)$ und $\tilde{G}_1(\hat{\mu}, \hat{\mu}) = G_1(\hat{\mu}, \hat{\mu}) = 0$, haben die Funktionsgraphen in der Umgebung von $\mu = \frac{1}{4}$ einen sehr geringen Abstand. Der Approximationsfehler ist im linken Teil des Intervalls $[0, \frac{1}{4})$ am größten. Dort ist auch die Abschätzung aus Lemma 3.3.2 ungenau. □

Lemma 3.3.7 ermöglicht die folgende Wahl von $\omega_2^{(1)}$ in (3.3.29):

$$\omega_2^{(1)} = \hat{\omega}^2, \tag{3.3.31}$$

wobei

$$\hat{\omega} = \hat{\omega}(\hat{\mu}) := \max_{0 \leq \mu \leq \mu_{\max}} |\tilde{G}_1(\mu, \hat{\mu})| = \max\left\{-\tilde{G}_1(0, \hat{\mu}), \tilde{G}_1(\mu_{\max}, \hat{\mu})\right\} \tag{3.3.32}$$

ist. Wir suchen nun ein $\hat{\mu}$, für das $\hat{\omega}$ minimal ist.

Dazu betrachten wir als Parameter statt $\hat{\mu}$ den Wert $t := \tau_\infty(\hat{\mu})$. Es gilt $\hat{\mu} = t - t^2$. Wir bezeichnen daher

$$\tilde{G}_1(\mu, t) := \frac{\mu - t + t^2}{t^2 - \frac{1}{2}t + \frac{1}{4} - 2t\mu},$$

wenn dabei kein Missverständnis entstehen kann. Den optimalen Parameter t finden wir nach dem folgenden Lemma:

Lemma 3.3.9 Es gelte $\mu_{\max} \in [0, \frac{1}{4})$. Es existiert ein Parameter $t_{\text{opt}} \in (\frac{1}{2}, 1]$, für den die Bedingung

$$-\tilde{G}_1(0, t_{\text{opt}}) = \tilde{G}_1(\mu_{\max}, t_{\text{opt}}) \tag{3.3.33}$$

erfüllt ist. Für diesen Parameter hat $\hat{\omega}$ aus (3.3.32) den kleinsten möglichen Wert für $\frac{1}{2} < t \leq 1$. Gleichung (3.3.33) bestimmt t_{opt} eindeutig, und es gilt: $t_{\min} \leq t_{\text{opt}} \leq 1$ mit $t_{\min} = \frac{1}{2} + \sqrt{\frac{1}{4} - \mu_{\max}}$.

Beweis: Wir bemerken sofort, dass der Fall $\frac{1}{2} < t < t_{\min}$ den Werten $\hat{\mu} \in [\mu_{\max}, \frac{1}{4})$ entspricht. Dabei sind $\tilde{G}_1(0, t)$ und $\tilde{G}_1(\mu_{\max}, t)$ beide negativ. Also kann (3.3.33) nicht gelten. Somit ist der Fall $\frac{1}{2} < t < t_{\min}$ ausgeschlossen.

Wir betrachten nun die Funktion $\phi(t) = \tilde{G}_1(0, t) + \tilde{G}_1(\mu_{\max}, t)$. Sie ist stetig auf $[t_{\min}, 1]$. Da $\tilde{G}_1(0, 1) = \tilde{G}_1(\mu_{\max}, t_{\min}) = 0$, erhalten wir:

$$\phi(1) \geq 0, \quad \phi(t_{\min}) \leq 0.$$

Folglich existiert ein $t_{\text{opt}} \in [t_{\min}, 1]$ mit $\phi(t_{\text{opt}}) = 0$, d.h. Bedingung (3.3.33) ist erfüllt.

Für $\mu \in [0, \frac{1}{4})$ und $t \in (\frac{1}{2}, 1]$ erhalten wir:

$$\frac{\partial}{\partial t}\tilde{G}_1(\mu, t) = \frac{2(\mu - \frac{1}{4})(\mu + \frac{3}{4} - (t + \frac{1}{2})^2)}{\left(t^2 - \frac{1}{2}t + \frac{1}{4} - 2t\mu\right)^2} > 0.$$

Also ist $\tilde{G}(\mu, t)$ für jedes μ eine steigende Funktion von t. Folglich ist $\phi(t)$ monoton auf $[t_{\min}, 1]$, und somit ist t_{opt} durch (3.3.33) eindeutig bestimmt.

Des Weiteren gilt für $t \in (t_{\text{opt}}, 1]$: $\tilde{G}_1(\mu_{\max}, t) > \tilde{G}_1(\mu_{\max}, t_{\text{opt}}) \geq 0$. Mit der gleichen Argumentation erhalten wir für $t \in (\frac{1}{2}, t_{\text{opt}})$: $\tilde{G}_1(0, t) < \tilde{G}_1(0, t_{\text{opt}}) \leq 0$. Wegen (3.3.33) ist also für alle $t \in (\frac{1}{2}, 1] \setminus \{t_{\text{opt}}\}$ das von t abhängige $\hat{\omega}$ aus (3.3.32) größer als für t_{opt}. □

Bemerkung 3.3.10 Für Satz 3.3.4 ist es wichtig, dass $\omega_2^{(1)}$ aus (3.3.31–3.3.32) im Intervall $[0, 1)$ liegt. In unserem Fall ist diese Bedingung für jedes $t \in (\frac{1}{2}, 1]$ erfüllt:

$$-1 = \tilde{G}_1(0, \tfrac{1}{2}) < \tilde{G}_1(0, t) \leq \tilde{G}_1(\mu_{\max}, t) < \tilde{G}_1(\tfrac{1}{4}, t) = 1,$$

also ist $\hat{\omega} < 1$. Hier haben wir die Monotonie von $\tilde{G}_1$ bzgl. μ (siehe Lemma 3.3.7) und t (siehe Beweis von Lemma 3.3.9) benutzt. □

Wir interessieren uns nun für den optimalen Parameter t_{opt} und die optimale Abschätzung $\hat{\omega}_{\text{opt}}$ als Funktionen von $\mu_{\max}$. Gleichung (3.3.33) kann in das folgende System umgeschrieben werden:

$$\begin{aligned} \tilde{G}_1(0, t_{\text{opt}}) &= -\hat{\omega}_{\text{opt}}, \\ \tilde{G}_1(\mu_{\max}, t_{\text{opt}}) &= \hat{\omega}_{\text{opt}}. \end{aligned} \tag{3.3.34}$$

Nach Multiplikation mit den Nennern der gebrochen-rationalen Funktionen auf den linken Seiten erhalten wir daraus ein algebraisches System:

$$\begin{aligned} &-t_{\text{opt}} + t_{\text{opt}}^2 + \hat{\omega}_{\text{opt}} \left(t_{\text{opt}}^2 - \tfrac{1}{2} t_{\text{opt}} + \tfrac{1}{4}\right) = 0, \\ &\mu_{\max} - t_{\text{opt}} + t_{\text{opt}}^2 - \hat{\omega}_{\text{opt}} \left(t_{\text{opt}}^2 - \tfrac{1}{2} t_{\text{opt}} + \tfrac{1}{4} - 2 t_{\text{opt}} \mu_{\max}\right) = 0. \end{aligned} \tag{3.3.35}$$

Wir behandeln (3.3.35) mit Hilfe der Eliminationstheorie (siehe [23], [36], [52]). Zunächst eliminieren wir t_{opt}, und untersuchen die Abhängigkeit der Variablen $\hat{\omega}_{\text{opt}}$ von $\mu_{\max}$. Dazu betrachten wir die linken Seiten in (3.3.35) als Polynome von t_{opt}:

$$\begin{aligned} &(1 + \hat{\omega}_{\text{opt}}) t_{\text{opt}}^2 - (1 + \tfrac{1}{2}\hat{\omega}_{\text{opt}}) t_{\text{opt}} + \tfrac{1}{4}\hat{\omega}_{\text{opt}} = 0, \\ &(1 - \hat{\omega}_{\text{opt}}) t_{\text{opt}}^2 - (1 - \hat{\omega}_{\text{opt}}(\tfrac{1}{2} + 2\mu_{\max})) t_{\text{opt}} + \mu_{\max} - \tfrac{1}{4}\hat{\omega}_{\text{opt}} = 0. \end{aligned} \tag{3.3.36}$$

Die Resultante dieses Systems ist

$$\begin{vmatrix} 1 + \hat{\omega}_{\text{opt}} & -(1 + \frac{1}{2}\hat{\omega}_{\text{opt}}) & \frac{1}{4}\hat{\omega}_{\text{opt}} & 0 \\ 0 & 1 + \hat{\omega}_{\text{opt}} & -(1 + \frac{1}{2}\hat{\omega}_{\text{opt}}) & \frac{1}{4}\hat{\omega}_{\text{opt}} \\ 1 - \hat{\omega}_{\text{opt}} & -(1 - \hat{\omega}_{\text{opt}}(\frac{1}{2} + 2\mu_{\max})) & \mu_{\max} - \frac{1}{4}\hat{\omega}_{\text{opt}} & 0 \\ 0 & 1 - \hat{\omega}_{\text{opt}} & -(1 - \hat{\omega}_{\text{opt}}(\frac{1}{2} + 2\mu_{\max})) & \mu_{\max} - \frac{1}{4}\hat{\omega}_{\text{opt}} \end{vmatrix}$$

$$= (\hat{\omega}_{\text{opt}}^4 + 2\hat{\omega}_{\text{opt}}^3 + 4\hat{\omega}_{\text{opt}}^2 + 4\hat{\omega}_{\text{opt}} + 1)\mu_{\max}^2 - \hat{\omega}_{\text{opt}}(\tfrac{3}{2}\hat{\omega}_{\text{opt}}^2 + \tfrac{5}{2}\hat{\omega}_{\text{opt}} + 2)\mu_{\max} + \tfrac{3}{4}\hat{\omega}_{\text{opt}}^2.$$

Also ist System (3.3.35) für $\hat{\omega}_{\text{opt}}$ und $\mu_{\max}$ zur Gleichung

$$(\hat{\omega}_{\text{opt}}^4 + 2\hat{\omega}_{\text{opt}}^3 + 4\hat{\omega}_{\text{opt}}^2 + 4\hat{\omega}_{\text{opt}} + 1)\mu_{\max}^2 - \hat{\omega}_{\text{opt}}(\tfrac{3}{2}\hat{\omega}_{\text{opt}}^2 + \tfrac{5}{2}\hat{\omega}_{\text{opt}} + 2)\mu_{\max} + \tfrac{3}{4}\hat{\omega}_{\text{opt}}^2 = 0 \tag{3.3.37}$$

äquivalent. Dies ist eine quadratische Gleichung für $\mu_{\max}$ und eine Gleichung vierten Grades in $\hat{\omega}_{\text{opt}}$. Daher ist es bequemer, erst $\mu_{\max}$ als Funktion von $\hat{\omega}_{\text{opt}}$ zu betrachten und dann ihre Inverse zu untersuchen. Die Gleichung (3.3.37) hat zwei Lösungen:

$$\begin{aligned} \mu_{\max,1} &= \hat{\omega}_{\text{opt}} \frac{3\hat{\omega}_{\text{opt}}^2 + 5\hat{\omega}_{\text{opt}} + 4 - (1 - \hat{\omega}_{\text{opt}})\sqrt{4 - 3\hat{\omega}_{\text{opt}}^2}}{4(\hat{\omega}_{\text{opt}} + 1)(\hat{\omega}_{\text{opt}}^3 + \hat{\omega}_{\text{opt}}^2 + 3\hat{\omega}_{\text{opt}} + 1)}, \\ \mu_{\max,2} &= \hat{\omega}_{\text{opt}} \frac{3\hat{\omega}_{\text{opt}}^2 + 5\hat{\omega}_{\text{opt}} + 4 + (1 - \hat{\omega}_{\text{opt}})\sqrt{4 - 3\hat{\omega}_{\text{opt}}^2}}{4(\hat{\omega}_{\text{opt}} + 1)(\hat{\omega}_{\text{opt}}^3 + \hat{\omega}_{\text{opt}}^2 + 3\hat{\omega}_{\text{opt}} + 1)}. \end{aligned} \tag{3.3.38}$$

Für $\hat{\omega}_{\text{opt}} \in [0, 1)$ liegen die beiden Werte $\mu_{\max,i}$ auf $[0, \frac{1}{4})$. Aber nur $\mu_{\max,2}$ entspricht einem zulässigen Parameter t_{opt}.

Um das zu beweisen, lösen wir die erste Gleichung in (3.3.36) nach t_{opt}. Daraus erhalten wir nur zwei mögliche Werte dieses Parameters:

$$\begin{aligned} t_{\text{opt},1} &= \frac{\hat{\omega}_{\text{opt}} + 2 - \sqrt{4 - 3\hat{\omega}_{\text{opt}}^2}}{4(\hat{\omega}_{\text{opt}} + 1)}, \\ t_{\text{opt},2} &= \frac{\hat{\omega}_{\text{opt}} + 2 + \sqrt{4 - 3\hat{\omega}_{\text{opt}}^2}}{4(\hat{\omega}_{\text{opt}} + 1)}. \end{aligned}$$

Einsetzen von Paaren $(\mu_{\max,i}, t_{\text{opt},j})$ in die zweite Gleichung von (3.3.36) zeigt, dass nur $(\mu_{\max,1}, t_{\text{opt},1})$ und $(\mu_{\max,2}, t_{\text{opt},2})$ diesem System genügen. Aber der Wert $t_{\text{opt},1}$ für $\hat{\omega}_{\text{opt}} \in [0, 1)$ liegt nicht im Intervall $(\frac{1}{2}, 1]$.

Deswegen betrachten wir im Folgenden nur $\mu_{\max} = \mu_{\max,2}$. Dies ist eine differenzierbare Funktion von $\hat{\omega}_{\text{opt}}$, und ihre Taylor-Entwicklung in diesem Punkt lautet:

$$\mu_{\max} = \tfrac{1}{4} + \tfrac{1}{8}(\hat{\omega}_{\text{opt}} - 1)^3 + O((\hat{\omega}_{\text{opt}} - 1)^4).$$

Daraus erhalten wir die Ordnung der Abhängigkeit des Wertes $\hat{\omega}_{\text{opt}}$ von $\mu_{\max}$:

$$\hat{\omega}_{\text{opt}} = 1 - 2(\tfrac{1}{4} - \mu_{\max})^{\frac{1}{3}} + o\left((\tfrac{1}{4} - \mu_{\max})^{\frac{1}{3}}\right).$$

Nach (3.3.31) bedeutet dies, dass

$$\omega_2^{(1)} = 1 - 4(\tfrac{1}{4} - \mu_{\max})^{\frac{1}{3}} + o\left((\tfrac{1}{4} - \mu_{\max})^{\frac{1}{3}}\right). \tag{3.3.39}$$

Wir interessieren uns nun für den Wert des optimalen Parameters t_{opt} als eine Funktion von $\mu_{\max}$. Wir eliminieren deswegen $\hat{\omega}_{\text{opt}}$ aus (3.3.35). Die beiden Gleichungen sind linear in $\hat{\omega}_{\text{opt}}$, und die Resultante dieses Systems ist

$$\begin{aligned} &\begin{vmatrix} -t_{\text{opt}} + t_{\text{opt}}^2 & t_{\text{opt}}^2 - \frac{1}{2}t_{\text{opt}} + \frac{1}{4} \\ \mu_{\max} - t_{\text{opt}} + t_{\text{opt}}^2 & -(t_{\text{opt}}^2 - \frac{1}{2}t_{\text{opt}} + \frac{1}{4} - 2t_{\text{opt}}\mu_{\max}) \end{vmatrix} \\ &= (\tfrac{1}{4} - \tfrac{1}{2}t_{\text{opt}} + 3t_{\text{opt}}^2 - 2t_{\text{opt}}^3)\mu_{\max} + \tfrac{1}{2}t_{\text{opt}}(t_{\text{opt}} - 1)(1 - 2t_{\text{opt}} + 4t_{\text{opt}}^2). \end{aligned}$$

Für jedes vorgegebene $\mu_{\max}$ genügt also das entsprechende t_{opt} der Gleichung

$$(\tfrac{1}{4} - \tfrac{1}{2}t_{\text{opt}} + 3t_{\text{opt}}^2 - 2t_{\text{opt}}^3)\mu_{\max} + \tfrac{1}{2}t_{\text{opt}}(t_{\text{opt}} - 1)(1 - 2t_{\text{opt}} + 4t_{\text{opt}}^2) = 0. \tag{3.3.40}$$

Die linke Seite ist ein Polynom vierten Grades in t_{opt} und ersten Grades in $\mu_{\max}$. Wir untersuchen hier wieder die inverse Funktion $\mu_{\max}(t_{\text{opt}})$, die nach (3.3.40) durch die Gleichung

$$\mu_{\max} = \tfrac{1}{2}\frac{t_{\text{opt}}(1 - t_{\text{opt}})(1 - 2t_{\text{opt}} + 4t_{\text{opt}}^2)}{\frac{1}{4} - \frac{1}{2}t_{\text{opt}} + 3t_{\text{opt}}^2 - 2t_{\text{opt}}^3} \tag{3.3.41}$$

gegeben ist. Nach Lemma 3.3.9 besitzt diese Gleichung für jedes $\mu_{\max} \in [0, \frac{1}{4})$ mindestens eine Lösung $t_{\text{opt}} \in (\frac{1}{2}, 1]$. Da die Ableitung der Funktion (3.3.41) negativ ist,

$$\mu'_{\max}(t_{\text{opt}}) = \frac{(8t_{\text{opt}}^4 - 16t_{\text{opt}}^3 - 2t_{\text{opt}} + 1)(2t_{\text{opt}} - 1)^2}{8(\frac{1}{4} - \frac{1}{2}t_{\text{opt}} + 3t_{\text{opt}}^2 - 2t_{\text{opt}}^3)^2} < 0, \quad t_{\text{opt}} \in (\tfrac{1}{2}, 1],$$

ist diese Funktion auf $(\frac{1}{2}, 1]$ monoton. Also hat (3.3.41) für jedes $\mu_{\max} \in [0, \frac{1}{4})$ genau eine Lösung auf $(\frac{1}{2}, 1]$, den dem Wert $\mu_{\max}$ entsprechenden optimalen Parameter t_{opt}. Diese Ergebnisse fassen wir im folgenden Satz zusammen.

Satz 3.3.11 Im Fall des Modellproblems (3.1.1–3.1.4) mit $LD^{-1}L^T \leq \mu_{\max} D$ gelten für die GIBLU(1)-Zerlegung mit Diagonalblöcken (3.1.34), (3.2.6) bei der optimalen Wahl des Parameters $\hat{\mu}_{\text{opt}} = t_{\text{opt}} - t_{\text{opt}}^2$ die folgenden Abschätzungen für den Spektralradius der Iterationsmatrix und für die Konditionszahl der Matrix $(\mathbf{W}^{(l)})^{-1}\mathbf{A}$:

$$\rho(I - (\mathbf{W}^{(1)})^{-1}\mathbf{A}) \leq \rho_1 = 1 - 4(\tfrac{1}{4} - \mu_{\max})^{\frac{1}{3}} + o\left((\tfrac{1}{4} - \mu_{\max})^{\frac{1}{3}}\right), \quad (3.3.42)$$

$$\kappa((\mathbf{W}^{(1)})^{-1}\mathbf{A}) \leq \kappa_1 = (\tfrac{1}{4} - \mu_{\max})^{-\frac{1}{3}}\left(\tfrac{1}{4} + o\left((\tfrac{1}{4} - \mu_{\max})^{\frac{1}{3}}\right)\right). \quad (3.3.43)$$

Der Wert t_{opt} ist die Lösung der Gleichung (3.3.41) auf $(\frac{1}{2}, 1]$. Für jedes $\mu_{\max} \in [0, \frac{1}{4})$ ist diese Gleichung immer eindeutig lösbar.

Beweis: Diese Aussage folgt aus Satz 3.3.4, den Gleichungen (3.3.28) und (3.3.39) sowie den oben gemachten Überlegungen. □

Bei der praktischen Implementierung der Zerlegungen entsteht hier das Problem der Berechnung des optimalen Parameters $\hat{\mu}_{\text{opt}}$. Das kann z.B. durch die numerische Lösung der Gleichung (3.3.41) nach der Variablen t_{opt} auf $[t_{\min}, 1]$ gemacht werden. Eine andere Möglichkeit ist die Approximation der Funktion $t_{\text{opt}}(\mu_{\max})$ durch deren asymptotische Entwicklung im Punkt $\mu_{\max} = \frac{1}{4}$. Dazu stellen wir zunächst die inverse, durch (3.3.41) gegebene Funktion $\mu_{\max}(t_{\text{opt}})$ nach der Taylor-Formel in der Form

$$\mu_{\max} = \tfrac{1}{4} - (t_{\text{opt}} - \tfrac{1}{2})^3 + o\left((t_{\text{opt}} - \tfrac{1}{2})^3\right)$$

dar. Daraus erhalten wir

$$t_{\text{opt}} = \tfrac{1}{2} + (\tfrac{1}{4} - \mu_{\max})^{\frac{1}{3}} + o\left((\tfrac{1}{4} - \mu_{\max})^{\frac{1}{3}}\right).$$

Daher können wir die folgende Annäherung an den optimalen Parameter benutzen:

$$\widetilde{\hat{\mu}_{\text{opt}}} = \tilde{t}_{\text{opt}} - \tilde{t}_{\text{opt}}^2, \quad \text{wobei } \tilde{t}_{\text{opt}} = \tfrac{1}{2} + (\tfrac{1}{4} - \mu_{\max})^{\frac{1}{3}}.$$

Obwohl die Konvergenzrate bei dieser Wahl des Parameters schlechter ist, bleibt die Konvergenzordnung unverändert. Dies folgt aus (3.3.32), da sowohl $-\tilde{G}_1(0, \widetilde{\hat{\mu}_{\text{opt}}})$ als auch $\tilde{G}_1(\mu_{\max}, \widetilde{\hat{\mu}_{\text{opt}}})$ gleich $1 - 2(\frac{1}{4} - \mu_{\max})^{\frac{1}{3}} + o\big((\frac{1}{4} - \mu_{\max})^{\frac{1}{3}}\big)$ sind.

Bemerkung 3.3.12 Um die GIBLU(1)-Zerlegung mit der TFF-Zerlegung (siehe Abschnitt 2.2) zu vergleichen, betrachten wir wieder das Modellproblem (2.1.16–2.1.17). Das ist ein Spezialfall von (3.1.2–3.1.4), und das Eigenwertproblem

$$D = \lambda L$$

ist hier äquivalent zu

$$LD^{-1}L = \frac{1}{\lambda^2} D.$$

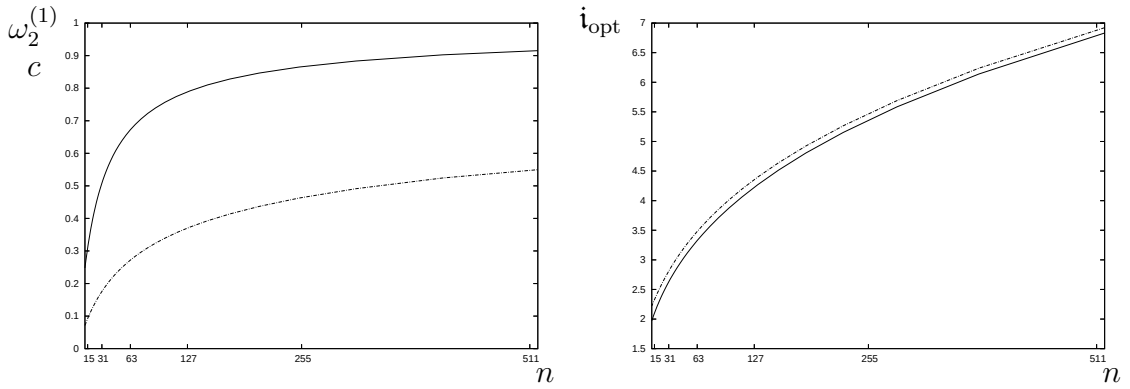

Abbildung 3.3: Konvergenzeigenschaften der GIBLU(1)-Zerlegung mit $\hat{\mu}_0 = \hat{\mu}_1 =: \hat{\mu}$ für die 5-Punkt-Stern-Diskretisierung (1.1.12–1.1.13) des Laplace-Operators (d.h. $a = b = 1$) als Funktionen der Blockgröße n. Links: Die Abschätzung $\rho_1 = \hat{\omega}_{\text{opt}}^2$ der Konvergenzrate der linearen Iteration nach (3.3.38) (durchgezeichnete Kurve) und die Abschätzung c der linearen Konvergenzrate des CG-Verfahrens (gestrichelt), Rechts: Der Eigenwertindex $\mathfrak{i}_{\text{opt}} = \frac{2(n+1)}{\pi} \sqrt{\arcsin \frac{\hat{\mu}_{\text{opt}}^{-1/2}-2}{4}}$ des optimalen Parameters $\hat{\mu}_{\text{opt}} = t_{\text{opt}} - t_{\text{opt}}^2$ nach (3.3.41) (durchgezeichnet) und seiner Annäherung $\widetilde{\hat{\mu}_{\text{opt}}}$ (gestrichelt).

Folglich entspricht jede Ungleichung $\rho(L^{-\frac{1}{2}} D L^{-\frac{1}{2}}) \geq \lambda_{\min}$ der Abschätzung $\rho(BB^T) \leq \dfrac{1}{\lambda_{\min}^2} =: \mu_{\max}$. Dabei besitzt dieses $\mu_{\max}$ die folgende asymptotische Entwicklung:

$$\mu_{\max} = \tfrac{1}{4} - \tfrac{1}{4}(\lambda_{\min} - 2) + O\left((\lambda_{\min} - 2)^2\right). \tag{3.3.44}$$

Einsetzen von (3.3.44) in (3.3.42) liefert:

$$\rho((\mathbf{W}^{(1)})^{-1}\mathbf{R}^{(1)}) \leq 1 - 2\sqrt[3]{2}(\lambda_{\min} - 2)^{\frac{1}{3}} + o\left((\lambda_{\min} - 2)^{\frac{1}{3}}\right).$$

Wie man nach Satz 2.2.7 sieht, ist nicht nur die Ordnung der Konvergenz von diesen Zerlegungen gleich, sondern sogar die konstanten Faktoren vor $(\lambda_{\min} - 2)^{\frac{1}{3}}$. (Siehe auch Bemerkung 3.2.3.) In Abbildung 3.3 sind für diesen Sonderfall die Konvergenzrate und der optimale Parameter $\hat{\mu}$ der GIBLU(1)-Zerlegung als Funktionen der Blockgröße dargestellt. □

Durch die Wahl von zwei nicht notwendig gleichen Parametern $\hat{\mu}_0, \hat{\mu}_1 \in [0, \frac{1}{4})$ können die Konvergenzeigenschaften der GIBLU(1)-Zerlegung noch verbessert werden. Satz 3.3.11 zeigt dann nur eine mögliche Abschätzung, die in diesem Fall nicht mehr optimal ist. Wie aber die Untersuchungen in [11] zeigen, ist dabei keine weitere Verbesserung der Ordnung zu erwarten. Deswegen führen wir eine detailliertere Analyse hier nicht durch und begnügen uns mit Abschätzungen (3.3.42–3.3.43).

3.3.2 Konvergenz der GIBLU(2)-Zerlegung

In diesem Abschnitt zeigen wir, dass die GIBLU(2)-Zerlegungen bei der optimalen Wahl der Parameter $\hat{\mu}_0$, $\hat{\mu}_1$ und $\hat{\mu}_2$ eine noch bessere Konvergenzordnung hat als GIBLU(1). Wie oben beginnen wir mit der Untersuchung des asymptotischen Verhaltens der Folge $\left\{\tilde{\tau}_k^{(2)}(\mu)\right\}_k$.

Lemma 3.3.13 Die Folge der durch die Bedingungen (3.1.19–3.1.20) für $l = 2$ definierten Funktionen $\tilde{\tau}_k^{(2)}$ (siehe (3.2.8)) konvergiert punktweise gegen die Funktion

$$\tilde{\tau}_\infty^{(2)}(\mu) = \frac{(\tau_\infty(\hat{\mu}_0) + \tau_\infty(\hat{\mu}_1) + \tau_\infty(\hat{\mu}_2) - 1) \cdot \mu - \tau_\infty(\hat{\mu}_0)\tau_\infty(\hat{\mu}_1)\tau_\infty(\hat{\mu}_2)}{\mu - (1 - \tau_\infty(\hat{\mu}_0))(1 - \tau_\infty(\hat{\mu}_1))(1 - \tau_\infty(\hat{\mu}_2)) - \tau_\infty(\hat{\mu}_0)\tau_\infty(\hat{\mu}_1)\tau_\infty(\hat{\mu}_2)}, \tag{3.3.45}$$

mit τ_∞ wie in Satz 3.2.1 definiert. Die Funktion $\tilde{\tau}_\infty^{(2)}$ ist für $\hat{\mu}_0, \hat{\mu}_1, \hat{\mu}_2 \in [0, \frac{1}{4})$ auf dem Intervall $[0, \frac{1}{4}]$ wohldefiniert und es gilt:

$$\tilde{\tau}_\infty^{(2)}(\mu) > \tfrac{1}{4} + \mu \geq \sqrt{\mu}, \quad 0 \leq \mu \leq \tfrac{1}{4}. \tag{3.3.46}$$

Beweis: Wie wir in Abschnitt 3.2 erwähnt haben, schreibt sich $\tilde{\tau}_k^{(2)}$ in Form (3.2.9), wobei die Koeffizienten a_k, b_k und d_k durch die Formeln (3.2.24) oder (3.2.27) gegeben sind. Nach Satz 3.2.1 und wegen $\tau_\infty'(\mu) \neq 0$, $\tau_\infty''(\mu) \neq 0$, $\mu \in [0, \frac{1}{4})$, haben diese für $k \to \infty$ endliche Grenzwerte a_∞, b_∞ und d_∞. Folglich gilt für jedes $\mu \in [0, \frac{1}{4})$ die Gleichheit

$$\lim_{k\to\infty} \tilde{\tau}_k^{(2)}(\mu) = \frac{a_\infty \mu + b_\infty}{\mu + d_\infty} =: \tilde{\tau}_\infty^{(2)}(\mu).$$

Wir wollen nun a_∞, b_∞ und d_∞ als Funktionen von $\tau_\infty(\hat{\mu}_i)$ schreiben. Wir betrachten hier nur den Fall $0 \leq \hat{\mu}_0 < \hat{\mu}_1 < \hat{\mu}_2 < \frac{1}{4}$, da alle Überlegungen aus Beweis hierfür sich offensichtlich auf den allgemeinen Fall übertragen lassen. Da nach der Definition von $\tau_\infty(\mu)$ die Gleichheit $\tau_\infty^2(\mu) - \tau_\infty(\mu) + \mu = 0$ gilt, erhalten wir $\hat{\mu}_i = \tau_\infty(\hat{\mu}_i) - \tau_\infty^2(\hat{\mu}_i)$. Einsetzen dieser Gleichung in die formalen Grenzwerte Δ_∞, $\Delta_\infty^{(a)}$, $\Delta_\infty^{(b)}$ und $\Delta_\infty^{(d)}$ von Δ_k, $\Delta_k^{(a)}$, $\Delta_k^{(b)}$ und $\Delta_k^{(d)}$ (siehe den Beweis von Satz 3.2.4), respektive, liefert:

$$\begin{aligned}
\Delta_\infty &= -(\tau_\infty(\hat{\mu}_1) - \tau_\infty(\hat{\mu}_0))(\tau_\infty(\hat{\mu}_2) - \tau_\infty(\hat{\mu}_1))(\tau_\infty(\hat{\mu}_2) - \tau_\infty(\hat{\mu}_0)),\\
\Delta_\infty^{(a)} &= \Delta_\infty \cdot \Big(\tau_\infty(\hat{\mu}_0) + \tau_\infty(\hat{\mu}_1) + \tau_\infty(\hat{\mu}_2) - 1\Big),\\
\Delta_\infty^{(b)} &= -\Delta_\infty \cdot \tau_\infty(\hat{\mu}_0) \cdot \tau_\infty(\hat{\mu}_1) \cdot \tau_\infty(\hat{\mu}_2),\\
\Delta_\infty^{(d)} &= -\Delta_\infty \cdot \Big((1 - \tau_\infty(\hat{\mu}_0))(1 - \tau_\infty(\hat{\mu}_1))(1 - \tau_\infty(\hat{\mu}_2)) + \tau_\infty(\hat{\mu}_0)\tau_\infty(\hat{\mu}_1)\tau_\infty(\hat{\mu}_2)\Big).
\end{aligned}$$

Nach (3.2.16) und (3.2.14) gilt:

$$a_\infty = \frac{\Delta_\infty^{(a)}}{\Delta_\infty}, \quad b_\infty = \frac{\Delta_\infty^{(b)}}{\Delta_\infty}, \quad d_\infty = \frac{\Delta_\infty^{(d)}}{\Delta_\infty}.$$

Daraus erhalten wir (3.3.45).

Die direkte Untersuchung von d_∞ zeigt, dass für $\hat{\mu}_0, \hat{\mu}_1, \hat{\mu}_2 \in [0, \frac{1}{4})$ dieser Wert der Ungleichung $d_\infty < -\frac{1}{4}$ genügt. Also ist $\tilde{\tau}_\infty^{(2)}$ auf $[0, \frac{1}{4}]$ wohldefiniert und stetig.

Da nach Satz 3.2.4 die Funktionen $\tilde{\tau}_k^{(2)}$ auf $[0, \frac{1}{4})$ monoton fallend sind, ist $\tilde{\tau}_\infty^{(2)}$ auf diesem Intervall nicht-steigend. Da außerdem $\tilde{\tau}_\infty^{(2)}$ im Punkt $\mu = \frac{1}{4}$ stetig ist, gilt diese Aussage auf $[0, \frac{1}{4}]$. Die Funktion $\frac{1}{4} + \mu$ ist hingegen steigend. Folglich brauchen wir zum Beweis von (3.3.46), diese Ungleichung nur für $\mu = \frac{1}{4}$ zu prüfen. In diesem Punkt erhalten wir nach (3.3.45):

$$\tilde{\tau}_\infty^{(2)}(\tfrac{1}{4}) - \tfrac{1}{2} = \frac{(\frac{1}{2} - \tau_\infty(\hat{\mu}_0))(\frac{1}{2} - \tau_\infty(\hat{\mu}_1))(\frac{1}{2} - \tau_\infty(\hat{\mu}_2))}{\frac{1}{4} - d_\infty} > 0.$$

Also gilt $\tilde{\tau}_\infty^{(2)}(\mu) > \frac{1}{4} + \mu$ für alle $\mu \in [0, \frac{1}{4}]$, und (3.3.46) ist bewiesen. □

Im Hinblick auf Satz 3.3.4 müssen wir $f_k^{(2)}$ durch die Grenzfunktion $f_\infty^{(2)}$ ersetzen, und es muss

$$\left| f_k^{(2)}(\mu) \right| \le \left| f_\infty^{(2)}(\mu) \right|, \quad \mu \in [0, \tfrac{1}{4}) \tag{3.3.47}$$

gelten. Es ist uns nicht gelungen, die Ungleichung (3.3.47) theoretisch zu beweisen. Allerdings weisen viele Beobachtungen auf die Gültigkeit von (3.3.47) hin, wie wir in der folgenden Bemerkung genauer erläutern. Deswegen behaupten wir hier ohne Beweis, dass diese Ungleichung gilt.

