Algebraische Mehrgitterverfahren mit F-Glättung

vorgelegt von
Diplom-Mathematiker
Florian Goßler
aus München

Von der Fakultät II - Mathematik und Naturwissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
- Dr. rer. nat. -

genehmigte Dissertation

Vorsitzender:	Prof. Dr. Michael Scheutzow	(TU Berlin)
Gutachter:	Prof. Dr. Reinhard Nabben	(TU Berlin)
	Prof. Dr. Jörg Liesen	(TU Berlin)
	Prof. Dr. Andreas Frommer	(Bergische Universität Wuppertal)

Tag der wissenschaftlichen Aussprache: 5. März 2013

Berlin 2013

D 83

Danksagung

Mein besonderer Dank gilt Prof. Dr. Reinhard Nabben für die hervorragende Betreuung während der letzten fünf Jahre. Eine konstruktive und interessante Diskussion, wie aber auch ein kleiner Plausch bei einer Tasse Kaffee prägen die angenehme Arbeitsatmosphäre. Dazu trägt ebenfalls Olivier Sète, mein Büropartner, bei; dafür vielen Dank. Er sorgt nicht nur für den besten Kaffee in unserem Flur, sondern hat immer ein offenes Ohr für schwierige Fragen und auftretende Probleme.

Ebenfalls gilt mein Dank Prof. Dr. Jörg Liesen, meinem BMS-Mentor. Er zeigt mir stets durch kritische Fragen eventuelle Ungenauigkeiten in meinen Argumentationen und Arbeiten auf. Und damit möchte ich gleich einen Dank an die gesamte Arbeitsgruppe anschließen, die ebenfalls zu der hervorragenden Atmosphäre beisteuert.

Zu einer erfolgreichen Promotion gehören aber nicht nur Personen im direkten Umfeld. Im wissenschaftlichen Alltag findet stets ein reger Austausch statt. In diesem Sinn möchte ich insbesondere Dr. Karsten Kahl für die umfangreiche Unterstützung, Prof. Dr. Kees Vuik und seiner gesamten Arbeitsgruppe, die mich 7 Wochen als Gast in Delft willkommen hieß, Prof. Dr. Scott MacLachlan für die Unterstützung bei dem Greedy-Coarser und Prof. Dr. Andreas Frommer für die Einladung nach Wuppertal und die Bereitschaft, ein Gutachten in solch einer kurzen Zeit anzufertigen, bedanken. Auch vielen Dank an die Berlin Mathematical School, die mir nicht nur die Teilnahme an der Copper Mountain Konferenz ermöglichte, sondern auch meinen Forschungsaufenthalt in Delft finanzierte.

Aus dem persönlichen Umfeld möchte ich zunächst meinen Eltern danken. Ein ganz großer Dank geht an meine Verlobte, die mich beim Schreiben der Arbeit unterstützt hat. Insbesondere in den letzten Wochen hat sie dazu beigetragen, dass die vorliegende Arbeit ihre derzeitige Form und Gestalt erhielt. Sie wurde nicht müde, die Dissertation auch noch zum zehnten Mal zu lesen (was ich mir als Nichtmathmatiker äußerst schwierig vorstelle).

Daher nochmal an alle Erwähnten:
VIELEN DANK!

Inhaltsverzeichnis

Kapitel 1

Einleitung

In vielen Bereichen der Wissenschaft, Industrie und Wirtschaft treten lineare Gleichungssysteme bei den verschiedensten Anwendungen auf, sei es beim Lösen von partiellen Differentialgleichungen [62, 91], bei linearen Ausgleichsproblemen [15, 20], in der Bildbearbeitung [2, 96] oder bei der Simulation von stochastischen Prozessen [94]. Diese Gleichungssysteme besitzen die Gestalt $Ax = b$ mit einer gegebenen Matrix A, einer rechten Seite b und einem gesuchten Vektor x. Zu den Aufgaben der numerischen linearen Algebra gehört sowohl das Lösen solch eines linearen Gleichungssystems, also die Entwicklung schneller und effizienter Verfahren, als auch die theoretische Analyse der verschiedenen Lösungsstrategien.

Die linearen Gleichungssysteme aus den angesprochenen Anwendungen sind nicht selten mit über 1 Million Unbekannter und Gleichungen sehr groß. Des Weiteren ist die zu dem System assoziierte Matrix A in der Regel schlecht konditioniert und dünn besetzt, d.h. die meisten Einträge in der Matrix sind gleich Null. Obwohl die Systemmatrix A wenig Nichtnulleinträge im Vergleich zu ihrer Größe besitzt, ist ihre Inverse A^{-1}, sofern diese existiert, üblicherweise voll besetzt, was häufig zu einem hohen numerischen Aufwand bei der Anwendung sogenannter direkter Verfahren, wie z.B. dem Gauß-Eliminations-Verfahren führt. In dieser Arbeit wird vorausgesetzt, dass A invertierbar ist.

Durch die Struktur der dünnbesetzten Matrix A können sehr effiziente iterative Verfahren entwickelt werden, die die Lösung $x = A^{-1}b$ nicht exakt bestimmen, sie aber durch eine Folge $\{x^{[k]}\}$ zufriedenstellend approximieren. In der Regel basieren diese Art von Verfahren auf einer schnellen Berechnung eines Matrix-Vektor-Produkts $A \cdot r$, da nur wenige Einträge in der Matrix A ungleich Null sind. Zu den iterativen Verfahren gehört unter anderem die Klasse der stationären Verfahren, für die eine Approximation B^{-1} von A^{-1} definiert und die Folge durch $x^{[k+1]} = x^{[k]} + B^{-1}(b - Ax^{[k]})$ und einen Startvektor $x^{[0]}$ rekursiv gebildet wird. Der Vektor $r^{[k]} := b - Ax^{[k]}$ wird als das *k-te Residuum* und $T := I - B^{-1}A$ als *Iterationsmatrix* bezeichnet. Eine Möglichkeit, die Güte eines solchen iterativen Verfahrens zu messen, bietet der Spektralradius, also der betragsmäßig größte Eigenwert von T, $\rho(T)$, der die Gleichung $\lim_{k\to\infty} \|T^k\|^{\frac{1}{k}} = \rho(T)$ erfüllt und daher als *asymptotische Konvergenzgeschwindigkeit* betitelt wird, siehe z.B. [40, Satz 1.8.9]. Außerdem gilt für den Fehler $e^{[k]} = x - x^{[k]}$, dass $e^{[k+1]} = e^{[k]} - B^{-1}Ae^{[k]} = Te^{[k]}$ ist, und daher der gewichtete Fehler (gemessen in einer geeigneten Norm $\|\cdot\|$), also $\|e^{[k]}\|/\|e^{[0]}\|$, pro Iterationsschritt durchschnittlich um den Faktor $\rho(T)$ abnimmt. Des Weiteren kann gezeigt werden, siehe z.B. [75, Satz 4.5], dass ein stationäres Verfahren genau dann gegen die Lösung $x = A^{-1}b$ konvergiert, wenn $\rho(T) < 1$ ist. Beispiele für stationäre Verfahren sind das Jacobi-Verfahren mit $B = \mathrm{diag}(A)$, als auch das Gauß-Seidel-Verfahren mit $B = \mathrm{triu}(A)$. Hierbei bezeichnen $\mathrm{diag}(A)$ und $\mathrm{triu}(A)$ die Diagonale bzw. das obere Dreieck, einschließlich der Diagonalen, von A.

Als weitaus effektivere iterative Verfahren gelten die sogenannten *Krylov-Unterraum-Verfahren*, die z.B. in [65] erläutert werden. Zu der Klasse dieser Verfahren gehört das *conjugate gradient Verfahren (CG-Verfahren)* von Hestenes und Stiefel [56]. Für dieses Verfahren muss $A \in \mathbb{R}^{n,n}$ symmetrisch positiv definit (spd) oder $A \in \mathbb{C}^{n,n}$ hermitesch positiv definit (hpd) sein. Das CG-Verfahren minimiert das Funktional $f(x) = \frac{1}{2}x^H Ax - x^H b$ auf einem Suchraum $V_k = x_0 + K_k(r_0, A)$, wobei $K_k(r_0, A) = \text{Spann}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}$ ein sogenannter Krylov-Unterraum ist. Dabei gilt, dass in $\widetilde{x}$ genau dann das Minimum auf $\mathbb{R}^n$ bzw. $\mathbb{C}^n$ angenommen wird, wenn $\widetilde{x} = A^{-1}b$ ist. Es kann für den Fehler die Ungleichung $\|e^{[k+1]}\|_A^2 \leq 2C^k \|e^{[0]}\|_A^2$ mit $C := \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ gezeigt werden, siehe z.B. [47] oder auch [54, Satz 8.3]. Hierbei ist $\|\cdot\|_A$ die Energie-Norm und κ die Konditionszahl bzgl. der 2-Norm von A. Diese Fehlerschranke bedeutet, dass eine kleine Konditionszahl eine schnelle Konvergenz garantiert, wodurch die Konditionszahl in der Literatur häufig als Maß für die Konvergenzgeschwindigkeit des CG-Verfahrens verwendet wird. Es sei angemerkt, dass es sich „nur“ um eine obere Schranke des Fehlers handelt und das CG-Verfahren trotz einer großen Konditionszahl schnell konvergieren kann. Ungeachtet dessen führt eine *Vorkonditionierung* in der Regel zu einer Beschleunigung des CG-Verfahrens. Dazu wird, wie auch bei den stationären Verfahren, eine Approximation B^{-1} an A^{-1} bestimmt, wobei B als spd bzw. hpd vorausgesetzt werden muss. Anstatt des Systems $Ax = b$ wird dann das System $B^{-\frac{1}{2}}AB^{-\frac{1}{2}}y = B^{-\frac{1}{2}}b$ mit $x = B^{-\frac{1}{2}}y$ gelöst, wobei die Konditionszahl von $B^{-\frac{1}{2}}AB^{-\frac{1}{2}}$ deutlich kleiner, als die von A sein sollte. Eine mögliche Wahl eines Vorkonditionierers B ist gegeben durch $B = \text{diag}(A)$; eine Übersicht über Vokonditionierer liefert die Arbeit von Benzi [16]. Häufig werden sogenannte *Mehrgitterverfahren (MGV)* als Vorkonditionierer verwendet.