Bemerkung 3.3.14 Die explizite Form von $f_k^{(2)}$ können wir aus (3.3.14) und (3.2.9) ermitteln: Da in diesem Fall $\tilde{P}_k^{(2)}(\mu) = a_k\mu + b_k$ und $\tilde{Q}_k^{(2)}(\mu) = \mu + d_k$ ist, erhalten wir

$$f_k^{(2)}(\mu) = \frac{(\mu - \hat{\mu}_0)(\mu - \hat{\mu}_1)(\mu - \hat{\mu}_2)}{(\mu + d_k)(a_{k-1}\mu + b_{k-1})} = \frac{(\mu - \hat{\mu}_0)(\mu - \hat{\mu}_1)(\mu - \hat{\mu}_2)}{a_{k-1}\left(\mu + \frac{b_{k-1}}{a_{k-1}}\right)(\mu + d_k)}, \quad k \ge 4.$$

Unsere numerischen Experimente für verschiedene Parameter $\hat{\mu}_0, \hat{\mu}_1, \hat{\mu}_2 \in [0, \frac{1}{4})$ zeigen, dass die negativen Werte $\dfrac{b_k}{a_k}$ und d_k mit k wachsen. Sie sind also von oben durch ihre Grenzwerte $\dfrac{b_\infty}{a_\infty}$ bzw. d_∞ beschränkt (siehe Beweis von Lemma 3.3.13), woraus folgt, dass für jedes $\mu \in [0, \frac{1}{4})$ die Ungleichungen

$$\mu + \frac{b_k}{a_k} \le \mu + \frac{b_\infty}{a_\infty} < 0, \quad \mu + d_k \le \mu + d_\infty < 0$$

gelten. Diese Experimente zeigen auch, dass die positive Folge $\{a_k\}_k$ fallend ist. Also sind die Werte a_k von unten mit a_∞ beschränkt. Aus diesen Aussagen können wir beschließen, dass für jedes $\mu \in [0, \frac{1}{4})$ die Folge $f_k^{(2)}(\mu)$ konvergiert, und die punktweise Grenzfunktion $f_\infty^{(2)}$ Ungleichung (3.3.47) genügt. □

Im Folgenden bezeichnen wir die Werte $\tau_\infty(\hat{\mu}_i)$, $i \in \{0, 1, 2\}$, mit t_i. Nach der Definition von $\tau_\infty(\mu)$ gilt dann

$$t_i = \tfrac{1}{2} + \sqrt{\tfrac{1}{4} - \hat{\mu}_i}, \qquad \hat{\mu}_i = t_i - t_i^2 \qquad (0 \le i \le 2). \tag{3.3.48}$$

Also ist $\hat{\mu}_i$ durch t_i eindeutig bestimmt, und umgekehrt. Deswegen können die t_i statt der $\hat{\mu}_i$ als Zerlegungsparameter betrachtet werden. Wir weisen darauf hin, dass τ_∞ eine fallende Funktion ist, und für $0 \le \hat{\mu}_0 \le \hat{\mu}_1 \le \hat{\mu}_2 < \frac{1}{4}$ das Verhältnis $\frac{1}{2} < t_2 \le t_1 \le t_0 \le 1$ gilt. Die Polynome $\tilde{P}_\infty^{(2)}$ und $\tilde{Q}_\infty^{(2)}$ schreiben sich nach (3.3.45) in dieser Notation wie folgt:

$$\tilde{P}_\infty^{(2)}(\mu) = (t_0 + t_1 + t_2 - 1) \cdot \mu - t_0 t_1 t_2,$$
$$\tilde{Q}_\infty^{(2)}(\mu) = \mu - (1 - t_0)(1 - t_1)(1 - t_2) - t_0 t_1 t_2.$$

Nach (3.3.17) erhalten wir dann die Funktion $g_2(\mu)$ in der Form

$$g_2(\mu) = \frac{(\mu - t_0 + t_0^2)(\mu - t_1 + t_1^2)(\mu - t_2 + t_2^2)}{\left(\tilde{P}_\infty^{(2)}(\mu) - \sqrt{\mu}\tilde{Q}_\infty^{(2)}(\mu)\right)^2}.$$

Für theoretische Zwecke werden wir zuweilen die Parameter t_0, t_1 und t_2 nicht nur auf $(\frac{1}{2}, 1]$ betrachten, sondern auf dem größeren Intervall $(\frac{1}{2}, +\infty)$. Solche t_i entsprechen nach (3.3.48) negativen Werten von $\hat{\mu}_i$. Diese Erweiterung ist bei der Analyse von in diesem Abschnitt entstehenden Funktionen, die in diesem Fall wohldefiniert sind, hilfreich. Da die Existenz der Zerlegungen unter diesen Bedingungen nicht bewiesen wurde, formulieren wir den die Ergebnisse dieses Abschnitts zusammenfassenden Satz 3.3.20 (S. 73) nur für zulässige Werte von Parametern, also für $\hat{\mu}_i \in [0, \frac{1}{4})$. Die Benutzung von $t_i > 1$ führt hier also nicht zu Widersprüchen.

Wie in Abschnitt 3.3.1 vereinfachen wir die Funktion g_2 zur Umgehung der Schwierigkeiten bei der direkten Untersuchung: wir ersetzen im Nenner die Wurzel $\sqrt{\mu}$ durch $\frac{1}{4} + \mu$ und verwenden folgendes Lemma.

Lemma 3.3.15 Für beliebige Wahl der Parameter $t_0, t_1, t_2 \in (\frac{1}{2}, 1]$ gilt

$$|g_2(\mu)| \le |\tilde{g}_2(\mu)|, \quad 0 \le \mu < \tfrac{1}{4}, \tag{3.3.49}$$

wobei

$$\tilde{g}_2(\mu) = \frac{(\mu - t_0 + t_0^2)(\mu - t_1 + t_1^2)(\mu - t_2 + t_2^2)}{\left(\tilde{P}_\infty^{(2)}(\mu) - (\frac{1}{4} + \mu)\tilde{Q}_\infty^{(2)}(\mu)\right)^2} = \frac{\Psi(\mu)}{(\Phi(\mu))^2}$$

mit

$$\Psi(\mu) := (\mu - t_0 + t_0^2)(\mu - t_1 + t_1^2)(\mu - t_2 + t_2^2),$$
$$\Phi(\mu) := \tilde{P}_\infty^{(2)}(\mu) - (\tfrac{1}{4} + \mu)\tilde{Q}_\infty^{(2)}(\mu).$$

Bei $t_0, t_1, t_2 \in (\frac{1}{2}, +\infty)$ hat $\Phi(\mu)$ keine Nullstellen in $(-\infty, \frac{1}{4}]$, also ist $\tilde{g}_2(\mu)$ auf diesem Intervall wohldefiniert und stetig.

Beweis: 1. Die Abschätzung (3.3.49) folgt aus der Ungleichung

$$\frac{\tilde{P}_\infty^{(2)}(\mu)}{\tilde{Q}_\infty^{(2)}(\mu)} - \sqrt{\mu} \ge \frac{\tilde{P}_\infty^{(2)}(\mu)}{\tilde{Q}_\infty^{(2)}(\mu)} - (\tfrac{1}{4} + \mu) > 0.$$

(Siehe Lemma 3.3.13.)

2. Für $t_0, t_1, t_2 > \frac{1}{2}$ gilt

$$\Phi'(\mu) = -2\mu + t_0t_1 + t_0t_2 + t_1t_2 - \tfrac{1}{4} > 0, \quad \mu \leq \tfrac{1}{4}.$$

Also ist Φ auf $(-\infty, \frac{1}{4}]$ monoton steigend. Folglich gilt für jedes $\mu \leq \frac{1}{4}$:

$$\Phi(\mu) \leq \Phi(\tfrac{1}{4}) = -(t_0 - \tfrac{1}{2})(t_1 - \tfrac{1}{2})(t_2 - \tfrac{1}{2}) < 0.$$

Somit hat $\Phi(\mu)$ keine Nullstellen auf $(-\infty, \frac{1}{4}]$. Insbesondere ist $\tilde{g}_2$ auf diesem Intervall wohldefiniert und stetig. □

Wir betrachten nun $\tilde{g}_2$ auf dem Intervall $(-\infty, \mu_{\max}]$. Nach Lemmata 3.3.15 und 3.3.13 sowie Hypothese (3.3.47) können wir in der Situation von Satz 3.3.4 die folgenden Werte für $\omega_1^{(2)}$ und $\omega_2^{(2)}$ nehmen:

$$\omega_1^{(2)} = \min_{\mu < \mu_{\max}} \tilde{g}_2(\mu), \qquad \omega_2^{(2)} = \max_{\mu < \mu_{\max}} \tilde{g}_2(\mu). \tag{3.3.50}$$

Da der Zähler von $\tilde{g}_2(\mu)$ einen kleineren Grad als der Nenner hat, gilt

$$\lim_{\mu \to -\infty} \tilde{g}_2(\mu) = 0. \tag{3.3.51}$$

Insbesondere existieren die in (3.3.50) eingeführten Werte $\omega_1^{(2)}$ und $\omega_2^{(2)}$. Wir weisen darauf hin, dass Funktion $\tilde{g}_2$ sogar auf $[0, \frac{1}{4})$ nicht monoton ist. Die nächste Aussage beschreibt die Positionen ihrer Extrema:

Lemma 3.3.16 Für beliebige Wahl von $t_0, t_1, t_2 > \frac{1}{2}$ hat die Funktion $\tilde{g}_2$ in $\mathbb{R}$ genau vier Extrema. Jedes der Intervalle $(-\infty, \hat{\mu}_0]$, $[\hat{\mu}_0, \hat{\mu}_1]$, $[\hat{\mu}_1, \hat{\mu}_2]$ und $(\frac{1}{4}, +\infty)$ enthält genau ein Extremum, wobei ein gemeinsames Extremum nur einem Intervall zugeordnet sein soll.

Beweis: Die Ableitung von $\tilde{g}_2$ bzgl. μ ist

$$\tilde{g}_2'(\mu) = \frac{\Psi'(\mu)\Phi(\mu) - 2\Psi(\mu)\Phi'(\mu)}{(\Phi(\mu))^3}. \tag{3.3.52}$$

Da $\deg \Phi(\mu) = 2$ und $\deg \Psi(\mu) = 3$, erhalten wir für den Zähler von $\tilde{g}_2'$: $\deg\left(\Psi'(\mu)\Phi(\mu) - 2\Psi(\mu)\Phi'(\mu)\right) \leq 4$. Dieser Zähler hat also höchstens vier reelle Nullstellen.

Da $\hat{\mu}_0$, $\hat{\mu}_1$ und $\hat{\mu}_2$ Nullstellen von $\tilde{g}_2$ sind und dazu noch (3.3.51) gilt, enthält jedes der Intervalle $(-\infty, \hat{\mu}_0]$, $[\hat{\mu}_0, \hat{\mu}_1]$ und $[\hat{\mu}_1, \hat{\mu}_2]$ mindestens ein Extremum dieser Funktion. Dies bleibt auch bei $\hat{\mu}_i = \hat{\mu}_{i+1}$ wahr, da in diesem Fall $\tilde{g}_2'(\hat{\mu}_i) = 0$ gilt: siehe (3.3.52). Die Gesamtzahl der auf $(-\infty, \hat{\mu}_2]$ liegenden Extrema ist dabei mindestens 3.

Die Funktion $\tilde{g}_2$ hat für $t_0, t_1, t_2 > \frac{1}{2}$ zwei verschiedene Pole $\mathring{\mu}_0$ und $\mathring{\mu}_1$, die nach Lemma 3.3.15 im Intervall $(\frac{1}{4}, +\infty)$ liegen. Da auf diesem Intervall die Funktion $\tilde{g}_2$ positiv ist (außer natürlich an den Polstellen, wo sie nicht definiert ist, da alle Nullstellen ihres Zählers in dem anderen Intervall liegen), konvergiert $\tilde{g}_2(\mu)$ gegen $+\infty$, wenn μ gegen $\mathring{\mu}_0$ oder $\mathring{\mu}_1$ geht. Folglich hat $\tilde{g}_2$ auf $(\mathring{\mu}_0, \mathring{\mu}_1) \subseteq (\frac{1}{4}, +\infty)$ noch ein Extremum.

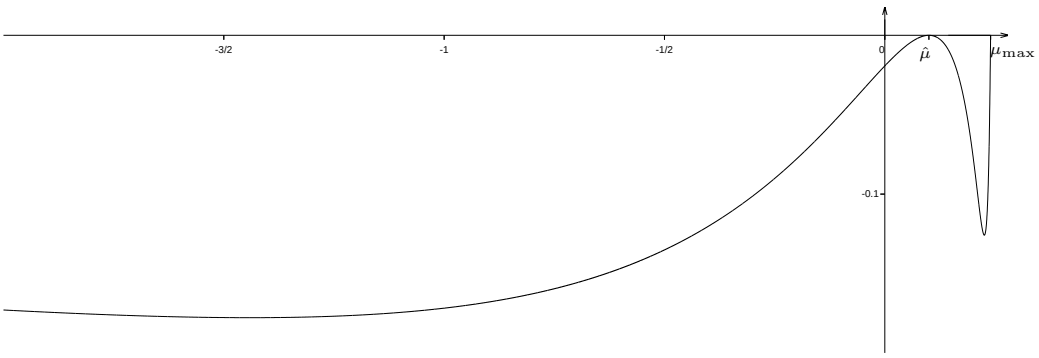

Abbildung 3.4: Die Funktion $\tilde{g}_2(\mu)$ für $\hat{\mu}_0 = \hat{\mu}_1 = \hat{\mu} = 0.1$, $\hat{\mu}_2 = \mu_{\max} = 0.24$ auf $[-2, \mu_{\max}]$.

Insgesamt hat $\tilde{g}_2$ in $\mathbb{R}$ also vier Extrema, deren Positionen in der Behauptung des Lemmas beschrieben sind. □

Um zu zeigen, dass die GIBLU(2)-Zerlegung eine bessere Konvergenzordnung als GIBLU(1) hat, betrachten wir den Spezialfall

$$\hat{\mu}_0 = \hat{\mu}_1 =: \hat{\mu}, \qquad \hat{\mu}_2 = \mu_{\max}, \qquad (\hat{\mu} < \mu_{\max}), \tag{3.3.53}$$

d.h.

$$t_0 = t_1 =: t = \tfrac{1}{2} + \sqrt{\tfrac{1}{4} - \hat{\mu}}, \qquad t_2 = t_{\min} := \tfrac{1}{2} + \sqrt{\tfrac{1}{4} - \mu_{\max}}, \qquad t > t_{\min}. \tag{3.3.54}$$

Die Funktion $\tilde{g}_2$ schreibt sich dann in der Form

$$\tilde{g}_2(\mu) = \frac{(\mu - t + t^2)^2(\mu - t_{\min} + t_{\min}^2)}{\left(\mu^2 - (t^2 + 2tt_{\min} - \frac{1}{4})\mu + \frac{3}{4}t^2 t_{\min} - \frac{1}{4}(1-t)^2(1-t_{\min})\right)^2}.$$

Auf $(-\infty, \mu_{\max}]$ ist sie nicht-positiv, also gilt

$$\omega_2^{(2)} = 0. \tag{3.3.55}$$

Des Weiteren hat $\tilde{g}_2$ auf diesem Intervall nach Lemma 3.3.16 drei lokale Extrema: ein lokales Maximum in $\mu = \hat{\mu}$ und zwei lokale Minima, je eins auf $(-\infty, \hat{\mu})$ und $(\hat{\mu}, \mu_{\max})$ (siehe Abbildung 3.4). Eines von diesen Minima ist nach (3.3.51) auf diesem Intervall global und soll deswegen gleich $\omega_1^{(2)}$ sein. Nach Satz 3.3.1 und (3.3.55) erhalten wir also

$$\begin{aligned} \rho(I - (\mathbf{W}^{(2)})^{-1}\mathbf{A}) &\leq \rho_2 = -\omega_1^{(2)}, \\ \kappa((\mathbf{W}^{(2)})^{-1}\mathbf{A}) &\leq \kappa_2 = 1 - \omega_1^{(2)} = 1 + \rho_2. \end{aligned} \tag{3.3.56}$$

Bemerkung 3.3.17 Wir weisen darauf hin, dass für die Wahl (3.3.53) der Parameter die GIBLU(2)-Zerlegung im allgemeinen nicht konvergent ist: Der Wert $\rho_2 = -\omega_1^{(2)}$ kann größer als oder gleich 1 sein. Wir betrachten diese Zerlegung also nur zur Vorkonditionierung des CG-Verfahrens. □

Bedingung (3.3.54) lässt noch den Parameter t unbestimmt. Wir betrachten in diesem Abschnitt nur den Fall solcher $t =: t_{\text{opt}}$ (d.h. $\hat{\mu} = \hat{\mu}_{\text{opt}}$), für welche die beiden lokalen Minima von $\tilde{g}_2$ auf $(-\infty, \mu_{\max}]$ gleich sind. Dieser gemeinsame Wert ist dann also $\omega_1^{(2)} = -\rho_2$. Wie wir nun zeigen, kann ρ_2 in dieser Situation explizit als eine Funktion von $t_{\min}$ dargestellt werden.

Die Bedingung, dass μ eine Extremalstelle und $\omega = -\rho$ das entsprechende Extremum ist, schreibt sich in der Form des Systems

$$\begin{aligned} \tilde{g}_2(\mu) &= -\rho, \\ \tilde{g}_2'(\mu) &= 0. \end{aligned} \tag{3.3.57}$$

Die zweite Gleichung lässt sich vereinfachen zu

$$\tilde{g}_2'(\mu) = \frac{\Psi'(\mu)}{(\Phi(\mu))^2} - 2\tilde{g}_2(\mu)\frac{\Phi'(\mu)}{\Phi(\mu)} = \frac{\Psi'(\mu)}{(\Phi(\mu))^2} + 2\rho\frac{\Phi'(\mu)}{\Phi(\mu)}$$

(siehe (3.3.52)). Nach Multiplikation beider Gleichungen in (3.3.57) mit $(\Phi(\mu))^2$ erhalten wir ein algebraisches System

$$\begin{aligned} \Psi(\mu) + \rho(\Phi(\mu))^2 &= 0, \\ \Psi'(\mu) + 2\rho\Phi'(\mu)\Phi(\mu) &= 0. \end{aligned} \tag{3.3.58}$$

Bemerkung 3.3.18 Das System (3.3.58) könnte mehr Lösungen haben als (3.3.57). Dies ist aber nicht der Fall: Die Systeme (3.3.58) und (3.3.57) sind für $t, t_{\min} > \frac{1}{2}$ einander äquivalent. In der Tat ist nämlich $\Phi(\mu)$ ungleich Null: Bei $\Phi(\mu) = 0$ wäre die erste Gleichung in (3.3.58) nicht erfüllt, da die Nullstellen von Φ auf $(\frac{1}{4}, +\infty)$ liegen, wo Ψ keine Nullstellen hat (siehe Lemma 3.3.15). Also entsprechen alle Lösungen von (3.3.58) genau den vier in Lemma 3.3.16 beschriebenen Extrema. □

Wir eliminieren nun die Variable μ aus dem System (3.3.58) und erhalten nur eine Gleichung in ρ. Dazu betrachten wir die linken Seiten in (3.3.58) als Polynome von μ, schreiben also die Gleichungen als

$$\begin{aligned} a_4\mu^4 + a_3\mu^3 + a_2\mu^2 + a_1\mu + a_0 &= 0, \\ b_3\mu^3 + b_2\mu^2 + b_1\mu + b_0 &= 0, \end{aligned}$$

wobei

$$\begin{aligned} a_4 &= \rho, \\ a_3 &= (\tfrac{1}{2} - 4tt_{\min} - 2t^2)\rho + 1, \\ a_2 &= \left(-\tfrac{7}{16} + t + \tfrac{1}{2}t_{\min} - 2tt_{\min} - t^2 + 2t^2t_{\min} + 4t^2t_{\min}^2 + 4t^3t_{\min} + t^4\right)\rho \\ &\quad -2t - t_{\min} + 2t^2 + t_{\min}^2, \\ a_1 &= \left(\tfrac{1}{8} - tt_{\min} + \tfrac{1}{2}t^2\right)(-1 + 2t + t_{\min} - 2tt_{\min} - t^2 + 4t^2t_{\min})\rho \\ &\quad +t(1-t)(t + 2t_{\min} - t^2 - 2t_{\min}^2), \\ a_0 &= \left(\tfrac{1}{4}(1-t)^2(1-t_{\min}) - \tfrac{3}{4}t^2t_{\min}\right)^2\rho - (t-t^2)^2(t_{\min} - t_{\min})^2 \end{aligned}$$

und

$$
\begin{aligned}
b_3 &= 4\rho,\\
b_2 &= (\tfrac{3}{2} - 12tt_{\min} - 6t^2)\rho + 3,\\
b_1 &= \left(-\tfrac{7}{8} + 2t + t_{\min} - 4tt_{\min} - 2t^2 + 4t^2t_{\min} + 8t^2t_{\min}^2 + 8t^3t_{\min} + 2t^4\right)\rho\\
&\quad -4t - 2t_{\min} + 4t^2 + 2t_{\min}^2,\\
b_0 &= (\tfrac{1}{8} - tt_{\min} - \tfrac{1}{2}t^2)(-1 + 2t + t_{\min} - 2tt_{\min} - t^2 + 4t^2t_{\min})\rho\\
&\quad +2(t-t^2)(t_{\min} - t_{\min}^2) + (t-t^2)^2.
\end{aligned}
$$

Daraus erhalten wir die Resultante des Systems (3.3.58):

$$
\begin{vmatrix}
a_4 & a_3 & a_2 & a_1 & a_0 & 0 & 0\\
0 & a_4 & a_3 & a_2 & a_1 & a_0 & 0\\
0 & 0 & a_4 & a_3 & a_2 & a_1 & a_0\\
b_3 & b_2 & b_1 & b_0 & 0 & 0 & 0\\
0 & b_3 & b_2 & b_1 & b_0 & 0 & 0\\
0 & 0 & b_3 & b_2 & b_1 & b_0 & 0\\
0 & 0 & 0 & b_3 & b_2 & b_1 & b_0
\end{vmatrix}
= C(t, t_{\min}) \cdot \rho^2 R(\rho, t, t_{\min}),
$$

wobei $C(t, t_{\min}) = \frac{1}{1024}(2t+1)^2(2t-1)^4(t+t_{\min}-1)^4 > 0$ eine von der Unbekannten ρ unabhängige Zahl ist, und

$$R(\rho, t, t_{\min}) = c_3\rho^3 + c_2\rho^2 + c_1\rho + c_0$$

mit

$$
\begin{aligned}
c_3 &= (4t_{\min}^2 - 1)\\
&\quad \times (17 - 32t - 16t_{\min} + 16tt_{\min} + 8t^2 - 64t^2t_{\min} + 64t^2t_{\min}^2 + 64t^3t_{\min} + 16t^4)^2,\\
c_2 &= -145 - 400t - 200t_{\min} + 3248t^2 + 1376t_{\min}^2 + 3312tt_{\min} - 2784t^2t_{\min}\\
&\quad -3840tt_{\min}^2 - 5248t^3 + 896t_{\min}^3 - 9408t^3t_{\min} - 17920t^2t_{\min}^2 - 3968t_{\min}^4 + 160t^4\\
&\quad -11776tt_{\min}^3 + 17024t^4t_{\min} + 18432tt_{\min}^4 + 20992t^2t_{\min}^2 + 44032t^3t_{\min}^2 + 5888t^5\\
&\quad +2048t_{\min}^5 - 16896t^4t_{\min}^2 + 23552t^2t_{\min}^4 - 6144tt_{\min}^5 + 27648t^3t_{\min}^3 - 8960t^5t_{\min}\\
&\quad -3328t^6 - 32768t^2t_{\min}^5 - 20480t^5t_{\min}^2 - 98304t^3t_{\min}^4 + 1536t^6t_{\min} - 69632t^4t_{\min}^3\\
&\quad +12288t^6t_{\min}^2 - 1024t^7t_{\min}^2 + 40960t^3t_{\min}^5 + 8192t^2t_{\min}^6 + 36864t^5t_{\min}^3\\
&\quad +55296t^4t_{\min}^4 - 256t^8,\\
c_1 &= -108 + 288t + 144t_{\min} - 80t^2 + 352t_{\min}^2 + 160tt_{\min} + 256t^2t_{\min}\\
&\quad -2816tt_{\min}^2 - 768t^3 - 128t_{\min}^3 + 4608t^3t_{\min} - 2304t^2t_{\min}^2 + 768t_{\min}^4\\
&\quad -1088t^4 - 256tt_{\min}^3 - 9472t^4t_{\min} + 2048tt_{\min}^4 + 8704t^2t_{\min}^3 + 3072t^3t_{\min}^2\\
&\quad +4608t^5 - 2048t_{\min}^5 + 2560t^4t_{\min}^2 - 4096t^2t_{\min}^4 - 7168t^3t_{\min}^3 + 3584t^5t_{\min}\\
&\quad -2816t^6 + 1024t_{\min}^6,\\
c_0 &= -256(t + t_{\min} - 1)(t - t_{\min})^3.
\end{aligned}
$$

Dies bedeutet, dass ρ dann und nur dann eine Lösung der Gleichung

$$\rho^2 R(\rho, t, t_{\min}) = 0$$

ist, wenn es ein μ gibt, für das (μ, ρ) eine Lösung vom System (3.3.58) ist. Die Wurzel $\rho = 0$ entspricht hier der Extremalstelle $\mu = \hat{\mu}$ der Funktion $\tilde{g}_2$. Die drei Wurzeln der Gleichung

$$R(\rho, t, t_{\min}) = 0 \tag{3.3.59}$$

müssen dann den Minimalstellen dieser Funktion auf $(-\infty, \hat{\mu})$, $(\hat{\mu}, \mu_{\max}]$ und $(\frac{1}{4}, +\infty)$ entsprechen. Wenn die Extrema auf $(-\infty, \hat{\mu})$ und $(\hat{\mu}, \mu_{\max}]$ gleich sind, hat (3.3.59) (betrachtet als eine algebraische Gleichung in ρ) eine zweifache Nullstelle. Dies ist dann eine gemeinsame Nullstelle von R und R'_ρ, d.h. eine Lösung des Systems

$$\begin{aligned} &c_3\rho^3 + c_2\rho^2 + c_1\rho + c_0 = 0, \\ &3c_3\rho^2 + 2c_2\rho + c_1 = 0. \end{aligned} \tag{3.3.60}$$

Die Resultante dieser Gleichungen (die Diskriminante von R), d.h.

$$\begin{vmatrix} c_3 & c_2 & c_1 & c_0 & 0 \\ 0 & c_3 & c_2 & c_1 & c_0 \\ 3c_3 & 2c_2 & c_1 & 0 & 0 \\ 0 & 3c_3 & 2c_2 & c_1 & 0 \\ 0 & 0 & 3c_3 & 2c_2 & c_1 \end{vmatrix},$$

ist durch $4t^2 - 8t_{\min}t + 1$ teilbar. Folglich hat $\tilde{g}_2$ unter der Bedingung

$$4t^2 - 8t_{\min}t + 1 = 0 \tag{3.3.61}$$

zwei gleiche Extrema. Gleichung (3.3.61) ist quadratisch in t. Wir setzen nun t_{opt} gleich der größten Lösung von (3.3.61):

$$t_{\text{opt}} = t_{\min} + \sqrt{t_{\min}^2 - \tfrac{1}{4}}. \tag{3.3.62}$$

Dieser Wert genügt der Ungleichung $t_{\text{opt}} > t_{\min} > \frac{1}{2}$. Für solche Parameter hat $\tilde{g}_2$ (neben dem Extremum in $\mu = \hat{\mu}$) zwei negative Extrema auf $(-\infty, \mu_{\max}]$ und ein positives auf $(\frac{1}{4}, +\infty)$. Folglich sind für $t = t_{\text{opt}}$ die zwei negativen Minima auf $(-\infty, \mu_{\max}]$ gleich, genau wie wir im Auswahlkriterium für t_{opt} gefordert haben.

Bemerkung 3.3.19 Für $t_{\min} \in (\frac{1}{2}, 1]$ liegt t_{opt} in $(\frac{1}{2}, 1+\frac{\sqrt{3}}{2}]$ und kann somit auch negativen Werten von $\hat{\mu}_{\text{opt}}$ entsprechen. Wenn aber $t_{\min}$ nahe genug bei $\frac{1}{2}$ ist, liegt t_{opt} in $(\frac{1}{2}, 1]$. Also kann Formel (3.3.62) für die Untersuchung der asymptotischen Konvergenzrate der Zerlegung verwendet werden. □

Wir geben nun die Konvergenzrate ρ_1 explizit an. Wie wir oben gezeigt haben, ist ρ_1 eine Lösung von den beiden Gleichungen in (3.3.60) für $t = t_{\text{opt}}$. Nach (3.3.62) (oder (3.3.61)) erhalten wir

$$t_{\min} = \tfrac{1}{2}\left(t_{\text{opt}} + \frac{1}{4t_{\text{opt}}}\right). \tag{3.3.63}$$

Einsetzen davon in die zweite Gleichung von (3.3.60) reduziert diese auf

$$\begin{aligned}&\frac{2(t_{\text{opt}}+\frac{1}{2})^2(2t_{\text{opt}}-\frac{1}{2})^4}{t_{\text{opt}}^6}\\&\times\Big(3t_{\text{opt}}(2t_{\text{opt}}-1)(-1+6t_{\text{opt}}+4t_{\text{opt}}^2+8t_{\text{opt}}^3)^2\rho_2\\&\quad-\tfrac{1}{8}+\tfrac{3}{2}t_{\text{opt}}-\tfrac{7}{2}t_{\text{opt}}^2+28t_{\text{opt}}^3-206t_{\text{opt}}^4-136t_{\text{opt}}^5+706t_{\text{opt}}^6\Big)\\&\times(8t_{\text{opt}}(2t_{\text{opt}}-1)\rho_2-1)=0.\end{aligned}\tag{3.3.64}$$

Nur der letzte Faktor in (3.3.64) liefert eine positive Wurzel ρ_2 für $t_{\text{opt}} > \frac{1}{2}$. Folglich gilt:

$$\rho_2 = \frac{1}{8t_{\text{opt}}(2t_{\text{opt}}-1)}.\tag{3.3.65}$$

Nach (3.3.62) erhalten wir daraus:

$$\rho_2 = \frac{1}{8\left(t_{\min}+\sqrt{t_{\min}^2-\frac{1}{4}}\right)\left(2t_{\min}+2\sqrt{t_{\min}^2-\frac{1}{4}}-1\right)}.\tag{3.3.66}$$

Damit gelangen wir zu folgendem Ergebnis:

Satz 3.3.20 Im Fall der Matrix (3.1.2–3.1.4) mit $LD^{-1}L^T \le \mu_{\max}D$, wobei $\mu_{\max} \in [0, \frac{1}{4})$ nahe genug bei $\frac{1}{4}$ liegt, gilt für die GIBLU(2)-Zerlegung zu den Koeffizienten (3.2.23–3.2.26) für die Wahl der Parameter

$$\hat{\mu}_{\text{opt}} := \hat{\mu}_0 = \hat{\mu}_1 = t_{\min}+\sqrt{t_{\min}^2-\tfrac{1}{4}} - \left(t_{\min}+\sqrt{t_{\min}^2-\tfrac{1}{4}}\right)^2,\tag{3.3.67}$$

$t_{\min} = \frac{1}{2}+\sqrt{\frac{1}{4}-\mu_{\max}}$, und

$$\hat{\mu}_2 = \mu_{\max}\tag{3.3.68}$$

die folgende Abschätzung:

$$\kappa((\mathbf{W}^{(2)})^{-1}\mathbf{A}) \le \kappa_2 = 1+\frac{1}{8\left(t_{\min}+\sqrt{t_{\min}^2-\frac{1}{4}}\right)\left(2t_{\min}+2\sqrt{t_{\min}^2-\frac{1}{4}}-1\right)}.\tag{3.3.69}$$

Für $\mu_{\max} \to \frac{1}{4}$ gilt

$$\kappa_2 = 1+(\tfrac{1}{4}-\mu_{\max})^{-\frac{1}{4}}\left(\tfrac{1}{8}+o\left((\tfrac{1}{4}-\mu_{\max})^{\frac{1}{4}}\right)\right).\tag{3.3.70}$$

Beweis: Die Aussagen (3.3.67–3.3.69) folgen aus den oben gemachten Überlegungen (siehe (3.3.53–3.3.54), (3.3.56), (3.3.62) und (3.3.66)).

Für die Abschätzung (3.3.70) stellen wir zunächst $\mu_{\max}$ als eine Funktion von $\dfrac{1}{\rho_2} =: \sigma$ dar. Nach (3.3.65) erhalten wir

$$t_{\text{opt}} = \tfrac{1}{4}(1+\sqrt{1+\sigma})$$

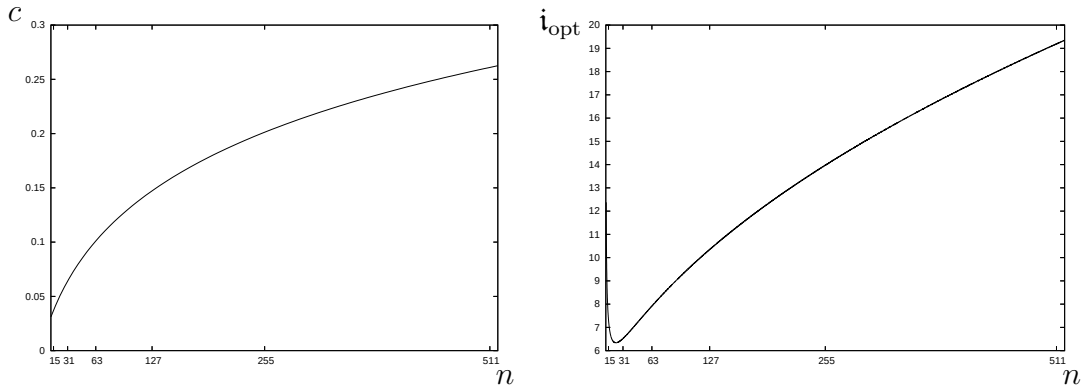

Abbildung 3.5: Konvergenzeigenschaften der GIBLU(2)-Zerlegung mit $\hat{\mu}_0 = \hat{\mu}_1 =: \hat{\mu}_{\text{opt}}$ und $\hat{\mu}_2 = \mu_{\max}$ für die 5-Punkt-Stern-Diskretisierung (1.1.12–1.1.13) des Laplace-Operators (d.h. $a = b = 1$) als Funktionen der Blockgröße n. Links: Die Abschätzung c der linearen Konvergenzrate des CG-Verfahrens, Rechts: Der Eigenwertindex $\mathrm{i}_{\text{opt}} = \frac{2(n+1)}{\pi}\sqrt{\arcsin\frac{\hat{\mu}_{\text{opt}}^{-1/2}-2}{4}}$ des optimalen Parameters $\hat{\mu}_{\text{opt}}$ nach (3.3.67).

(da $t_{\text{opt}} > 0$). Zusammen mit (3.3.63) und $\mu_{\max} = t_{\min} - t_{\min}^2$ liefert dies:

$$\mu_{\max} = \frac{(6 + 2\sqrt{1+\sigma} + \sigma)(2 + 6\sqrt{1+\sigma} - \sigma)}{\left(1 + \sqrt{1+\sigma}\right)^2}.$$

Die Taylor-Entwicklung davon ist

$$\mu_{\max} = \tfrac{1}{4} - \tfrac{1}{4096}\sigma^4 + O(\sigma^5).$$

Nach (3.3.56) erhalten wir daraus (3.3.70). □

Bemerkung 3.3.21 Wir betrachten nun die Fünf-Punkt-Stern-Diskretisierung (1.1.8) des Laplace-Operators (d.h. $a = b = 1$) auf dem strukturierten Gitter (1.1.5). Dann gilt $L = I$, $D = \text{tridiag}\{-1, 4, -1\}$ und

$$\begin{aligned}\mu_{\max} &= \rho(D^{-2}) = \frac{1}{\left(2 + 4\sin^2\dfrac{\pi}{2(n+1)}\right)^2} \\ &= \tfrac{1}{4} - \left(\frac{\pi}{2(n+1)}\right)^2 + O\left(\left(\frac{\pi}{2(n+1)}\right)^4\right),\end{aligned}$$

wobei n die Blockgröße ist (siehe (1.1.12–1.1.13) und (1.1.14)). Nach (3.3.70) erhalten wir also

$$\kappa_2 = O(n^{\frac{1}{2}}).$$

Das bedeutet, dass das mit dieser Zerlegung vorkonditionierte CG-Verfahren die Konvergenzrate der Ordnung $O(n^{\frac{1}{4}})$ hat (siehe [18]). Die Konvergenzrate dieses Verfahrens sowie der Parameter $\hat{\mu}_{\text{opt}}$ werden für den Fall dieser Diskretisierung auf Abbildung 3.5 als Funktionen von n gezeigt.

Wir weisen auch darauf hin, dass Formel (3.3.67) für diese Diskretisierung schon für $n = 15$ positive Werte von $\hat{\mu}_{\text{opt}}$ liefert. Negative Werte können also nur bei sehr groben Gittern entstehen. Für diese Gitter können z.B. die gleichen Parameter wie bei $n = 15$ genommen werden.

Für $\mu_{\max} \to 0$ wird der Wert $\hat{\mu}_{\text{opt}}$ kleiner als $\mu_{\min}$. Er ist daher nicht durch Eigenwertindizes $\mathfrak{i} \in [0, n]$ darstellbar. Deswegen kann sogar für positive, aber kleine Werte von $\hat{\mu}_{\text{opt}}$ die Zahl $\mathfrak{i}_{\text{opt}} = \frac{2(n+1)}{\pi} \sqrt{\arcsin \frac{\hat{\mu}_{\text{opt}}^{-1/2} - 2}{4}}$ größer als n werden, und die rechte Kurve in Abbildung 3.5 hat ein unerwartetes Minimum neben $n = 15$. □

Obwohl wir hier nur den Spezialfall (3.3.53) betrachtet haben, zeigen diese Überlegungen, dass die GIBLU(2)-Zerlegungen eine Verbesserung der Ordnung der Konvergenzrate im Vergleich mit GIBLU(1) liefern. Wir weisen auch darauf hin, dass die numerischen Experimente mit der Funktion g_2 für die optimale Wahl nur eines Parameters $\hat{\mu} := \hat{\mu}_0 = \hat{\mu}_1 = \hat{\mu}_2$ die Abschätzung $\kappa_2 = O(n)$ zeigen, also eine schlechtere als in Satz 3.3.20. Daher sollten bei praktischen Anwendungen für GIBLU(2)-Zerlegungen mindestens zwei Parameter (oder Testvektoren) angegeben werden.

3.4 Numerische Experimente für das Modellproblem

In diesem Abschnitt stellen wir die Ergebnisse der numerischen Experimente mit GIBLU(1)- und GIBLU(2)-Zerlegungen vor. Wir betrachten die knotenzentrierte Finite-Volumen-Diskretisierung des Randwertproblems

$$\begin{aligned} -\Delta u &= 1, \quad (u : \Omega \to \mathbb{R}) \\ u|_{\partial\Omega} &= 1, \end{aligned} \tag{3.4.1}$$

wobei $\Omega = (0,1)^2$ das Einheitsquadrat ist, auf dem strukturierten Dreiecksgitter (siehe Abbildung 1.1). Wie wir oben erwähnt haben (siehe Bemerkung 1.1.1 und (1.1.6–1.1.7), (1.1.12–1.1.13)), hat das diskrete Problem (bis auf den Skalierungsfaktor) die Form (3.1.1–3.1.2) mit

$$D = \text{tridiag}\,\{-1, 4, -1\}, \quad L = I, \tag{3.4.2}$$

und die Bedingungen (3.1.3–3.1.4) sind erfüllt. Wir wollen hier erst zeigen, wie die oben hergeleiteten Abschätzungen mit den experimentellen Daten übereinstimmen, und dann die numerische Stabilität und den Rechenaufwand betrachten, also auch Fragen, die bei der theoretischen Untersuchung nicht behandelt wurden. Die Experimente wurden mit der Hilfe des Programmpakets UG durchgeführt (siehe Anhang A).

Wir betrachten zunächst die oben untersuchten vereinfachten Fälle des optimalen Parameters. Als Abbruchskriterium wählen wir hier die Reduktion der Norm $\|r^{(i)}\|_2$ des Residuums $r^{(i)} = f - \mathbf{A}u^{(i)}$ um einen Faktor von mindestens 10^{10}.

Die GIBLU(1)-Zerlegung ist in diesem Fall wegen $\mathbf{R}^{(1)} \geq 0$ konvergent und kann als Vorkonditionierer für die lineare Iteration benutzt werden. Tabelle 3.1 zeigt die Konvergenzrate

$$\rho_{\text{num}}^{(i)} := \frac{\|r^{(i)}\|_2}{\|r^{(i-1)}\|_2} \tag{3.4.3}$$

im letzten Schritt dieses Verfahrens für den Parameter $\hat{\mu}_0 = \hat{\mu}_1 = \hat{\mu}_{\text{opt}}$ aus Satz 3.3.11.

Tabelle 3.1: Konvergenzrate der mit GIBLU(1)-Zerlegung vorkonditionierten linearen Iteration im Fall $\hat{\mu}_0 = \hat{\mu}_1 = \hat{\mu}_{\text{opt}}$

n	$\hat{\mu}_{\text{opt}}$	Schritte	letzte Rate $\rho_{\text{num}}^{(i)}$	Abschätzung ρ_1
15	0.2128710073	15	0.2186	0.3082
31	0.2342413354	26	0.4125	0.5084
63	0.2434770039	43	0.5927	0.6725
127	0.2473350525	72	0.7334	0.7889
255	0.2489207796	121	0.8315	0.8659
511	0.2495657342	201	0.8955	0.9153

In Tabelle 3.2 stellen wir die bzgl. der Schrittnummer i gemittelte Konvergenzrate (3.4.3) des mit GIBLU(1)-Zerlegung vorkonditionierten CG-Verfahrens unter der gleichen Bedingungen vor. Wir weisen darauf hin, dass die Angabe der optimalen Parameter $\hat{\mu}_{\text{opt}}$ mit der Genauigkeit von 6 Nachkommastellen zu den gleichen 5 in dieser Tabelle angezeigten Ziffern der Konvergenzrate führt.

Tabelle 3.2: Konvergenzrate des mit GIBLU(1)-Zerlegung vorkonditionierten CG-Verfahrens im Fall $\hat{\mu}_0 = \hat{\mu}_1 = \hat{\mu}_{\text{opt}}$

n	$\hat{\mu}_{\text{opt}}$	Schritte	Konvergenzrate	Abschätzung $c = \frac{\sqrt{\kappa_1}-1}{\sqrt{\kappa_1}+1}$
15	0.212871	9	0.0649	0.0918
31	0.234241	12	0.1360	0.1757
63	0.243477	16	0.2246	0.2720
127	0.247335	21	0.3225	0.3703
255	0.248921	27	0.4197	0.4639
511	0.249566	35	0.5114	0.5492

Die Anwendung der GIBLU(2)-Zerlegung reduziert die Anzahl der Schritte wesentlich. Die Tabelle 3.3 zeigt die gemittelte Konvergenzrate (3.4.3) des mit GIBLU(2)-Zerlegung vorkonditionierten CG-Verfahrens für die Parameter $\hat{\mu}_0 = \hat{\mu}_1 = \hat{\mu}_{\text{opt}}$, $\hat{\mu}_2 = \mu_{\max}$ aus Satz 3.3.20 (siehe auch Bemerkung 3.3.21).

Bemerkung 3.4.1 Diese Ergebnisse bedürfen der zwei folgenden Bemerkungen. Erstens ist es bei den praktischen Anwendungen bequemer, die Parameter $\hat{\mu}_i$ nicht

direkt, sondern durch ihre Eigenwertindizes $\mathfrak{i}_i$ bzgl. des Problems $BB^T e = \mu e$, $B = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$, anzugeben:

$$\hat{\mu}_i = \frac{1}{\left(2 + 4\sin^2 \frac{\pi \mathfrak{i}_i}{2(n+1)}\right)^2}$$

für D, L aus (3.4.2) (siehe (1.1.14) und (1.1.13)). Die Abhängigkeit dieser Indizes von n im Fall des optimalen theoretischen Parameters ist in den Abbildungen 3.3 und 3.5 vorgestellt. Die Experimente zeigen, dass die Anzahl der CG-Iterationen sich in etwa gleich bleibt, wenn nur die ganzzahligen Werte $\mathfrak{i}_i$ benutzt werden. Daher muss man nicht für jeden Spezialfall die Parameter mit der oben angezeigten Genauigkeit bestimmen. Wir haben dies hier nur zur Überprüfung der theoretischen Aussagen getan.

Zweitens wurde bei den oben durchgeführten theoretischen Überlegungen nicht die Konvergenzrate, sondern ihre Abschätzung optimiert. Wir können für das in diesem Abschnitt betrachtete Modellproblem bei etwas anderer Wahl der Parameter sogar noch bessere Konvergenz beobachten. Zum Beispiel macht das mit der GIBLU(1)-Zerlegung vorkonditionierte CG-Verfahren unter den gleichen Bedingungen wie im Experiment aus Tabelle 3.2 für $n = 127$ bei $\hat{\mu}_0 = \hat{\mu}_1 = 0.246282$ (d.h. $\mathfrak{i}_0 = \mathfrak{i}_1 = 5$) nur 19 Schritte mit der mittleren Konvergenzrate 0.2975. Eine weitere Verbesserung ist durch die Wahl von ungleichen Parametern möglich: Wählen wir in diesem Experiment (für $n = 127$) $\hat{\mu}_0 = 0.229785$ und $\hat{\mu}_1 = 0.246282$ (d.h. $\mathfrak{i}_0 = 12$, $\mathfrak{i}_1 = 5$), wird das Problem in 18 Schritten bei der mittleren Rate 0.2609 gelöst. Wie man sieht, sind die Unterschiede zwischen den Konvergenzraten nicht allzu groß. Für die GIBLU(2)-Zerlegung sind sie noch kleiner. Das Wesentliche bei der Wahl der Parameter ist also, deren Eigenwertindizes nahe den theoretisch optimalen zu nehmen. □

Als ein weiteres Beispiel betrachten wir das anisotrope Problem (1.1.3–1.1.4) mit $a = \epsilon$, $b = 1$ und $\mathfrak{f} = 1$. Die Diskretisierung durch Finite-Volumen-Verfahren liefert dann die Matrix (3.1.2) mit

$$D = \text{tridiag}\{-\epsilon, 2(\epsilon+1), -\epsilon\}, \qquad L = I \tag{3.4.4}$$

Tabelle 3.3: Konvergenzrate des mit GIBLU(2)-Zerlegung vorkonditionierten CG-Verfahrens im Fall $\hat{\mu}_0 = \hat{\mu}_1 = \hat{\mu}_{\text{opt}}$, $\hat{\mu}_2 = \mu_{\max}$

n	$\hat{\mu}_{\text{opt}}$	Schritte	Konvergenzrate	Abschätzung $c = \frac{\sqrt{\kappa_2}-1}{\sqrt{\kappa_2}+1}$
15	0.0717838	5	0.0086	0.0372
31	0.174185	7	0.0232	0.0641
63	0.216506	9	0.0624	0.1010
127	0.234696	11	0.1093	0.1473
255	0.242825	13	0.1688	0.2013
511	0.246573	16	0.2335	0.2610

(siehe (1.1.12–1.1.13)), die Bedingungen (3.1.3–3.1.4) genügt. Wir erhalten damit also auch einen Spezialfall des Modellproblems. Den Wert ϵ lassen wir zwischen 10^{-6} und 10^6 laufen, sodass wir die starke Anisotropie in beiden Richtungen beschreiben können. Wie wir in Bemerkung 2.2.8 schon erwähnt haben, ist die in diesem Kapitel durchgeführte Art der theoretischen Untersuchung ungenau für kleine ϵ. Die Tabelle 3.4 zeigt, dass für die starke Anisotropie, egal, in welcher Richtung, das mit der GIBLU(1)-Zerlegung vorkonditionierte CG-Verfahren sogar besser konvergiert, als für das isotrope Problem. Diese Experimente wurden mit dem gleichen Abbruchskriterium wie für das vorherige Problem durchgeführt. Wir betrachten hier nur den Fall von gleichen Parametern $\hat{\mu}_0 = \hat{\mu}_1 =: \hat{\mu}$. Der Wert $\hat{\mu}$ ist durch den Eigenwertindex $\mathfrak{i}$ gegeben, der für diese Ergebnisse per Hand gewählt wurde. Die Experimente zeigen, dass je stärker die Anisotropie ist, desto kleinere Rolle die Genauigkeit der Wahl des Parameters spielt: Sie ändert die Anzahl von Iterationen des CG-Verfahrens fast nicht.

Tabelle 3.4: Konvergenz des mit der GIBLU(1)-Zerlegung ($\hat{\mu}_0 = \hat{\mu}_1 =: \hat{\mu}$) vorkonditionierten CG-Verfahren für das anisotrope Problem (3.4.4). Die Ergebnisse sind in der Form „*Eigenwertindex für $\hat{\mu}$: gemittelte Konvergenzrate*“ dargestellt

ϵ	n				
	15	31	63	127	255
10^{-6}	8: $1.4 \cdot 10^{-10}$	11: $8.7 \cdot 10^{-10}$	15: $1.6 \cdot 10^{-8}$	19: $1.2 \cdot 10^{-7}$	24: $8.1 \cdot 10^{-7}$
10^{-3}	7: 0.0001	10: 0.0021	14: 0.0275	18: 0.1407	23: 0.3508
10^{-1}	6: 0.0357	9: 0.0992	13: 0.1835	17: 0.2771	22: 0.3826
1	3: 0.0479	4: 0.1025	5: 0.1817	6: 0.2732	7: 0.3830
10	2: 0.0167	2: 0.0441	2: 0.1013	3: 0.1747	4: 0.2695
10^3	2: $1.5 \cdot 10^{-8}$	2: $2.3 \cdot 10^{-6}$	2: 0.0005	2: 0.0074	2: 0.0327
10^6	1: $6.2 \cdot 10^{-16}$	1: $7.5 \cdot 10^{-16}$	1: $5.2 \cdot 10^{-15}$	1: $9.4 \cdot 10^{-13}$	1: $3.9 \cdot 10^{-11}$

Die Invertierung der Blöcke $\tilde{T}_k^{(l)}$ erfolgt hier durch die Lösung der Systeme (3.1.28–3.1.29). Da die Faktoren $\theta_k^{(i)}$ dabei oft klein sind (siehe Tabellen 3.5 und 3.6), kann die numerische Stabilität der Gaußschen Elimination nicht aus der Diagonaldominanz gefolgert werden. Bei den oben beschriebenen Experimenten ist aber keine Instabilität eingetreten. Der Grund dafür ist offensichtlich die passende Anordnung der Variablen der Systeme (3.1.28) (siehe Abbildung 3.1), bei der die großen und die kleinen Koeffizienten fast gleich oft bei der Elimination vorkommen.

Die Rechenzeit dieser Zerlegungen hängt sehr stark von den Datenstrukturen und somit von der Rechnerarchitektur ab. Da das Paket UG hauptsächlich für Mehrgitterverfahren geeignet ist, sollte man nicht erwarten, dass die für diese Experimente benutzte Implementierung in Konkurrenz mit den Standardlösern von UG treten könnte. (Wir weisen aber darauf hin, dass bei der starken Anisotropie die hier vorgestellten Zerlegungen schneller als Standardmehrgitterverfahren sind.) Wir vergleichen deswegen die Zerlegungen miteinander für zwei verschiedene Rechner: AMD

Athlon 1 GHz und SGI Indigo 2 (200 MHz Prozessor). Die Rechenzeiten für die in Tabellen 3.2 und 3.3 zusammengefassten Experimente sind in Tabelle 3.7 aufgelistet. Wie man sieht, ist der gesamte Aufwand für die beiden Zerlegungen fast gleich. Obwohl die GIBLU(2)-Zerlegung theoretisch eine bessere Konvergenzordnung hat, ist dies in den Experimenten nur bei der alten Rechnerarchitektur und bei großen Gittern feststellbar.

Tabelle 3.5: Grenzwerte von $\theta_k^{(i)}$ der GIBLU(1)-Zerlegung für $\hat{\mu}_0 = \hat{\mu}_1 = \hat{\mu}_{\text{opt}}$

n	$\hat{\mu}_{\text{opt}}$	$\theta_\infty^{(1)}$	$\theta_\infty^{(0)}$
15	0.2128710073	1.2451	0.38538
31	0.2342413354	1.5585	0.25107
63	0.2434770039	2.0881	0.16153
127	0.2473350525	2.9472	0.10325
255	0.2489207796	4.3214	0.065703
511	0.2495657342	6.5088	0.041678

Tabelle 3.6: Grenzwerte der Koeffizienten $\theta_k^{(i)}$ der GIBLU(2)-Zerlegung im Fall $\hat{\mu}_0 = \hat{\mu}_1 = \hat{\mu}_{\text{opt}}$, $\hat{\mu}_2 = \mu_{\max}$

n	$\hat{\mu}_{\text{opt}}$	$\theta_\infty^{(2)}$	$\theta_\infty^{(1)}$	$\theta_\infty^{(0)}$
15	0.07178375069	0.99521	1.1437	0.44574
31	0.1741847008	0.93546	2.1492	0.16413
63	0.2165063028	0.83664	5.4260	0.053902
127	0.2346956790	0.74264	15.739	0.017049
255	0.2428250386	0.67013	47.694	0.0054140
511	0.2465728857	0.61840	144.93	0.0017511

Tabelle 3.7: Rechenzeit für das GIBLU(l)-vorkonditionierte CG-Verfahren (Sekunden)

n	Athlon				SGI Indigo 2			
	GIBLU(1)		GIBLU(2)		GIBLU(1)		GIBLU(2)	
	gesamt	pro Iter.	gesamt	pro Iter.	gesamt	pro Iter.	gesamt	pro Iter.
15	0.12	0.013	0.1	0.020	1.04	0.12	1.05	0.21
31	0.28	0.023	0.27	0.039	1.34	0.11	1.47	0.21
63	1.50	0.094	1.22	0.14	7.32	0.46	7.13	0.79
127	6.91	0.33	5.57	0.51	47.37	2.3	39.88	3.6
255	41.81	1.5	42.59	3.3	261.6	9.7	203.3	15.6
511	307.1	8.8	328.4	20.5	—	—	—	—

Kapitel 4

GIBLU(l)-Zerlegungen für allgemeine schwachbesetzte Gleichungssysteme

In Definition 3.1.2 haben wir GIBLU(l)-Zerlegungen für jede Matrix der allgemeinen Form $\mathbf{A} = \text{blocktridiag}\,\{-L_k, D_k, -U_k\}$ definiert. Für die praktischen Anwendungen auf schwachbesetzte Gleichungssysteme bleiben noch zwei Fragen zu beantworten: die nach der Wahl der Koeffizienten $\theta_k^{(i)}$ und die der Darstellung der Steifigkeitsmatrix in blocktridiagonaler Form. In diesem Kapitel beschreiben wir mögliche Lösungen dieser Probleme.

4.1 Wahl der Koeffizienten

Zunächst beschreiben wir die möglichen Vorgehensweisen zur Wahl der Koeffizienten $\theta_k^{(i)}$. Wir betrachten eine Matrix der Form

$$\mathbf{A} = \begin{pmatrix} D_1 & -U_1 & & \\ -L_2 & D_2 & \ddots & \\ & \ddots & \ddots & -U_{N-1} \\ & & -L_N & D_N \end{pmatrix}, \tag{4.1.1}$$

wobei D_k, L_k und U_k Blöcke der Größe $n_k \times n_k$, $n_k \times n_{k-1}$ und $n_k \times n_{k+1}$, respektive, sind. Die Überlegungen aus Abschnitt 3.1 können wir nicht auf diesen Abschnitt übertragen, da sich das Problem im Allgemeinen nicht auf die Untersuchung von skalaren Funktionen reduzieren lässt. Wir betrachten daher andere Verfahren. Wie wir aber zeigen, genügen die so berechneten Koeffizienten im Fall des Modellproblems (3.1.1–3.1.4) den Gleichungen (3.1.19–3.1.20).

Beim ersten Verfahren wählen wir statt der Frequenzparameter $\hat{\mu}_0, \ldots, \hat{\mu}_l$ die *Testvektoren* $\mathbf{e}^{(0)}, \ldots, \mathbf{e}^{(l)}$, $\mathbf{e}^{(i)} = \text{blockvector}\,\{e_k^{(i)}\}$, und reduzieren die Blockmatrix $\mathbf{A}$ auf $l+1$ (skalare) tridiagonale Matrizen

$$\mathcal{A}^{(i)} = \text{tridiag}\,\{-a_k^{(i)}, d_k^{(i)}, -b_k^{(i)}\}, \quad i \in \{0, \ldots, l\},$$

mit

$$\begin{aligned} d_k^{(i)} &= (D_k e_k^{(i)}, e_k^{(i)}), \\ a_k^{(i)} &= (L_k e_{k-1}^{(i)}, e_k^{(i)}), \\ b_k^{(i)} &= (U_k e_{k+1}^{(i)}, e_k^{(i)}). \end{aligned} \tag{4.1.2}$$

Die vollständigen LU-Zerlegungen der Matrizen $\mathcal{A}^{(i)}$ seien

$$\mathcal{A}^{(i)} = (\mathcal{L}^{(i)} + \mathcal{T}^{(i)}) \left(\mathcal{T}^{(i)}\right)^{-1} (\mathcal{T}^{(i)} + \mathcal{U}^{(i)}),$$

wobei $\mathcal{T}^{(i)} := \operatorname{diag} \left\{t_k^{(i)}\right\}$, $\mathcal{L}^{(i)} := \operatorname{tridiag} \{-a_k^{(i)}, 0, 0\}$ und $\mathcal{U}^{(i)} := \operatorname{tridiag} \{0, 0, -b_k^{(i)}\}$. Die Koeffizienten $t_k^{(i)}$ genügen dann der Rekursionsformel

$$t_k^{(i)} = \begin{cases} d_1^{(i)}, & k = 1, \\ d_k^{(i)} - \dfrac{a_k^{(i)} b_{k-1}^{(i)}}{t_{k-1}^{(i)}}, & k \geq 2. \end{cases} \tag{4.1.3}$$

Sie spielen hier die gleiche Rolle wie die Werte $\tau_k(\hat{\mu}_i)$ in Abschnitt 3.1.

Für jede Matrix $\mathcal{A}^{(i)}$ konstruieren wir die GIBLU(l)-Zerlegung mit noch unbekannten Koeffizienten $\theta_k^{(0)}, \ldots, \theta_k^{(l)}$:

$$\mathcal{A}^{(i)} = \mathcal{W}^{(i)} - \mathcal{R}^{(i)}, \qquad \mathcal{W}^{(i)} = (\mathcal{L}^{(i)} + \tilde{\mathcal{T}}^{(i,l)}) \left(\tilde{\mathcal{T}}^{(i,l)}\right)^{-1} (\tilde{\mathcal{T}}^{(i,l)} + \mathcal{U}^{(i)}) \tag{4.1.4}$$

mit $\tilde{\mathcal{T}}^{(i,l)} = \operatorname{diag} \left\{\tilde{t}_k^{(i,l)}\right\}$, wobei

$$\begin{aligned} \tilde{t}_k^{(i,l)} &= t_k^{(i)}, \quad 1 \leq k \leq l+1, \\ \tilde{t}_k^{(i,l)} &= \theta_k^{(l)} d_k^{(i)} - \cfrac{a_k^{(i)} b_{k-1}^{(i)}}{\theta_{k-1}^{(l-1)} d_{k-1}^{(i)} - \cfrac{a_{k-1}^{(i)} b_{k-2}^{(i)}}{\ddots \; - \cfrac{a_{k-l+1}^{(i)} b_{k-l}^{(i)}}{\theta_{k-l}^{(0)} d_{k-l}^{(i)}}}}, \quad k \geq l+2. \end{aligned} \tag{4.1.5}$$

Diese Koeffizienten ersetzen hier in gewissem Sinn die für das Modellproblem eingeführten Größen $\tilde{\tau}_k^{(l)}(\hat{\mu}_i)$ (siehe (3.1.24)). Analog zu den Gleichungen (3.1.19) formulieren wir deswegen die folgenden Bedingungen an die Koeffizienten $\theta_k^{(0)}, \ldots, \theta_k^{(l)}$ für jedes $k \geq l+2$:

$$\tilde{t}_k^{(i,l)} = t_k^{(i)}, \quad 0 \leq i \leq l. \tag{4.1.6}$$

Das ist ein System von $l+1$ Gleichungen mit $l+1$ Unbekannten. Die Koeffizienten der zu konstruierenden GIBLU(l)-Zerlegung sollen damit also vollständig bestimmt sein.

Bemerkung 4.1.1 1. Die Größen $\tilde{t}_k^{(i,l)}$ können auch anders definiert werden. Die aus (3.1.31–3.1.32) entstehenden Matrizen

$$\begin{pmatrix} D_1 & -U_1 & & \\ -L_2 & D_2 & \ddots & \\ & \ddots & \ddots & -U_{k-1} \\ & & -L_k & D_k \end{pmatrix} \text{ und } \begin{pmatrix} \theta_{k-l}^{(0)} D_{k-l} & -U_{k-l} & & \\ -L_{k-l+1} & \theta_{k-l+1}^{(1)} D_{k-l+1} & \ddots & \\ & \ddots & \ddots & -U_{k-1} \\ & & -L_k & \theta_k^{(l)} D_k \end{pmatrix} \tag{4.1.7}$$

(siehe auch (3.1.29) und (3.1.28)) lassen sich nach der gleichen Vorgehensweise wie für $\mathbf{A}$ auf die tridiagonalen Matrizen

$$\begin{pmatrix} d_1^{(i)} & -b_1^{(i)} & & \\ -a_2^{(i)} & d_2^{(i)} & \ddots & \\ & \ddots & \ddots & -b_{k-1}^{(i)} \\ & & -a_k^{(i)} & d_k^{(i)} \end{pmatrix} \text{ und } \begin{pmatrix} \theta_{k-l}^{(0)} d_{k-l}^{(i)} & -b_{k-l}^{(i)} & & \\ -a_{k-l+1}^{(i)} & \theta_{k-l+1}^{(1)} d_{k-l+1}^{(i)} & \ddots & \\ & \ddots & \ddots & -b_{k-1}^{(i)} \\ & & -a_k^{(i)} & \theta_k^{(l)} d_k^{(i)} \end{pmatrix},$$

respektive, reduzieren. Die Werte $\tilde{t}_k^{(i,l)}$ sind die letzten Diagonalkoeffizienten der vollständigen LU-Zerlegungen dieser Matrizen, analog wie $\tilde{T}_k^{(l)}$ sind die letzten Diagonalblöcke der vollständigen Block-LU-Zerlegungen von (4.1.7).

2. Die Gleichungen (4.1.6) bedeuten, dass in (4.1.4) die vollständigen LU-Zerlegungen von $\mathcal{A}^{(i)}$ stehen. Die äquivalente Form dieser Bedingungen ist also

$$\mathcal{R}^{(i)} = 0, \quad 0 \le i \le l.$$

□

Bemerkung 4.1.2 Zur Matrix (3.1.2–3.1.4) wählen wir $\mathbf{e}^{(i)} = \text{blockvector}\{e^{(i)}\}$ mit $e^{(i)} \in \mathbb{R}^n$. Dann gilt

$$d_k^{(i)} = (De^{(i)}, e^{(i)}) =: d^{(i)}, \quad a_k^{(i)} = b_{k-1}^{(i)} = (Le^{(i)}, e^{(i)}) =: a^{(i)},$$

und wir erhalten nach (4.1.3):

$$\frac{t_k^{(i)}}{d^{(i)}} = \begin{cases} 1, & k = 1, \\ 1 - \dfrac{(a^{(i)}/d^{(i)})^2}{t_{k-1}^{(i)}/d^{(i)}}, & k \ge 2. \end{cases}$$

Falls wir also

$$\hat{\mu}_i = \left(\frac{a^{(i)}}{d^{(i)}}\right)^2, \quad 0 \le i \le l,$$

setzen, gilt $\dfrac{t_k^{(i)}}{d^{(i)}} = \tau_k(\hat{\mu}_i)$ (siehe (3.1.9)). Auf ähnliche Weise erhalten wir aus (4.1.5) für diese Wahl von $\hat{\mu}_i$ die Gleichheit $\dfrac{\tilde{t}_k^{(i,l)}}{d^{(i)}} = \tilde{\tau}_k^{(l)}(\hat{\mu}_i)$. Daraus folgt, dass die Bedingungen (4.1.6) und (3.1.19) im Fall des Modellproblems (3.1.1–3.1.4) einander äquivalent

sind, und die hier beschriebene Vorgehensweise definiert die gleichen Koeffizienten $\theta_k^{(i)}$ wie in Kapitel 3. □

In (4.1.6) setzt man die Existenz der Koeffizienten $t_k^{(i)}$ in (4.1.3) voraus. Diese hängt von den Eigenschaften der Matrix $\mathbf{A}$ sowie von der Wahl der Testvektoren ab. Beispielsweise können wir die folgenden Voraussetzungen machen: $\mathbf{A}$ ist symmetrisch und positiv definit; für alle Testvektoren $\mathbf{e}^{(i)}$, $0 \le i \le l$, gilt $e_k^{(i)} \neq 0$, $1 \le k \le N$. Dann sind die Matrizen $\mathcal{A}^{(i)}$ positiv definit, da für jeden Vektor $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)^T \in \mathbb{R}^N \setminus \{0\}$

$$(\mathcal{A}^{(i)}\boldsymbol{\alpha}, \boldsymbol{\alpha}) = (\mathbf{Ae}, \mathbf{e}) > 0$$

gilt, wobei $\mathbf{e} = \text{blockvector}\,\{\alpha_k e_k^{(i)}\} \neq 0$. Deswegen besitzt in diesem Fall jedes $\mathcal{A}^{(i)}$ eine vollständige LU-Zerlegung.

Wir betrachten nun die Fälle $l = 1$ und $l = 2$ ausführlicher. Für $l = 1$ lässt sich das System (4.1.6) für jedes $k \ge 3$ in der folgenden Form darstellen:

$$\begin{aligned} \theta_k^{(1)} d_k^{(0)} - \frac{a_k^{(0)} b_{k-1}^{(0)}}{\theta_{k-1}^{(0)} d_{k-1}^{(0)}} &= t_k^{(0)}, \\ \theta_k^{(1)} d_k^{(1)} - \frac{a_k^{(1)} b_{k-1}^{(1)}}{\theta_{k-1}^{(0)} d_{k-1}^{(1)}} &= t_k^{(1)}. \end{aligned} \tag{4.1.8}$$

Das ist ein lineares Gleichungssystem für die Unbekannten $\theta_k^{(1)}$ und $\left(\theta_k^{(0)}\right)^{-1}$. Im folgenden Satz untersuchen wir die Lösbarkeit dieses Systems:

Satz 4.1.3 Die Matrix $\mathbf{A}$ aus (4.1.1) sei symmetrisch und positiv definit. Für die Testvektoren $\mathbf{e}^{(0)}$ und $\mathbf{e}^{(1)}$ gelte

$$(D_k e_k^{(0)}, e_k^{(0)}) = (D_k e_k^{(1)}, e_k^{(1)}) = 1, \quad 1 \le k \le N. \tag{4.1.9}$$

Außerdem seien für alle $k \ge 3$ die folgenden Bedingungen erfüllt:

$$|a_k^{(0)}| \neq |a_k^{(1)}|, \tag{4.1.10}$$

$$t_k^{(0)} \neq t_k^{(1)}. \tag{4.1.11}$$

Dann sind die Koeffizienten $\theta_k^{(1)}$ und $\theta_{k-1}^{(0)}$ der GIBLU(1)-Zerlegung durch (4.1.8) für alle $k \ge 3$ wohldefiniert, und es gilt $\theta_{k-1}^{(0)} \neq 0$.

Beweis: Wir betrachten (4.1.8) für ein festgewähltes $k \ge 3$ als ein lineares System von Unbekannten $\xi = \theta_k^{(1)}$ und $\zeta = \left(\theta_k^{(0)}\right)^{-1}$. Nach (4.1.9) gilt $d_k^{(i)} = 1$, $i \in \{0, 1\}$, also lautet (4.1.8) wegen der Symmetrie von $\mathbf{A}$:

$$\begin{aligned} \xi - (a_k^{(0)})^2 \zeta &= t_k^{(0)}, \\ \xi - (a_k^{(1)})^2 \zeta &= t_k^{(1)}. \end{aligned} \tag{4.1.12}$$

Die Determinante dieses Systems ist $(a_k^{(1)})^2 - (a_k^{(0)})^2$, also nach (4.1.10) ungleich Null. Somit besitzt (4.1.12) genau eine Lösung (ξ^*, ζ^*). Da nach (4.1.11)

$$\begin{vmatrix} 1 & t_k^{(0)} \\ 1 & t_k^{(1)} \end{vmatrix} \neq 0$$

gilt, ist $\zeta^* \neq 0$. Folglich entspricht (ξ^*, ζ^*) der Lösung $\theta_k^{(1)} = \xi$, $\theta_k^{(0)} = 1/\zeta^* \neq 0$ des Systems (4.1.8). □

In einigen speziellen Fällen können (4.1.10) und (4.1.11) durch Bedingungen an die Variablen $a_k^{(i)}$ ersetzt werden, und die Ungleichheiten (4.1.11) folgen daraus: Falls für alle $k \geq 2$ Ungleichung

$$|a_k^{(0)}| < |a_k^{(1)}| \tag{4.1.13}$$

gilt, erhalten wir $t_k^{(0)} > t_k^{(1)}$, $k \geq 2$, also ist (4.1.11) automatisch erfüllt. Diese Aussage lässt sich durch Induktion beweisen: Es gilt $t_2^{(0)} = 1 - (a_2^{(0)})^2 > 1 - (a_2^{(1)})^2 = t_2^{(1)}$, und unter der Bedingung $t_{k-1}^{(0)} > t_{k-1}^{(1)}$ erhalten wir

$$t_k^{(0)} = 1 - \frac{(a_k^{(0)})^2}{t_{k-1}^{(0)}} > 1 - \frac{(a_k^{(1)})^2}{t_{k-1}^{(1)}} = t_k^{(1)}.$$

(Alle $t_k^{(i)}$ sind positiv, weil $\mathcal{A}^{(i)}$ positiv definit sind.)

Bedingung (4.1.13) kann oft durch die Wahl der Testvektoren erfüllt werden: Die Größen $a_k^{(i)}$ charakterisieren Spektraleigeschaften von L_k bezüglich D_k. Wenn alle $a_k^{(i)}$ positiv sind, bedeutet (4.1.13), dass $\mathbf{e}^{(0)}$ die „niedrigen“ Frequenzen dieser Matrix annähert, wobei $\mathbf{e}^{(1)}$ — die „höheren“.