Die ersten Ideen zu Mehrgitterverfahren wurden von Fedorenko [49] und Bakhvalov [12] vorgestellt. Nach einer Arbeit von Brandt [26] entwickelten sich Mehrgitterverfahren zu effektiven und robusten Verfahren [23–25], die sowohl als stationäre Verfahren sowie auch als Vorkonditionierer verwendet werden. Die Idee der Mehrgitterverfahren ist es, für das gegebene Problem eine Gitterhierarchie aufzubauen. Dafür benötigt es Transfer-Operatoren, die zwischen den verschiedenen Gittern für den nötigen Informationsaustausch sorgen und auf der jeweiligen Geometrie des Problems basieren. Daher ist die Anwendbarkeit klassischer Mehrgitterverfahren, auch geometrische MGV genannt, durch die jeweilig gegebene Geometrie eingeschränkt. Des Weiteren bestehen Mehrgitterverfahren in der Regel aus einer Kombination von zwei stationären Verfahren. Das erste ist ein sogenannter Glätter, der stark oszillierende Fehler auf glatte überführt. Diese glatten Fehler können ohne starken Informationsverlust auf das nächst gröbere Gitter mittels des Transfer-Operators restringiert werden. Dies geschieht durch die Anwendung einer *Grobgitterkorrektur*, welche das zweite Verfahren darstellt. Für eine ausführliche Beschreibung und Analyse von Mehrgitterverfahren sei hier auf die Arbeiten von W. Hackbusch [55], Wesseling [97], Trottenberg, Oosterlee, Schüller [90] und Köckler [59] verwiesen. Damit Mehrgitterverfahren auch bei Problemen ohne Geometrie bzw. Problemen mit schwer beherrschbarer Geometrie angewendet werden können, wurden algebraische Varianten, also algebraische Mehrgitterverfahren (AMG's), von Brandt, McCormick und Ruge [30] entwickelt. In weiteren Untersuchungen konnten Ruge und Stüben [84, 85] zeigen, dass AMG's für eine große Menge verschiedener Matrizen effiziente Verfahren darstellen.

Algebraische Mehrgitterverfahren wurden in den letzten 30 Jahren stetig weiterentwickelt und an den verschiedensten Problemen getestet [6, 19, 21, 22, 36, 41, 57, 58, 68, 77, 78, 80, 81, 89].

Für die Analyse von AMG's benötigt es gewisse Voraussetzungen an die Matrix A. In der Regel soll A spd bzw. hpd oder auch eine M-Matrix sein. In dieser Arbeit wird A, abhängig von dem jeweiligen Kontext, als symmetrisch positiv definit oder hermitesch positiv definit vorausgesetzt. Eine Analyse von AMG's für M-Matrizen kann in [76–78] gefunden werden. Die Möglichkeit, AMG's auf semidefinite Systeme zu erweitern, wird z.B. in [22] vorgestellt. Für die Konstruktion eines algebraischen Mehrgitterverfahrens benötigt es zunächst einen Coarser, der ähnlich wie bei den geometrischen MGV, Gitter bzw. Level in Abhängigkeit der Matrix A erzeugt. Ruge und Stüben [85] führten dazu den Begriff der *starken Nachbarn* ein. Die Einteilung in ein "grobes„ und ein "feines„ Gitter erfolgt mithilfe des Graphen der Matrix A. Innerhalb der letzten zehn Jahre wurden alternative Coarsing-Strategien entwickelt. Zu diesen zählen das CR-Coarsening [32, 34, 66] und der Greedy-Coarser [70, 71]. Beide haben gemeinsam, dass sie nicht auf einer Einteilung in starke Nachbarn basieren. Die CR-Coarsening-Strategie verwendet *compatible relaxation* [27] und teilt die Unbekannten mithilfe des Glätters in Grob- und Feingitterpunkte ein. Der Greedy-Coarser verfolgt das Ziel, dass die zu den Feingitterpunkten assoziierte Matrix durch eine Diagonalmatrix zufriedenstellend approximiert werden kann. In dieser Arbeit wird ausschließlich der Greedy-Coarser verwendet.

Um ein effizientes AMG zu konstruieren, benötigt es neben einem Coarser einen effektiven Glätter, eine Grobgittermatrix A_c und den zugehörigen Transferoperator bzw. einen Interpolationsoperator P. Häufig wird bei algebraischen Mehrgitterverfahren die Grobgittermatrix durch $A_c = P^H AP$, den *Galerkin-Ansatz*, definiert. Als Glätter kann z.B. das Jacobi- oder auch das Gauß-Seidel-Verfahren gewählt werden. Jedoch ist es in der Regel nicht notwendig, dass auf allen Unbekannten geglättet wird. Stattdessen reicht es aus, einen F-Glätter, der nur auf den Feingitterpunkten operiert, zu betrachten. Beispiele für derartige Verfahren sind das AMGr-Verfahren von MacLachlan, Manteuffel und McCormick [68] und das mathematisch dazu äquivalente MAMLI-Verfahren von Mense und Nabben [76–78].

Das Ziel dieser Arbeit liegt in einer umfangreichen Analyse algebraischer Mehrgitterverfahren bei der Verwendung einer F-Glättung. Es werden Zusammenhänge verschiedener Verfahren, basierend auf F-Glättung, wie dem *Multilevel-Block-Faktorisierungs-Verfahren (MBF)* [81] und dem *reduction-based AMG (AMGr)* [68] aufgezeigt. Obwohl diese Verfahren unterschiedlich hergleitet wurden, lassen sie sich mit den gleichen Methoden analysieren. Diese Methoden basieren insbesondere auf der Verwendung von *generalisierten hierarchischen Basen* [6, 81, 93] und der damit verbundenen Abschätzung der *Cauchy-Bunyakovski-Schwarz-Konstante*. Der Fokus der Arbeit liegt dabei auf dem AMGr-Verfahren, welches im Jahr 2006 von MacLachlan, Manteuffel und McCormick entwickelt wurde und seitdem viel Beachtung erhielt. In [68] konnte gezeigt werden, dass das AMGr-Verfahren für symmetrisch positiv definite Matrizen unter gewissen Voraussetzungen konvergiert. Diese Voraussetzungen werden im weiteren Verlauf dieser Arbeit detailliert untersucht; es wird insbesondere der Zusammenhang zur Cauchy-Bunyakovski-Schwarz-Konstante, als auch zu den Tschebysheff Polynomen vorgestellt, was zu einer neuen Interpretation der von MacLachlan, Manteuffel und McCormick zur Analyse des AMGr-Verfahrens verwendeten Parameter führt.

Diese Untersuchung ermöglicht es, das AMGr-Verfahren auf ein *polynombasiertes AMG (AMGp)* zu verallgemeinern. Es kann gezeigt werden, dass diese Verallgemeinerung auf eine bessere Konvergenzschranke im Vergleich zu dem AMGr-Verfahren ohne zusätzlichen numerischen Aufwand führt. Des Weiteren sind die Voraussetzungen an die verschiedenen Matrizen, die bei dem AMGr-Verfahren benötigt werden, sehr restriktiv. Zum einen können durch die Ver-

wendung der Tschebysheff Polynome allgemeinere Matrizen als Glätter verwendet werden, zum anderen erlaubt die Untersuchung des AMGr-Verfahrens, mittels generalisierter hierarchischer Basen, allgemeinere Voraussetzungen für die Interpolationsmatrix und Grobgittermatrix zu treffen. Diese detaillierten Untersuchungen bilden den Kern dieser Arbeit. Um eine vollständige Analyse des polynombasierten AMGs durchzuführen, benötigt es weitere Hilfsmittel. Diese basieren zum großen Teil auf einer Transformation des AMGp-Verfahrens auf ein Block-Faktorisierungs-Verfahren. Diese Transformation wurde erstmals in [76] vorgestellt. In [93] wird gezeigt, dass solch eine Transformation für eine sehr große Menge verschiedener AMGs möglich ist. Aus diesen Gründen beschäftigt sich ein Großteil dieser Arbeit mit den AMG-verwandten MBF-Verfahren, die z.B. in [10, 11] und [81] analysiert werden. In diesen Arbeiten werden MBF- und MBF-ähnliche Verfahren mithilfe gewisser Voraussetzungen untersucht. Einige dieser Ideen werden in dieser Arbeit aufgegriffen, um zu zeigen, dass analoge Ergebnisse unter allgemeineren und damit auch schwächeren Voraussetzungen bewiesen werden können.