Die praktische Berechnung der Koeffizienten der GIBLU(1)-Zerlegung aus System (4.1.8) kann nach Algorithmus 4.1.4 erfolgen:

Algorithmus 4.1.4 *(Berechnung der GIBLU(1)-Koeffizienten mit zwei Testvektoren)* Die Systemmatrix $\mathbf{A}$ habe die Form (4.1.1), $\mathbf{e}^{(0)} = \text{blockvector}\,\{e_k^{(0)}\}$ und $\mathbf{e}^{(1)} = \text{blockvector}\,\{e_k^{(1)}\}$ seien zwei verschiedene Testvektoren, und die Größen $d_k^{(i)}$, $a_k^{(i)}$ und $b_k^{(i)}$ ($i \in \{0,1\}$) seien durch (4.1.2) definiert. Die Koeffizienten $\theta_k^{(0)}$ und $\theta_k^{(1)}$ der GIBLU(1)-Zerlegung lassen sich in folgenden Schritten auswerten:

BEGIN

1: $t_2^{(0)} \leftarrow d_2^{(0)} - \dfrac{a_2^{(0)} b_1^{(0)}}{d_1^{(0)}}$; $\quad t_2^{(1)} \leftarrow d_2^{(1)} - \dfrac{a_2^{(1)} b_1^{(1)}}{d_1^{(1)}}$;

2: **for all** k **from** 3 **to** N **do**

3: $\quad t_k^{(0)} \leftarrow d_k^{(0)} - \dfrac{a_k^{(0)} b_{k-1}^{(0)}}{t_{k-1}^{(0)}}$; $\quad t_k^{(1)} \leftarrow d_k^{(1)} - \dfrac{a_k^{(1)} b_{k-1}^{(1)}}{t_{k-1}^{(1)}}$;

4: $\quad \Delta_k \leftarrow \dfrac{a_k^{(0)} b_{k-1}^{(0)} d_k^{(1)}}{d_{k-1}^{(0)}} - \dfrac{a_k^{(1)} b_{k-1}^{(1)} d_k^{(0)}}{d_{k-1}^{(1)}}$;

5: $\quad \theta_k^{(1)} \leftarrow \dfrac{a_k^{(0)} b_{k-1}^{(0)} t_k^{(1)}/d_{k-1}^{(0)} - a_k^{(1)} b_{k-1}^{(1)} t_k^{(0)}/d_{k-1}^{(1)}}{\Delta_k}$;

6: $\quad \theta_{k-1}^{(0)} \leftarrow \dfrac{\Delta_k}{d_k^{(0)} t_k^{(1)} - d_k^{(1)} t_k^{(0)}}$

7: **end for**

END □

In der oben beschriebenen Vorgehensweise zur Berechnung der Koeffizienten der GIBLU(1)-Zerlegungen ist es wesentlich, dass die Vektoren $\mathbf{e}^{(0)}$ und $\mathbf{e}^{(1)}$ nicht identisch sind. Aber bei praktischen Anwendungen ist es oft zu kompliziert, zwei solche Testvektoren zu wählen. Deswegen untersuchen wir die Möglichkeit, die Konstruktion der GIBLU(1)-Zerlegung für $\hat{\mu}_0 = \hat{\mu}_1$ auf den allgemeineren Fall zu übertragen. Wir betrachten dazu den Fall, dass der zweite Testvektor $\mathbf{e}^{(1)}$ gegen den ersten geht: Wir haben also den Testvektor $\mathbf{e}^{(0)} = \text{blockvector}\,\{e_k^{(0)}\}$ und eine gegen ihn konvergierende Folge $\left\{\mathbf{e}^{(1,j)}\right\}_{j\geq 1}$, $\mathbf{e}^{(1,j)} = \text{blockvector}\,\{e_k^{(1,j)}\}$. Wir bezeichnen

$$\begin{aligned} d_k^{(1,j)} &= (D_k e_k^{(1,j)}, e_k^{(1,j)}), \\ a_k^{(1,j)} &= (L_k e_{k-1}^{(1,j)}, e_k^{(1,j)}), \\ b_k^{(1,j)} &= (U_k e_{k+1}^{(1,j)}, e_k^{(1,j)}). \end{aligned} \tag{4.1.14}$$

Der folgende Satz zeigt, dass unter gewissen Bedingungen die nach Algorithmus 4.1.4 für Paare $(\mathbf{e}^{(0)}, \mathbf{e}^{(1,j)})$ berechneten Koeffizienten auch einen Grenzwert haben:

Satz 4.1.5 Es sei $\mathbf{A}$ eine symmetrische, positiv definite Matrix der Form (4.1.1), $\mathbf{e}^{(0)} = \text{blockvector}\,\{e_k^{(0)}\}$ ein Testvektor und $\mathbf{e}^{(1,j)} = \text{blockvector}\,\{e_k^{(1,j)}\}$ $(j \geq 1)$ eine für $j \to \infty$ gegen $\mathbf{e}^{(0)}$ konvergierende Folge von Testvektoren. Dabei gelte

$$\begin{aligned} &(D_k e_k^{(0)}, e_k^{(0)}) = 1, \quad 1 \leq k \leq N, \\ &(D_k e_k^{(1,j)}, e_k^{(1,j)}) = 1, \quad 1 \leq k \leq N, \quad j \geq 1. \end{aligned} \tag{4.1.15}$$

Wenn

$$\lim_{j\to\infty} \frac{(a_{k-1}^{(0)})^2 - (a_{k-1}^{(1,j)})^2}{(a_k^{(0)})^2 - (a_k^{(1,j)})^2} = 1 \tag{4.1.16}$$

gilt, und für jedes Paar $(\mathbf{e}^{(0)}, \mathbf{e}^{(1,j)})$ die Koeffizienten $\theta_k^{(1)}$ und $\theta_{k-1}^{(0)}$ als eindeutige Lösungen von (4.1.8) existieren, konvergieren sie für $j \to \infty$ gegen

$$\begin{aligned} \theta_k^{(1)} &= t_k^{(0)} - (a_k^{(0)})^2 t'_k, \\ \theta_{k-1}^{(0)} &= \frac{1}{t'_k} \end{aligned} \tag{4.1.17}$$

$(k \geq 3)$, respektive. Hierbei sind die Größen t'_k durch die Rekursionsformel

$$t'_k = \begin{cases} 0, & k = 1, \\ -\dfrac{1}{t_{k-1}^{(0)}} + \left(\dfrac{a_k^{(0)}}{t_{k-1}^{(0)}}\right)^2 t'_{k-1}, & k \geq 2 \end{cases} \tag{4.1.18}$$

gegeben.

Beweis: Da nach der Symmetrie von $\mathbf{A}$ und (4.1.15) die Gleichungen $d^{(0)} = d^{(1,j)} = 1$, $b_{k-1}^{(0)} = a_k^{(0)}$, $b_{k-1}^{(1,j)} = a_k^{(1,j)}$ gelten, lässt sich die Lösung des Systems (4.1.8) für die

Testvektoren $\mathbf{e}^{(0)}$ und $\mathbf{e}^{(1,j)}$ in der folgenden Form schreiben:

$$\begin{aligned}\theta_k^{(1)} &= \frac{(a_k^{(0)})^2 t_k^{(1,j)} - (a_k^{(1,j)})^2 t_k^{(0)}}{(a_k^{(0)})^2 - (a_k^{(1,j)})^2} = t_k^{(1,j)} - (a_k^{(1,j)})^2 \Delta t_k^{(j)}, \\ \theta_{k-1}^{(0)} &= \frac{(a_k^{(0)})^2 - (a_k^{(1,j)})^2}{t_k^{(1,j)} - t_k^{(0)}} = -\frac{1}{\Delta t_k^{(j)}},\end{aligned} \tag{4.1.19}$$

wobei

$$\Delta t_k^{(j)} = \frac{t_k^{(0)} - t_k^{(1,j)}}{(a_k^{(0)})^2 - (a_k^{(1,j)})^2} \tag{4.1.20}$$

und

$$t_k^{(1,j)} = \begin{cases} 1, & k = 1, \\ 1 - \dfrac{(a_k^{(1,j)})^2}{t_{k-1}^{(1,j)}}, & k \geq 2. \end{cases} \tag{4.1.21}$$

Für $\Delta t_k^{(j)}$ erhalten wir aus (4.1.20), (4.1.3) und (4.1.21):

$$\begin{aligned}\Delta t_k^{(j)} &= \frac{\dfrac{(a_k^{(1,j)})^2}{t_{k-1}^{(1,j)}} - \dfrac{(a_k^{(0)})^2}{t_{k-1}^{(0)}}}{(a_k^{(0)})^2 - (a_k^{(1,j)})^2} \\ &= -\frac{1}{t_{k-1}^{(1,j)}} + \frac{(a_k^{(1,j)})^2}{t_{k-1}^{(0)} t_{k-1}^{(1,j)}} \cdot \frac{t_{k-1}^{(0)} - t_{k-1}^{(1,j)}}{(a_k^{(0)})^2 - (a_k^{(1,j)})^2} \\ &= -\frac{1}{t_{k-1}^{(1,j)}} + \frac{(a_k^{(1,j)})^2}{t_{k-1}^{(0)} t_{k-1}^{(1,j)}} \cdot \Delta t_{k-1}^{(j)} \cdot \frac{(a_{k-1}^{(0)})^2 - (a_{k-1}^{(1,j)})^2}{(a_k^{(0)})^2 - (a_k^{(1,j)})^2}.\end{aligned}$$

Außerdem gilt $\Delta t_1^{(j)} = 0$. Da für $j \to \infty$ die Größen $a_k^{(1,j)}$ gegen $a^{(0)}$ gehen, können wir sukzessive für alle k zeigen, dass $\lim\limits_{j\to\infty} t_k^{(1,j)} = t_k^{(0)}$ und wegen (4.1.16) $\lim\limits_{j\to\infty} \Delta t_k^{(j)} = t'_k$ gilt (siehe (4.1.18)). Also sind (4.1.17) Grenzwerte für (4.1.19). □

Bemerkung 4.1.6 Bedingung (4.1.16) bedeutet einfach, dass alle Blöcke des Testvektors $\mathbf{e}^{(1,j)}$ „gleich schnell“ gegen die entsprechenden $\mathbf{e}^{(0)}$ konvergieren. Wenn für jedes $k \in \{1, \ldots, N\}$ die Gleichungen $e_k^{(0)} = e$, $e_k^{(1,j)} = e^j$ gelten, ist (4.1.16) immer erfüllt. □

Die Formeln (4.1.17) können zur Berechnung der Koeffizienten mit nur einem Testvektor $\mathbf{e}^{(0)}$ benutzt werden. Nach der Rückskalierung, wenn also Voraussetzung (4.1.15) nicht mehr gilt, erhalten wir

$$t'_k = \begin{cases} 0, & k = 1, \\ -\dfrac{d_{k-1}^{(0)}}{t_{k-1}^{(0)}} + \dfrac{d_{k-1}^{(0)} (a_k^{(0)})^2}{d_k^{(0)} (t_{k-1}^{(0)})^2} \cdot t'_{k-1}, & k \geq 2. \end{cases} \tag{4.1.22}$$

Das führt zu folgendem Algorithmus:

Algorithmus 4.1.7 *(Berechnung der GIBLU(1)-Koeffizienten mit einem Testvektor)* Es seien $\mathbf{A}$ eine symmetrische Matrix der Form (4.1.1) und $\mathbf{e}^{(0)} = \text{blockvector}\{e_k^{(0)}\}$ ein Testvektor. Mit den nach (4.1.2) definierten Größen $d_k^{(0)}$ und $a_k^{(0)}$ lassen sich die Koeffizienten $\theta_k^{(0)}$ und $\theta_k^{(1)}$ der GIBLU(1)-Zerlegung in folgenden Schritten berechnen:

BEGIN

1: $t_2^{(0)} \leftarrow d_2^{(0)} - \dfrac{(a_2^{(0)})^2}{d_1^{(0)}}; \qquad t_2' \leftarrow -1;$

2: **for all** k **from** 3 **to** N **do**

3: $\quad t_k^{(0)} \leftarrow d_k^{(0)} - \dfrac{(a_k^{(0)})^2}{t_{k-1}^{(0)}}; \qquad t_k' \leftarrow -\dfrac{d_{k-1}^{(0)}}{t_{k-1}^{(0)}} + \dfrac{d_{k-1}^{(0)}(a_k^{(0)})^2}{d_k^{(0)}(t_{k-1}^{(0)})^2} \cdot t_{k-1}';$

4: $\quad \theta_k^{(1)} \leftarrow \dfrac{t_k^{(0)}}{d_k^{(0)}} - \dfrac{(a_k^{(0)})^2}{d_{k-1}^{(0)} d_k^{(0)}} \cdot t_k'; \qquad \theta_{k-1}^{(0)} \leftarrow -\dfrac{1}{t_k'}$

5: **end for**

END □

Bemerkung 4.1.8 Wir vergleichen nun Algorithmus 4.1.7 mit dem in Abschnitt 3.2 beschriebenen Fall eines einzigen Parameters $\hat{\mu} := \hat{\mu}_0 = \hat{\mu}_1$. Dazu betrachten wir wieder das Modellproblem (3.1.2–3.1.4). Wie schon in Bemerkung 4.1.2 gezeigt wurde, gilt für die Matrix $\mathbf{A}$ bei der Wahl $\hat{\mu} = \left(\dfrac{a^{(0)}}{d^{(0)}}\right)^2$ die Gleichheit $\dfrac{t_k^{(0)}}{d^{(0)}} = \tau_k(\hat{\mu})$. Einsetzen dieser in (4.1.22) liefert:

$$t_k' = \begin{cases} 0, & k = 1, \\ -\dfrac{1}{\tau_{k-1}(\hat{\mu})} + \left(\dfrac{\hat{\mu}}{\tau_{k-1}(\hat{\mu})}\right)^2 t_{k-1}', & k \geq 2. \end{cases}$$

Durch Induktion erhalten wir daraus $t_k' = \tau_k'(\hat{\mu})$ (siehe (3.2.2)). Folglich sind in diesem Spezialfall die Formeln in Zeile 4 des Algorithmus 4.1.7 und (3.2.6) gleich. □

Wir betrachten nun den Fall $l = 2$. Dabei sind drei Testvektoren $\mathbf{e}^{(0)}$, $\mathbf{e}^{(1)}$ und $\mathbf{e}^{(2)}$ zu wählen. Wir beschränken uns auf die symmetrische Matrix $\mathbf{A}$ und fordern die folgende Skalierung von $\mathbf{e}^{(i)} = \text{blockvector}\{e_k^{(i)}\}$:

$$(D_i e_k^{(i)}, e_k^{(i)}) = 1, \quad 0 \leq i \leq 2, \quad 1 \leq k \leq N. \tag{4.1.23}$$

Da jetzt $a_k^{(i)} = b_{k-1}^{(i)}$ ist, schreibt sich (4.1.6) wie folgt:

$$\theta_k^{(2)} - \cfrac{(a_k^{(i)})^2}{\theta_{k-1}^{(1)} - \cfrac{(a_{k-1}^{(i)})^2}{\theta_{k-2}^{(0)}}} = t_k^{(i)}, \quad i \in \{0, 1, 2\}, \tag{4.1.24}$$

Dies ist kein lineares System. Wegen der Schwierigkeiten bei der Untersuchung der Eigenschaften der Lösungen von (4.1.24) beschränken wir uns hier bei unseren Betrachtungen auf die Methode, sie analytisch zu finden. Wir weisen auch darauf hin,

dass nach Bemerkung 4.1.2 die aus (4.1.24) berechneten Koeffizienten $\theta_k^{(2)}$, $\theta_{k-1}^{(1)}$ und $\theta_{k-2}^{(0)}$ im Fall des Modellproblems (3.1.2–3.1.4) mit den Formeln (3.2.23–3.2.26) übereinstimmen.

Nach elementaren Umformungen lässt sich System (4.1.24) in der Form

$$\theta_k^{(2)}\theta_{k-1}^{(1)}\theta_{k-2}^{(0)} - \theta_k^{(2)}(a_{k-1}^{(i)})^2 - \theta_{k-2}^{(0)}(a_k^{(i)})^2 - \theta_{k-1}^{(1)}\theta_{k-2}^{(0)}t_k^{(i)} = -(a_{k-1}^{(i)})^2 t_k^{(i)}, \quad 0 \le i \le 2, \tag{4.1.25}$$

schreiben. Wir führen die Variablen

$$x_k = \theta_{k-2}^{(0)}, \quad y_k = -\theta_{k-1}^{(1)}\theta_{k-2}^{(0)}, \quad z_k = -\theta_k^{(2)}\theta_{k-1}^{(1)}\theta_{k-2}^{(0)} \tag{4.1.26}$$

ein und transformieren (4.1.25) in ein System mit insgesamt vier Unbekannten x_k, y_k, z_k und $\theta_k^{(2)}$, drei linearen Gleichungen

$$\begin{aligned}
(a_k^{(0)})^2 x_k - t_k^{(0)} y_k + z_k + (a_{k-1}^{(0)})^2\theta_k^{(2)} &= (a_{k-1}^{(0)})^2 t_k^{(0)},\\
(a_k^{(1)})^2 x_k - t_k^{(1)} y_k + z_k + (a_{k-1}^{(1)})^2\theta_k^{(2)} &= (a_{k-1}^{(1)})^2 t_k^{(1)},\\
(a_k^{(2)})^2 x_k - t_k^{(2)} y_k + z_k + (a_{k-1}^{(2)})^2\theta_k^{(2)} &= (a_{k-1}^{(2)})^2 t_k^{(2)}
\end{aligned} \tag{4.1.27}$$

und einer zusätzlichen nichtlinearen Bedingung

$$\theta_k^{(2)} y_k = z_k. \tag{4.1.28}$$

Wir lösen nun (4.1.27) nur bzgl. der Variablen x_k, y_k und z_k. Dann erhalten wir y_k und z_k als affin-lineare Funktionen von $\theta_k^{(2)}$. Einsetzen von diesen in (4.1.28) liefert eine algebraische Gleichung in $\theta_k^{(2)}$ höchstens zweiten Grades. Diese kann also leicht analytisch gelöst werden. Mit dem daraus ermittelten $\theta_k^{(2)}$ erhalten wir wiederum aus (4.1.27) die Werte x_k und y_k, die durch die zu (4.1.26) inverse Transformation zu den zwei restlichen Koeffizienten $\theta_{k-1}^{(1)}$ und $\theta_{k-2}^{(0)}$ führen.

Die Determinante des Gleichungssystems (4.1.27) für die Unbekannten x_k, y_k und z_k ist

$$\begin{aligned}
\Delta_k &= \begin{vmatrix} (a_k^{(0)})^2 & -t_k^{(0)} & 1\\ (a_k^{(1)})^2 & -t_k^{(1)} & 1\\ (a_k^{(2)})^2 & -t_k^{(2)} & 1 \end{vmatrix}\\
&= \left((a_k^{(1)})^2 - (a_k^{(0)})^2\right)\left((a_k^{(2)})^2 - (a_k^{(1)})^2\right)\\
&\qquad \times \left[\frac{t_k^{(1)} - t_k^{(0)}}{(a_k^{(1)})^2 - (a_k^{(0)})^2} - \frac{t_k^{(2)} - t_k^{(1)}}{(a_k^{(2)})^2 - (a_k^{(1)})^2}\right].
\end{aligned} \tag{4.1.29}$$

Wir nehmen des Weiteren an, dass Δ_k für alle $k \ge 3$ ungleich Null ist. Dann gilt

$$y_k = \frac{\Delta_{y,k} - \hat{\Delta}_{y,k}\theta_k^{(2)}}{\Delta_k}, \quad z_k = \frac{\Delta_{z,k} - \hat{\Delta}_{z,k}\theta_k^{(2)}}{\Delta_k}, \tag{4.1.30}$$

wobei

$$\Delta_{y,k} = \begin{vmatrix} (a_k^{(0)})^2 & (a_{k-1}^{(0)})^2 t_k^{(0)} & 1\\ (a_k^{(1)})^2 & (a_{k-1}^{(1)})^2 t_k^{(1)} & 1\\ (a_k^{(2)})^2 & (a_{k-1}^{(2)})^2 t_k^{(2)} & 1 \end{vmatrix}, \qquad \hat{\Delta}_{y,k} = \begin{vmatrix} (a_k^{(0)})^2 & (a_{k-1}^{(0)})^2 & 1\\ (a_k^{(1)})^2 & (a_{k-1}^{(1)})^2 & 1\\ (a_k^{(2)})^2 & (a_{k-1}^{(2)})^2 & 1 \end{vmatrix} \tag{4.1.31}$$

und

$$\Delta_{z,k} = \begin{vmatrix} (a_k^{(0)})^2 & -t_k^{(0)} & (a_{k-1}^{(0)})^2 t_k^{(0)} \\ (a_k^{(1)})^2 & -t_k^{(1)} & (a_{k-1}^{(1)})^2 t_k^{(1)} \\ (a_k^{(2)})^2 & -t_k^{(2)} & (a_{k-1}^{(2)})^2 t_k^{(2)} \end{vmatrix}, \qquad \hat{\Delta}_{z,k} = \begin{vmatrix} (a_k^{(0)})^2 & -t_k^{(0)} & (a_{k-1}^{(0)})^2 \\ (a_k^{(1)})^2 & -t_k^{(1)} & (a_{k-1}^{(1)})^2 \\ (a_k^{(2)})^2 & -t_k^{(2)} & (a_{k-1}^{(2)})^2 \end{vmatrix}. \tag{4.1.32}$$

Nach Multiplikation mit Δ_k schreibt sich Gleichung (4.1.28) also in der Form

$$-\hat{\Delta}_{y,k}(\theta_k^{(2)})^2 + (\Delta_{y,k} - \hat{\Delta}_{z,k})\theta_k^{(2)} - \Delta_{z,k} = 0. \tag{4.1.33}$$

An dieser Stelle weisen wir darauf hin, dass wir im Fall des Modellproblems (3.1.2–3.1.4) wegen $a_k^{(i)} = a_{k-1}^{(i)}$ die Gleichungen

$$\hat{\Delta}_{y,k} = \hat{\Delta}_{z,k} = 0$$

erhalten. Damit reduziert sich (4.1.33) auf die lineare Gleichung $\Phi(\theta_k^{(2)}) := \Delta_{y,k}\theta_k^{(2)} - \Delta_{z,k} = 0$, die eindeutig nach $\theta_k^{(2)}$ lösbar ist. Die linke Seite $\tilde{\Phi}(\theta) := -\hat{\Delta}_{y,k}\theta^2 + (\Delta_{y,k} - \hat{\Delta}_{z,k})\theta - \Delta_{z,k}$ der Gleichung (4.1.33) können wir für kleine Werte von $\hat{\Delta}_{y,k}$ und $\hat{\Delta}_{z,k}$ als eine Abweichung von Φ betrachten. Es ist also zu erwarten, dass neben der Wurzel $\theta_k^{(2)}$, die wir akzeptieren sollen, die Funktion $\tilde{\Phi}$ das gleiche Monotonieverhalten hat wie Φ. Das kann als ein Auswahlkriterium für $\theta_k^{(2)}$ dienen: Wenn $\hat{\Delta}_{y,k}\Delta_{y,k} \geq 0$ gilt, soll $\theta_k^{(2)}$ die kleinste Lösung von (4.1.33) sein. Ist umgekehrt $\hat{\Delta}_{y,k}\Delta_{y,k} < 0$, dann soll die größte Wurzel von (4.1.33) als $\theta_k^{(2)}$ genommen werden.

Nach der Berechnung von $\theta_k^{(2)}$, erhalten wir y_k aus (4.1.30). Außerdem gilt

$$\theta_{k-2}^{(0)} = x_k = \frac{\Delta_{x,k} - \hat{\Delta}_{x,k}\theta_k^{(2)}}{\Delta_k}, \tag{4.1.34}$$

wobei

$$\Delta_{x,k} = \begin{vmatrix} (a_{k-1}^{(0)})^2 t_k^{(0)} & -t_k^{(0)} & 1 \\ (a_{k-1}^{(1)})^2 t_k^{(1)} & -t_k^{(1)} & 1 \\ (a_{k-1}^{(2)})^2 t_k^{(2)} & -t_k^{(2)} & 1 \end{vmatrix}, \qquad \hat{\Delta}_{x,k} = \begin{vmatrix} (a_{k-1}^{(0)})^2 & -t_k^{(0)} & 1 \\ (a_{k-1}^{(1)})^2 & -t_k^{(1)} & 1 \\ (a_{k-1}^{(2)})^2 & -t_k^{(2)} & 1 \end{vmatrix}. \tag{4.1.35}$$

Der verbleibende Koeffizient $\theta_{k-1}^{(1)}$ lässt sich aus $\theta_{k-1}^{(0)}$ und y_k berechnen: $\theta_{k-1}^{(1)} = -\dfrac{y_k}{\theta_{k-2}^{(0)}}$. (Siehe (4.1.26).) Wir fassen dies im folgenden Algorithmus zusammen:

Algorithmus 4.1.9 *(Berechnung der GIBLU(2)-Koeffizienten)* Die Matrix $\mathbf{A}$ sei positiv definit und habe die Form (4.1.1). Die Testvektoren $\mathbf{e}^{(0)}$, $\mathbf{e}^{(1)}$ und $\mathbf{e}^{(2)}$ genügen den Bedingungen (4.1.23). Die Größen $a_k^{(i)}$, $0 \leq i \leq 2$, seien durch (4.1.2) definiert. Der Algorithmus berechnet die Koeffizienten $\theta_k^{(0)}$, $\theta_k^{(1)}$ und $\theta_k^{(2)}$ der GIBLU(2)-Zerlegung. Es wird angenommen, dass für jedes $k \geq 4$ die Determinante Δ_k (siehe (4.1.29)) von Null verschieden ist, und die Gleichung (4.1.33) reelle Lösungen hat.

BEGIN

Berechne $t_3^{(0)}$, $t_3^{(1)}$ und $t_3^{(2)}$ nach (4.1.3);
for all k **from** 4 **to** N **do**
  $t_k^{(i)} \leftarrow 1 - \dfrac{(a_k^{(i)})^2}{t_{k-1}^{(i)}}, \qquad i \in \{0,1,2\};$
  Berechne Δ_k, $\Delta_{x,k}$, $\hat{\Delta}_{x,k}$, $\Delta_{y,k}$, $\hat{\Delta}_{y,k}$ und $\Delta_{z,k}$ nach (4.1.29), (4.1.35) und (4.1.31–4.1.32);
  if $\hat{\Delta}_{y,k} = 0$ **then**
    $\theta_k^{(2)} \leftarrow \dfrac{\Delta_{z,k}}{\Delta_{y,k}}; \qquad y_k \leftarrow \dfrac{\Delta_{y,k}}{\Delta_k}$
  else
    Berechne $\hat{\Delta}_{z,k}$ nach (4.1.32);
    Löse die quadratische Gleichung (4.1.33);
    if $\hat{\Delta}_{y,k}\Delta_{y,k} \geq 0$ **then**
      Setze $\theta_k^{(2)}$ gleich der kleinsten Lösung von (4.1.33)
    else
      Setze $\theta_k^{(2)}$ gleich der größten Lösung von (4.1.33)
    end if
    Berechne y_k nach (4.1.30)
  end if
  Berechne $\theta_{k-2}^{(0)}$ nach (4.1.34);
  $\theta_{k-1}^{(1)} \leftarrow -\dfrac{y_k}{\theta_{k-2}^{(0)}}$
end for

END □

Es können auch andere Vorgehensweisen zur Berechnung der Parameter betrachtet werden. Wir beschreiben hier noch eine für $l = 1$. Statt der Reduktion auf tridiagonale Matrizen stellen wir an die Restmatrix die schwache Filterbedingung (2.1.29):

$$(\mathbf{R}^{(1)}\mathbf{e}^{(i)}, \mathbf{e}^{(i)}) = 0, \quad i \in \{0,1\}. \tag{4.1.36}$$

Wegen der blockdiagonalen Struktur der Matrix $\mathbf{R}^{(1)}$ (siehe Satz 2.1.5) ist (4.1.36) äquivalent zu

$$(R_k^{(1)} e_k^{(i)}, e_k^{(i)}) = 0, \quad i \in \{0,1\}, \quad 3 \leq k \leq N, \tag{4.1.37}$$

wobei

$$R_k^{(1)} = \tilde{T}_k^{(1)} - \left(D_k - L_k \left(\tilde{T}_{k-1}^{(1)} \right)^{-1} U_{k-1} \right). \tag{4.1.38}$$

Einsetzen von (4.1.38) in (4.1.37) führt zur Gleichung

$$(\tilde{T}_k^{(1)} e_k^{(i)}, e_k^{(i)}) = (D_k e_k^{(i)}, e_k^{(i)}) - \left(L_k \left(\tilde{T}_{k-1}^{(1)} \right)^{-1} U_{k-1} e_k^{(i)}, e_k^{(i)} \right).$$

Wegen (3.1.34) erhalten wir daraus für jedes $k \in \{3, \ldots, N\}$ ein lineares System für

die Größen $\theta_k^{(1)}$ und $\left(\theta_{k-1}^{(0)}\right)^{-1}$:

$$\theta_k^{(1)} - \frac{(L_k D_{k-1}^{-1} U_{k-1} e_k^{(i)}, e_k^{(i)})}{(D_k e_k^{(i)}, e_k^{(i)})} \frac{1}{\theta_{k-1}^{(0)}} = 1 - \frac{(L_k (\tilde{T}_{k-1}^{(1)})^{-1} U_{k-1} e_k^{(i)}, e_k^{(i)})}{(D_k e_k^{(i)}, e_k^{(i)})}. \tag{4.1.39}$$

Die Berechnung der Koeffizienten kann rekursiv über k erfolgen. Dabei sollte das Produkt der Inversen der schon bekannten Matrix $\tilde{T}_{k-1}^{(1)}$ mit dem Vektor $U_{k-1} e_k^{(i)}$ auf der rechten Seite auf die gleiche Weise wie in Abschnitt 3.1 berechnet werden. Damit kommen wir zum folgenden Algorithmus:

Algorithmus 4.1.10 *(Berechnung der Koeffizienten der GIBLU(1)-Zerlegung nach (4.1.36))* Gegeben seien eine Matrix $\mathbf{A}$ der Form (4.1.1) und zwei Testvektoren $\mathbf{e}^{(0)}$ und $\mathbf{e}^{(1)}$. Der Algorithmus habe bereits die Koeffizienten $\theta_k^{(1)}$ und $\theta_k^{(0)}$ einer GIBLU(1)-Zerlegung berechnet, die Bedingung (4.1.36) genügt:
BEGIN

1: **for all** k **from** 3 **to** N **do**

2: $t_k^{(i)} \leftarrow 1 - \dfrac{(L_k (\tilde{T}_{k-1}^{(1)})^{-1} U_{k-1} e_k^{(i)}, e_k^{(i)})}{(D_k e_k^{(i)}, e_k^{(i)})}, \quad i \in \{0,1\};$

3: $\mu_k^{(i)} \leftarrow \dfrac{(L_k D_{k-1}^{-1} U_{k-1} e_k^{(i)}, e_k^{(i)})}{(D_k e_k^{(i)}, e_k^{(i)})}, \quad i \in \{0,1\};$

4: $\Delta_k \leftarrow \mu_k^{(0)} - \mu_k^{(1)};$

5: $\theta_k^{(1)} \leftarrow \dfrac{\mu_k^{(0)} t_k^{(1)} - \mu_k^{(1)} t_k^{(0)}}{\Delta_k}; \quad \theta_{k-1}^{(0)} \leftarrow \dfrac{\Delta_k}{t_k^{(1)} - t_k^{(0)}}$

6: **end for**

END □

Diese Vorgehensweise zeigt deutliche Nachteile gegenüber der am Anfang dieses Kapitels betrachteten: Algorithmus 4.1.10 ist wegen der Invertierung der Blöcke viel aufwendiger als Algorithmus 4.1.4. Auch die Untersuchung der berechneten Koeffizienten ist hier schwieriger, wenn überhaupt möglich. Dazu lässt sich diese Methode nicht unmittelbar auf die Fälle $l \geq 2$ verallgemeinern: Die kompliziertere Struktur der Diagonalblöcke führt nicht mehr zu skalaren Gleichungen.

Bemerkung 4.1.11 Wir betrachten wieder Modellproblem (3.1.2–3.1.4) und wählen zwei Eigenvektoren $\hat{e}^{(0)}$ und $\hat{e}^{(1)}$ der Matrix BB^T zu verschiedenen Eigenwerten $\hat{\mu}_0$ und $\hat{\mu}_1$, respektive. Als Testvektoren nehmen wir $\mathbf{e}^{(i)} = \text{blockvector}\{D^{-\frac{1}{2}} \hat{e}^{(i)}\}$. Dann erhalten wir:

$$\frac{(L_k D_{k-1}^{-1} U_{k-1} e_k^{(i)}, e_k^{(i)})}{(D_k e_k^{(i)}, e_k^{(i)})} = \frac{(BB^T \hat{e}^{(i)}, \hat{e}^{(i)})}{(\hat{e}^{(i)}, \hat{e}^{(i)})} = \hat{\mu}_i$$

und analog

$$\frac{(L_k (\tilde{T}_{k-1}^{(1)})^{-1} U_{k-1} e_k^{(i)}, e_k^{(i)})}{(D_k e_k^{(i)}, e_k^{(i)})} = \frac{\hat{\mu}_i}{\tilde{\tau}_{k-1}^{(1)}(\hat{\mu}_i)}.$$

Dies zeigt, dass (4.1.39) zu System (3.1.19) äquivalent ist. Algorithmus 4.1.10 liefert also in diesem Spezialfall die Koeffizienten (3.2.5). □

In diesem Abschnitt haben wir die Methoden zur Berechnung der Koeffizienten der GIBLU(l)-Zerlegungen für $l = 1$ und $l = 2$ beschrieben. Die Wahl einer geeigneten Methode ist aber noch zu treffen. Die Zerlegungen mit diesen Parametern auf unstrukturierten Gittern zeigen viel schlechteren Konvergenzeigenschaften als für das Modellproblem (siehe Abschnitt 4.3 unten). Dies könnte an der ungünstigen Wahl der Parameter liegen oder auch generell an Unzulänglichkeiten der Zerlegungen in diesem Fall. Diese Probleme brauchen noch weitere Untersuchung.

Bemerkung 4.1.12 1. Wie wir in Abschnitt 3.3.1 gesehen haben, ist die Restmatrix $\mathbf{R}^{(1)}$ der GIBLU(1)-Zerlegung mit einem Parameter $\hat{\mu} = \hat{\mu}_0 = \hat{\mu}_1$ im Fall unseres Modellproblems positiv semidefinit. Im allgemeinen Fall, wenn $\mathbf{A}$ die Form (4.1.1) hat, ist dies nicht mehr zutreffend. Und es scheint nicht möglich zu sein, diese Eigenschaft für den allgemeinen Fall zu erhalten. Doch selbst wenn die Konvergenz der linearen Iteration nicht garantiert ist, kann man die Zerlegung als Vorkonditionierer in CG-Verfahren benutzen.

2. Es ist auch im allgemeinen unmöglich, die GIBLU(l)-Zerlegung mit skalaren Koeffizienten $\theta_k^{(i)}$ so zu konstruieren, dass die Restmatrix auf $l+1$ vorgegebenen Testvektoren Null ist. Die Bedeutung eines „Testvektors“ unterscheidet sich hier also z.B. von der im Fall von TFF-Zerlegungen. Sie ist hier mit der für die tangentialen Zerlegungen verwandt: Die Testvektoren geben nur die Eigenmoden an, die gedämpft werden sollen.

3. Die Existenz der Koeffizienten bedeutet leider nicht, dass die Diagonalblöcke der Zerlegung wohldefiniert wären. Da Existenzaussagen der Diagonalblöcke im allgemeinen Fall aber sehr kompliziert zu beweisen sind, betrachten wir sie in dieser Arbeit nicht. □

4.2 Zerlegung des Gitters

Die Vorgehensweise zur Zerlegung des Gitters in Blöcke haben wir schon in Abschnitt 2.3 besprochen. Bei dieser Prozedur wird das Gitter Ω_h so als Vereinigung der Knotenmengen $\Omega_h^{(1)}, \ldots, \Omega_h^{(N)}$ dargestellt, dass

$$\Omega_h^{(1)} \cup \cdots \cup \Omega_h^{(N)} = \Omega_h \quad \text{und} \quad \Omega_h^{(1)} \cap \cdots \cap \Omega_h^{(N)} = \emptyset \tag{4.2.1}$$

und bei der Anordnung aller Gitterknoten, die mit der Anordnung der Blöcke $\Omega_h^{(i)}$ konsistent ist, die Matrix $\mathbf{A}$ die blocktridiagonale Struktur (4.1.1) hat. Für die in dieser Arbeit eingeführten GIBLU(l)-Zerlegungen gelten im Prinzip die gleichen Forderungen. Es gibt allerdings noch die folgenden Besonderheiten:

1. Die Blöcke $\tilde{T}_k^{(l)}$ müssen jetzt nicht unbedingt tridiagonal sein. Für die filternden und TFF-Zerlegungen könnte dies die Struktur der weiteren Diagonalblöcke schädigen. Im Fall der GIBLU(l)-Zerlegungen erfordert die Behandlung von $\tilde{T}_k^{(l)}$ fast den gleichen Aufwand bei jedem schwachbesetzten Muster der Matrixblöcke von $\mathbf{A}$.

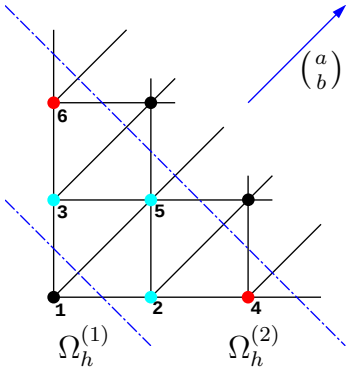

Abbildung 4.1: Gitterzerlegung mit Geraden (4.2.2). $\Omega_h^{(1)}$ besteht aus einem Knoten 1. Die blauen Kreise bezeichnen die Knoten aus $\tilde{\Omega}_h^{(2)}$. Zusammen mit den roten Kreisen, die mit $\Omega_h^{(1)}$ nicht verbunden sind, bilden sie $\Omega_h^{(2)}$.

2. Zur Zerlegungsprozedur muss auch eine Methode existieren, die die Knotenmengen $\Omega_h^{(\max\{k-l,0\})} \cup \cdots \cup \Omega_h^{(k)}$ für die Lösung der Systeme $\tilde{T}_k^{(l)} x_k = \tilde{f}_k$ ordnet. (Siehe Abschnitt 3.1.) Ohne die zulässige Anordnung kann die Invertierung von $\tilde{T}_k^{(l)}$ sehr ineffizient sein.

In diesem Abschnitt beschreiben wir eine einfache Prozedur, die basierend auf geometrischen Überlegungen eine solche Zerlegung für jedes unstrukturierte Gitter liefert. Sie spaltet die Knotenmenge mit einem vorgewählten System $\{\phi_s\}_s$ von zueinander „parallelen“ Kurven ϕ_s. Die Blöcke $\Omega_h^{(i)}$ sind dann die Teilmengen, die zwischen den Kurven liegen.

Weiter unten betrachten wir nur ebene Gitter Ω_h. Jeder Gitterknoten α hat Koordinaten (x, y). Den Graph der Matrix $\mathbf{A} = (a_{\alpha,\beta})_{\alpha,\beta\in\Omega_h}$ bezeichnen wir mit $G(\mathbf{A})$:

$$G(\mathbf{A}) = \{(\alpha, \beta) \in \Omega_h \times \Omega_h : a_{\alpha,\beta} \neq 0 \text{ oder } a_{\beta,\alpha} \neq 0\}\,.$$

Wir demonstrieren nun die Idee der Gitterzerlegung zunächst am Beispiel der parallelen Geradenschar

$$\phi_s := \{(x, y) : ax + by = s\} \tag{4.2.2}$$

mit vorgegebenen $a, b \in \mathbb{R}$. Wenn s wächst, werden die Geraden (4.2.2) parallel in die Richtung $(a, b)^T$ geschoben. Wir wählen einen Anfangswert $s_1 \in \mathbb{R}$ so, dass

$$\Omega_h^{(1)} := \{(x, y) \in \Omega_h : ax + by \leq s_1\}$$

nicht-leer ist. (Außerdem soll $\Omega_h^{(1)}$ aber auch „nicht zu groß“ sein, da es einer der Blöcke ist.) Die weiteren Schritte erfolgen rekursiv: Seien die Blöcke $\Omega_h^{(1)}, \ldots, \Omega_h^{(i)}$ schon konstruiert und es gelte für alle Knoten in diesen Mengen $ax + by \leq s_i$. Im nächsten Schritt betrachten wir die Menge der mit $\Omega_h^{(i)}$ verbundenen Knoten, die nicht in $\Omega_h^{(1)} \cup \cdots \cup \Omega_h^{(i)}$ liegen:

$$\tilde{\Omega}_h^{(i+1)} = \left\{\alpha = (x, y) \in \Omega_h : ax + by > s_i, \exists \beta \in \Omega_h^{(i)} : (\alpha, \beta) \in G(\mathbf{A})\right\}.$$

Als neuen Geradenparameter wählen wir

$$s_{i+1} = \max\{s = ax + by : (x,y) \in \tilde{\Omega}_h^{(i+1)}\}.$$

Alle Punkte aus $\tilde{\Omega}_h^{(i+1)}$ liegen also auf der selben Seite von ϕ_s, d.h. $\forall (x,y) \in \tilde{\Omega}_h^{(i+1)} \quad ax + by \leq s_{i+1}$. Zum Block $\Omega_h^{(i+1)}$ gehören nun alle Gitterknoten, die auf dieser Seite der Geraden liegen, aber noch in keinem der Blöcke $\Omega_h^{(1)}, \ldots, \Omega_h^{(i)}$ sind:

$$\Omega_h^{(i+1)} = \{(x,y) \in \Omega_h : s_i < ax + by \leq s_{i+1}\}$$

(siehe Abbildung 4.1). Diese Methode lässt sich auf nebenliegende Weise auf überschneidungsfreie Kurvenscharen, die Ω_h ganz überdecken, verallgemeinern.

Algorithmus 4.2.1 *(Geometrische Zerlegung des Gitters mit Hilfe einer Kurvenschar)* Seien Ω_h ein ebenes Gitter, $\mathbf{A} = (a_{\alpha,\beta})_{\alpha,\beta \in \Omega_h}$ eine auf diesem Gitter definierte, schwachbesetzte Matrix und $\{\phi_s\}_s$ ein System von überschneidungsfreien Kurven $\phi(x,y) = s$, die Ω_h ganz überdecken. Der Algorithmus liefert das System von Blöcken $\Omega_h^{(1)}, \ldots, \Omega_h^{(N)}$ mit der Eigenschaft (4.2.1), für das die Matrix $\mathbf{A}$ die Form (4.1.1) hat. Der Anfangsparameter s_1 muss so vorgewählt werden, dass $\Omega_h^{(1)} = \{(x,y) \in \Omega_h : \phi(x,y) \leq s_1\} \neq \emptyset$ gilt.

BEGIN
  $\hat{\Omega}_h, \Omega_h^{(1)} \leftarrow \{(x,y) \in \Omega_h : \phi(x,y) \leq s_1\}; \quad i \leftarrow 1;$
  while $\hat{\Omega}_h \neq \Omega_h$ **do**
    $i \leftarrow i + 1;$
    $s_i \leftarrow \max\{\phi(x,y) : \alpha = (x,y) \in \Omega_h, \phi(x,y) > s_{i-1}, \exists \beta \in \Omega_h^{(i-1)} : (\alpha,\beta) \in G(\mathbf{A})\};$
    $\Omega_h^{(i)} \leftarrow \{(x,y) \in \Omega_h : s_{i-1} < \phi(x,y) \leq s_i\};$
    $\hat{\Omega}_h \leftarrow \hat{\Omega}_h \cup \Omega_h^{(i)}$
  end while
END □

Nachdem die Blockstruktur mit Algorithmus 4.2.1 festgelegt wurde, können im Fall der Geraden (4.2.2) die Knoten $\alpha = (x,y)$ in jeder Vereinigung $\Omega_h^{(\max\{k-l,0\})} \cup \cdots \cup \Omega_h^{(k)}$ z.B. unter Verwendung der Größe $ax - by$, also in der zu $(a,b)^T$ perpendikulären Richtung, geordnet werden. Eine ähnliche Idee ist auch für viele allgemeinere Kurven anwendbar. Eine bessere Vorgehensweise wäre, die Bandbreite zu optimieren, obwohl die oben beschriebenen Methoden oft die gleichen Ergebnisse liefern.

Im Fall des strukturierten Gitters (1.1.9) (siehe auch (1.1.5)) und Fünf-Punkt-Stern-Diskretisierung (1.1.8) liefert Algorithmus 4.2.1 für Geraden $y = s$ die Standardblockstruktur (1.1.10). Die Anwendung dieses Algorithmus auf ein unstrukturiertes Gitter ist in Abbildung 4.2 gezeigt.

4.3 Numerische Experimente

Wir wenden nun die in diesem Kapitel beschriebenen Verfahren auf drei verschiedene Beispiele von Randwertproblemen an: ein elliptisches Problem mit variierenden Koeffizienten auf dem Einheitsquadrat, eine Poisson-Gleichung auf einem komplexeren

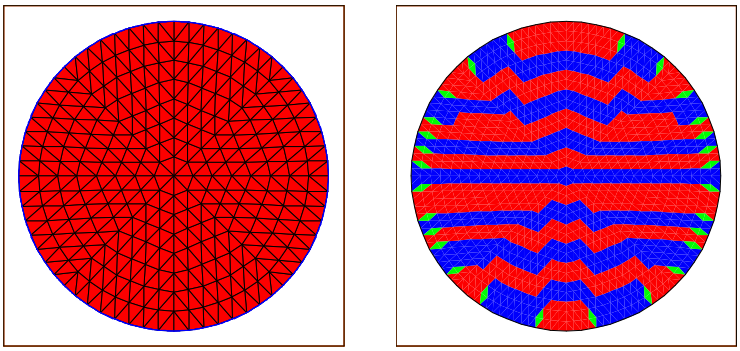

Abbildung 4.2: Anwendung von Algorithmus 4.2.1 auf ein unstrukturiertes Gitter. Links: Das Gitter. Rechts: Die Blockstruktur. Zur Gitterzerlegung wurden die Geraden $y = s$ verwendet. Die Blöcke mit ungeraden Nummern sind rot, mit geraden — blau. Der erste Block ist unten.

Gebiet mit strukturiertem Gitter und eine Anwendung auf ein unstrukturiertes Gitter. Die Koeffizienten $\theta_k^{(i)}$ in allen hier unten beschriebenen Experimenten berechnen wir mit Testvektoren $\mathbf{e}(\mathfrak{i}) = \text{blockvector}\,\{e_k(\mathfrak{i})\}$, wobei

$$\left(e_k(\mathfrak{i})\right)_j = \sin\frac{\pi j \min\{\mathfrak{i}, n_k\}}{n_k + 1}, \tag{4.3.1}$$

$j \in \{1, \ldots, n_k\}$ die Nummer des Knotens im Block ist.

Für das erste Beispiel wählen wir das Problem

$$\begin{aligned} -\nabla \cdot (P(x,y)\nabla u) &= 1, \quad (x,y) \in \Omega \\ u|_{\partial\Omega} &= 1, \end{aligned} \tag{4.3.2}$$

wobei $\Omega = (0,1)^2$ das Einheitsquadrat und

$$P(x,y) = 1 - \exp(-xy) \tag{4.3.3}$$

eine skalare Funktion ist. Dieses Problem wird mit dem Finite-Volumen-Verfahren auf dem strukturierten Dreiecksgitter diskretisiert. Wir benutzen dabei die Standardblockstruktur (1.1.10).

Tabelle 4.1 stellt die Ergebnisse der Experimente mit der GIBLU(1)-Zerlegung für Problem (4.3.2–4.3.3) dar. Die Koeffizienten werden nach Algorithmus 4.1.7 mit einem Testvektor (4.3.1) berechnet. Die Wellenzahl $\mathfrak{i}$ wird experimentell neben den in Abbildung 3.3 gezeigten Werten so bestimmt, dass die Anzahl der Schritte des CG-Verfahrens minimal ist. Wie man sieht, genügt dieses $\mathfrak{i}$ als Funktion der Blockgröße n einer einfachen Regel. Weiter unten in diesem Abschnitt bezeichnen wir mit $\overline{\rho_{\text{num}}}$ die über die Schrittzahl gemittelte Konvergenzrate (3.4.3) der CG-Iteration. Wie dieses Experiment zeigt, sind die Konvergenzeigenschaften der GIBLU(1)-Zerlegung im Fall des Modellproblems (3.4.1) und des Problems (4.3.2–4.3.3) für das strukturierte Gitter fast gleich (siehe Tabelle 3.2 und Bemerkung 3.4.1).

Die Ergebnisse des gleichen Experiments mit GIBLU(2)-Zerlegung sind in Tabelle 4.2 dargestellt. Hier wählen wir drei Testvektoren $\mathbf{e}^{(0)} = \mathbf{e}(\mathfrak{i}_0)$, $\mathbf{e}^{(1)} = \mathbf{e}(\mathfrak{i}_1)$,

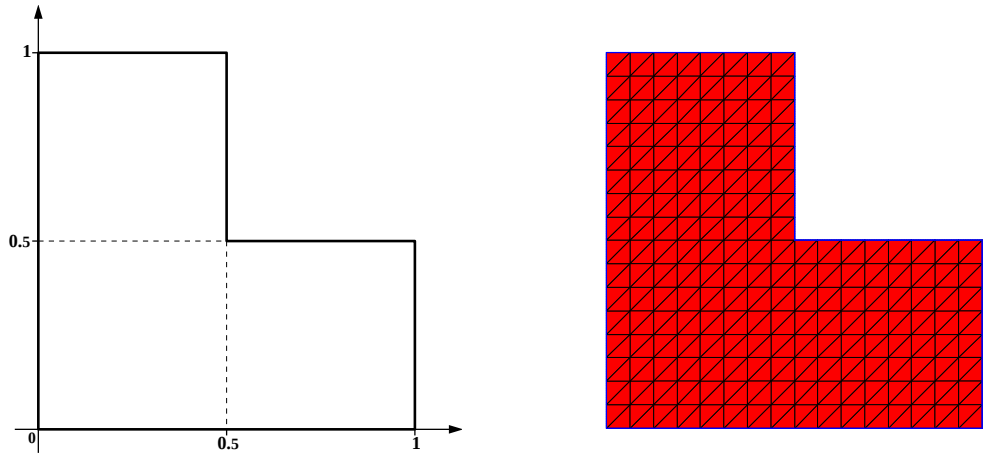

Abbildung 4.3: Das Gebiet und ein Beispiel des strukturierten Gitters für das Problem (4.3.4)

$\mathbf{e}^{(2)} = \mathbf{e}(\mathbf{i}_2)$ und berechnen die Koeffizienten nach Algorithmus 4.1.9. Diese Ergebnisse sind mit denen aus Tabelle 3.3 vergleichbar, aber etwas schlechter. Wir weisen auch darauf hin, dass die Wahl der drei Frequenzparameter $\mathbf{i}_i$ hier nicht so offensichtlich wie die des einzigen Parameters $\mathbf{i}$ bei der GIBLU(1)-Zerlegung ist.

Im zweiten Beispiel betrachten wir das Randwertproblem

$$\begin{aligned} -\Delta u &= 1, \\ u|_{\partial\Omega} &= 1 \end{aligned} \tag{4.3.4}$$

auf dem auf Abbildung 4.3 (links) dargestellten Gebiet Ω. Wir diskretisieren (4.3.4) mit dem Finite-Volumen-Verfahren auf strukturierten Gittern, dessen gröbstes rechts in Abbildung 4.3 gezeigt. Die feineren Gitter erhalten wir durch die reguläre Verfeinerung. Die Blockstruktur definieren wir wie im ersten Beispiel: Jeder Block besteht aus den auf einer horizontalen Linie liegenden Knoten. Die Knoten in jedem Block nummerieren wir von links nach rechts. Mit n bezeichnen wir hier die maximale Anzahl von inneren Knoten in einem Block. Da die Blöcke nun zwei verschiedene

Tabelle 4.1: Konvergenz des mit GIBLU(1)-Zerlegung vorkonditionierten CG-Verfahrens für Problem (4.3.2–4.3.3)

n	Wellenzahl $\mathbf{i}$	Schritte	$\overline{\rho_{\text{num}}}$
15	3	8	0.0543
31	4	11	0.1198
63	5	15	0.1999
127	6	19	0.2861
255	7	25	0.3879
511	8	33	0.4973

Größen haben, hängen die Konvergenzeigenschaften der Zerlegungen auch von der jeweiligen Anordnung der Blöcke ab. Für jedes n führen wir deswegen zwei Experimente durch: Im ersten werden die Linien von unten nach oben nummeriert und im zweiten in umgekehrter Reihenfolge. Die Ergebnisse dieser Experimente mit der GIBLU(1)-Zerlegung stellen wir in Tabelle 4.3 dar. Die Koeffizienten $\theta_k^{(i)}$ berechnen wir nach Algorithmus 4.1.7 mit einem Testvektor (4.3.1). Der Parameter $\mathfrak{i}$ wird wie im ersten Experiment gewählt. Wie man sieht, unterscheiden sich diese Ergebnisse von denen des ersten Experiments (siehe Tabelle 4.1) nicht wesentlich.

Allerdings ist die Anwendung von der GIBLU(2)-Zerlegung mit den nach Algorithmus 4.1.9 berechneten Koeffizienten für dieses Problem nicht möglich, da die quadratische Gleichung (4.1.33) am Blockgrössenübergang keine Wurzeln in $\mathbb{R}$ hat. Daher betrachten wir diese Art der Zerlegungen weiter unten nicht mehr.

Im letzten Beispiel betrachten wir das Randwertproblem (4.3.4) auf dem Gebiet Ω, das die Form der Oberfläche des Bodensees hat. Die Triangulierung von Ω ist in Abbildung 4.4 oben gezeigt. Dieses Gitter wird einmal regulär verfeinert. Es hat dann insgesamt 13445 Knoten (mit Randknoten). Die Blockstruktur erhalten wir nach Algorithmus 4.2.1 mit der Geradenschar $\phi = \{x = s\}$. Diese Struktur ist in Abbildung 4.4 unten dargestellt. Sie besteht aus 264 Blöcken, deren größter 103 innere Knoten enthält. Die Blöcke werden von links nach rechts angeordnet. Die Nummerierung in Blöcken entspricht aufsteigender Reihenfolge der Ordinaten. Wir verwenden hier die GIBLU(1)-Zerlegung, deren Koeffizienten nach Algorithmus 4.1.7 mit einem Testvektor (4.3.1) berechnet werden. Das CG-Verfahren mit diesem Vorkonditionierer für $\mathfrak{i} = 10$ reduziert die euklidische Norm des Residuums um einen Faktor von mindestens 10^{10} in 48 Schritten mit der mittleren Konvergenzrate $\overline{\rho_{\text{num}}} = 0.6113$. Für dieses Problem sind solche Vorkonditionierer also leider nicht effizient: Bei $\theta_k^{(0)} = \theta_k^{(1)} = 1$ (d.h. im Fall des Schwarz-Verfahrens, siehe Abschnitt 3.1) macht die CG-Iteration 53 Schritte mit der mittleren Konvergenzrate 0.6390.

Wir fassen nun unsere Ergebnisse zusammen. Die Experimente zeigen, dass für strukturierte Gitter die GIBLU(1)-Zerlegung mit einem Testvektor ein effizienter Vorkonditionierer für das Verfahren der konjugierten Gradienten ist. Variierende Koeffizienten des Randwertproblems oder die nicht-triviale Geometrie des Problemgebiets verschlechtern die Konvergenzeigenschaften nicht wesentlich. Die Anwendung dieser Zerlegungen auf unstrukturierte Gitter weist sich als nicht sehr effizient. Hier

Tabelle 4.2: Konvergenz des mit GIBLU(2)-Zerlegung vorkonditionierten CG-Verfahrens für Problem (4.3.2–4.3.3)

n	$\mathfrak{i}_0$, $\mathfrak{i}_1$, $\mathfrak{i}_2$	Schritte	$\overline{\rho_{\text{num}}}$
15	1, 4, 5	6	0.0180
31	1, 6, 7	8	0.0509
63	1, 8, 9	10	0.0986
127	1, 13, 14	13	0.1593
255	1, 19, 20	17	0.2372
511	1, 26, 27	22	0.3500

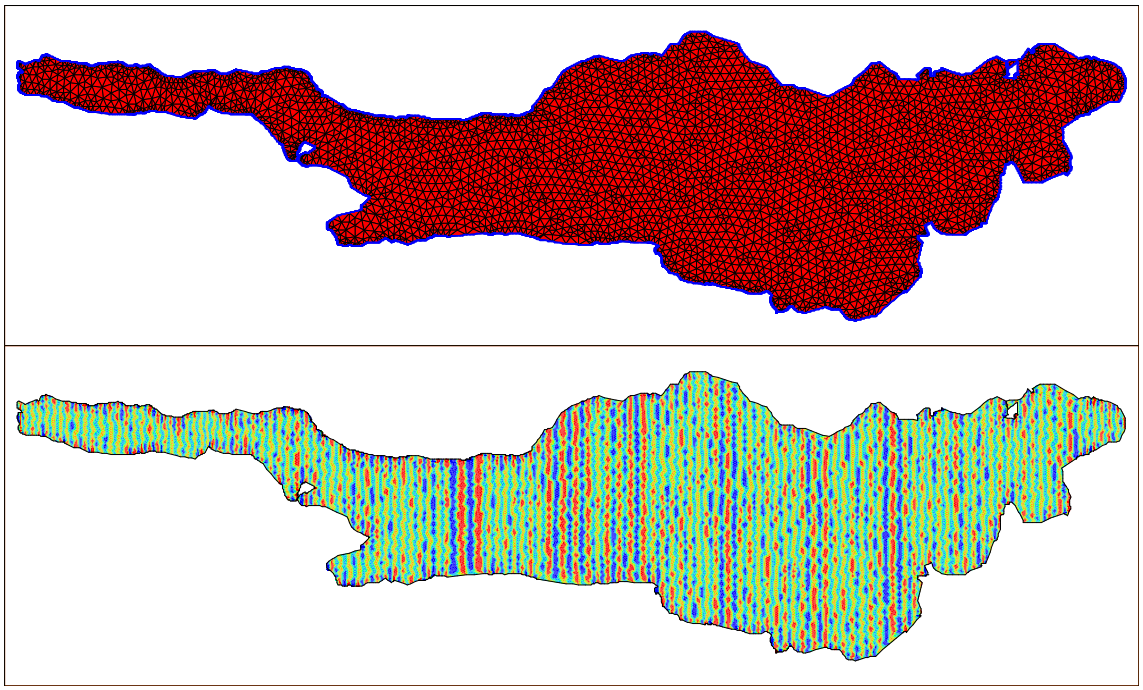

Abbildung 4.4: Das Gebiet und die Blockstruktur für das Beispiel „Bodensee“

wäre eine Verbesserung des Verfahrens nötig. Die GIBLU(2)-Zerlegungen leiden für die allgemeinen Probleme, anderseits, die Schwierigkeiten bei der Berechnung von Koeffizienten. Obwohl die Konvergenzeigenschaften dieser Zerlegungen theoretisch besser sind, wurden ihre Vorteile auf den in dieser Arbeit benutzten Gittergrößen nicht beobachtet.

Tabelle 4.3: Konvergenz des mit GIBLU(1)-Zerlegung vorkonditionierten CG-Verfahrens für Problem (4.3.4) auf dem Gebiet aus Abbildung 4.3

n	Wellenzahl $\mathfrak{i}$	Ordnung der Blöcke			
		v. u. n. o.		v. o. n. u.	
		Schritte	$\overline{\rho_{\text{num}}}$	Schritte	$\overline{\rho_{\text{num}}}$
15	3	8	0.0469	8	0.0527
31	4	12	0.1267	11	0.1220
63	5	17	0.2397	16	0.2231
127	6	23	0.3585	22	0.3381
255	7	33	0.4914	31	0.4651
511	8	45	0.5984	42	0.5766

Kapitel 5

Frequenzfilternde Vorkonditionierer für Eigenwertprobleme

In diesem Kapitel beschreiben wir die Anwendung von GIBLU(l)-Zerlegungen als Vorkonditionierer für das Gradientenverfahren bei symmetrischen Eigenwertproblemen. Dabei wird der Vorkonditionierer bei jedem Schritt des Verfahrens neu gewählt. In Abschnitt 5.1 beschreiben wir das Problem und erläutern die Grundidee der betrachteten Methode. In Abschnitt 5.2 stellen wir das Verfahren vor und beweisen dessen Konvergenz. Die Ergebnisse der numerischen Experimente mit GIBLU(l)-Zerlegungen werden in Abschnitt 5.3 vorgestellt.

5.1 Symmetrische Eigenwertprobleme

Eigenwertprobleme der Form

$$\begin{aligned} &\mathfrak{A}\mathfrak{u} = \lambda\mathfrak{B}\mathfrak{u}, \\ &\mathfrak{C}\left(\mathfrak{u}|_{\partial\Omega}\right) = 0, \end{aligned} \tag{5.1.1}$$

wobei $\mathfrak{u} : \Omega \to \mathbb{R}$ eine auf dem Gebiet $\Omega \subseteq \mathbb{R}^n$ definierte Funktion, $\mathfrak{A}$, $\mathfrak{B}$ Hermitesche Differentialoperatoren und $\mathfrak{C}$ eine Randbedingung sind, spielen eine wichtige Rolle in technischen Anwendungen. Eine typische Klasse der durch (5.1.1) beschreibbaren Phänomene sind die Eigenschwingungen von zwei- und dreidimensionalen physikalischen Objekten (siehe z.B. [2]). In vielen Fällen ist wegen der komplexen Geometrie von Ω oder der Struktur der Operatoren $\mathfrak{A}$ und $\mathfrak{B}$ eine analytische Behandlung nicht möglich, und die Lösungen $(\lambda, \mathfrak{u})$ müssen numerisch approximiert werden.

Weiter unten betrachten wir solche Probleme (5.1.1), die eine abzählbare, von unten durch 0 beschränkte Menge von Eigenwerten haben (siehe [2]). Für viele praktische Anwendungen (siehe z.B. [31], [14]) ist nur eine gewisse Anzahl von kleinsten Eigenwerten und entsprechenden Eigenfunktionen interessant. Die Aufgabe lautet dann also: Finde R kleinste Eigenwerte von (5.1.1) und die entsprechenden Eigenfunktionen, wobei R eine vorgegebene Zahl ist.

In der numerischen Behandlung von (5.1.1) werden die Operatoren $\mathfrak{A}$ und $\mathfrak{B}$ mit Finite-Differenzen-, Finite-Elemente- oder Finite-Volumen-Verfahren diskretisiert (siehe Kapitel 1 und [2]). Dabei ergibt sich ein algebraisches Eigenwertproblem

$$\mathbf{A}u = \lambda \mathbf{B}u, \tag{5.1.2}$$

mit positiv semidefiniter Matrix $\mathbf{A}$ und positiv definiter Matrix $\mathbf{B}$ der Größe $M \times M$, wobei M die Anzahl der Freiheitsgraden des diskretisierten Problems ist. Das Problem besitzt ein System $\mathbf{B}$-orthonormaler Eigenvektoren, die wir mit

$$u^{(1)}, u^{(2)}, \ldots, u^{(M)} \in \mathbb{R}^M \tag{5.1.3}$$

bezeichnen, wobei $u^{(i)}$ dem Eigenwert $\lambda^{(i)}$ entspricht und die Nummerierung in aufsteigender Reihenfolge der Eigenwerte vorgenommen ist. Für wachsende M konvergieren diese Gitterfunktionen gegen Eigenfunktionen des kontinuierlichen Problems (5.1.1), und das gleiche gilt für die Eigenwerte. Die Situation ist aber komplizierter als bei den üblichen Randwertproblemen für die partiellen Differentialgleichungen: Die Genauigkeit jeder Gitterfunktion $u^{(k)}$ hängt dabei nicht nur von M (also von der Gitterlänge h), sondern auch vom Index k der Eigenfunktion ab. Es kann gezeigt werden (siehe z.B. [2]), dass für zweidimensionale Probleme nur die $\sim\sqrt{M}$ ersten Eigenvektoren eine sinnvolle Approximation an die Eigenfunktionen von (5.1.1) bilden. Für die Berechnung von weiteren Eigenfunktionen benötigt man also noch feinere Gitter.

In diesem Abschnitt betrachten wir eine Klasse von Lösungsverfahren für die aus den Diskretisierungen entstandenen algebraischen Eigenwertprobleme (5.1.2). Eine wichtige Eigenschaft solcher Probleme ist, dass die Matrizen $\mathbf{A}$ und $\mathbf{B}$ groß und schwachbesetzt sind. Das macht Verfahren, die auf Transformationen dieser Matrizen beruhen, ineffizient. Daher spielen hier, wie auch für die linearen Gleichungssysteme, iterative Methoden die wichtigste Rolle, die nur Vektor-Skalar-, Vektor-Vektor- und Matrix-Vektor-Operationen benötigen. Einen kurzen aber guten Überblick über solche Algorithmen findet man in [7].

Wir weisen darauf hin, dass man bei algebraischen Eigenwertproblemen die Aufgabe auf die Berechnung von nur einem Eigenvektor reduzieren kann: Wenn $u^{(0)}$, $\ldots$, $u^{(r-1)}$ bereits berechnet sind, kann $u^{(r)}$ durch die Anwendung eines Verfahrens zum Problem (5.1.2) auf dem Raum $\{u : (u, u^{(i)})_{\mathbf{B}} = 0,\ 0 \le i \le r-1\}$ berechnet werden. Deswegen formulieren wir zunächst die Idee zum Gradientenverfahren für den Fall eines Eigenvektors u^* ($\|u^*\|_{\mathbf{B}} = 1$) mit dem Eigenwert λ^* (siehe auch [50]).

Ein iteratives Verfahren konstruiert eine gegen u^* konvergierende Folge

$$\tilde{u}_0, \tilde{u}_1, \ldots, \tilde{u}_k, \ldots \tag{5.1.4}$$

von $\mathbf{B}$-normierten Vektoren aus $\mathbb{R}^M$. Standardbeispiele für solche Algorithmen sind das Von-Mises-Verfahren und die inverse Iteration von Wielandt (siehe z.B. [28], [15] und [50]).

Zu dieser Klasse gehört auch das Mehrgitterverfahren von W. Hackbusch (siehe [17]). Dieser Algorithmus ist im Wesentlichen analog zum (geometrischen) Mehrgitterverfahren für lineare Gleichungssysteme (siehe [17], [18]). Auf dem gröbsten Gitter muss das Problem mit einem anderen Algorithmus behandelt werden. Da es für die Konvergenz der Mehrgitteriteration wichtig ist, dass alle zu berechnenden Eigenvektoren auf dem gröbsten Gitter gut aufgelöst werden, sollte dieses sehr fein

sein, und für die Lösung des entsprechenden Problems werden iterative Verfahren benötigt.

Da Problem (5.1.2) stark nicht-linear ist, unterscheidet sich die Konvergenzanalyse für Löser für Eigenwertprobleme prinzipiell von dieser für die linearen Löser. Die theoretische Untersuchung der Konvergenzeigenschaften in diesem Fall findet man z.B. in [28], [25], [26], [22], [17], [53].

Wir beschäftigen uns hier mit dem vorkonditionierten Gradientenverfahren. Dieses Verfahren basiert auf der Minimierung des Rayleigh-Quotienten

$$\lambda(u) = \frac{(\mathbf{A}u, u)}{(\mathbf{B}u, u)}, \quad u \in \mathbb{R}^M \setminus \{0\}.$$

Da für jeden Vektor $u \in \mathbb{R}^M \setminus \{0\}$

$$\nabla\lambda(u) = \frac{2}{\|u\|_{\mathbf{B}}^2}(\mathbf{A}u - \lambda(u)\mathbf{B}u)$$

gilt, sind die Eigenvektoren des Problems (5.1.2) die kritischen Punkte des Funktionals λ. Wir stellen also die Optimierungsaufgabe

$$\lambda(u) \to \min, \quad u \in \mathbb{R}^M \setminus \{0\}, \tag{5.1.5}$$

und wenden das Quasi-Newton-Verfahren (siehe z.B. [27]) an. In der vorliegenden Situation lässt sich dieser Algorithmus aber in einer speziellen Form schreiben.

Des Weiteren nennen wir den Vektor $r = \lambda(u)\mathbf{B}u - \mathbf{A}u$, der antiparallel zu $\nabla\lambda(u)$ ist, *Residuum von u bezüglich Problem (5.1.2)*. Im k-ten Schritt des Quasi-Newton-Verfahrens nimmt man als neuen Näherungswert $\tilde{u}_{k+1}$ die Minimalstelle von $\lambda(\cdot)$ auf der Menge $\{\tilde{u}_k + \beta c_k : \beta \in \mathbb{R}\}$, wobei $c_k = \mathbf{W}_k^{-1} r_k$ ist, r_k das Residuum von $\tilde{u}_k$ und $\mathbf{W}_k$ ein symmetrischer, positiv definiter *Vorkonditionierer*. Da die Skalierung von $\tilde{u}_{k+1}$ frei wählbar ist, können wir den neuen Näherungswert aus dem Raum

$$U_k = \{\alpha\tilde{u}_k + \beta c_k : \alpha, \beta \in \mathbb{R}\} \setminus \{0\} = \text{span}\,\{\tilde{u}_k, c_k\} \setminus \{0\} \tag{5.1.6}$$

wählen. Die einzige Ausnahme ist der Fall $\alpha = 0$, der keinem Newton-Schritt entspricht. Dies stellt aber keine Schwierigkeiten dar sondern verallgemeinert das Verfahren nur. Die Lösung der Optimierungsaufgabe (5.1.5) auf U_k ist der Ritz-Vektor des Problems (5.1.2) bzgl. des Raums span $\{\tilde{u}_k, c_k\}$ mit dem kleinsten Ritz-Wert (siehe [28]), d.h. der Vektor $\tilde{u}_{k+1} = \alpha_k\tilde{u}_k + \beta_k c_k$, wobei $\boldsymbol{\alpha}_k = (\alpha_k, \beta_k) \in \mathbb{R}^2$ der Eigenvektor des Problems

$$A_k\boldsymbol{\alpha} = \mu B_k\boldsymbol{\alpha}, \tag{5.1.7}$$

$$A_k = \begin{pmatrix} (\mathbf{A}\tilde{u}_k, \tilde{u}_k) & (\mathbf{A}\tilde{u}_k, c_k) \\ (\mathbf{A}\tilde{u}_k, c_k) & (\mathbf{A}c_k, c_k) \end{pmatrix}, \quad B_k = \begin{pmatrix} (\mathbf{B}\tilde{u}_k, \tilde{u}_k) & (\mathbf{B}\tilde{u}_k, c_k) \\ (\mathbf{B}\tilde{u}_k, c_k) & (\mathbf{B}c_k, c_k) \end{pmatrix} \tag{5.1.8}$$

zum kleinsten Eigenwert ist. Da die Matrizen dieses Problems klein sind, lassen sich die Lösungen sehr leicht analytisch berechnen. Für den Rayleigh-Quotient $\mu_k(\boldsymbol{\alpha})$ bzgl. (5.1.7–5.1.8) gilt:

$$\mu_k(\boldsymbol{\alpha}) := \frac{(A_k\boldsymbol{\alpha}, \boldsymbol{\alpha})}{(B_k\boldsymbol{\alpha}, \boldsymbol{\alpha})} = \frac{(\mathbf{A}(\alpha\tilde{u}_k + \beta c_k), \alpha\tilde{u}_k + \beta c_k)}{(\mathbf{B}(\alpha\tilde{u}_k + \beta c_k), \alpha\tilde{u}_k + \beta c_k)} = \lambda(\alpha\tilde{u}_k + \beta c_k). \tag{5.1.9}$$

Dieses Verfahren können wir letztlich in der folgenden Form darstellen:

Algorithmus 5.1.1 *(Gradientenverfahren für Eigenwertprobleme)* Wir betrachten Problem (5.1.2). Als Startvektor nehmen wir ein beliebiges $\tilde{u}_0 \in \mathbb{R}^M \setminus \{0\}$.
BEGIN

1: $k \leftarrow 0$;
2: **while** die Genauigkeit nicht erreicht **do**
3: $r_k \leftarrow \lambda(\tilde{u}_k)\mathbf{B}\tilde{u}_k - \mathbf{A}\tilde{u}_k$;
4: Finde c_k aus $\mathbf{W}_k c_k = r_k$;
5: Finde einen Eigenvektor $\boldsymbol{\alpha}_k = (\alpha_k, \beta_k)^T$ des Problems (5.1.7–5.1.8) mit dem kleinsten Eigenwert;
6: $\tilde{u}_{k+1} \leftarrow \dfrac{\hat{u}_{k+1}}{\|\hat{u}_{k+1}\|_{\mathbf{B}}}$, wobei $\hat{u}_{k+1} = \alpha_k \tilde{u}_k + \beta_k c_k$;
7: $k \leftarrow k + 1$
8: **end while**

END □

Die Eigenschaften des Gradientenverfahrens ohne Vorkonditionierer werden in [50] betrachtet. Die vorkonditionierte Variante wurde erst in [53] beschrieben. Wir betrachten hier die frequenzfilternde Vorkonditionierung, bei der die Wahl des Vorkonditionierers von der Schrittnummer (aber nicht vom Näherungswert u_k) abhängt. Als $\mathbf{W}_k$ nehmen wir die oben eingeführten GIBLU(l)-Zerlegungen in einer Reihe, so dass alle Frequenzanteile des Residuums gedämpft werden.