Eine dieser Verallgemeinerungen bezieht sich auf ein Resultat von Notay [81]. In [81] werden algebraische Mehrgitterverfahren sowie Block-Faktorisierungs-Verfahren verglichen, wozu eine Abschätzung der C.B.S. Konstante einer Matrix $\widehat{A}$, die durch eine Kongruenztransformation der Matrix A hervorgeht, benötigt wird (siehe [81, Theorem 7]). Das Besondere an der Matrix $\widehat{A}$ ist ihre 2×2-Blockstruktur, da der (2×2)-Block identisch zu der Grobgittermatrix A_c bei der Verwendung des Galerkin-Ansatzes ist und dadurch eine spektrale Äquivalenz zwischen der optimalen Grobgittermatrix und der Matrix A_c, abhängig von der C.B.S. Konstante von $\widehat{A}$, erzielt werden kann. Eine Besonderheit der Abschätzung von Notay ist ihre Unabhängigkeit von der C.B.S. Konstante von A. Stattdessen werden nur algebraische Voraussetzungen an die Matrix A und die Interpolationsmatrix P benötigt. Dieses Resultat scheint bis zum heutigen Tag die einzige bekannte Abschätzung der C.B.S. Konstante von $\widehat{A}$ unabhängig von der C.B.S. Konstante von A zu sein (siehe [93, Seite 86]), wobei die Voraussetzungen nicht für jede Matrix A erfüllbar sind. In Kapitel 6 werden weitere Abschätzungen in ähnlicher Form von Theorem 7 aus [81] vorgestellt. Eine dieser Abschätzungen (Satz 6.11) stellt insbesondere eine echte Verallgemeinerung von Notays Theorem dar. Des Weiteren wird gezeigt, dass die Voraussetzungen von Satz 6.11 immer erfüllbar sind, sobald die C.B.S. Konstante von A kleiner als $\sqrt{\frac{4}{5}}$ ist, was wiederum bei vielen Anwendungen unter der Verwendung von einer sogenannten hierarchischen finiten Elemente Basis der Fall ist [18, 63, 83]. Des Weiteren kann für generelle elliptische 2D Probleme mit stückweisen linearen Basisfunktionen gezeigt werden, dass die C.B.S. Konstante von A kleiner als $\sqrt{\frac{3}{4}}$ ist [5, 72]. In [73] wurde dieselbe Schranke für 2D Elastizitäts-Probleme gefunden, siehe dazu [6].

Die theoretische Analyse von AMG's mit F-Glättung, die den Hauptbestandteil dieser Arbeit bildet, wird durch numerische Berechnungen ergänzt. Diese Berechnungen orientieren sich an den Modellproblemen aus den Arbeiten [33] und [68]. In [68] wird das AMGr-Verfahren auf die Poisson-Gleichung angewendet. Mit diesem Modellproblem beschäftigt sich der erste Teil der Berechnungen. Dabei werden die Vorteile des AMGp- im Vergleich zum AMGr-Verfahren verdeutlicht. Als zweites Modellproblem wird die Gauge-Laplace-Matrix verwendet, die in vielen Arbeiten als Modellproblem für die Quantenelektrodynamik und die Quantenchromodynamik verwendet wird, [1, 31, 33, 43, 43, 57, 58]. In diesen numerischen Berechnungen wird eine adaptive Version des AMGp-Verfahrens (αAMGp), welches sich an αAMGr [68] orientiert, vorgestellt. Adaptive Verfahren stellen eine Möglichkeit dar, lokale Eigenschaften des

Problems auszunutzen, um die Konvergenzgeschwindigkeit zu beschleunigen [33, 35, 37, 68]. Häufig benötigen adaptive Verfahren Informationen über Kern-nahe Vektoren von A. Es wird das Ziel verfolgt, Interpolationsmatrizen zu erzeugen, die auf die Kern-nahen Vektoren dieselbe Wirkung haben, wie die „optimale“ Interpolationsmatrix. Dies bedeutet jedoch, dass eine Approximation eines Eigenvektors zum kleinsten Eigenwert von A benötigt wird. In den letzten Jahren wurden verschiedene Möglichkeiten zur Konstruktion einer, auf das jeweilige Problem angepassten, Interpolationsmatrix vorgestellt. Neben den erwähnten adaptiven Verfahren und der Verwendung von Eigenvektorapproximationen ist die Verwendung von Ausgleichsproblemen ein vielversprechender Ansatz, der z.B. bei dem Bootstrap AMG [21, 28, 29] Anwendung findet.

Diese Arbeit ist in 10 Kapitel und 2 Kapitel im Anhang unterteilt. In Kapitel 2 werden zwei Modellprobleme, die Poisson-Gleichung und die Gauge-Laplace-Gleichung, eingeführt und ihre wichtigsten Eigenschaften für die spätere Analyse vorgestellt. Kapitel 3 und 4 dienen zur Darstellung der Grundlagen algebraischer Mehrgitterverfahren. Dazu werden zunächst in Kapitel 3 stationäre Verfahren und anschließend Mehrgitterverfahren im Allgemeinen besprochen. Das Kapitel 4 befasst sich detaillierter mit den algebraischen Mehrgitterverfahren bei der Verwendung einer F-Glättung und führt das Multilevel-Block-Faktorisierungs- sowie das AMGr-Verfahren ein. Das Multilevel-Block-Faktorisierungsverfahren wird schließlich in den Kapiteln 5 und 6 analysiert. Dabei wird in Kapitel 5 dargestellt, wie gewisse spektrale Eigenschaften des Systems von einem zum nächsten Level transportiert werden. Kapitel 6 bildet einen Hauptbestandteil dieser Arbeit. Die in diesem Kapitel angesprochene C.B.S. Konstante wird sowohl bei der Analyse von Block-Faktorisierungs,- als auch bei algebraischen Mehrgitter-Verfahren benötigt. Es wird zum einen erläutert, warum diese Konstante bei der Untersuchung von AMG's nicht vernachlässigt werden darf, aber auch, wie die Konstante abgeschätzt werden kann, um effektive AMG's zu konstruieren. In Kapitel 7 werden als weiterer Hauptbestandteil dieser Arbeit AMG's mit F-Glättung und deren Beschleunigung durch Tschebysheff Polynome untersucht. Es wird gezeigt, dass das AMGr-Verfahren in einer allgemeineren Form, als der bisher bekannten, dargestellt und untersucht werden kann. Außerdem wird in diesem Kapitel verdeutlicht, wie hilfreich ein gutes Verständnis der Block-Faktorisierungs-Verfahren bei der Analyse algebraischer Mehrgitterverfahren mit F-Glättung ist. Ergänzend wird in Kapitel 8 gezeigt, dass einige der vorgestellten Ideen auch bei voller Glättung Anwendung finden. Kapitel 9 verifiziert die theoretischen Ergebnisse anhand numerischer Versuche und der Modellprobleme aus Kapitel 2. Hierbei sei darauf hingewiesen, dass das Ziel dieser Arbeit nicht die Konstruktion eines neuen AMGs für die jeweiligen Modellprobleme verfolgt. Vielmehr steht die theoretische Analyse der wichtigsten AMG's mit F-Glättung und das Finden neuer Zusammenhänge und Interpretationsmöglichkeiten bekannter Resultate im Mittelpunkt, was zu Verallgemeinerungen von Resultaten von Notay sowie von MacLachlan, Manteuffel und McCormick führt.

Zum Verständnis dieser Arbeit wird eine gewisse Vorkenntnis der numerischen linearen Algebra vorausgesetzt. Begriffe wie *Norm, Eigenwerte* oder auch *Singulärwerte* werden daher nicht definiert. Des Weiteren werden die üblichen Notationen aus der Literatur verwendet, so bezeichnet z.B. $\lambda(B)$ einen Eigenwert einer Matrix $B \in \mathbb{R}^{n,n}(\mathbb{C}^{n,n})$ und $\mathbb{R}^{n,n}(\mathbb{C}^{n,n})$ den Vektorraum aller $n \times n$ Matrizen mit reellen (komplexen) Koeffizienten. Die Matrix I ist die Einheitsmatrix aus dem in dem Kontext betrachteten Vektorraum.

Kapitel 2

Die Modellprobleme

Hier werden die verwendeten Modellprobleme erläutert. In Abschnitt 2.1 wird zunächst auf ein Standardproblem in der numerischen Mathematik, die Poisson-Gleichung[1], eingegangen. Anhand dieser Gleichung können wichtige Ideen und Grundkonzepte numerischer Verfahren erläutert werden, siehe Kapitel 3. Des Weiteren eignet sich die Poisson-Gleichung für numerische Berechnungen, da eine fortgeschrittene Analyse dieser Gleichung vorhanden ist, die es erlaubt, die Konzentration bei numerischen Berechnungen auf das Wesentliche zu beschränken. Dies wird im ersten Teil von Kapitel 9 genutzt, um die theoretischen Ergebnisse dieser Arbeit zu verifizieren.