Bei iterativen Methoden, die nur einen (den ersten) Eigenvektor berechnen, entsteht das Problem, dass die Konvergenzrate sehr stark von $\lambda^{(2)}/\lambda^{(1)}$ abhängt. Wenn die ersten Eigenwerte sehr nahe beieinander liegen, konvergieren die Iterationsverfahren nur sehr langsam. Noch ungünstiger für die Konvergenz ist es, wenn $\lambda^{(1)}$ ein vielfacher Eigenwert ist. Denn die entsprechenden Eigenwerte $\lambda^{(1)}$ und $\lambda^{(2)}$ des diskreten Problems könnten zwar selbst numerisch verschieden sein, aber sehr nahe beieinander liegen.

Diese Schwierigkeit kann man damit umgehen, dass man die Näherungen zu den $m > 1$ ersten Eigenvektoren $u^{(1)}, \ldots, u^{(m)}$ gleichzeitig iteriert und die Rayleigh-Ritz-Prozedur (siehe [28]) für die gegenseitige Korrektur anwendet. Auf dieser Idee beruhen Blockvarianten der iterativen Algorithmen. Diese Verfahren haben zwei große Vorteile gegenüber einfachen: Selbst wenn die Näherungswerte für die letzten zu berechnenden Eigenvektor $u^{(m)}$ nur langsam konvergieren, können die Näherungen an die anderen Vektoren $u^{(1)}, \ldots, u^{(m-1)}$ sehr schnell konvergieren, und zwar schneller als bei einem Verfahren, bei dem $u^{(1)}, \ldots, u^{(m)}$ separat berechnet werden. Des Weiteren ist manchmal die Verteilung von Eigenwerten a-priori bekannt. In diesem Fall kann man die Menge der zu berechnenden Eigenvektoren so auswählen, dass der kleinste Eigenwert der restlichen Eigenvektoren, vom größtem Eigenwert der iterierenden möglichst weit entfernt ist. Dadurch wird eine gute Konvergenz für alle Näherungswerte erzielt. Obwohl die Anzahl der arithmetischen Operationen pro Iteration und Eigenvektor bei Block-Methoden meistens größer als bei gewöhnlichen Verfahren ist, sind diese Verfahren oft schneller dank ihrer Konvergenzeigenschaften. Besonders lohnt sich die Anwendung solcher Algorithmen im Fall von vielfachen

Eigenwerten.

5.2 Block-Gradientenverfahren mit variierenden Vorkonditionierern

Wir betrachten nun die Blockvariante zu Algorithmus 5.1.1, die gleichzeitig m Näherungswerte iteriert. Das m-Tupel der k-ten Näherungen bezeichnen wir mit $\mathbf{u}_k = (\tilde{u}_k^{(1)}, \ldots, \tilde{u}_k^{(m)})$. Der Vorkonditionierer hängt nicht nur vom Schrittindex k ab, sondern kann für jedes $\tilde{u}_k^{(q)}$ individuell gewählt werden. Diesen Vorkonditionierer bezeichnen wir mit $\mathbf{W}_{q,k}$. Wir nehmen an, dass alle $\mathbf{W}_{q,k}$ symmetrisch und positiv definit sind.

Dieses Verfahren konstruiert eine Folge

$$\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_k, \ldots \tag{5.2.1}$$

von m-Tupeln von Näherungswerten. Die Struktur dieses Algorithmus ist der von Algorithmus 5.1.1 ähnlich, auch wenn nun die zu lösenden Hilfsprobleme größer sind.

Algorithmus 5.2.1 *(Block-Gradientenverfahren für Eigenwertprobleme)* Es wird Problem (5.1.2) betrachtet. Die Anfangsbedingung $\mathbf{u}_0$ bestehe aus m $\mathbf{B}$-orthogonalen Vektoren $\tilde{u}_0^{(q)} \neq 0$ $(1 \leq q \leq m)$.

BEGIN

1: $k \leftarrow 0$;

2: **while** die Genauigkeit nicht erreicht **do**

3: $r_k^{(q)} \leftarrow \lambda(\tilde{u}_k^{(q)})\mathbf{B}\tilde{u}_k^{(q)} - \mathbf{A}\tilde{u}_k^{(q)}$, $1 \leq q \leq m$;

4: Berechne $c_k^{(q)}$, $1 \leq q \leq m$, aus $\mathbf{W}_{q,k}c_k^{(q)} = r_k^{(q)}$;

5: Wähle aus allen $c_k^{(q)}$, $1 \leq q \leq m$, eine solche Teilmenge $\left\{c_k'^{(i)}\right\}_{1 \leq i \leq m'}$, für die das System $\left\{\tilde{u}_k^{(1)}, \ldots, \tilde{u}_k^{(m)}, c_k'^{(1)}, \ldots, c_k'^{(m')}\right\}$ linear unabhängig ist;

6: Berechne m B_k-orthonormale Eigenvektoren $\boldsymbol{\alpha}_k^{(1)}, \ldots, \boldsymbol{\alpha}_k^{(m)}$, $\boldsymbol{\alpha}_k^{(q)} = \left(\alpha_{k,1}^{(q)}, \ldots, \alpha_{k,m+m'}^{(q)}\right)^T \in \mathbb{R}^m \times \mathbb{R}^{m'}$, *des reduzierten Eigenwertproblems*

$$A_k\boldsymbol{\alpha} = \mu B_k\boldsymbol{\alpha} \tag{5.2.2}$$

mit

$$A_k = \mathbf{H}_k^T\mathbf{A}\mathbf{H}_k, \qquad B_k = \mathbf{H}_k^T\mathbf{B}\mathbf{H}_k, \tag{5.2.3}$$

wobei $\mathbf{H}_k \in \mathbb{R}^{M \times (m+m')}$ die Matrix mit Spalten $\tilde{u}_k^{(1)}, \ldots, \tilde{u}_k^{(m)}, c_k'^{(1)}, \ldots, c_k'^{(m')}$ ist. Dabei nehmen wir an, dass die den Eigenvektoren $\boldsymbol{\alpha}_k^{(q)}$ entsprechenden Eigenwerte $\mu_k^{(1)}, \ldots, \mu_k^{(m+m')}$ in aufsteigender Reihenfolge geordnet sind;

7: $\tilde{u}_{k+1}^{(q)} \leftarrow \dfrac{\hat{u}_{k+1}^{(q)}}{\|\hat{u}_{k+1}^{(q)}\|_{\mathbf{B}}}$, $1 \leq q \leq m$, wobei $\hat{u}_{k+1}^{(q)} = \displaystyle\sum_{i=1}^{m} \alpha_{ki}^{(q)}\tilde{u}_k^{(i)} + \sum_{i=1}^{m'} \alpha_{k,m+i}^{(q)}c_k'^{(i)}$;

8: $k \leftarrow k + 1$

9: **end while**

END □

Die in Algorithmus 5.2.1 entstehenden Matrizen A_k und B_k können Größen von $m \times m$ bis $2m \times 2m$ haben. Die Anzahl m der zu berechnenden Eigenvektoren sollte hier klein genug sein, damit Hilfsproblem (5.2.2) effizient lösbar ist.

Algorithmus 5.2.1 beruht auf dem gleichen Prinzip wie Algorithmus 5.1.1: $\tilde{u}_{k+1}^{(q)}$, $1 \leq q \leq m$ sind die Ritz-Vektoren von Problem (5.1.2) bzgl. des Raums span $\left\{\tilde{u}_k^{(1)}, \ldots, \tilde{u}_k^{(m)}, c_k'^{(1)}, \ldots, c_k'^{(m')}\right\}$ (siehe [28]). Daher sind $\tilde{u}_{k+1}^{(q)}$, $1 \leq q \leq m$, die Punkte der m kleinsten lokalen Extrema von $\lambda(\cdot)$ auf diesem Raum. Für $m = 1$ sind die beiden Algorithmen gleich. Im Folgenden betrachten wir die qualitativen Konvergenzeigenschaften von Algorithmus 5.2.1. Wir konzentrieren uns hier auf den Fall von variierenden $\mathbf{W}_{q,k}$.

Bemerkung 5.2.2 1. Da $\mathbf{B}$ positiv definit ist und

$$(B_k\boldsymbol{\alpha}, \boldsymbol{\alpha}) = (\mathbf{B}u, u) \tag{5.2.4}$$

für $u = \sum_{i=1}^{m} \alpha_i \tilde{u}_k^{(i)} + \sum_{i=1}^{m'} \alpha_{m+i} c_k'^{(i)}$ gilt, ist bei linear unabhängigen $\tilde{u}_k^{(1)}, \ldots, \tilde{u}_k^{(m)}, c_k'^{(1)}, \ldots, c_k'^{(m')}$ die Matrix B_k immer positiv definit. Somit ist Problem (5.2.2) immer lösbar und besitzt $m + m'$ B_k-orthonormale Eigenvektoren. Insbesondere sind also alle Operationen in Algorithmus 5.2.1 ausführbar.

Für das reduzierte Problem (5.2.2–5.2.3) führen wir den Rayleigh-Quotient ein: $\mu_k(\boldsymbol{\alpha}) := \dfrac{(A_k\boldsymbol{\alpha}, \boldsymbol{\alpha})}{(B_k\boldsymbol{\alpha}, \boldsymbol{\alpha})}$. Für jedes $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{m+m'})^T \in \mathbb{R}^m \times \mathbb{R}^{m'}$ gilt dann

$$\mu_k(\boldsymbol{\alpha}) = \frac{(A_k\boldsymbol{\alpha}, \boldsymbol{\alpha})}{(B_k\boldsymbol{\alpha}, \boldsymbol{\alpha})} = \lambda\left(\sum_{i=1}^{m} \alpha_i \tilde{u}_k^{(i)} + \sum_{i=1}^{m'} \alpha_{m+i} c_k'^{(i)}\right) \tag{5.2.5}$$

(siehe (5.2.4)). Insbesondere erhalten wir daraus:

$$\lambda(\tilde{u}_{k+1}^{(q)}) = \mu_k(\boldsymbol{\alpha}_k^{(q)}) = \mu_k^{(q)}, \quad 1 \leq q \leq m. \tag{5.2.6}$$

2. Wir weisen darauf hin, dass durch die Beschreibung von Punkt 6 in Algorithmus 5.2.1 die Eigenvektoren $\boldsymbol{\alpha}_k^{(q)}$ nicht eindeutig festgelegt sind, sondern von dem speziellen verwendeten Lösungsverfahren für (5.2.2) abhängen können, und zwar, wenn (5.2.2) vielfache Eigenwerte hat. Algorithmus 5.2.1, die Folge (5.2.1) und deren Asymptotik hängen also von dieser Wahl ab. Da unsere weiteren Betrachtungen aber von der speziellen Wahl unabhängig sind, gehen wir auf diese nicht genauer ein.

Für $m = 1$, d.h. im Fall von Algorithmus 5.1.1, kann diese Situation nicht auftreten. Vielfachheit größer 1 eines Eigenwertes des Problems (5.1.7) würde bedeuten, dass $\lambda(\cdot)$ auf U_k (siehe (5.1.6)) konstant ist. Dies ist aber ausgeschlossen, da wegen $\mathbf{W}_k > 0$ für $r_k \neq 0$ der Vektor $c_k = \mathbf{W}_k^{-1} r_k$ eine Abstiegsrichtung dieser Funktion in Punkt $\tilde{u}_k \in U_k$ ist. Wenn also $\tilde{u}_k$ kein Eigenvektor von (5.1.2) ist, sind die Eigenwerte von (5.1.7) einfach. □

Wir benötigen das folgende Lemma:

Lemma 5.2.3 Das m-Tupel $\mathbf{u}_0$ bestehe aus $\mathbf{B}$-orthonormalen Vektoren $\tilde{u}_0^{(q)}$ mit $\lambda(\tilde{u}_0^{(q_1)}) \leq \lambda(\tilde{u}_0^{(q_2)})$ für $q_1 < q_2$. Dann gilt in jedem Schritt von Algorithmus 5.2.1:

1. Die Vektoren $\tilde{u}_k^{(q)}$, $1 \leq q \leq m$, sind wohldefiniert, ungleich Null und $\mathbf{B}$-orthonormal.

2. $\lambda(\tilde{u}_k^{(q_1)}) \leq \lambda(\tilde{u}_k^{(q_2)})$ für $q_1 < q_2$.

Beweis: Die erste Behauptung wird zusammen mit der folgenden Aussage durch Induktion über k bewiesen: Die Vektoren $\hat{u}_{k+1}^{(q)}$, $1 \leq q \leq m$, sind wohldefiniert, ungleich Null und $\mathbf{B}$-orthogonal. Da nach der Induktionsannahme die $\tilde{u}_k^{(q)}$ orthonormal sind, kann man für beliebige Vektoren $c_k^{(q)}$ ein linear unabhängiges System $\left\{\tilde{u}_k^{(1)}, \ldots, \tilde{u}_k^{(m)}, c_k'^{(1)}, \ldots, c_k'^{(m')}\right\}$ konstruieren. Deswegen (siehe auch Bemerkung 5.2.2) besitzt das Problem (5.2.2) mindestens m B_k-orthonormalen Eigenvektoren $\boldsymbol{\alpha}_k^{(1)}$, ..., $\boldsymbol{\alpha}_k^{(m)}$. Da $\boldsymbol{\alpha}_k^{(q)} \neq 0$ für jedes q gilt, ist $\hat{u}_{k+1}^{(q)}$ ungleich Null als eine nicht-triviale Linearkombination von linear unabhängigen Vektoren $\tilde{u}_k^{(1)}$, ..., $\tilde{u}_k^{(m)}$, $c_k'^{(1)}$, ..., $c_k'^{(m')}$. Die Vektoren $\hat{u}_{k+1}^{(1)}$, ..., $\hat{u}_{k+1}^{(m)}$ sind $\mathbf{B}$-orthonormal, da $(B_k\boldsymbol{\alpha}_k^{(q)}, \boldsymbol{\alpha}_k^{(q)}) = (\mathbf{B}\hat{u}_{k+1}^{(q)}, \hat{u}_{k+1}^{(q)})$ analog zu (5.2.4) gilt.

Die zweite Behauptung folgt aus (5.2.6) und der Anordnung der Eigenwerte des Problems (5.2.2) im Algorithmus. □

Bemerkung 5.2.4 Wenn im k-ten Schritt ($k \geq 1$) $m' = 0$ ist, wird die Folge der Näherungen stationär, d.h. $\mathbf{u}_k = \mathbf{u}_{k+1} = \ldots$: Die Matrizen A_k und B_k sind in diesem Fall diagonal, B_k wegen der $\mathbf{B}$-Orthogonalität von $\tilde{u}_k^{(q)}$, und A_k wegen

$$(\mathbf{A}\tilde{u}_k^{(q_1)}, \tilde{u}_k^{(q_2)}) = \frac{(\mathbf{A}\hat{u}_k^{(q_1)}, \hat{u}_k^{(q_2)})}{\|\hat{u}_k^{(q_1)}\|_{\mathbf{B}} \cdot \|\hat{u}_k^{(q_2)}\|_{\mathbf{B}}} = \frac{\mu_{k-1}^{(q_1)}(B_{k-1}\boldsymbol{\alpha}_{k-1}^{(q_1)}, \boldsymbol{\alpha}_{k-1}^{(q_2)})}{\|\hat{u}_k^{(q_1)}\|_{\mathbf{B}} \cdot \|\hat{u}_k^{(q_2)}\|_{\mathbf{B}}} = 0$$

für $q_1 \neq q_2$ (siehe (5.2.4)). Die Eigenvektoren $\boldsymbol{\alpha}_k^{(q)}$ sind also die Standardbasisvektoren, und daher $\tilde{u}_k^{(q)} = \tilde{u}_{k-1}^{(q)}$. Also bleiben im k-ten Schritt die Vektoren und ihre Ordnung gleich.

Eine Ausnahme ist der Fall, dass manche Eigenwerte $\mu_k^{(q)}$ vielfach sind. Dann hängen die $u_k^{(q)}$ noch von dem Lösungsverfahren für das Problem (5.2.2) ab. Wenn für $\boldsymbol{\alpha}_k^{(q)}$ eine nicht-kanonische Basis ausgewählt ist, gilt im Allgemeinen nicht $\mathbf{u}_k = \mathbf{u}_{k+1}$. □

Im Folgenden beweisen wir, dass die Vektoren $\tilde{u}_k^{(q)}$ aus (5.2.1) für $k \to \infty$ gegen Eigenvektoren des Problems (5.1.2) konvergieren. Wir weisen aber darauf hin, dass die $\mathbf{B}$-normierten Eigenvektoren (5.1.3) nur dann eindeutig definiert sind, wenn die entsprechenden Eigenwerte $\lambda^{(i)}$ einfach sind. Sind s Eigenwerte $\lambda^{(i)}$, ..., $\lambda^{(i+s-1)}$ gleich, ist jedes $u \in \text{span}\left\{u^{(i)}, \ldots, u^{(i+s-1)}\right\} \setminus \{0\}$ ein Eigenvektor von (5.1.2). In solchen Situationen sollen wir nicht erwarten, dass die Folgen $\{\tilde{u}_k^{(q)}\}_k$ im üblichen Sinn gegen die Vektoren aus der Liste (5.1.3) konvergieren. Wir benutzen daher im

Folgenden eine andere Bedingung, welche die qualitativen Konvergenzeigenschaften eines solchen Verfahrens charakterisiert:

$$\begin{aligned}&\forall \epsilon > 0 \quad \exists k^* \in \mathbb{N}_0 : \quad \forall k > k^* \quad \exists u_*^{(1)}, \ldots, u_*^{(m)} \in \mathbb{R}^M : \\ &\quad u_*^{(1)}, \ldots, u_*^{(m)} \text{ sind Eigenvektoren von (5.1.2)}, \\ &\quad \|u_*^{(1)}\|_{\mathbf{B}} = \cdots = \|u_*^{(m)}\|_{\mathbf{B}} = 1, \\ &\quad \max_{1 \le q \le m} \|\tilde{u}_k^{(q)} - u_*^{(q)}\| < \epsilon, \end{aligned} \tag{5.2.7}$$

wobei $\|\cdot\|$ eine beliebige Norm in $\mathbb{R}^M$ ist. Also gibt es für jedes $\epsilon > 0$ einen Index, ab dem die Folgenglieder in (5.2.1) Eigenvektoren des Problems mit der Genauigkeit nicht schlechter als ϵ approximieren. Das Verfahren kann also das Problem (5.1.2) mit jeder vorgegebenen Genauigkeit lösen, d.h. für praktische Anwendungen reicht die Berechnung der Folge (5.2.1) bis zu einem bestimmten Index k aus. Eigenschaft (5.2.7) impliziert die Konvergenz gegen bestimmte Eigenvektoren nicht. Im Fall $m = 1$ (d.h. Algorithmus 5.1.1) mit einem von k unabhängigen Vorkonditionierer wurde aber die Konvergenz im üblichen Sinn in [53] bewiesen.

Zur Vereinfachung der Notation führen wir folgende Norm auf $\left(\mathbb{R}^M\right)^m$, aufgefasst als den Raum der m-Tupel $\mathbf{u} = (u^{(1)}, \ldots, u^{(m)})$, ein:

$$\|\mathbf{u}\| := \max_{1 \le q \le m} \|u^{(q)}\|_{\mathbf{B}}. \tag{5.2.8}$$

Wir führen zwecks besserer Verständlichkeit den Beweis für den Fall $m = 2$ vor. Der Beweis für den allgemeinen Fall ist analog. Aussage (5.2.7) beweisen wir in mehreren Schritten. Zunächst zeigen wir in Satz 5.2.7 und Bemerkung 5.2.8, wie sich die Asymptotik der Folge (5.2.1) durch deren Häufungspunkte beschreiben lässt. Dann beweisen wir in Satz 5.2.9 unter weiteren Annahmen an die Vorkonditionierer, dass alle Häufungspunkte Paare der Eigenvektoren sind.

Lemma 5.2.5 Seien V, W Untervektorräume von $\mathbb{R}^M$ und $W \subseteq V$. Wir bezeichnen die Extrema (in aufsteigender Reihenfolge) des Funktionals $\lambda(\cdot)$ auf $V \setminus \{0\}$ mit $\mu_1, \ldots, \mu_s$ und auf $W \setminus \{0\}$ mit $\nu_1, \ldots, \nu_t$. Dann gilt:

$$\mu_i \le \nu_i, \quad 1 \le i \le t. \tag{5.2.9}$$

Beweis: Wir wählen in W eine $\mathbf{B}$-orthogonale Basis $\{v_1, \ldots, v_t\}$ und ergänzen sie zu einer $\mathbf{B}$-orthogonalen Basis $\{v_1, \ldots, v_s\}$ von V. Dann sind $\mu_1, \ldots, \mu_s$ Eigenwerte der Matrix $A = ((\mathbf{A}v_i, v_j))_{1 \le i,j \le s} \in \mathbb{R}^{s \times s}$ und $\nu_1, \ldots, \nu_t$ Eigenwerte ihrer Teilmatrix $\hat{A} = ((\mathbf{A}v_i, v_j))_{1 \le i,j \le t} \in \mathbb{R}^{t \times t}$. Daraus folgt (5.2.9) (siehe „Cauchy's interlace theorem“ in [28]). □

Mit der Hilfe dieses Lemmas erhalten wir folgende Aussage:

Korollar 5.2.6 Für $m = 2$ betrachten wir die durch Algorithmus 5.2.1 für den Anfangswert $\mathbf{u}_0$ definierte Folge (5.2.1). Es gelten für alle $k \ge 1$ die Monotonieungleichungen

$$\lambda(\tilde{u}_{k+1}^{(1)}) \le \lambda(\tilde{u}_k^{(1)}), \quad \lambda(\tilde{u}_{k+1}^{(2)}) \le \lambda(\tilde{u}_k^{(2)}). \tag{5.2.10}$$

Beweis: Es sei $k \geq 1$. Wir betrachten den k-ten Schritt von Algorithmus 5.2.1 und wenden Lemma 5.2.5 auf die Räume $V = \text{span}\left\{\tilde{u}_k^{(1)}, \tilde{u}_k^{(2)}, c_k'^{(1)}, \ldots, c_k'^{(m')}\right\}$ und $W = \text{span}\left\{\tilde{u}_k^{(1)}, \tilde{u}_k^{(2)}\right\}$. Vektoren $\tilde{u}_k^{(1)}$ und $\tilde{u}_k^{(2)}$ sind lokale Extremstellen von $\lambda(\cdot)$ auf der Menge $\text{span}\left\{\tilde{u}_{k-1}^{(1)}, \tilde{u}_{k-1}^{(2)}, c_k'^{(1)}, \ldots, c_k'^{(m')}\right\} \setminus \{0\}$ (siehe (5.2.5)), und somit auch solche auf der Teilmenge $\text{span}\left\{\tilde{u}_k^{(1)}, \tilde{u}_k^{(2)}\right\} \setminus \{0\}$. Also sind $\lambda(\tilde{u}_k^{(1)})$ und $\lambda(\tilde{u}_k^{(2)})$ die Extrema von $\lambda(\cdot)$ auf $W \setminus \{0\}$. Da die zwei kleinsten Extrema von $\lambda(\cdot)$ auf $V \setminus \{0\}$ die Werte $\lambda(\tilde{u}_{k+1}^{(1)})$ und $\lambda(\tilde{u}_{k+1}^{(2)})$ sind (siehe (5.2.6)), folgt (5.2.10) aus (5.2.9). □

Dieses Ergebnis verwenden wir zum Beweis von Satz 5.2.7:

Satz 5.2.7 Für $m = 2$ sei Folge (5.2.1) durch Algorithmus 5.2.1 und den Anfangswert $\mathbf{u}_0$ mit $\lambda(\tilde{u}_0^{(1)}) \leq \lambda(\tilde{u}_0^{(2)})$ definiert. Dann gilt:

1. Folge (5.2.1) hat mindestens einen Häufungspunkt.
2. Für jeden Häufungspunkt $\mathbf{u}_* = (u_*^{(1)}, u_*^{(2)})$ von (5.2.1) gilt: $\|u_*^{(1)}\|_{\mathbf{B}} = \|u_*^{(2)}\|_{\mathbf{B}} = 1$, $(u_*^{(1)}, u_*^{(2)})_{\mathbf{B}} = 0$ und $\lambda(u_*^{(1)}) \leq \lambda(u_*^{(2)})$.
3. Für jedes $q \in \{1, 2\}$ ist die Folge $\left\{\lambda(\tilde{u}_k^{(q)})\right\}_k$ konvergent, $\lim_{k\to\infty} \lambda(\tilde{u}_k^{(q)}) =: \lambda_*^{(q)}$. Dabei gilt für alle $k \geq 1$

$$\lambda(\tilde{u}_k^{(1)}) \geq \lambda_*^{(1)}, \qquad \lambda(\tilde{u}_k^{(2)}) \geq \lambda_*^{(2)}. \tag{5.2.11}$$

   Für jeden Häufungspunkt $\mathbf{u}_* = (u_*^{(1)}, u_*^{(2)})$ gilt $(\lambda(u_*^{(1)}), \lambda(u_*^{(2)})) = (\lambda_*^{(1)}, \lambda_*^{(2)})$. Also ist das Paar $(\lambda(u_*^{(1)}), \lambda(u_*^{(2)}))$ für alle Häufungspunkte $\mathbf{u}_* = (u_*^{(1)}, u_*^{(2)})$ gleich.

Beweis: 1. Da die Vektoren $\tilde{u}_k^{(q)}$ **B**-normiert sind, ist Folge (5.2.1) bzgl. der Norm (5.2.8) beschränkt und hat daher mindestens einen Häufungspunkt.

2. Wir betrachten nun eine Teilfolge $\{\mathbf{u}_{k_n}\}_n$, die gegen $\mathbf{u}_*$ konvergiert. Da das Skalarprodukt $(\cdot,\cdot)_{\mathbf{B}}$ und die Norm $\|\cdot\|_{\mathbf{B}}$ stetige Funktionen sind, und für jedes $k_n \in \mathbb{N}_0$ die Aussagen $\|\tilde{u}_{k_n}^{(1)}\|_{\mathbf{B}} = \|\tilde{u}_{k_n}^{(2)}\|_{\mathbf{B}} = 1$ und $(\tilde{u}_{k_n}^{(1)}, \tilde{u}_{k_n}^{(2)})_{\mathbf{B}} = 0$ gelten, erhalten wir: $\|u_*^{(1)}\|_{\mathbf{B}} = \|u_*^{(2)}\|_{\mathbf{B}} = 1$, $(u_*^{(1)}, u_*^{(2)})_{\mathbf{B}} = 0$. Insbesondere folgt daraus, dass $u_*^{(1)}, u_*^{(2)} \neq 0$. Also ist $\lambda(\cdot)$ in einer Umgebung von diesen Punkten wohldefiniert und stetig. Da $\lambda(\tilde{u}_{k_n}^{(1)}) \leq \lambda(\tilde{u}_{k_n}^{(2)})$ für jedes $k_n \in \mathbb{N}_0$, gilt die entsprechende Ungleichung für die Grenzwerte $u_*^{(1)}$ und $u_*^{(2)}$.

3. Es sei $q \in \{1, 2\}$. Nach Korollar 5.2.6 ist die Folge $\left\{\lambda(\tilde{u}_k^{(q)})\right\}_k$ monoton fallend. Außerdem ist $\lambda(\cdot)$ von unten mit dem kleinsten Eigenwert von Problem (5.1.2) beschränkt. Daraus folgt, dass $\left\{\lambda(\tilde{u}_k^{(q)})\right\}_k$ konvergent ist und für ihren Grenzwert $\lambda_*^{(q)}$ die Ungleichungen (5.2.11) gelten.

Es sei nun $\{\mathbf{u}_{k_n}\}_n$ eine Teilfolge von (5.2.1), die gegen deren Häufungspunkt $\mathbf{u}_*$ konvergiert. Dann gilt $\lim_{n\to\infty} \lambda(\tilde{u}_{k_n}^{(q)}) = \lambda_*^{(q)}$. Da $\lambda(\cdot)$ stetig ist, erhalten wir daher: $\lambda(u_*^{(q)}) = \lambda_*^{(q)}$. □

Bemerkung 5.2.8 Satz 5.2.7 beschreibt das Grenzverhalten der von Algorithmus 5.2.1 hergestellten Folge (5.2.1) vollständig, da **für jedes $\epsilon > 0$ es ein solches $k^* \in \mathbb{N}_0$ gibt, dass für jedes $k \geq k^*$ das Glied $\mathbf{u}_k$ in der ϵ-Umgebung eines der Häufungspunkte dieser Folge liegt**. In der Tat kann außerhalb der Vereinigung von ϵ-Umgebungen (für ein beliebiges $\epsilon > 0$) *aller* Häufungspunkte keine unendliche Menge von Folgengliedern liegen: Diese Menge wäre beschränkt, und hätte daher einen eigenen Häufungspunkt, der dann auch Häufungspunkt der Folge wäre. Das ist aber ein Widerspruch, weil dieser Häufungspunkt nicht zu der Menge aller Häufungspunkte dieser Folge gehört. □

Wie diese Bemerkung zeigt, erhält man Aussage (5.2.7), wenn man beweist, dass alle Häufungspunkte der Folge (5.2.1) nur aus Eigenvektoren des Problems (5.1.2) bestehen. Die Voraussetzungen von Satz 5.2.7 sind aber dafür nicht ausreichend. Die Situation ist analog zu der beim Quasi-Newton-Verfahren: Für die Konvergenz gegen einen Punkt des lokalen Extremums sollen die angenäherten Jacobi-Matrizen nicht nur positiv definit sein, sondern auch asymptotisch nicht singulär werden (siehe [27]). Wir fordern im folgenden eine restriktivere Bedingung, die aber für unsere Zwecke reicht: Die Anzahl der verwendeten Vorkonditionierer soll endlich sein. Der folgende Satz 5.2.9 beweist die Konvergenz (im Sinne von (5.2.7)) von Algorithmus 5.2.1 in diesem Fall.

Satz 5.2.9 Alle Vorkonditionierer $\mathbf{W}_{q,k}$ in Algorithmus 5.2.1 seien Elemente einer endlichen Menge $\{\mathbf{W}_{(1)}, \ldots, \mathbf{W}_{(P)}\}$. Unter den Voraussetzungen von Satz 5.2.7 besteht jeder Häufungspunkt $\mathbf{u}_* = (u_*^{(1)}, u_*^{(2)})$ der Folge (5.2.1) aus Eigenvektoren des Problems (5.1.2).

Zum Beweis dieses Satzes benötigen wir die folgenden Hilfsaussagen:

Bemerkung 5.2.10 Seien $u \in \mathbb{R}^M$ ein beliebiger Vektor und $r = \lambda(u)\mathbf{B}u - \mathbf{A}u$ sein Residuum. Es gilt:
1. r und u sind zueinander orthogonal: $(r, u) = (\lambda(u)\mathbf{B}u - \mathbf{A}u, u) = \lambda(u)(\mathbf{B}u, u) - (\mathbf{A}u, u) = 0$.
2. Sei u_* ein Eigenvektor von (5.1.2) mit $(u, u_*)_{\mathbf{B}} = 0$. Dann sind r und u_* orthogonal zueinander: $(r, u_*) = (\lambda(u)\mathbf{B}u - \mathbf{A}u, u_*) = \lambda(u)(\mathbf{B}u, u_*) - (u, \mathbf{A}u_*) = (\lambda(u) - \lambda_*)(\mathbf{B}u, u_*) = 0$, wobei λ_* der dem Eigenvektor u_* entsprechende Eigenwert ist. □

Bemerkung 5.2.11 Sei $n \in \mathbb{N}$ eine beliebige Zahl. Die Eigenwerte $\nu_1, \ldots, \nu_n$ des Problems

$$A\boldsymbol{\omega} = \nu B\boldsymbol{\omega}, \tag{5.2.12}$$

wobei $\boldsymbol{\omega} \in \mathbb{R}^n$,

$$A = \mathbf{H}^T\mathbf{A}\mathbf{H}, \quad B = \mathbf{H}^T\mathbf{B}\mathbf{H},$$

$\mathbf{H}$ die Matrix mit Spalten $u_1, \ldots, u_n$ ist, sind stetige Funktionen von $(u_1, \ldots, u_n)$, wenn die Vektoren $u_1, \ldots, u_n \in \mathbb{R}^M$ linear unabhängig sind. In diesem Fall ist die Matrix B nämlich positiv definit. Deswegen ist (5.2.12) auf das symmetrische Problem $\hat{A}\hat{\boldsymbol{\omega}} = \nu\hat{\boldsymbol{\omega}}$ mit den gleichen Eigenwerten (z.B. durch die Cholesky-Zerlegung von B) reduzierbar, wobei die Einträge von $\hat{A}$ stetig von denen der Matrizen A und

B und somit der Vektoren u_i abhängen. Die Aussage folgt jetzt aus der Wielandt-Hoffmann-Ungleichung (siehe [45]), nach der die Eigenwerte von $\hat{A}$ stetige Funktionen dieser Matrix sind. □

Beweis von Satz 5.2.9: Sei $\mathbf{u}_* = (u_*^{(1)}, u_*^{(2)})$ ein Häufungspunkt der Folge (5.2.1). Im Folgenden zeigen wir, dass

$$r_*^{(q)} := \lambda(u_*^{(q)})\mathbf{B}u_*^{(q)} - \mathbf{A}u_*^{(q)} = 0, \quad q \in \{1, 2\}.$$

1. Wir nehmen an, dass $r_*^{(1)} \neq 0$, und bezeichnen

$$c_{*,p}^{(1)} := \mathbf{W}_{(p)}^{-1} r_*^{(1)} = \mathbf{W}_{(p)}^{-1} \left(\lambda(u_*^{(1)})\mathbf{B}u_*^{(1)} - \mathbf{A}u_*^{(1)}\right) \neq 0.$$

Da nach Bemerkung 5.2.10 $(r_*^{(1)}, u_*^{(1)}) = 0$ gilt, sind die $c_{*,p}^{(1)}$ $(p \in \{1, \ldots, P\})$ $\mathbf{W}_{(p)}$-orthogonal zu $u_*^{(1)}$. Insbesondere sind die Systeme $\{u_*^{(1)}, c_{*,p}^{(1)}\}$ für alle p linear unabhängig.

Da alle Vorkonditionierer $\mathbf{W}_{(p)}$ positiv definit sind, ist jeder Vektor $c_{*,p}^{(1)}$ eine Abstiegsrichtung von $\lambda(\cdot)$ im Punkt $u_*^{(1)}$, also gilt

$$\mu_{(p)} := \min_{u \in \operatorname{span}\left\{u_*^{(1)}, c_{*,p}^{(1)}\right\} \setminus \{0\}} \lambda(u) < \lambda(u_*^{(1)}).$$

Nach Bemerkung 5.2.11 (für zwei Vektoren) und den Eigenschaften von Extrema des Rayleigh-Quotienten $\lambda(\cdot)$ erhalten wir, dass

$$\mu(u, c) := \min_{\tilde{u} \in \operatorname{span}\{u, c\}} \lambda(\tilde{u}) \tag{5.2.13}$$

eine stetige Funktion in $(u, c) = (u_*^{(1)}, c_{*,p}^{(1)})$ ist. Dann existieren für jedes $\epsilon_p = \lambda(u_*^{(1)}) - \mu_{(p)} > 0$ Umgebungen $N_{u_*^{(1)}}$ und $N_{c_{*,p}^{(1)}}$ von $u_*^{(1)}$ bzw. $c_{*,p}^{(1)}$, sodass $\mu(u, c) \in (\mu_{(p)} - \epsilon_p, \mu_{(p)} + \epsilon_p)$ für jede $u \in N_{u_*^{(1)}}$ und $c \in N_{c_{*,p}^{(1)}}$ gilt. D.h.

$$\forall u \in N_{u_*^{(1)}} \quad \forall c \in N_{c_{*,p}^{(1)}} \quad \mu(u, c) < \lambda(u_*^{(1)}). \tag{5.2.14}$$

Da zusätzlich der Vektor $\mathbf{W}_{(p)}^{-1}(\lambda(u)\mathbf{B}u - \mathbf{A}u)$ stetig von u (bei $u \neq 0$) abhängt, gibt es für jedes $p \in \{1, \ldots, P\}$ eine solche Umgebung $N_{u_*^{(1)}}^{(p)}$ von $u_*^{(1)}$, dass $\mathbf{W}_{(p)}^{-1}(\lambda(u)\mathbf{B}u - \mathbf{A}u) \in N_{c_{*,p}^{(1)}}$ für alle $u \in N_{u_*^{(1)}}^{(p)}$ gilt. Also erhalten wir nach (5.2.14) die Ungleichung

$$\forall u \in \hat{N}_{u_*^{(1)}} \quad \mu(u, \mathbf{W}_{(p)}^{-1}(\lambda(u)\mathbf{B}u - \mathbf{A}u)) < \lambda(u_*^{(1)}), \tag{5.2.15}$$

wobei $\hat{N}_{u_*^{(1)}} := N_{u_*^{(1)}} \cap N_{u_*^{(1)}}^{(1)} \cap \cdots \cap N_{u_*^{(1)}}^{(P)}$.

Da aber $\hat{N}_{u_*^{(1)}}$ auch eine Umgebung von $u_*^{(1)}$ ist, ist $\hat{N}_{u_*^{(1)}} \times \mathbb{R}^M$ eine Umgebung des Häufungspunktes $\mathbf{u}_*$. Daher existiert ein solches $k \geq 0$, für das $\mathbf{u}_k$ in $\hat{N}_{u_*^{(1)}} \times \mathbb{R}^M$ liegt. Nach (5.2.15) und (5.2.13) gilt dann:

$$\min_{u \in \operatorname{span}\left\{\tilde{u}_k^{(1)}, c_k^{(1)}\right\} \setminus \{0\}} \lambda(u) < \lambda(u_*^{(1)}). \tag{5.2.16}$$

Nach (5.2.6) und der Wahl von $\mu_k^{(1)}$ gilt aber

$$\lambda(\tilde{u}_{k+1}^{(1)}) = \min_{u \in \operatorname{span}\left\{\tilde{u}_k^{(1)}, \tilde{u}_k^{(2)}, c_k^{(1)}, c_k^{(2)}\right\} \setminus \{0\}} \lambda(u).$$

Da $\operatorname{span}\left\{\tilde{u}_k^{(1)}, c_k^{(1)}\right\} \setminus \{0\}$ eine Teilmenge von $\operatorname{span}\left\{\tilde{u}_k^{(1)}, \tilde{u}_k^{(2)}, c_k^{(1)}, c_k^{(2)}\right\} \setminus \{0\}$ ist, folgt aus (5.2.16), dass

$$\lambda(\tilde{u}_{k+1}^{(1)}) < \lambda(u_*^{(1)}),$$

was der Aussage von Satz 5.2.7 widerspricht. Also folgt $r_*^{(1)} = 0$, und $u_*^{(1)}$ ist ein Eigenvektor von (5.1.2).

2. Wir nehmen nun an, dass $r_*^{(1)} = 0$, aber $r_*^{(2)} \neq 0$. Dann ist $u_*^{(1)}$ ein Eigenvektor von (5.1.2). Außerdem sind nach Satz 5.2.7 die Vektoren $u_*^{(1)}$ und $u_*^{(2)}$ **B**-orthogonal zueinander. Nach Bemerkung 5.2.10 erhalten wir dann:

$$(r_*^{(2)}, u_*^{(1)}) = 0, \quad (r_*^{(2)}, u_*^{(2)}) = 0. \tag{5.2.17}$$

Daraus folgt, dass

$$c_{*,p}^{(2)} := \mathbf{W}_{(p)}^{-1} r_*^{(2)} = \mathbf{W}_{(p)}^{-1} \left(\lambda(u_*^{(2)})\mathbf{B}u_*^{(2)} - \mathbf{A}u_*^{(2)}\right)$$

$\mathbf{W}_{(p)}$-orthogonal zu den Vektoren $u_*^{(1)}$ und $u_*^{(2)}$ ist. Somit ist das System $\{u_*^{(1)}, u_*^{(2)}, c_{*,p}^{(2)}\}$ linear unabhängig für jedes $p \in \{1, \ldots, P\}$.

Für beliebige Vektoren $u^{(1)}, u^{(2)}, c \in \mathbb{R}^M$ definieren wir

$$A(u^{(1)}, u^{(2)}, c) := \begin{pmatrix} (\mathbf{A}u^{(1)}, u^{(1)}) & (\mathbf{A}u^{(1)}, u^{(2)}) & (\mathbf{A}u^{(1)}, c) \\ (\mathbf{A}u^{(2)}, u^{(1)}) & (\mathbf{A}u^{(2)}, u^{(2)}) & (\mathbf{A}u^{(2)}, c) \\ (\mathbf{A}c, u^{(1)}) & (\mathbf{A}c, u^{(2)}) & (\mathbf{A}c, c) \end{pmatrix},$$

$$B(u^{(1)}, u^{(2)}, c) := \begin{pmatrix} (\mathbf{B}u^{(1)}, u^{(1)}) & (\mathbf{B}u^{(1)}, u^{(2)}) & (\mathbf{B}u^{(1)}, c) \\ (\mathbf{B}u^{(2)}, u^{(1)}) & (\mathbf{B}u^{(2)}, u^{(2)}) & (\mathbf{B}u^{(2)}, c) \\ (\mathbf{B}c, u^{(1)}) & (\mathbf{B}c, u^{(2)}) & (\mathbf{B}c, c) \end{pmatrix}$$

und betrachten für jedes $p \in \{1, \ldots, P\}$ das Eigenwertproblem

$$A(u_*^{(1)}, u_*^{(2)}, c_{*,p}^{(2)})\boldsymbol{\alpha} = \mu B(u_*^{(1)}, u_*^{(2)}, c_{*,p}^{(2)})\boldsymbol{\alpha}. \tag{5.2.18}$$

Da die Vektoren $u_*^{(1)}$, $u_*^{(2)}$ und $c_{*,p}^{(2)}$ linear unabhängig sind, ist die Matrix $B(u_*^{(1)}, u_*^{(2)}, c_{*,p}^{(2)})$ positiv definit, also hat (5.2.18) drei linear unabhängige Eigenvektoren. Es seien $\mu_{*,p}^{(1)}$, $\mu_{*,p}^{(2)}$ und $\mu_{*,p}^{(3)}$ ihre Eigenwerte in aufsteigender Reihenfolge. Wir zeigen nun folgende Aussage:

$$\mu_{*,p}^{(1)} < \lambda(u_*^{(1)}) \quad \text{oder} \quad \mu_{*,p}^{(2)} < \lambda(u_*^{(2)}). \tag{5.2.19}$$

Wir bezeichnen mit

$$\mu_{*,p}(\boldsymbol{\alpha}) := \frac{(A(u_*^{(1)}, u_*^{(2)}, c_{*,p}^{(2)})\boldsymbol{\alpha}, \boldsymbol{\alpha})}{(B(u_*^{(1)}, u_*^{(2)}, c_{*,p}^{(2)})\boldsymbol{\alpha}, \boldsymbol{\alpha})}$$

den mit dem Problem (5.2.18) assoziierten Rayleigh-Quotienten. Analog zu (5.2.5) erhalten wir dann für $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$:

$$(B(u_*^{(1)}, u_*^{(2)}, c_{*,p}^{(2)})\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\mathbf{B}(\alpha_1 u_*^{(1)} + \alpha_2 u_*^{(2)} + \alpha_3 c_{*,p}^{(2)}), \beta_1 u_*^{(1)} + \beta_2 u_*^{(2)} + \beta_3 c_{*,p}^{(2)}) \quad (5.2.20)$$

und

$$\mu_{*,p}(\boldsymbol{\alpha}) = \lambda(\alpha_1 u_*^{(1)} + \alpha_2 u_*^{(2)} + \alpha_3 c_{*,p}^{(2)}). \quad (5.2.21)$$

Aus (5.2.21) folgt, dass jeder Eigenvektor $\boldsymbol{\alpha}$ von (5.2.18) dem lokalen Extremum $u = \alpha_1 u_*^{(1)} + \alpha_2 u_*^{(2)} + \alpha_3 c_{*,p}^{(2)}$ von $\lambda(\cdot)$ auf der Menge

$$U_{*,p} := \text{span}\left\{u_*^{(1)}, u_*^{(2)}, c_{*,p}^{(2)}\right\} \setminus \{0\}$$

entspricht. Da außerdem $u_*^{(1)}$ nach Voraussetzung ein Eigenvektor von (5.1.2) ist, hat $\lambda(\cdot)$ in $u_*^{(1)}$ ein lokales Extremum auf einer größeren Menge. Deswegen ist $\boldsymbol{\alpha} = (1, 0, 0)^T$ ein Eigenvektor von (5.2.18), und $\lambda(u_*^{(1)})$ ist einer der Eigenwerte $\mu_{*,p}^{(1)}$, $\mu_{*,p}^{(2)}$ oder $\mu_{*,p}^{(3)}$.

Wir nehmen an, die erste Ungleichung in (5.2.19) wäre falsch, also $\mu_{*,p}^{(1)} \geq \lambda(u_*^{(1)})$. Dann gilt $\mu_{*,p}^{(1)} = \lambda(u_*^{(1)})$. Die Eigenvektoren von (5.2.18) zum Eigenwert $\mu_{*,p}^{(2)}$ sind in diesem Fall $B(u_*^{(1)}, u_*^{(2)}, c_{*,p}^{(2)})$-orthogonal zu $(1, 0, 0)^T$. Nach (5.2.20) entsprechen sie also solchen Vektoren aus $U_{*,p}$, die $\mathbf{B}$-orthogonal zu $u_*^{(1)}$ sind. Einer von diesen ist $u_*^{(2)}$. Das gleiche gilt auch für die Projektion

$$\tilde{c}_{*,p}^{(2)} = c_{*,p}^{(2)} - \frac{(\mathbf{B}c_{*,p}^{(2)}, u_*^{(1)})}{(\mathbf{B}u_*^{(1)}, u_*^{(1)})} u_*^{(1)}.$$

Dieser Vektor ist ungleich Null, da $u_*^{(1)}$ und $c_{*,p}^{(2)}$ linear unabhängig sind. Da nach (5.2.17)

$$(\tilde{c}_{*,p}^{(2)}, r_*^{(2)}) = (c_{*,p}^{(2)}, r_*^{(2)}) < 0$$

gilt, ist $\tilde{c}_{*,p}^{(2)}$ eine Abstiegsrichtung von $\lambda(\cdot)$ in $u_*^{(2)}$. Folglich ist das Minimum von $\lambda(\cdot)$ auf $\{u \in U_{*,p} : (\mathbf{B}u, u_*^{(1)}) = 0\}$ kleiner als $\lambda(u_*^{(2)})$. Also $\mu_{*,p}^{(2)} < \lambda(u_*^{(2)})$, und (5.2.19) ist bewiesen.

Nach Bemerkung 5.2.11 hängen die Eigenwerte des Problems

$$A(u^{(1)}, u^{(2)}, c)\boldsymbol{\alpha} = \mu B(u^{(1)}, u^{(2)}, c)\boldsymbol{\alpha} \quad (5.2.22)$$

mit $c = \mathbf{W}_{(p)}^{-1}(\lambda(u^{(2)})\mathbf{B}u^{(2)} - \mathbf{A}u^{(2)})$ stetig von den Vektoren $u^{(1)}$ und $u^{(2)}$ ab, wenn das System $\{u^{(1)}, u^{(2)}, c\}$ linear unabhängig ist. Analog zum ersten Teil des Beweises können wir daraus schließen, dass es für jedes $p \in \{1, \ldots, P\}$ eine solche Umgebung $\mathbf{N}_{\mathbf{u}_*}^{(p)}$ des Punktes $\mathbf{u}_*$ gibt, dass für jedes $\mathbf{u} = (u^{(1)}, u^{(2)}) \in \mathbf{N}_{\mathbf{u}_*}^{(p)}$ entweder $\mu^{(1)} < \lambda(u^{(1)})$ oder $\mu^{(2)} < \lambda(u^{(2)})$ gilt, wobei $\mu^{(1)}$ und $\mu^{(2)}$ die zwei kleinsten Eigenwerte von (5.2.22) sind. Die Menge $\hat{\mathbf{N}}_{\mathbf{u}_*} := \mathbf{N}_{\mathbf{u}_*}^{(1)} \cap \cdots \cap \mathbf{N}_{\mathbf{u}_*}^{(P)}$ ist eine Umgebung von $\mathbf{u}_*$. Da $\mathbf{u}_*$ ein Häufungspunkt ist, gibt es mindestens ein Folgenglied $\mathbf{u}_{k^*}$, das in $\hat{\mathbf{N}}_{\mathbf{u}_*}$ liegt. Unabhängig von den Vorkonditionierern $\mathbf{W}_{1,k^*}$ und $\mathbf{W}_{2,k^*}$ gilt dann

$$\tilde{\mu}_{k^*}^{(1)} < \lambda(u_*^{(1)}) \quad \text{oder} \quad \tilde{\mu}_{k^*}^{(2)} < \lambda(u_*^{(2)}), \quad (5.2.23)$$

wobei $\tilde{\mu}_{k^*}^{(1)}$ und $\tilde{\mu}_{k^*}^{(2)}$ die zwei kleinsten Eigenwerte des Problems

$$A(\tilde{u}_{k^*}^{(1)}, \tilde{u}_{k^*}^{(2)}, c_{k^*}^{(2)})\boldsymbol{\omega} = \mu B(\tilde{u}_{k^*}^{(1)}, \tilde{u}_{k^*}^{(2)}, c_{k^*}^{(2)})\boldsymbol{\omega} \tag{5.2.24}$$

sind. Analog zu (5.2.5) sind dann $\tilde{\mu}_{k^*}^{(1)}$ und $\tilde{\mu}_{k^*}^{(2)}$ die kleinsten Extrema von $\lambda(\cdot)$ auf $W \setminus \{0\}$ mit $W = \text{span}\left\{\tilde{u}_{k^*}^{(1)}, \tilde{u}_{k^*}^{(2)}, c_{k^*}^{(2)}\right\}$, einem Untervektorraum von $V = \text{span}\left\{\tilde{u}_{k^*}^{(1)}, \tilde{u}_{k^*}^{(2)}, c_{k^*}^{\prime(1)}, \ldots, c_{k^*}^{\prime(m')}\right\}$. Nach (5.2.6) liefert die Anwendung von Lemma 5.2.5 auf die so definierten V und W

$$\lambda(\tilde{u}_{k^*+1}^{(1)}) \leq \tilde{\mu}_{k^*}^{(1)}, \quad \lambda(\tilde{u}_{k^*+1}^{(2)}) \leq \tilde{\mu}_{k^*}^{(2)},$$

woraus nach (5.2.23)

$$\lambda(\tilde{u}_{k^*+1}^{(1)}) < \lambda(u_*^{(1)}) \quad \text{oder} \quad \lambda(\tilde{u}_{k^*+1}^{(2)}) < \lambda(u_*^{(2)})$$

folgt. Das widerspricht aber Satz 5.2.7. Damit haben wir bewiesen, dass auch $u_*^{(2)}$ ein Eigenvektor von Problem (5.1.2) ist. □

5.3 Folge der frequenzfilternden Vorkonditionierer

Hier versuchen wir, als die Vorkonditionierer $\mathbf{W}_{q,k}$ die oben in dieser Arbeit eingeführten GIBLU(1)-Zerlegungen der Matrix $\mathbf{A}$ zu benutzen. Wir berechnen die Koeffizienten dieser Zerlegungen nach Algorithmus 4.1.7 und wählen den Testvektor abhängig von der Schrittnummer k, sodass diese Vorkonditionierer während des Lösungsprozesses die Matrix $\mathbf{A}$ in verschiedenen Teilen deren Spektrums annähern.

In der Literatur findet man die Anwendung von Folgen von frequenzfilternden Zerlegungen mit schrittabhängigen Parametern auf lineare Gleichungssysteme. C. Wagner zeigt in [41] für das Modellproblem, dass für eine spezielle Wahl der Folge der Parameter die TFF-Zerlegungen eine gitterunabhängige Konvergenzrate haben. Das gleiche beweist A. Buzdin in [9] für die tangentialen Zerlegungen. Wir beschränken uns hier mit einer einfacheren Wahl der Parameter und untersuchen die Konvergenz des so vorkonditionierten Block-Gradienten-Verfahren durch numerische Experimente.

Wir betrachten also Algorithmus 5.2.1, wobei $\mathbf{W}_{q,k}$ gleich $\mathbf{W}^{(1)}$ von den GIBLU(1)-Zerlegungen der Matrix $\mathbf{A}$ sind. In jedem Schritt wählen wir für die Näherungen an alle Eigenvektoren den gleichen Vorkonditionierer, sodass $\mathbf{W}_{q,k} = \mathbf{W}_k$ nur von k abhängt. Als Testvektoren wählen wir Sinus-Funktionen, wie in Abschnitt 4.3: Abhängig von der Wellenzahl $\mathfrak{i}$ ist der Testvektor $\mathbf{e}(\mathfrak{i}) = \text{blockvector}\{e_j(\mathfrak{i})\}$, wobei

$$(e_j(\mathfrak{i}))_i = \sin \frac{\pi i \cdot \min\{\mathfrak{i}, n_j\}}{n_j + 1}, \tag{5.3.1}$$

j die Nummer des Matrixblocks und $i \in \{1, \ldots, n_j\}$ die Nummer des Knotens in diesem Block ist. Im k-ten Schritt des Gradienten-Verfahrens wählen wir dann den Vorkonditionierer mit dem Testvektor $\mathbf{e}(\mathfrak{i}_k)$, wobei

$$\mathfrak{i}_k = 2^{(k-1) \bmod S} \tag{5.3.2}$$

mit einer vorgegebenen Zahl $S \in \mathbb{N}$ ist. Die Folge der Vorkonditionierer wird also mehrmals angewendet. Den Wert S wählen wir so, dass

$$2^{S-1} \leq \max_j n_j < 2^S. \tag{5.3.3}$$

Die entsprechenden Testvektoren stellen dann die Frequenzmoden in allen Teilen des Spektrums der Diagonalblöcke der Matrix dar. Nach Satz 5.2.9 konvergiert das so vorkonditionierte Block-Gradientenverfahren gegen die Eigenvektoren des Problems (5.1.2).

Bemerkung 5.3.1 Die nach Regel (5.3.2) gewählten Frequenzindizes $\mathbf{i}_k$ sind nicht gleichmäßig im Intervall $[1, \max_j n_j]$ verteilt. Dies ist nach den folgenden Überlegungen auch so vorgesehen: Die Konstruktion der Funktion g_1 (siehe Abschnitt 3.3.1) zeigt, dass im Fall des Modellproblems die GIBLU(1)-Zerlegung bei dem einer niedrigen Frequenz entsprechenden Parameter $\hat{\mu} := \hat{\mu}_0 = \hat{\mu}_1$ eine breite Reihe von glatten Frequenzmoden im Fehler dämpft. Wenn umgekehrt dieser Parameter einer höheren Frequenz entspricht, werden nur die Moden gedämpft, deren Wellenzahl sich vom durch den Parameter angegebenen nicht sehr stark unterscheidet. Um alle Frequenzanteile des Fehlers gleich zu dämpfen, müssen wir mehr Vorkonditionierer für die höheren Frequenzen benutzen als für die niedrigen. □

Obwohl dieses Verfahren mehrere Vorkonditionierer benutzt, ist der gesamte Aufwand vergleichsweise klein. Da alle Vorkonditionierer für die gleiche Blockstruktur konstruiert werden, muss die Zerlegung des Gitters nur einmal vor dem Lösungsprozess durchgeführt werden. Ebenso werden auch die Koeffizienten jeder GIBLU(1)-Zerlegung nur einmal vor dem Prozess berechnet. In jedem Schritt des Gradientenverfahrens wird einfach der Algorithmus 2.1.3 mit den entsprechenden Koeffizienten aufgerufen. Im Lauf des Lösungsprozesses sind also die Varianten des Block-Gradientenverfahren mit einem und mehreren Vorkonditionierern vom Aufwand her gleich. Da die Anzahl der Vorkonditionierer bei diesem Verfahren logarithmisch mit der Blockgröße wächst und die GIBLU(1)-Zerlegungen nur zwei Koeffizienten pro Block benötigen, ist der Speicherbedarf bei dieser Vorgehensweise auch nicht groß.

Wir betrachten zunächst das kontinuierliche Eigenwertproblem

$$\begin{aligned} -\Delta u &= \lambda u, \\ u|_{\partial\Omega} &= 0, \end{aligned} \tag{5.3.4}$$

wobei $\Omega = (0,1)^2$ das Einheitsquadrat ist. Die Diskretisierung von (5.3.4) mit dem Finite-Volumen-Verfahren auf dem regulären Dreiecksgitter (siehe Abbildung 1.1) führt dann zum algebraischen Problem

$$\mathbf{A}u = \lambda \mathbf{B}u \tag{5.3.5}$$

mit

$$\mathbf{A} = \text{blocktridiag}\,\{-L, D, -L\}, \qquad \mathbf{B} = h^2 I, \tag{5.3.6}$$

wobei $D = \text{tridiag}\,\{-1, 4, -1\}$, $L = I$, $h = \frac{1}{n+1}$ und n die Anzahl der inneren Knoten in jedem Matrixblock ist.

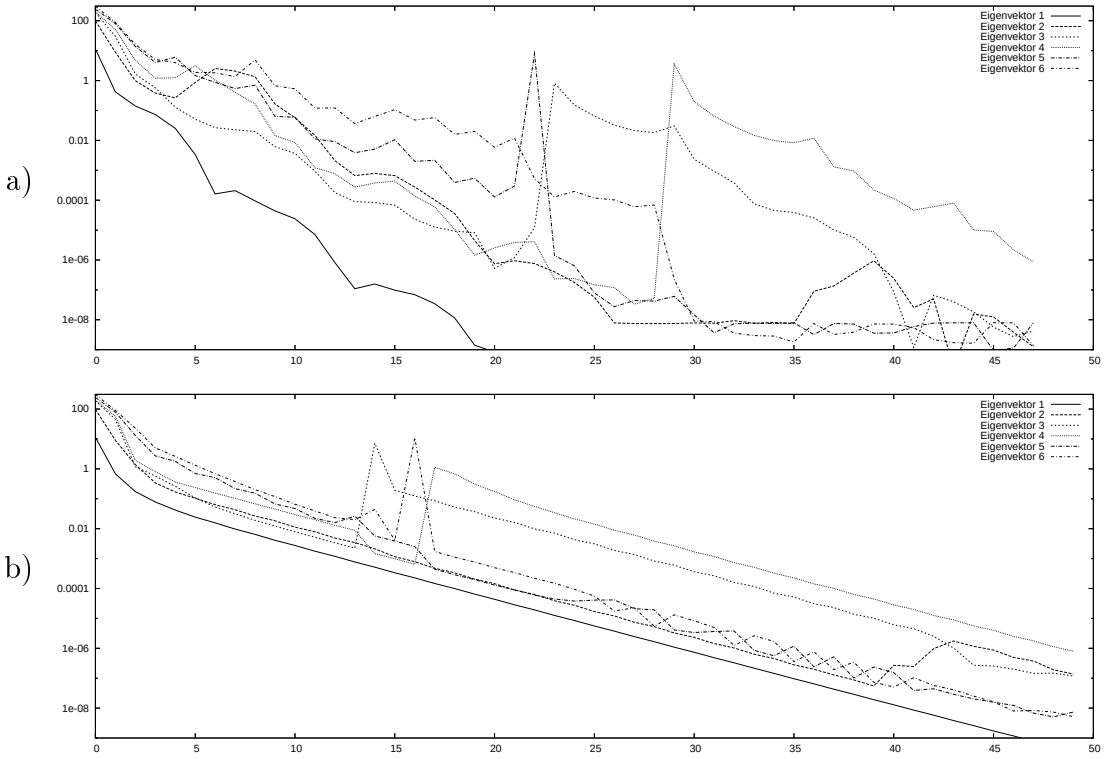

Abbildung 5.1: Die Konvergenz des Block-Gradientenverfahren mit mehreren (a) und einem (b) GIBLU(1)-Vorkonditionierer im Fall des Problems (5.3.5–5.3.6) für $n = 127$. Die Kurven zeigen die Norm $\|\mathbf{A}u - \lambda(u)\mathbf{B}u\|_2$ als Funktion der Schrittnummer.

Als Eingabewerte für Algorithmus 5.2.1 braucht man m **B**-orthogonale Anfangsnäherungen. In allen Experimenten weiter unten erhalten wir diese $\tilde{u}_0^{(q)}$ $(1 \leq q \leq m)$ in einem Schritt des einfachen Gradientenverfahrens (Algorithmus 5.1.1) ohne Vorkonditionierer für den Vektor $\mathbf{1}$, der in allen inneren Knoten des Gitters gleich 1 ist. Das machen wir sukzessiv in aufsteigender Reihenfolge von q. Dabei wird für $q > 1$ diese Iteration im zu den Vektoren $\tilde{u}_0^{(1)}, \ldots, \tilde{u}_0^{(q-1)}$ **B**-orthogonalen Untervektorraum angewendet.

Als erstes Beispiel berechnen wir 6 Eigenvektoren von (5.3.5–5.3.6) für $n = 127$. Die 6 ersten Eigenwerte dieses Problems sind 19.73822, 49.33960, 49.33960, 78.94098, 98.65542, 98.65542. Wir weisen darauf hin, dass der 2. und 3. sowie 5. und 6. Eigenwert in diesem Fall zweifach sind. Wir wenden nun das oben beschriebene Verfahren mit 7 Testvektoren an (siehe (5.3.1–5.3.3)). Die Ergebnisse dieses Iterationsprozesses sind in Abbildung 5.1 (a) dargestellt. Die Kurven zeigen die Norm $\|\mathbf{A}\tilde{u}_k^{(q)} - \lambda(\tilde{u}_k^{(q)})\mathbf{B}\tilde{u}_k^{(q)}\|_2$ des Residuums als Funktion der Schrittzahl. Das Verfahren zeigt im Mittel lineare Konvergenz. Die Sprünge in den Kurven zwischen Schrittzahl 20 und 30 entsprechen dem Entstehen von den zweiten Eigenvektoren bei den zweifachen Eigenwerten. Vor und nach diesen Stellen approximieren die Näherungen $\tilde{u}_k^{(q)}$ verschiedene Eigenvektoren, deren Residuen miteinander nicht verbunden sind.

Hier und in anderen Experimenten führen wir so viele Schritte aus, bis die

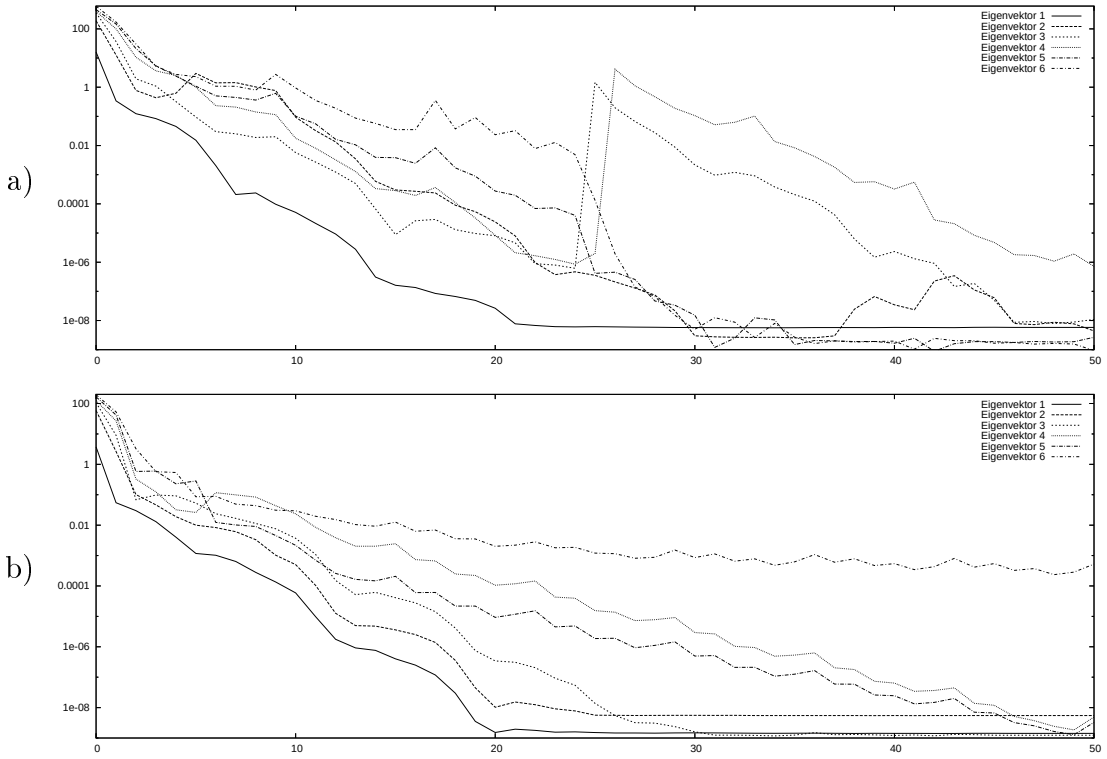

Abbildung 5.2: Die Konvergenz des Block-Gradientenverfahren mit mehreren Vorkonditionierern für (a) das Problem (5.3.5–5.3.6) bei $n = 255$ und (b) das Problem (5.3.8–5.3.9) bei $n = 127$. Die Kurven zeigen die Norm $\|\mathbf{A}u - \lambda(u)\mathbf{B}u\|_2$ als Funktion der Schrittnummer.

euklidischen Normen der Residuen zu *allen* Eigenvektoren höchstens 10^{-6} sind:

$$\|\mathbf{A}\tilde{u}_k^{(q)} - \lambda(\tilde{u}_k^{(q)})\mathbf{B}\tilde{u}_k^{(q)}\|_2 \leq 10^{-6}, \quad 1 \leq q \leq m. \tag{5.3.7}$$

Auf AMD Athlon 1 GHz benötigen wir 47 Schritte, was zusammen mit dem Präprozess (Gitterzerlegung und Berechnung der GIBLU(1)-Koeffizienten) 153 Sekunden gedauert hat. Wir weisen darauf hin, dass die einzelnen Vektorfolgen viel schneller die geforderte Genauigkeit erreicht haben.

Zum Vergleich stellen wir in Abbildung 5.1 (b) die Ergebnisse des Block-Gradientenverfahrens für nur einen der im vorherigen Beispiel benutzten Vorkonditionierern (mit $\mathfrak{i} = 3$ — siehe (5.3.1)) dar. Dieses Verfahren hat unter gleichen Bedingungen 49 Schritte und 192 Sekunden benötigt. Obwohl die Anzahl der Schritte bei Bedingung (5.3.7) nicht viel größer ist, sind die Konvergenzraten für die meisten Vektorfolgen wesentlich schlechter.

Bemerkung 5.3.2 Bei der Ausführung von Algorithmus 5.2.1 betrachten wir in den hier vorgestellten Experimenten die Residuen, deren euklidische Norm kleiner gleich 10^{-8} ist, aus Stabilitätsgründen als Null. Deswegen wird durch das Verfahren für manche Näherungswerte, die diese Genauigkeit schon erreicht haben, das Residuum nicht weiter verkleinert. (Siehe z.B. Abbildung 5.1 (a)). Es kann aber

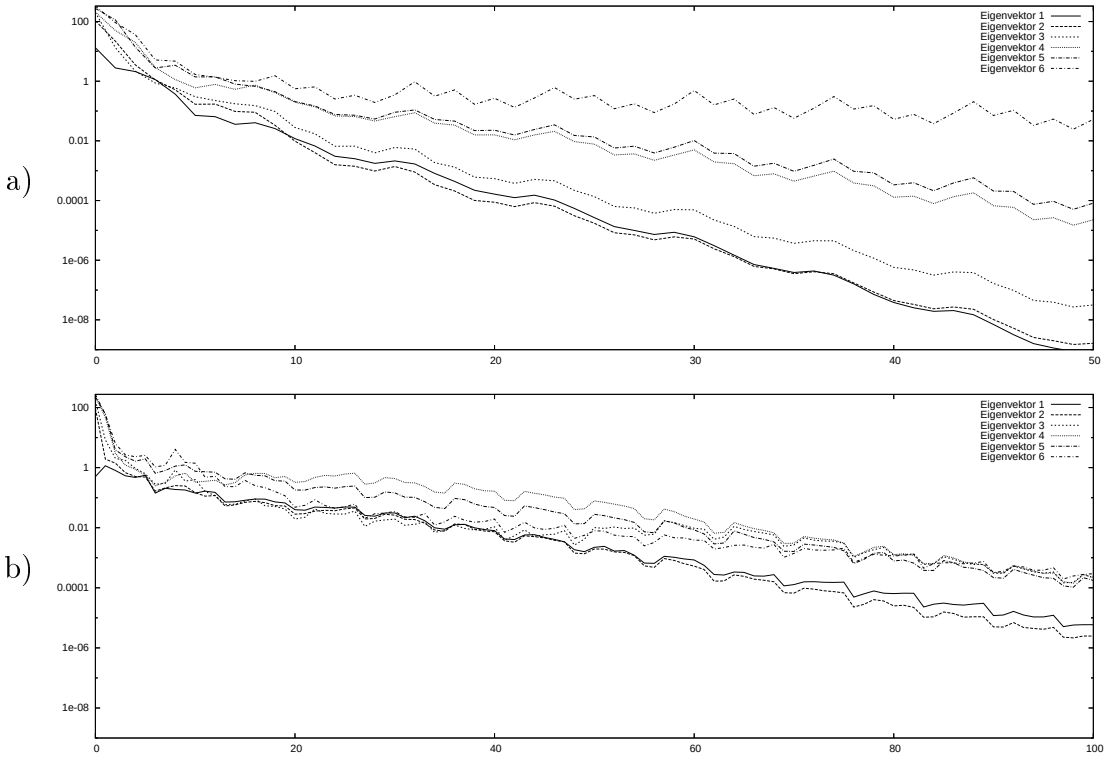

Abbildung 5.3: Die Konvergenz des Block-Gradientenverfahren mit 7 Testvektoren für den Laplace-Operator auf dem L-förmigen Gebiet mit (a) Dirichlet-Randbedingungen ($\max_j n_j = 127$) und (b) Neumann-Randbedingungen ($\max_j n_j = 129$).

vorkommen, dass die Fehler durch die anderen Näherungswerte und deren Residuen weiter gedämpft werden. Dann wird sogar eine höhere Genauigkeit erreicht. □

In Abbildung 5.2 (a) sind die Ergebnisse für das Problem (5.3.5–5.3.6) für $n = 255$ dargestellt. Die Eigenwerte sind dann 19.73896, 49.34592, 49.34592, 78.95287, 98.68589, 98.68589. Wir berechnen wieder 6 Eigenvektoren, benutzen aber 8 Testvektoren. Wie man im Vergleich mit Abbildung 5.1 (a) sieht, ist die Konvergenz in etwa so wie bei Problem (5.3.5–5.3.6) für $n = 127$.