Diese ersten numerischen Resultate helfen, die vorgestellten Varianten von Mehrgitterverfahren numerisch zu analysieren. Das Problem bei der Verwendung der Poisson-Gleichung liegt jedoch darin, dass die betrachteten Verfahren nur selten an ihre Grenzen geführt werden. Um eine effizientere Analyse durchführen zu können, wird ein komplexeres Problem, die Gauge-Laplace-Gleichung, eingeführt, siehe Abschnitt 2.2. Diese Gleichung dient des Öfteren zur Analyse neu entwickelter algebraischer Mehrgitterverfahren [33, 57, 58]. Um die Herleitung der Gauge-Laplace-Gleichung verstehen zu können, werden in Anhang B die wichtigsten Grundbausteine der Quantenelektrodynamik (QED) sowie der Quantenchromodynamik (QCD) vorgestellt.

2.1

Die Poisson-Gleichung - 1D und 2D. In diesem Abschnitt wird die Poisson-Gleichung

$$-\Delta u = f \tag{2.1}$$

untersucht. Hierbei sei $f : \mathbb{R}^m \to \mathbb{R}$ eine stetige Funktion, wobei $m \in \{1, 2\}$ sein soll und Δ der Laplace-Operator[2]. Für $f = 0$ heißt die Gleichung (2.1) auch die Laplace-Gleichung. In diesem Zusammenhang wird angenommen, dass $u : \mathbb{R}^m \to \mathbb{R}$ genügend oft differenzierbar ist. Andernfalls müsste zur schwachen Formulierung der Gleichung (2.1) übergegangen werden, dazu siehe [86].

Die Gleichung (2.1) findet vor allem in Teilgebieten der Physik, wie der Wärmeleitung oder der Elektrostatik, ihre Anwendung. In dieser Arbeit wird sie verwendet, um zum einen die fundamentalen Ideen der Mehrgitterverfahren zu erläutern und zum anderen, um die theoretischen Ergebnisse experimentell zu bestätigen.

Die Darstellung und Entwicklung der Ergebnisse orientiert sich an [39] bzw. [86].

[1] Benannt nach Siméon Denis Poisson (1781-1840), einem französischen Mathematiker und Physiker.

[2] Eingeführt von Pierre-Simon Laplace (1749-1827), französischer Mathematiker, Physiker und Astronom, der sich insbesondere mit Wahrscheinlichkeitstheorie und Differentialgleichungen beschäftigte.

Die Gleichung (2.1) kann auf verschiedenen Mengen Ω unterschiedlicher Dimensionen betrachtet werden, wobei hier das Einheitsintervall $\Omega = [0,1]$ bzw. das Einheitsquadrat $\Omega = [0,1]^2$ betrachtet wird. Im eindimensionalen Fall erhält man $-u''(x) = f(x)$ und im zweidimensionalen $-\frac{\partial u(x,y)}{\partial x^2} - \frac{\partial u(x,y)}{\partial y^2} = f$. Des Weiteren werden Dirichlet-Randbedingungen[3] , also $u = 0$ auf dem Rand $\Gamma = \partial\Omega$ gefordert.

Um dieses Randwertproblem zu untersuchen, wird (2.1) durch finite Differenzen diskretisiert.

2.1.1 Der eindimensionale Fall.

Zunächst wird der eindimensionale Fall betrachtet. Das Intervall $\Omega = [0,1]$ wird in $n+2$ gleich verteilte Punkte $x_j = jh\,, j = 0,\ldots,n+1$ für $n \in \mathbb{N}$ aufgeteilt, wobei $h = \frac{1}{n+1}$ ist. Dies führt auf das diskretisierte Gebiet $\Omega_h = \{x_j \,|\, j = 0,\ldots,n+1\}$.

Wird die zweite Ableitung in (2.1) durch den zentralen Differenzenquotienten ersetzt, so erhält man für die Elemente aus Ω_h

$$-u''(x_j) = \frac{-u(x_{j-1})+2u(x_j)-u(x_{j+1})}{h^2} + \tau_j = f(x_j), \quad i = 1,\ldots,n, \tag{2.2}$$

wobei die Randbedingung $u(x_0) = u(x_{n+1}) = 0$ verwendet wurde. Der Term τ_j stellt den Diskretisierungsfehler dar, der im Weiteren vernachlässigt wird. Definiert man die Vektoren $x, b \in \mathbb{R}^n$ durch

$$x = \begin{bmatrix} u(x_1) \\ \vdots \\ u(x_n) \end{bmatrix} \quad \text{und} \quad b = h^2 \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix},$$

so erhält man aus (2.2) das lineare Gleichungssystem

$$Ax = b$$

mit

$$A = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{n,n}. \tag{2.3}$$

Bemerkung 2.1. Die Art der Diskretisierung durch die Wahl des Differenzenquotienten (2.2) wird als zentrierter Differenzenquotient bezeichnet und kann abkürzend als 3-Punkte-Stern wie in Abbildung 2.1 geschrieben werden.

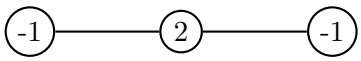


Abbildung 2.1.: 3-Punkte-Stern mit zentrierten Differenzen für $-u'' = f$.

[3] Benannt nach Peter Gustav Lejeune Dirichlet (1805-1859), deutscher Mathematiker, der in Berlin und Göttingen lehrte und im Wesentlichen in den Gebieten der Analysis und der Zahlentheorie arbeitete.

In Kapitel 3 wird gezeigt, dass die Eigenwerte und Eigenvektoren dieser Matrix eine wichtige Rolle spielen. Dazu sei bemerkt, dass die Matrix A aus (2.3) symmetrisch positiv definit (spd) ist. Dies bedeutet, dass $A = A^T$ und $y^T Ay \geq 0$ für alle $y \in \mathbb{R}^n$ erfüllt ist und $y^T Ay = 0$ nur für $y = 0$ gilt. Insbesondere hat A nur reelle Eigenwerte $\lambda(A) > 0$.

Um die Eigenvektoren der Matrix A zu bestimmen, betrachtet man für $\theta_k = \frac{k\pi}{n+1}$, $k \in \mathbb{N}$ die Vektoren $v_k = [v_j^k] \in \mathbb{R}^n$ mit den Einträgen $v_j^k = \sin(j\theta_k)$, $j = 1, \ldots, n$. Aus der Identität $\sin(x) + \sin(y) = 2\sin(\frac{x+y}{2})\cos(\frac{x-y}{2})$, die für alle x, y erfüllt ist und der speziellen Wahl $x = (j+1)\theta_k$ als auch $y = (j-1)\theta_k$ kann die Gleichung $\sin((j+1)\theta_k) + \sin((j-1)\theta_k) = 2\sin(j\theta)\cos(i\theta_k)$ gefolgert werden. Also sind die Einträge von Av_k gegeben durch

$$(Av_k)_j = -\sin\big((j-1)\theta_k\big) + 2\sin\big(j\theta_k\big) - \sin\big((j+1)\theta_k\big) = 2\big(1 - \cos\big(\theta_k\big)\big)\sin\big(j\theta_k\big).$$

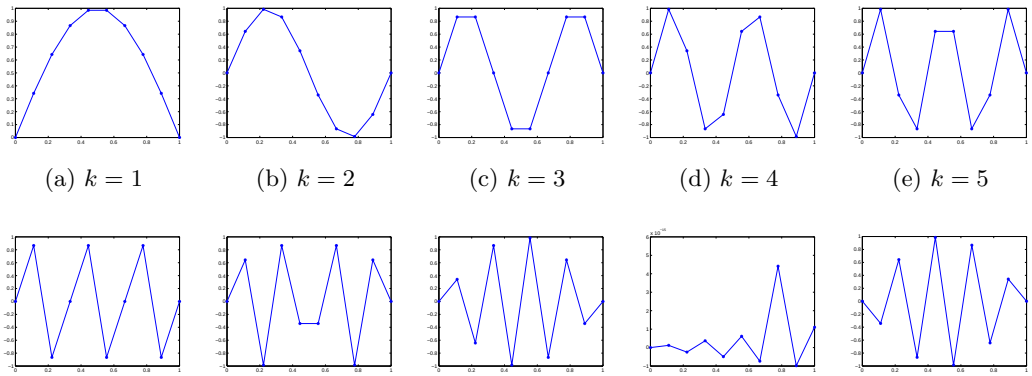
Damit erhält man die Eigenwerte von A

$$\lambda_k(A) = 2\big(1 - \cos(\theta_k)\big) = 4\sin^2(\tfrac{\theta_k}{2}), \quad k = 1, \ldots, n \tag{2.4}$$

mit den zugehörigen Eigenvektoren

$$v_k = \begin{bmatrix} \sin(\theta_k) \\ \vdots \\ \sin(n\theta_k) \end{bmatrix} = \begin{bmatrix} \sin(\frac{k\pi}{2(n+1)}) \\ \vdots \\ \sin(\frac{nk\pi}{2(n+1)}) \end{bmatrix}, \quad k = 1, \ldots, n. \tag{2.5}$$

Die Komponenten der Eigenvektoren können zu $\sin(j\theta_k) = \sin(k\pi x_j)$ umgeschrieben werden und stellen damit eine Diskretisierung der Eigenfunktionen $f_k(x) = \sin(k\pi x)$ auf Ω_h dar. Für $n = 10$ können die 10 Eigenfunktionen auf dem diskretisierten Gebiet Ω_h in Abbildung 2.2 gefunden werden.



(a) $k = 1$ (b) $k = 2$ (c) $k = 3$ (d) $k = 4$ (e) $k = 5$

(f) $k = 6$ (g) $k = 7$ (h) $k = 8$ (i) $k = 9$ (j) $k = 10$

Abbildung 2.2.: Die zehn Eigenfunktionen $u_k(x) = \sin(k\pi x)$ des 1D Laplace-Operators für $n = 10$ mit $x \in \Omega_h$.

Anhand Abbildung 2.2 erkennt man, dass die zu den kleinen Eigenwerten zugehörigen Eigenfunktionen bzw. die Einträge der zugehörigen Eigenvektoren von A nur leicht oszillieren und die zu den größeren Eigenwerten stark oszillieren.

2.1.2 Der zweidimensionale Fall. Die 2D Poisson-Gleichung mit homogener Dirichlet-Randbedingung ist gegeben durch

$$-\left(\frac{\partial u}{\partial x^2} + \frac{\partial u}{\partial y^2}\right) = f \quad \text{in} \quad \Omega,$$
$$u = 0 \quad \text{auf} \quad \Gamma.$$

Hier wird das Gebiet $\Omega = [0,1] \times [0,1]$ in $n+2$ Gitterpunkte pro Richtung aufgeteilt, also

$$x_j = jh,\ y_j = jh, \quad j = 0, \ldots, n+1.$$

Der Einfachheit halber wird in beide Richtungen der gleiche Gitterabstand h gewählt. Nach einer analogen Diskretisierung wie in (2.2) erhält man für die 2D Poisson-Gleichung das lineare Gleichungssystem

$$A = \begin{bmatrix} T & -I & & & \\ -I & T & -I & & \\ & \ddots & \ddots & \ddots & \\ & & -I & T & -I \\ & & & -I & T \end{bmatrix} \in \mathbb{R}^{n^2,n^2},\ T = \begin{bmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{bmatrix} \in \mathbb{R}^{n,n}. \quad (2.6)$$

Der zugehörige 5-Punkte-Stern für die Diskretisierung ist in Abbildung 2.3 dargestellt.

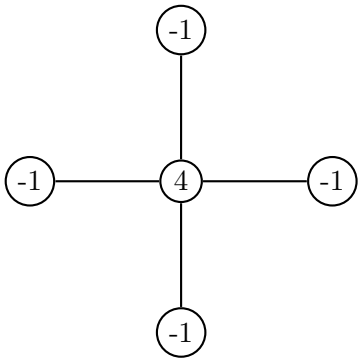
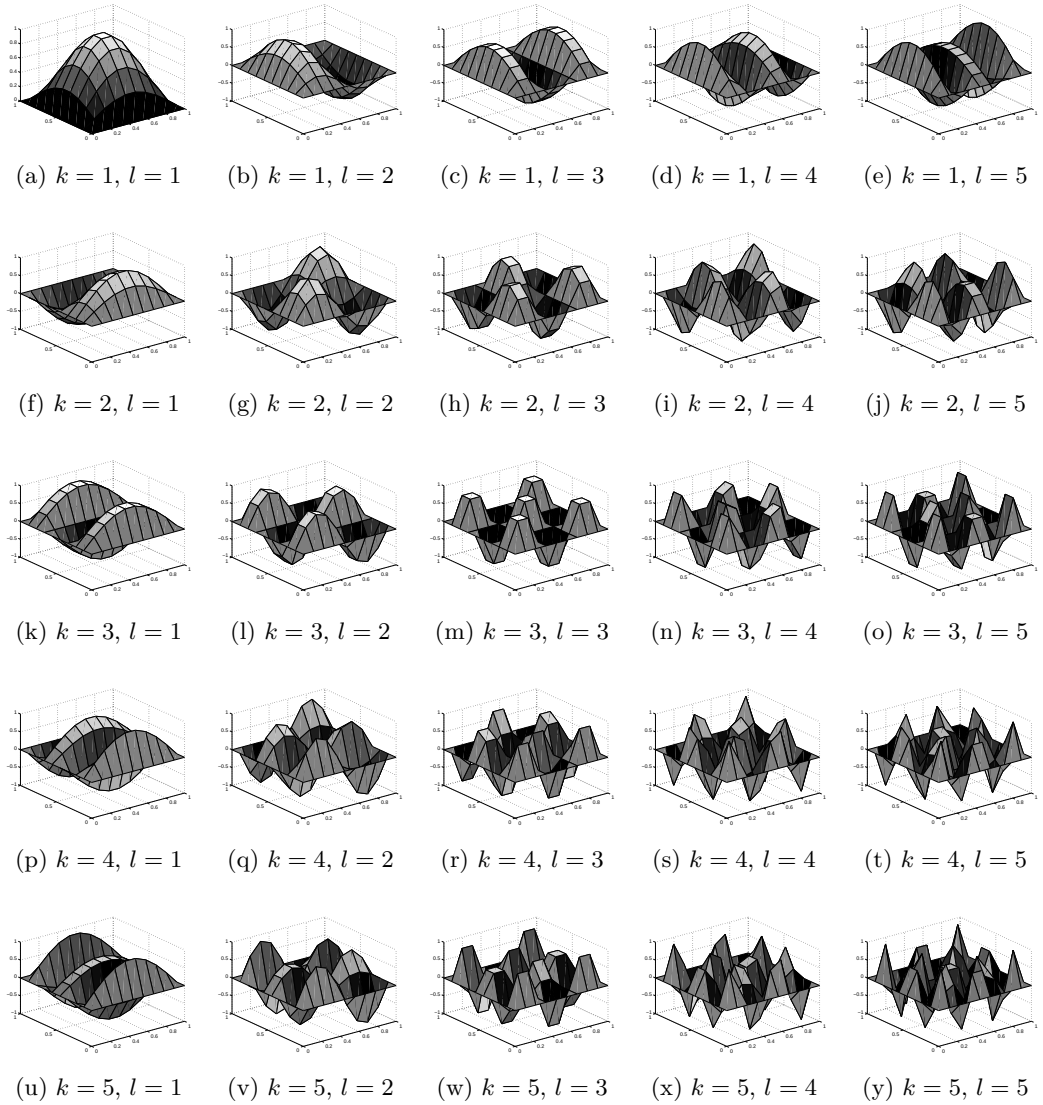


Abbildung 2.3.: 5-Punkte-Stern mit zentrierten Differenzen für $-\Delta u = f$.

Eine elegante und nützliche Variante, um die obige Matrix zu beschreiben, nutzt das Tensorprodukt bzw. Kroneckerprodukt[4]. Dies ist für zwei Matrizen $B = [b_{ij}] \in \mathbb{R}^{n,n}, C \in \mathbb{R}^{m,m}$ durch $B \otimes C = [b_{ij}C]$ definiert. Damit erhält man $A = I \otimes K_x + K_y \otimes I$, wobei K_x und K_y durch die Matrix aus (2.3) gegeben sind. Diese Schreibweise hilft bei der Charakterisierung des Spektrums von A. Seien v_k^x und v_l^y Eigenvektoren zu den jeweils n Eigenwerten $\lambda_k^x(K_x)$ und $\lambda_l^y(K_y)$ von K_x bzw. K_y, $k, l = 1, \ldots, n$, dann sind $v_k^x \otimes v_l^y$ Eigenvektoren von A zu den n^2 Eigenwerten $\lambda_{kl}(A)) = 4\left(\sin^2(\frac{k\pi}{2(n+1)}) + \sin^2(\frac{l\pi}{2(n+1)})\right)$, $k, l = 1, \ldots, n$, (vgl. (2.5)), was aus dem Satz von Stephanos folgt, siehe z.B. in [64, Satz 20.9]. Die Eigenvektoren entsprechen

[4]Benannt nach Leopold Kronecker (1823-1891), deutscher Mathematiker. Doch vermutlich wurde das Kroneckerprodukt zuerst durch Johann Georg Zehfuss (1832-1901) im Jahre 1858 definiert.

erneut einer Diskretisierung der Eigenfunktionen $f_{kl}(x, y) = \sin(k\pi x)\sin(l\pi y)$ auf Ω_h. Für den Fall $n = 10$ sind diese in Abbildung 2.4 dargestellt.



(a) $k = 1$, $l = 1$ (b) $k = 1$, $l = 2$ (c) $k = 1$, $l = 3$ (d) $k = 1$, $l = 4$ (e) $k = 1$, $l = 5$

(f) $k = 2$, $l = 1$ (g) $k = 2$, $l = 2$ (h) $k = 2$, $l = 3$ (i) $k = 2$, $l = 4$ (j) $k = 2$, $l = 5$

(k) $k = 3$, $l = 1$ (l) $k = 3$, $l = 2$ (m) $k = 3$, $l = 3$ (n) $k = 3$, $l = 4$ (o) $k = 3$, $l = 5$

(p) $k = 4$, $l = 1$ (q) $k = 4$, $l = 2$ (r) $k = 4$, $l = 3$ (s) $k = 4$, $l = 4$ (t) $k = 4$, $l = 5$

(u) $k = 5$, $l = 1$ (v) $k = 5$, $l = 2$ (w) $k = 5$, $l = 3$ (x) $k = 5$, $l = 4$ (y) $k = 5$, $l = 5$

Abbildung 2.4.: Die ersten 25 Eigenfunktionen $u_{kl}(x, y) = \sin(k\pi x)\sin(l\pi y)$ des 2D Laplace-Operators für $n = 10$ mit $(x, y) \in \Omega_h$

2.2 Gauge-Laplace.

Die Gauge-Laplace-Matrix ist ein Modellproblem, um Verfahren für numerisch anspruchsvolle Probleme aus der Quantenelektro- und Quantenchromodynamik zu testen. In diesen physikalischen Gebieten werden Bewegungsgleichungen in Form der *Dirac-Gleichung* untersucht. Da sich die analytische Behandlung, insbesondere auf dem Gebiet der Quantenchromodynamik, als äußerst schwierig erwies, hat sich der Ansatz der numerischen Betrachtung durch eine Diskretisierung durchgesetzt [98].