Als nächstes Beispiel betrachten wir das Problem

$$\begin{aligned} -\nabla \cdot (P(x,y)\nabla u) &= \lambda u, \qquad (x,y) \in \Omega, \\ u|_{\partial\Omega} &= 0 \end{aligned} \tag{5.3.8}$$

mit

$$P(x,y) = 1 - \exp(-xy) \tag{5.3.9}$$

auf dem Einheitsquadrat $\Omega = (0,1)^2$. Wie im letzten Beispiel wird die Gleichung (5.3.8) auf dem strukturierten Dreiecksgitter (siehe Abbildung 1.1) mit dem Finite-Volumen-Verfahren diskretisiert. In Abbildung 5.2 (b) stellen wir die Konvergenz der beschriebenen Variante des Block-Gradientenverfahren für die Matrixblockgröße

$n = 127$ mit 7 Testvektoren (5.3.1) dar. Alle Eigenwerte sind in diesem Fall einfach: 0.9084394, 1.407032, 2.058528, 2.891445, 3.115707, 3.776564. Außer der letzten konvergieren alle Vektorfolgen gegen Testvektoren nicht schlechter, als in den vorherigen Beispielen. Zur Berechnung der letzten Näherung empfiehlt es sich, diese separat vorzunehmen, nachdem die anderen Vektoren schon mit hinreichender Genauigkeit berechnet worden sind.

Wir stellen noch ein Beispiel mit einem komplizierteren Gebiet vor. Wir betrachten das Problem (5.3.4) auf dem L-förmigen Gebiet aus Abbildung 4.3. Das kontinuierliche Gleichung (5.3.4) wird auf dem strukturierten Gitter mit Finite-Volumen-Verfahren diskretisiert. Die Zerlegung des Gitters erfolgt wie in Abschnitt 4.3. Die maximale Anzahl der Knoten in einem Block ist in diesem Experiment 127. Wie oben, berechnen wir 6 Eigenvektoren auf einmal. Die Eigenwerte des diskreten Problems sind in diesem Fall 38.58809, 60.77660, 78.94098, 118.0410, 127.6799 und 165.8494. Wir verwenden hier das oben beschriebene vorkonditionierte Block-Gradientenverfahren mit 7 Testvektoren (5.3.1–5.3.2). Die Konvergenzergebnisse in den ersten 50 Schritten sind in Abbildung 5.3 (a) dargestellt.

Wir betrachten nun das gleiche Beispiel mit Neumann-Randbedingungen. Das Problem ist dann

$$\begin{aligned} -\Delta u &= \lambda u, \\ \tfrac{\partial u}{\partial \vec{n}}\big|_{\partial\Omega} &= 0, \end{aligned} \tag{5.3.10}$$

wobei $\vec{n}$ die äußere Normale und Ω das Gebiet aus Abbildung 4.3 ist. Dieses Problem wird auf dem gleichen Dreiecksgitter wie im vorherigen Beispiel mit dem Finite-Volumen-Verfahren diskretisiert. Da die Randknoten in diesem Fall auch die Freiheitsgrade enthalten, ist die maximale Blockgröße in diesem Beispiel 129. Die ersten Eigenwerte des diskreten Problems sind 0, 5.907671, 14.13476, 39.46835, 39.47263, 45.54558, 50.30187. Der Eigenwert 0 entspricht dem konstanten Eigenvektor $\mathbf{1}$. Wir berechnen die 6 weiteren Eigenvektoren im zu $\mathbf{1}$ $\mathbf{B}$-orthogonalen Raum mit dem oben beschriebenen Verfahren für 7 Testvektoren $\mathbf{e}(\mathbf{i}) = \text{blockvector}\,\{e_j(\mathbf{i})\}$, wobei

$$\left(e_j(\mathbf{i})\right)_i = \cos\frac{\pi i \cdot \min\{\mathbf{i}, n_j - 1\}}{n_j - 1}. \tag{5.3.11}$$

Die Indizes $\mathbf{i}_k$ werden nach Regel (5.3.2) gewählt. Die Anfangswerte $\tilde{u}_0^{(q)}$ berechnen wir hier auch in einem Schritt von Algorithmus 5.1.1 ohne Vorkonditionierer, aber für den Vektor $\mathbf{X}$, der im Gitterpunkt (x, y) gleich x ist. Die Ergebnisse stellen wir in Abbildung 5.3 (b) dar. Wir weisen darauf hin, dass mit den Testvektoren (5.3.1) dieses Verfahren eine ähnliche Konvergenzrate wie mit (5.3.11) zeigt.

Bemerkung 5.3.3 In dieser Arbeit haben wir die GIBLU(1)-Zerlegungen nur für Probleme mit Dirichlet-Randbedingungen betrachtet. Es stellt aber kein Problem dar, diese Verfahren mit Neumann-Randbedingungen zu benutzen. Die Diagonalblöcke der Zerlegungen und die entsprechenden Vorkonditionierer sind in diesem Fall regulär. □

Wir fassen nun die Ergebnisse zusammen. Eine Folge von GIBLU(1)-Zerlegungen kann effizient zum Vorkonditionieren vom Block-Gradientenverfahren

verwendet werden. Der Aufwand der Zerlegung des Gitters in Blöcke und der Berechnung der Koeffizienten ist hier im Vergleich mit dem des gesamten Prozesses klein. In dieser Arbeit wurde das Verfahren nur experimentell geprüft. Es bedarf einer quantitativen analytischen Untersuchung sowie einer weiteren Verbesserung im Fall komplizierter Gebiete und unstrukturierter Gitter.

Resümee und Ausblick

In dieser Arbeit stellen wir eine neue Klasse der Löser für die großen schwachbesetzten linearen Gleichungssysteme vor, die GIBLU(l)-Zerlegungen. Diese Verfahren beruhen auf der Approximation der Diagonalblöcke der vollständigen Block-LU-Zerlegung mit Kettenbrüchen und besitzen in gewissem Sinn die Eigenschaft der schwachen Frequenzfilterung. Sie können als Grobgitterlöser in geometrischen Mehrgitterverfahren sowie auch als Vorkonditionierer für das Block-Gradientenverfahren bei Eigenwertproblemen verwendet werden.

Die Konvergenz dieser Verfahren wird im Fall des Modellproblems analytisch untersucht. Insbesondere wird es gezeigt, dass das mit den GIBLU(1)- und GIBLU(2)-Zerlegungen vorkonditionierte CG-Verfahren im Fall der Fünf-Punkt-Diskretisierung des Laplace-Operators auf dem Einheitsquadrat die Konvergenzordnungen $\frac{1}{3}$ und $\frac{1}{4}$, respektive, haben. Da aber die Konstruktion der GIBLU(2)-Zerlegung komplizierter ist, ist der Aufwand für beide Zerlegungen ungefähr gleich.

Wir haben die vorgestellten Zerlegungen noch für eine größere Klasse von Problemen verallgemeinert. Die Zerlegungsparameter ersetzen wir in diesem Fall durch Testvektoren, sodass die Verfahren besser an Probleme mit variierenden Koeffizienten angepasst sind. Für die GIBLU(1)-Zerlegungen entwickeln wir die Variante mit nur einem Testvektor. Die Existenz der in dieser Zerlegung auftretenden Koeffizienten beweisen wir unter sehr allgemeinen Bedingungen. Wir stellen auch eine mögliche Vorgehensweise zur Erzeugung der zulässigen blocktridiagonalen Struktur der Matrix im Fall der unstrukturierten Gitter vor.

Die GIBLU(1)- und GIBLU(2)-Zerlegungen wurden mit Hilfe des Simulationssoftware-Pakets UG implementiert. In unseren numerischen Experimenten erhalten wir für die GIBLU(1)-Zerlegung auch für komplizierte Gebiete mit einem strukturierten Gitter gute Konvergenzeigenschaften.

Des Weiteren wenden wir Folgen von GIBLU(1)-Zerlegungen als Vorkonditionierer von Block-Gradientenverfahren auf Eigenwertprobleme an. Unsere numerischen Experimente zeigen eine hohe Effizienz dieser Methode.

Schließlich formulieren wir noch einige in dieser Arbeit nicht beantwortete interessante Fragen zu den GIBLU(l)-Zerlegungen, die als Ausgangspunkt für weitere Forschung dienen können. Die erste bezieht sich auf die Darstellung der Matrix in der zulässigen blocktridiagonalen Form. Unsere numerischen Experimente zeigen, dass diese Algorithmen für die großen Gitter einen sehr hohen Aufwand haben können. Einer weiteren Entwicklung bedarf auch die Behandlung der mit den Diagonalblöcken assoziierten linearen Probleme. Bei völlig unstrukturierten Gittern ist die Konstruktion der Zerlegung und die Berechnung ihrer Koeffizienten noch weitergehend zu un-

tersuchen, da in diesen Fällen die experimentell beobachtete Konvergenzrate stark von der theoretisch erwarteten abweichen kann.

Des Weiteren stellen sich folgende theoretische Fragen. Es gibt noch keine Vorgehensweise zur Untersuchung der Existenz und den Konvergenzeigenschaften der Betrachteten Zerlegungen im allgemeinen Fall. Auch die quantitativen Abschätzungen der Konvergenzeigenschaften des mit den Folgen der frequenzfilternden Zerlegungen vorkonditionierten Block-Gradientenverfahren müssen noch gefunden werden.

Anhang A

Simulationssoftware-Paket UG

Das Paket UG wurde seit 1990 in der Arbeitsgruppe Prof. Dr. G. Wittum an der Universität Stuttgart und (später) an der Universität Heidelberg zur numerischen Simulation in verschiedenen wissenschaftlichen und technischen Bereichen entwickelt. Dieses Programm bietet ein flexibles Werkzeug zur Implementierung und Kopplung von verschiedenen Verfahren in allen Schritten der Simulation: Darstellung des Gebiets, Gittergenerierung, Diskretisierung der Modellgleichnungen, Lösung der daraus entstandenen Systeme und weiterer Bearbeitung oder Ausgabe der Ergebnisse. Simulationen sind für zwei- und dreidimensionalen Gebiete möglich. Die Benutzung von Parallelrechnern wird dabei in besonderen Maße unterstützt. (Siehe [5].) UG wurde erfolgreich in Bereichen Strömungsmechanik, Strukturmechanik, poröse Medien, Parameterschätzung für Bingham'sche Fluide u.s.w. angewandt. In diesem Anhang beschreiben wir kurz die Struktur dieses Software-Pakets und die für die GIBLU(l)-Zerlegungen implementierten Module.

A.1 Struktur des Pakets UG

Die Struktur des Pakets UG ist schematisch auf Abbildung A.1 dargestellt. Die zentrale Rolle spielt hier der Kern. Er enthält u.a. die elementaren Funktionen zur Behandlung von Standarddatenstrukturen, die Schnittstellen für die Beschreibung des Gebiets und der Randbedingungen, die Mechanismen zur Gitterbeschreibung und Steuerung der numerischen Verfahren, die graphische Ausgabe, das Benutzerinterface. Dazu gehören auch eine umfangreiche Bibliothek von verschiedenen Standardlösern (die linearen Löser inklusive Mehrgitterverfahren; das Newton-Verfahren; die Zeitschrittiterationen für instationäre Probleme u.s.w.) und Hilfsmittel zur Implementierung von Finite-Volumen- und Finite-Elemente-Diskretisierungen. Die Implementierung von Diskretisierungen und problemspezifischen Verfahren gehört zur Problem-Klassen-Ebene. Auf deren Basis werden dann die Applikationen geschrieben, welche die Beschreibung des Gebiets und der Randbedingungen sowie auch andere problemspezifische Daten enthalten. Jede Applikation ist ein ausführbares Programm, das den gesamten UG-Kern enthält und bei Aufruf das UG-Benutzerinterface startet. Die weitere Steuerung dieses Programms erfolgt durch die UG-eigene Skriptsprache, mit deren Hilfe z.B. das Gitter definiert werden kann,

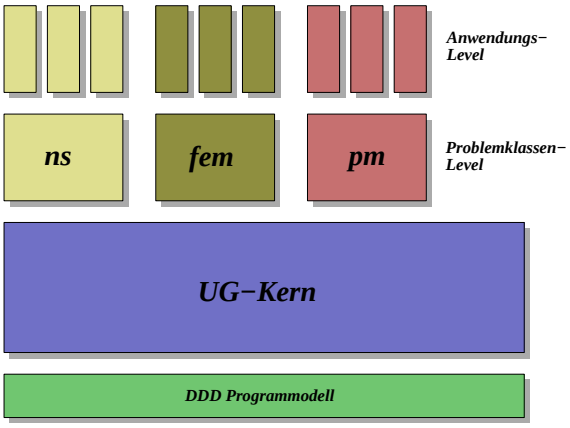

Abbildung A.1: Struktur des Pakets UG

sowie die numerischen Verfahren in einer zulässigen Ordnung verbunden und aufgerufen werden.

Zur Benutzung von parallelen Rechnerarchitekturen befindet sich unter dem Kern noch ein Mechanismus zur Datenverteilung. Da unsere Anwendungen nicht parallelisiert wurden, gehen wir hierauf nicht weiter ein.

Wir beschreiben nun kurz die für uns wichtigen Aspekte von UG. Diese sind

- die Datenstrukturen zur Darstellung des Gitters,
- die Beschreibung des Problemgebiets und der Randbedingungen,
- der Mechanismus zur Steuerung der numerischen Verfahren
- und die Skriptsprache.

Wir erläutern hier nur die wichtigsten Ideen und Begriffe. Ausführlichere Informationen findet man im in der UG-Distribution enthaltenen Tutorial.

Die Datenstrukturen zur Darstellung des Gitters sind in UG für 2- und 3D Rechnungen gleich. Die Gitter bestehen aus *geometrischen Elementen* mit gemeinsamen Knoten. In 2D können diese Elemente Dreiecke oder Vierecke sein, in 3D Tetraeder, Pyramiden mit viereckiger Basis, Prismen und Hexaeder. Dabei wird eine Hierarchie von Gittern gespeichert: Das am Anfang angelegte Gitter entspricht dem *Gitterlevel* 0, und bei dessen sukzessiver Verfeinerung werden weitere Gitterlevels erzeugt. Diese Hierarchie kann dann in Mehrgitterverfahren und geschachtelten Iterationen benutzt werden.

Die Elemente und deren Knoten bilden die geometrische Struktur des Gitters. Für jeden Gitterlevel werden Listen von Elementen und Knoten angelegt. Die Freiheitsgrade (die Werte der Gitterfunktionen) können den Knoten sowie den Elementen, ihren Kanten und Seiten zugeordnet werden. Die Anzahl der Freiheitsgrade pro Knoten und Element wird in einem Format-Befehl angegeben. Unabhängig davon wird für jeden Gitterlevel eine Liste von *allen* Freiheitsgraden angelegt. Diese beschreibt die algebraische Struktur des Gitters.

Für die Darstellung des Problemgebiets gibt es in UG zwei Möglichkeiten. Je nach der Konfiguration kann entweder die Standard- oder LGM-Schnittstelle benutzt werden. In der Standardschnittstelle werden die Randsegmente als Funktionen im Code angegeben. Dies ermöglicht eine einfache und genaue Beschreibung in vielen Fällen, wenn das Gebiet durch wenige einfache geometrischen Formen beschreibbar ist. Der Nachteil dieses Verfahren ist, dass jede Änderung des Gebiets eine Übersetzung der gesamten Applikation verlangt. Bei der LGM-Schnittstelle werden das Gebiet und (manchmal) das Anfangsgitter in separaten Files vom Benutzer gespeichert und erst im Programmverlauf von der Applikation gelesen. Dies erlaubt die Beschreibung von sehr komplizierten, für technische Anwendungen relevanten Geometrien. Die Randsegmente werden dann mit stückweise linearen Kurven, bzw. Flächen approximiert. Die Schnittstellen selber sind im Wesentlichen gleich, und viele Diskretisierungen können ohne weitere Abänderungen mit beiden kombiniert werden.

Die Randbedingungen müssen als Funktionen im Code dargestellt werden. Das stellt aber grundsätzlich keine Einschränkungen dar, da sie so implementiert werden können, dass sie entweder über das Skript steuerbar sind oder aus externen Files die Information lesen können.

Ein wichtiger Begriff in UG ist der der *Numproc* („Numerical Procedure“). Dabei handelt es sich um eine Klasse von Objekten, die durch Kommandos in der Skriptsprache erzeugt, initialisiert und dann, je nach dem Zweck, ausgeführt oder als Argument an andere Numprocs übergeben werden können. Im Hauptspeicher existiert ein solches Objekt als eine von `NP_BASE` ausgeleitete Struktur, die auch weitere Daten enthalten kann, wie Zeiger an Verfahren-spezifische Funktionen. Von der Basisklasse `NP_BASE` werden z.B. alle Diskretisierungs- und Lösungsmethoden abgeleitet. Die Ausleitung kann geschachtelt erfolgen, wodurch eine große Bibliothek von Basisklassen erstellt werden kann.

Als Beispiel betrachten wir hier die Implementierung des CG-Verfahrens. Dieses wird von der Klasse `NP_LINEAR_SOLVER` abgeleitet, die selbst von `NP_BASE` abgeleitet ist. Dieses Verfahren benötigt einen Vorkonditionierer der eine Numproc der Klasse `NP_ITER` sein soll. Dieser Vorkonditionierer soll dann im Initialisierungsbefehl im Shell-Script durch das Argument `$I` dem CG-Verfahren übergeben werden. (Zur Syntax von Kommandos s. weiter unten.) Wenn bei der Initialisierung noch die Systemmatrix und die rechte Seite spezifiziert werden, sowie weitere Parameter wie z.B. die Genauigkeit u.s.w., kann man dieses Objekt von CG-Verfahren direkt im Skript aufrufen. Es kann aber auch von anderen Verfahren benutzt werden, z.B. innerhalb der Newton-Iteration als linearer Löser. Die Objekte der Klasse `NP_ITER` — SSOR, ILU_ω, die Mehrgitterverfahren u.s.w. — können nicht ausgeführt werden. Da sie aber alle die gleiche Schnittstelle haben, können sie jeweils z.B. als Vorkonditionierer für das CG-Verfahren benutzt, also über dieses aufgerufen werden. Auch der Benutzer kann auf der Problem-Klassen-Ebene oder in der Applikation ein neues Verfahren dazu implementieren. Diese Vorkonditionierer können weitere Numprocs als Argumente haben: Die Mehrgitterverfahren brauchen Vor- und Nachglätter (die auch Objekte von `NP_ITER` sein müssen), einen Grobgitterlöser und eine Numproc für Prolongation/Restriktion.

Da die Numprocs über das Shell-Skript gesteuert werden, spielt in UG die Skriptsprache eine wichtige Rolle. Jedes UG-Skript besteht aus Operatoren und Kommandos. Es gibt eine im UG-Kern vordefinierte Menge von Standardoperatoren, welche zu solchen in vielen anderen algorithmischen Sprachen analog sind. Da UG auch Benutzer-definierte Variablen unterstützt, können dadurch Schleifen u.s.w. organisiert werden. Die Definition von neuen Kommandos ist auch auf der Problem-Klassen-Ebene oder in einer Applikation möglich. Der Aufruf eines Kommando erfordert dessen Namen und eine Liste von Argumenten, die mit dem Zeichen „$“ voneinander getrennt werden. Die Objekte von Numprocs werden mit dem Kommando „`npcreate`“ angelegt:

`npcreate` *<Objektname>* `$c` *<Klassenname>*;

Die Initialisierung erfolgt dann mit dem Kommando „`npinit`“:

`npinit` *<Objektname>* *<klassenabhängige Liste von Argumenten>*;

Mit dem Kommando „`npdisplay`“, das nur den Objektnamen als Argument benötigt, kann man die initialisierten Parameter eines Numproc-Objekts auflisten. Der Aufruf von ausführbaren Numproc-Objekten erfolgt durch das Kommando „`npexecute`“, dessen Syntax analog zu der von „`npinit`“ ist.

A.2 Implementierung von GIBLU(l)-Zerlegungen

In der Implementierung von GIBLU(l)-Zerlegungen gibt es zwei wichtige Klassen von Numprocs: die Blockgeneratoren (d.h. die Verfahren zur Zerlegung des Gitters in Blöcke) werden von `NP_BLOCK_GENERATOR` abgeleitet, und die auf dieser Blockstruktur operierenden Löser von `NP_BLOCK_LINEAR_SOLVER`. Die Klasse `NP_BLOCK_LINEAR_SOLVER` ist von `NP_ITER` abgeleitet. Also können diese Blocklöser standardweise als Vorkonditionierer (oder Glätter) angewandt werden.

Bei der Initialisierung sollen die Objekte dieser Klassen miteinander beidseitig verbunden werden. Man muss mindestens einen Blockgenerator anlegen. Damit kann man beliebig viele Blocklöser koppeln. Vor der Benutzung der Löser auf einem Gitter, muss für dieses der Blockgenerator aufgerufen werden. Im Skript schreibt sich dies wie folgt:

`npcreate` *<Blockgenerator>* `$c` *<Blockgeneratorsklasse>*;
`npcreate` *<Löser 0>* `$c` *<Blocklösersklasse 0>*;
`npcreate` *<Löser 1>* `$c` *<Blocklösersklasse 1>*;
...
`npcreate` *<Löser $n-1$>* `$c` *<Blocklösersklasse $n-1$>*;
`npinit` *<Blockgenerator>*
  `$lev <Gitterlevel zum Zerlegen>`
  `$n_solvers` n
  `$solver0` *<Löser 0>* ...`$solver<`$n-1$`>` *<Löser $n-1$>*
  <Blockgenerator-spezifische Parameter>;
...

```
npinit <Löser i> # für jedes i ∈ {0,...,n-1}
  $block <Blockgenerator>
  <Löser-spezifische Parameter>;
...
npexecute <Blockgenerator> $create;
```

Bei der Initialisierung eines Blockgenerators müssen die Löser nur angegeben werden, damit ihm die Größe der Datenstrukturen bekannt ist. Wenn alle Löser zur selben Blocklöserklasse gehören, reicht es, wenn nur einen von ihnen anzugeben. Dabei kann statt „`$n_solvers 1 $solver0` *<Löser>*“ die Konstruktion „`$solver` *<Löser>*“ benutzt werden.

Zum Zerlegen des Gitters mit Geraden (4.2.2) (siehe Abschnitt 4.2) für die GIBLU(l)-Zerlegungen mit $l \geq 1$ kann man die Numproc-Klasse `gdlnblockgen` anwenden. Als zusätzliche Parameter werden dann die Ordnung l (Parameter `$order`), die Koeffizienten a und b (Parameter `$A` und `$B`, respektive) und die erste rechte Seite s_0 (Parameter `$start_step`) übergeben. ($s_0 = 0$ entspricht dem kleinsten s, bei dem die Gerade ϕ_s gemeinsame Punkte mit dem Gitter hat.)

Die GIBLU(1)- und GIBLU(2)-Zerlegungen werden als Numproc-Klassen `gdIBLU_1` und `gdIBLU_2` implementiert. Sie unterscheiden sich nur in der Spezifikation von Zerlegungsparameter $\hat{\mu}_i$ oder Testvektoren. Für `gdIBLU_1` gibt es vier Möglichkeiten. Die Angaben „`$mu` $\hat{\mu}$“ und „`$mu0` $\hat{\mu}_0$ `$mu1` $\hat{\mu}_1$“ starten die Berechnung der Koeffizienten $\theta_k^{(1)}$ und $\theta_k^{(0)}$ nach den Formeln (3.2.6), bzw. (3.2.5). Im zweiten Fall müssen $\hat{\mu}_0$ und $\hat{\mu}_1$ ungleich sein. Bei der Kombination „`$e0` $\mathbf{e}^{(0)}$ `$e1` $\mathbf{e}^{(1)}$ `$tridiag`“ wird Algorithmus 4.1.4 dazu benutzt. Die Spezifikation „`$e` $\mathbf{e}^{(0)}$“ bedeutet die Verwendung von Algorithmus 4.1.7.

Bei `gdIBLU_2` muss entweder „`$mu0` $\hat{\mu}_0$ `$mu1` $\hat{\mu}_1$ `$mu2` $\hat{\mu}_2$“ oder „`$e0` $\mathbf{e}^{(0)}$ `$e1` $\mathbf{e}^{(1)}$ `$e2` $\mathbf{e}^{(2)}$“ angegeben werden. Dann werden Formeln (3.2.23–3.2.26) bzw. Algorithmus 4.1.9 verwendet. Der Fall $\hat{\mu}_0 = \hat{\mu}_1$ ist hier mitbetrachtet, aber $\hat{\mu}_1$ und $\hat{\mu}_2$ müssen ungleich sein.

Die Berechnung der Koeffizienten $\theta_k^{(i)}$ geschieht erst auf das Kommando

```
npexecute <Objektenname> $IBLU $A <Systemmatrix>;
```

Dieses wird nach der Assemblierung der Systemmatrix aufgerufen.

Nach der Anwendung der hier beschriebenen Numprocs sollte der von ihnen allozierte Speicherplatz wieder freigegeben werden. Bei `gdIBLU_1` und `gdIBLU_2` erfolgt dies durch Aufruf des Kommandos `npexecute` mit der Option `$remove`. Analog soll für die Blockgeneratoren das Kommando „`npexecute` *<Blockgenerator>* `$delete;`“ ausgeführt werden.

Literaturverzeichnis

[1] Axelsson, O. und B. Polman: *A robust preconditioner based on algebraic subsructuring and two-lewel grids.* In: Hackbusch, W. (Herausgeber): *Robust multigrid methods*, Seiten 1–26. Vieweg-Verlag, Braunschweig, 1989.

[2] Babuška, I. und J. Osborn: *Eigenvalue Problems.* In: Ciarlet, P. G. und J. L. Lions (Herausgeber): *Handbook of Numerical Analysis*, Band Volume II: Finite Element Methods (Part 1), Seiten 642–787. Elsevier Science Publishers B. V. (North-Holland), Amsterdam, London, New-York, Tokio, 1991.

[3] Baker jr., G. A.: *Essentials of Padé Approximants.* Academic Press, New York, 1975.

[4] Barrett, R., M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine und H. van der Vorst: *Templates for the solution of linear systems: building blocks for iterative methods.* SIAM, Philadelphia, PA, 1993.

[5] Bastian, P., K. Birken, K. Johannsen, S. Lang, N. Neuss, H. Rentz-Reichert und C. Wieners: *UG — A Flexible Software Toolbox for Solving Partial Differential Equations.* Computing and Visualization in Science, 1:27–40, 1997.

[6] Bey, J.: *Finite-Volumen- und Mehrgitterverfahren für elliptische Randwertprobleme.* Advances in Numerical Mathematics. Teubner-Verlag, Stuttgart, 1998.

[7] Bramble, J. H., A. V. Knyazev und J. E. Pasciak: *A Subspace Preconditioning Algorithm for Eigenvector/Eigenvalue Computations.* Adv. Comput. Math., 6:159–189, 1996.

[8] Brezzi, F. und M. Fortin: *Mixed and Hybrid Finite Element Methods.* Springer Series in Computational Mathematics. 15. Springer-Verlag, New York, 1991.

[9] Buzdin, A.: *Tangential Decompositions.* Computing, 61:257–276, 1998.

[10] Buzdin, A. und D. Logashenko: *Incomplete Block Triangular Decompositions of Order l.* Computing, 64:69–95, 2000.

[11] Buzdin, A. und G. Wittum: *Two-Frequency Decompositions.* Numer. Math., 97(2):269–295, 2004.

[12] Elsner, L. und V. Mehrmann: *Convergence of Block Iterative Methods for Linear Systems Arising in the Numerical Solution of Euler Equations.* Numer. Math., 59:541–559, 1991.

[13] Gander, M. J. und F. Nataf: *AILU: A Preconditioner Based on the Analytic Factorization of the Elliptic Operator.* Numerical Linear Algebra with Applications, 7(7–8):505–526, 2000.

[14] Geiser, J.: *Untersuchung der Schwingungen am Resonanzboden eines Cembalos: Reduktion auf 2 Raumdimensionen.* Diplomarbeit, Fakultät Mathematik der Universität Stuttgart, Oktober 1998.

[15] Golub, G. und C. F. van Loan: *Matrix computations. 3rd ed.* The Johns Hopkins Univ. Press, Baltimore, MD, 1996.

[16] Grossmann, Ch. und H.-G. Roos: *Numerik partieller Differentialgleichungen.* Teubner, Stuttgart, 1994.

[17] Hackbusch, W.: *Multigrid Methods and Applications.* Springer, New-York, 1985.

[18] Hackbusch, W.: *Iterative Solution of Large Sparse Systems of Equations.* Springer, New-York, 1994.

[19] Hackbusch, W.: *Integralgleichungen. Theorie und Numerik.* Leitfäden der Angewandten Mathematik und Mechanik (LAMM). 68. B. G. Teubner, Stuttgart, 1997.

[20] Il'in, V. P.: *Iterative Incomplete Factorization Methods.* Series on Soviet and East European Mathematics. 4. World Scientific, Singapore, 1992. Die russische ergänzte Variante: [49].

[21] Kettler, R.: *Analysis and comparison of relaxation schemes in robust multigrid and preconditioned conjugate gradient methods.* In: Hackbusch, W. und U. Trottenberg (Herausgeber): *Multigrid methods, Proceedings, Köln-Porz, 1981*, Seiten 502–534, Heidelberg, New York, Tokyo, 1981. Springer.

[22] Knyazev, A. V. und K. Neymeyr: *A geometric theory for preconditioned inverse iteration. III: A short and sharp convergence estimate for generalized eigenvalue problems.* Linear Algebra Appl., 358:95–114, 2003.

[23] Lang, S.: *Algebra. Third Edition.* Addison-Wesley Publishing Company, Massachusetts, 1993.

[24] Morton, K. W. und D. F. Mayers: *Numerical Solution of Partial Differential Equations.* Cambridge University Press, Cambridge, 1994.

[25] Neymeyr, K.: *A geometric theory for preconditioned inverse iteration, I: Extrema of the Rayleigh quotient.* Linear Algebra Appl., 332:61–85, 2001.

[26] NEYMEYR, K.: *A geometric theory for preconditioned inverse iteration, II: Convergence estimates.* Linear Algebra Appl., 332:87–104, 2001.

[27] NOCEDAL, J. und S. J. WRIGHT: *Numerical Optimization.* Springer-Verlag, New-York, 1999.

[28] PARLETT, B. N.: *The Symmetric Eigenvalue Problem.* Classics In Applied Mathematics, 20. SIAM, 1998.

[29] RUBINSTEIN, I. und L. RUBINSTEIN: *Partial Differential Equations in Classical Mathematical Physics.* Cambridge University Press, 1998.

[30] SAMARSKII, A. A. und E. S. NIKOLAEV: *Numerical Methods for Grid Equations*, Band 1: Direct Methods. Birkhäuser, Basel, 1989.

[31] SAUTER, S. und G. WITTUM: *A multigrid method for the computation of eigenmodes of closed water basins.* IMPACT Comput. Sci. Eng., 4, No. 2:124–152, 1992.

[32] SOMMERFELD, A.: *Elektrodynamik.* Vorlesungen über Theoretische Physik. Band III. Verlag Harri Deutsch, Thun, Frankfurt am Main, 1988.

[33] STOER, J. und R. BULIRSCH: *Introduction to numerical analysis.* Springer-Verlag, New York, 1992.

[34] STRACK, O. D. L.: *Groundwater Mechanics.* Prentice-Hall, Englewood Cliffs, 1989.

[35] TROTTENBERG, U., C. W. OOSTERLEE und A. SCHÜLLER: *Multigrid. With guest contributions by A. Brandt, P. Oswald, K. Stüben.* Academic Press, Orlando, FL, 2001.

[36] VAN DER WAERDEN, B. L.: *Algebra I.* Springer-Verlag, Berlin, 1993.

[37] VARGA, R. S.: *Matrix iterative analysis. 2nd revised and expanded edition.* Springer Series in Computational Mathematics. 27. Springer-Verlag, Berlin, 2000.

[38] VERFÜRTH, R.: *A review of a posteriori error estimation and adaptive mesh-refinement techniques.* B. G. Teubner, Stuttgart, 1996.

[39] WACHSPRESS, E. L.: *Iterative solution of elliptic systems and applications to the neutron diffusion equations of reactor physics.* Prentice-Hall International Series in Applied Mathematics. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1966.

[40] WAGNER, C.: *Ein robustes Mehrgitterverfahren für Diffusions-Transport-Probleme der Bodenphysik.* IWR-Report 93-70, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, Universität Heidelberg, 1993.

[41] Wagner, C.: *Frequenzfilternde Zerlegungen für unsymmetrische Matrizen und Matrizen mit stark variierenden Koeffizienten.* Doktorarbeit, Institut für Computeranwendungen der Universität Stuttgart, Juli 1995.

[42] Wagner, C.: *Tangential frequency filtering decompositions for symmetric matrices.* Numer. Math., 78:143–163, 1997.

[43] Wagner, C. und G. Wittum: *Adaptive Filtering.* Numer. Math., 78:305–328, 1997.

[44] Wesseling, P.: *An Introduction to Multigrid Methods.* Pure and Applied Mathematics. A Wiley-Interscience Series of Texts Monographs & Tracts. John Wiley & Sons Ltd., Chichester, 1992.

[45] Wilkinson, J. H.: *The Algebraic Eigenvalue Problem.* Oxford University Press, New York, 1965.

[46] Wittum, G.: *Distributive Iterationen für indefinite Systeme.* Preprint Nr. 454, SFB 123, Universität Heidelberg, 1988.

[47] Wittum, G.: *Filtende Zerlegungen — Schnelle Löser für große Gleichungssusteme.* Teubner Skripten zur Numerik. Band 1. Teubner, Stuttgart, 1992.

[48] Zulehner, W.: *Analysis of iterative methods for saddle point problems: A unified approach.* Math. Comput., 71(238):479–505, 2002.

[49] В. П. Ильин: *Методы неполной факторизации для решения алгебраических систем.* Физматлит, М., 1995. Die englische Variante: [20].

[50] А. В. Князев: *Вычисление собственных значений и векторов в сеточных задачах: алгоритмы и оценки погрешности.* Отдел вычислительной математики АН СССР, 1986.

[51] А. В. Колдоба, Ю. А. Повещенко, Е. А. Самарская, В. Ф. Тишкин: *Методы математического моделирования окружающей среды.* Наука, М., 2000.

[52] М. А. Красносельский, Г. М. Вайникко, П. П. Забрейко, Я. Б. Рутицкий, В. Я. Стеценко: *Приближенное решение операторных уравнений.* Наука, М., 1969.

[53] Б. А. Самокиш: *Метод наискорейшего спуска в задаче о собственных элементах полуограниченных операторов.* Известия высших учебных заведений, Математика, 5(6):105–114, 1958.

[54] А. Н. Тихонов, А. А. Самарский: *Уравнения математической физики.* Изд-во МГУ, М., 1999.