Die Quantenchromodynamik gehört in das weite Gebiet der Teilchenphysik und ist eine Eichfeldtheorie. Wie bei einer Modellierung von Bewegungsgleichungen üblich, treten partielle Ableitungen in den drei Ortsvariablen x, y, z und der Zeitvariablen t auf. Durch eine gewisse Eigenschaft, die ein physikalisches System erfüllen soll (Eichinvarianz), ergibt sich eine modifizierte Ableitung.

Definition 2.2. Seien für $\mu \in \{0, 1, 2, 3\}$ sogenannte *Eichfelder* A_μ gegeben. Für diese Eichfelder werden *kovariante Ableitungen* D_μ durch

$$D_\mu := \partial_\mu + iA_\mu, \quad \mu = 0, 1, 2, 3 \tag{2.7}$$

definiert. Hierbei entspricht ∂_0 der Ableitung nach der Zeit und $\partial_1, \partial_2, \partial_3$ den partiellen Ableitungen nach dem Ort und i ist die imaginäre Einheit.

Für die Gauge-Laplace-Matrix werden nur D_1 und D_2 benötigt. Diese Matrix entspricht einer Diskretisierung des 2D Laplace-Operators, wobei die gewöhnlichen Ableitungen durch die kovarianten Ableitungen ersetzt werden. Diese Matrix wurde unter anderem für numerische Berechnungen in [33, 45, 50, 57, 69] verwendet.

Weitere Darstellungen zur Quantenchromodynamik sowie eine Möglichkeit der Diskretisierung werden im Anhang B erläutert.

Für eine kurze Herleitung der Gauge-Laplace-Matrix sei Ω ein zweidimensionales Gebiet, über welches ein Gitter mit den Gitterpunkten $\{(k, l) \mid k, l = 1, \ldots, N\}$ gelegt wird. Hierbei sei der Abstand zwischen den Gitterpunkten $h = 1$, d.h. anstatt die Diskretisierung feiner zu machen, ist man an einer Vergrößerung des Gebietes interessiert. Jeder Kante wird ein sogenanntes *diskretes Eichfeld* aus

$$U(1) := \{e^{i\beta\theta_\nu^{(k,l)}} \mid k, l = 1, \ldots, N; \nu = 1, 2\}$$

zugewiesen. Dabei wird $\beta \in \mathbb{R}^+$ als Temperaturparameter bezeichnet. $\beta = 0$ wird als *kalte* und $\beta \to \infty$ als *warme Konfiguration* definiert. Es ergibt sich das zu lösende Problem

$$(-\Delta_D + m)u(x, y) = f(x, y). \tag{2.8}$$

Dabei ist

$$\Delta_D := \sum_{\mu=1}^{2} D_\mu^2$$

und $m \in \mathbb{R}$ ein Parameter, der physikalisch als Masse interpretiert werden kann. Mit einer, zu zentralen finiten Differenzen ähnlichen, Diskretisierung kann (2.8) in das diskrete Problem

$$(4 + m)u_{kl} - \left(e^{i\beta\theta_1^{(k,l)}} u_{k-1,l} + e^{i\beta\theta_2^{(k,l)}} u_{k,l-1} + e^{-i\beta\theta_1^{(k+1,l)}} u_{k+1,l} + e^{-i\beta\theta_2^{(k,l+1)}} u_{k,l+1}\right) = f_{kl}$$

überführt werden. Hierbei sind $u_{kl} := u((k,l))$, $f_{kl} := f((k,l))$ und es werden periodische Randbedingungen verwendet. Detaillierte Informationen zur Diskretisierung können im Anhang B gefunden werden.

In der Literatur [33, 45, 57, 69] wird die entstehende Matrix in der Form $A = I - \kappa D$ geschrieben, wobei D als *Hopping-Matrix* und die Konstante $\kappa = \frac{1}{4+m}$ als *Hopping-Parameter* bezeichnet werden.

In Abbildung 2.5 ist der zugehörige 5-Punkte-Stern angegeben.

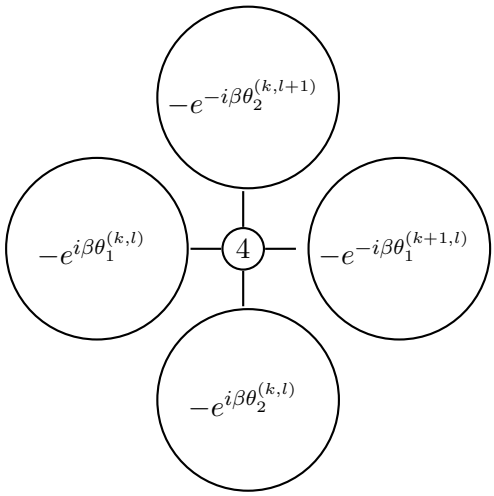


Abbildung 2.5.: Der 5-Punkte-Stern von $-\Delta_D u = f$.

Man erkennt, dass die Matrix $A \in \mathbb{C}^{n,n}$ hermitesch ist, also $A = A^H$. Der Hopping-Parameter kann verwendet werden, um für verschiedene physikalisch relevante Situationen gewünschte Eigenwerte der Matrix A zu erzeugen. In den numerischen Berechnungen sind vor allem Situationen mit $\lambda_{\min}(A) = 10^{-m}$, $m = 0, 1, \ldots, 6$ von Interesse.

Bemerkung 2.3. Für $m = 0$ und $\beta = 0$ (also eine kalte Konfiguration) ergibt sich $\{e^{i\beta\theta_\nu^{(k,l)}} \,|\, k, l = 1, \ldots, N; \nu = 1, 2\} = \{1\}$ und damit entspricht die Gauge-Laplace-Matrix der Laplace-Matrix aus (2.6).

Kapitel 3

Von einfachen stationären Verfahren zu Mehrgitterverfahren

Die Lösung eines linearen Gleichungssystems $Ax = b$ kann durch iterative Verfahren approximiert werden. Einfache stationäre Verfahren, wie das Jacobi- und das gewichtete Jacobi-Verfahren, gehören zu dieser Klasse. Bei der Analyse der Stärken und Schwächen solcher Verfahren werden die Grundideen der Mehrgitterverfahren veranschaulicht, die erstmals in [49] und [12] betrachtet wurden und seit Brandt [26] für die Verwendung als iterative Verfahren immer mehr Anwendungen fanden. Dies soll auf das darauffolgende Kapitel vorbereiten, in dem die Algebraischen Mehrgitterverfahren (AMG) eingeführt und erläutert werden, die in verschiedenen Variationen den Hauptteil der theoretischen Analyse dieser Arbeit einnehmen.

3.1 Stationäre Verfahren.

Im ersten Abschnitt dieses Kapitels werden stationäre Verfahren im Allgemeinen erläutert. Ziel dieser Verfahren ist es, für eine gegebene Matrix $A \in \mathbb{R}^{n,n}$ oder auch $A \in \mathbb{C}^{n,n}$ mit $\det(A) \neq 0$ und $b \in \mathbb{R}^n$ bzw. $b \in \mathbb{C}^n$ das Gleichungssystem

$$Ax = b, \tag{3.1}$$

zu lösen. Die Lösung ist gegeben durch $x = A^{-1}b$.

Bemerkung 3.1. In Abschnitt 2.2 wurden komplexwertige Matrizen $A \in \mathbb{C}^{n,n}$ betrachtet. Sowohl in diesem, als auch zu Beginn des nächsten Kapitels werden die verschiedenen Konzepte ausschließlich für reelle Matrizen besprochen. Diese Kapitel dienen zur Darstellung der wichtigsten Ideen der Mehrgitterverfahren. Ab Kapitel 5 wird die Beschränkung aufgehoben und komplexwertige Matrizen werden zugelassen.

Eine Alternative zu direkten Lösungsverfahren, wie der Gauß-Elimination oder der LU-Zerlegung, bieten iterative Verfahren, die die Lösung x durch eine Folge $\{x^{[k]}\}$ annähern. Definiert man eine Iterationsmatrix $T \in \mathbb{R}^{n,n}$, wird ausgehend von einem Startvektor $x^{[0]}$ die Folge durch die Vorschrift $x^{[k+1]} = Tx^{[k]} + c$ gebildet. Hierbei muss $c \in \mathbb{R}^n$ so gewählt werden, dass die Lösung von (3.1) ebenfalls eine Lösung von der Fixpunktgleichung $x = Tx + c$ ist. Es lässt sich zeigen, dass die so gebildete Folge $\{x^{[k]}\}$ genau dann gegen $x = A^{-1}b$ konvergiert, wenn der Spektralradius $\rho(T)$ kleiner Eins ist, siehe z.B. [75, Satz 4.5].

Eine Iterationsmatrix T kann durch eine Zerlegung von $A = M - N$ mit M invertierbar erzeugt werden. Das lineare Gleichungssystem $Ax = b$ ist mit dieser Zerlegung äquivalent zu $x = M^{-1}Nx + M^{-1}b = (I - M^{-1}A)x + M^{-1}b$. Somit ist die Iterationsmatrix durch

$T = I - M^{-1}A$ gegeben. Eine einfache Möglichkeit, solch eine Zerlegung zu finden, ergibt sich aus einer Aufteilung

$$A = D - R - L,$$

mit $D = \mathrm{diag}(A)$ und $-R, -L$ der obere rechte Block bzw. der untere linke Block der Matrix A. Wird außerdem gefordert, dass D nichtsingulär ist (z.B. wenn A spd ist), dann kann $M = D$ gewählt werden und es ergibt sich die Iteration

$$x^{[k+1]} = (I - D^{-1}A)x^{[k]} + D^{-1}b = x^{[k]} + D^{-1}(b - Ax^{[k]}).$$

Diese Iteration nennt man *Jacobi-Iteration.*

3.2

Das Jacobi- und gewichtete Jacobi-Verfahren. Für $D = \mathrm{diag}(A)$ bildet die Iterationsmatrix $T_1 = I - D^{-1}A$ bzw. die gewichtete Jacobi-Iterationsmatrix

$$T_\omega = I - \omega D^{-1}A$$

mit einem Gewicht $\omega \in \mathbb{R}$ eines der einfachsten stationären Verfahren. In diesem Abschnitt wird dieses Iterationsverfahren auf die 1D Laplace-Matrix (2.3) angewendet und ihre Wirkung analysiert. Sei A gegeben durch (2.3). Dann ist $D = \mathrm{diag}(2, \ldots, 2) \in \mathbb{R}^n$ und damit

$$T_\omega = I - \tfrac{\omega}{2}A.$$

Mit den Eigenwerten von A, gegeben durch (2.4), lassen sich die Eigenwerte von T_ω angeben:

$$\lambda_k(T\omega) = 1 - 2\omega \sin^2(\tfrac{k\pi}{2(n+1)}), \quad k = 1, \ldots, n.$$

Der Sinusterm $\sin(\frac{k\pi}{2(n+1)})$ bewegt sich zwischen den extremen Werten für $k = 1$ bzw. $k = n$. Setzt man $s := \sin(\frac{\pi}{2(n+1)})$, dann ist $\lambda_k(T_\omega) \leq 1 - 2\omega s^2$ für alle $k = 1, \ldots, n$ und

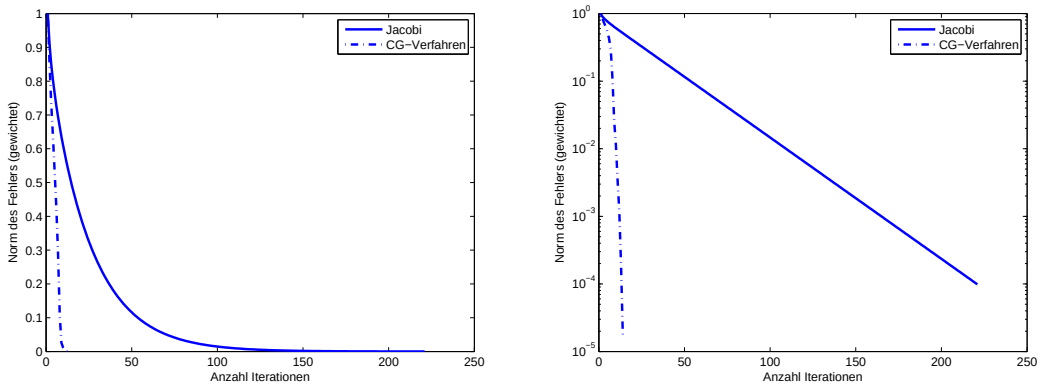
$$\begin{aligned}
\lambda_k(T_\omega) &\geq 1 - 2\omega \sin^2(\tfrac{n\pi}{2(n+1)}) = 1 - 2\omega \sin^2(\tfrac{\pi}{2} + \tfrac{\pi}{2(n+1}) \\
&= 1 - 2\omega \left(\cos(\tfrac{\pi}{2(n+1)}) - \sin(\tfrac{\pi}{2(n+1)})\cos(\tfrac{\pi}{2})\right)^2 = 1 - 2\omega(1 - s^2) \\
&= 1 - 2\omega + 2\omega s^2, \quad k = 1, \ldots, n.
\end{aligned}$$

Daher liegen die Eigenwerte von T_ω im Intervall $[1-2\omega+2\omega s^2, 1-2\omega s^2]$ und der Spektralradius von T_ω erfüllt die Gleichung

$$\rho(T_\omega) = \max\left\{ \left|\, (1 - 2\omega) + 2\omega s^2 \,\right|, \left|\, 1 - 2\omega s^2 \,\right| \right\}.$$

Somit ist das gewichtete Jacobi-Verfahren konvergent, wenn $\omega \in [0, 1]$ ist. Interessanterweise ist die optimale Wahl von ω für eine Minimierung des Spektralradius' gegeben durch $\omega = 1$. Diese Wahl von ω entspricht dem Jacobi-Verfahren, das im Folgenden auf die 1D Laplace-Matrix (2.3) angewendet wird.

Gegeben sei die Matrix aus (2.3) für $n = 10$. Als rechte Seite wird $b = 0 \in \mathbb{R}^n$ und als Startvektor $x^{[0]} = [\,1 \ldots 1\,]^T \in \mathbb{R}^n$ gewählt. Die Folge $x^{[k]}$ wird solange iteriert, bis $\|x^{[k]}\|_2$ kleiner als 10^{-3} ist.



(a) Jacobi und CG

(b) Jacobi und CG - logarithmische Darstellung

Abbildung 3.1.: Das Konvergenzverhalten des Jacobi-Verfahrens, angewandt auf die 1D Laplace-Matrix. Hier ist die Norm des Fehlers (gewichtet mit dem Anfangsfehler) über die Anzahl der benötigten Iterationen aufgetragen. Zum Vergleich ist das Konvergenzverhalten des CG-Verfahrens dargestellt.

In Abbildung 3.1 ist das typische Konvergenzverhalten des Jacobi-Verfahrens aufgetragen. Dieser Verlauf der Fehlerreduktion lässt sich bei vielen einfachen stationären Verfahren beobachten. Nach einer schnellen Reduktion des Fehlers nimmt die Konvergenzgeschwindigkeit ab. In der Abbildung erkennt man, dass nur 50 Iterationen benötigt werden, um den gewichteten Fehler bis zu einem Zehntel zu reduzieren. Jedoch werden weitere 150 Schritte gebraucht, um die Lösung zufriedenstellend zu approximieren.

Um dieses Phänomen zu verstehen, sind weitere Untersuchungen erforderlich. Da die betrachtete Matrix A spd ist, existiert eine Basis des $\mathbb{R}^n$ aus Eigenvektoren von A, die mit $v_1, \ldots, v_n$ bezeichnet werden und gegeben sind durch (2.5). Sei $e^{[m]} = x - x^{[m]}$ der Fehler nach der m-ten Iteration mit $x = A^{-1}b$. Dann existieren $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ mit

$$e^{[m]} = \sum_{k=1}^{n} \alpha_k v_k. \tag{3.2}$$

Betrachtet man die Wirkung der Iterationsmatrix T_1 auf $e^{[m]}$, so erhält man

$$e^{[m+1]} = T_1 e^{[m]} = \Big(I - \tfrac{1}{2}A\Big)\Big(\sum_{k=1}^{n} \alpha_k v_k\Big) = \sum_{k=1}^{n} \alpha_k \Big(1 - \tfrac{\lambda_k(A)}{2}\Big) v_k.$$

Dabei sind die Eigenwerte $\lambda_k(A)$ von A gegeben durch (2.4). Dies ergibt für den Fehler

$$e^{[m]} = T_1^m e^{[0]} = \sum_{k=1}^{n} \alpha_k \Big(1 - \tfrac{\lambda_k(A)}{2}\Big)^m v_k.$$

Das bedeutet, dass der Fehler in den Summanden mit $\lambda_k(A) \approx 0$ bzw. $\lambda_k(A) \approx 4$ nicht genügend reduziert wird, jedoch Eigenwerte $\lambda_k(A) \approx 2$ zu einer schnellen Verkleinerung der entsprechenden Fehleranteile führen. Nach genügend Iterationen wird der Fehler daher nur noch durch wenige Eigenvektoren dominiert.

Mit dem Wissen, dass Eigenwerte existieren, die zu einer langsamen Reduktion des Fehlers führen, als auch solche, die die Konvergenz beschleunigen, wird erneut das gewichtete Jacobi-Verfahren betrachtet. In Sachen Konvergenz wurde in dem Spezialfall der 1D Laplace-Matrix festgestellt, dass die optimale Wahl von ω gleich Eins ist. Im Folgenden soll jedoch nicht mehr der Spektralradius von T_ω minimiert, sondern eine Wahl von ω angegeben werden, die zu einer vernünftigen Aufteilung der Eigenvektoren von A führt. Diese Aufteilung soll so erfolgen, dass gewisse Anteile im Fehler reduziert werden, andere beinahe erhalten bleiben. Ferner wird ω so gewählt, dass die schwach oszillierenden Anteile (gegeben durch die Eigenvektoren zu den kleinen Eigenwerten, vgl. Abbildungen 2.2 und 2.4) dominant bleiben, die stark oszillierenden Anteile jedoch verschwinden. Dies führt zu der Möglichkeit, die Diskretisierung zu überdenken und den Gitterabstand h zu vergrößern.

In Abbildung 3.2 sind die Eigenfunktionen der 1D Laplace-Matrix aufgetragen. Im Gegensatz zur Abbildung 2.2 wurden hier zusätzlich die Funktionen auf einem gröberen Gitter mit Gitterabstand $2h$ aufgetragen. Man erkennt, dass die schwach oszillierenden Funktionen auch noch auf dem groben Gitter hinreichend repräsentiert werden. Diese Funktionen werden in Zukunft als *glatte Funktionen* bezeichnet. Die oszillierenden Funktionen verlieren dagegen jegliche Information.

Diese Beobachtung führt auf eine Wahl von ω, so dass die Reduktionsrate von T_ω bei den Eigenvektoren für $k > \frac{n}{2}$ möglich groß ist, d.h. $\lambda_k(T_\omega) = 1 - 2\omega \sin^2(\frac{k\pi)}{2(n+1)})$ ist möglichst klein. Für $k > \frac{n}{2}$ und $s = \sin(\frac{\pi}{2(n+1)})$ ist $s^2 \geq \frac{1}{2}$ und damit gilt

$$1 - 2\omega < 1 - 2\omega + 2\omega s^2 \leq \lambda_k(T_\omega) \leq 1 - 2\omega s^2 < 1 - \omega.$$

Daher erhält man die optimale Reduktion der oszillierenden Eigenvektoren, wenn $\left| 1 - 2\omega \right| = \left| 1 - \omega \right|$ erfüllt ist, also für $\omega = \frac{2}{3}$. Hier werden die Vektoren um den Faktor $\frac{1}{3}$ reduziert.

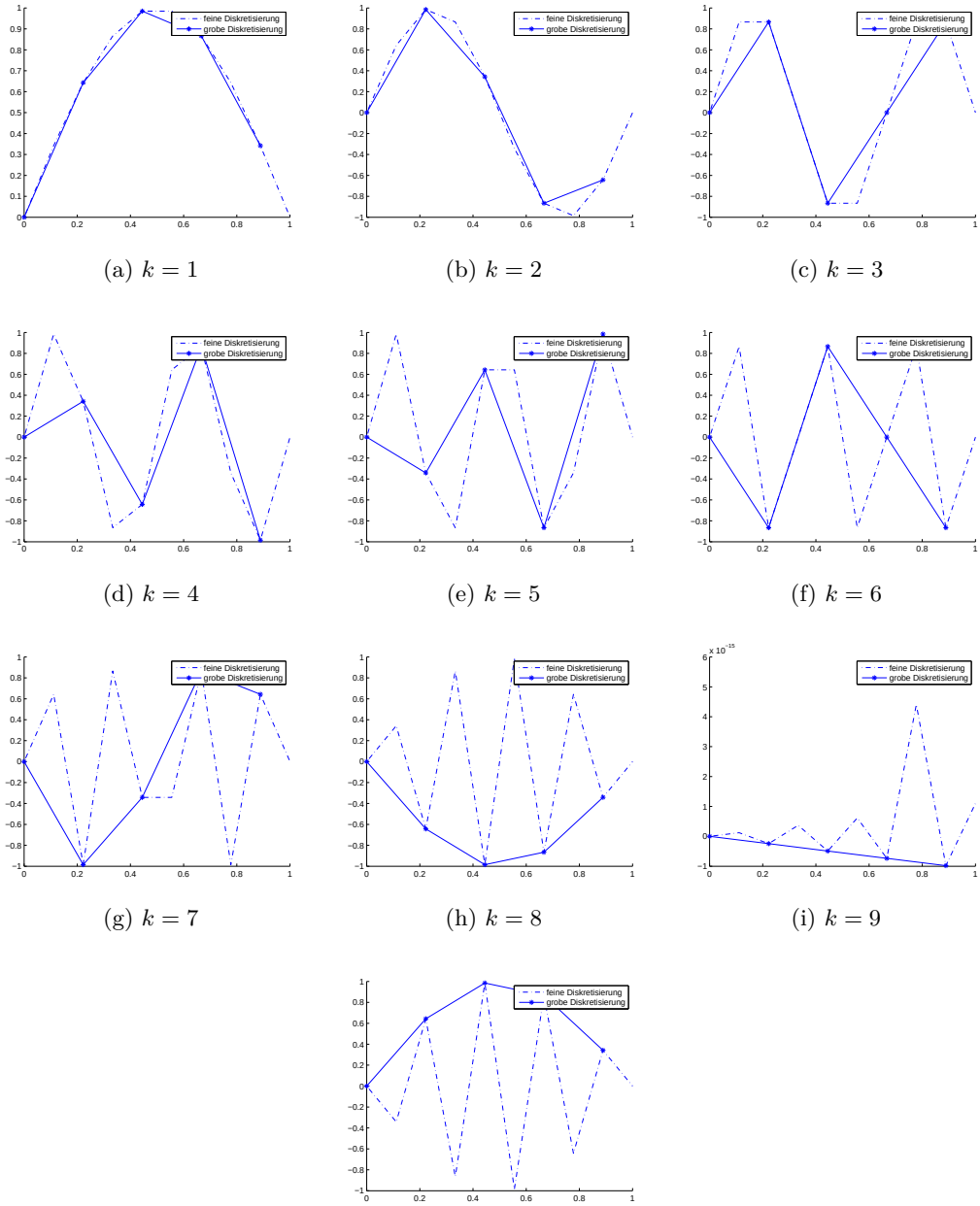
Im weiteren Verlauf werden solche einfachen stationären Verfahren wegen der gezeigten Eigenschaften als *Glätter* bezeichnet. Es werden abhängig vom Kontext sowohl M, als auch $I - M^{-1}A$ als Glätter betitelt.

Glätter werden durch den folgenden Algorithmus beschrieben. Dabei sei $M \in \mathbb{R}^{n,n}$ eine leicht zu invertierende Approximation an A.

Algorithmus 1: $u = \text{Glätter}^\nu(A, M, x_0, b)$

1 **for** $k = 1, \ldots, \nu$ **do**
2 $\quad r^{[k]} = b - Ax^{[k]}$,
3 $\quad x^{[k+1]} = x^{[k]} + M^{-1}r^{[k]}$.
4 **return** $x^{[\nu+1]}$

Obwohl bisher nur die 1D bzw. 2D Laplace-Gleichung betrachtet wurde, lassen sich viele der gezeigten Eigenschaften und Zusammenhänge auf andere Probleme übertragen. Dazu sei z.B. auf [38, 55] verwiesen.



(a) $k = 1$ (b) $k = 2$ (c) $k = 3$

(d) $k = 4$ (e) $k = 5$ (f) $k = 6$

(g) $k = 7$ (h) $k = 8$ (i) $k = 9$

(j) $k = 10$

Abbildung 3.2.: Die zehn diskretisierten Eigenfunktionen $u_k(x) = \sin(k\pi x)$ des 1D Laplace-Operators für $n = 10$ jeweils auf dem Originalgitter mit Gitterabstand h und dem gröberen Gitter mit Gitterabstand $2h$.

3.3

Mehrgitterverfahren. Um die wichtigsten Ideen von Mehrgitterverfahren zu erklären, werden die Beobachtungen aus Abschnitt 3.2 aufgegriffen. Es wurde festgestellt, dass gewisse Operationen den Fehler $e^{[k]} = x - x^{[k]}$ glätten und es daher erlaubt ist, auf einem gröberen Gitter zu operieren. Um die Hauptideen im Detail zu verstehen, werden zunächst nur zwei verschiedene Gitter betrachtet. Seien also Ω_h, Ω_H zwei diskretisierte Gebiete von Ω mit Gitterabstand h bzw. H. Für die Anschauung kann z.B. $H = 2h$ gewählt werden.

Geht man davon aus, dass der Fehler $e^{[k]} \in \Omega_h$ aus (3.2) von der Iterierten $x^{[k]}$ und der Lösung x genügend glatt ist, also wenig oszilliert, so kann die Größe des zu lösenden Problems maßgeblich reduziert, also ein Repräsentant von $e^{[k]}$ auf Ω_H gefunden werden. Es gibt viele denkbare Varianten, solch einen Repräsentanten anzugeben. Die wohl einfachste ist eine kanonische Einbettung (Injektion). Dafür sei auf dem feinen Gitter der Fehler mit $e^h := {e^{[k]}}^h$ und den Einträgen $[e_j^h]$ und analog auf dem groben Gitter mit $e^{2h} := {e^{[k]}}^{2h}$ und den Einträgen $[e_j^{2h}]$ bezeichnet. Dann soll gelten

$$e_j^{2h} = e_{2j}^h.$$

In Matrixform ergibt das die Gleichung

$$e^{2h} = R_h^{2h} e^h \tag{3.3}$$

mit

$$R_h^{2h} : \Omega_h \to \Omega_{2h}, \quad R_h^{2h} = [r_{jk}] = \begin{cases} 1, & k = 2j-1 \\ 0, & \text{sonst} \end{cases}.$$

Eine weitere und mehr gebräuchlichere Möglichkeit, genannt *full weighting (FW)*, kann wie folgt definiert werden

$$e_j^{2h} = \tfrac{1}{4}(e_{2j-1}^h + 2e_{2j}^h + e_{2j+1}^h)$$

bzw. erneut in der Matrixschreibweise

$$R_h^{2h} = \frac{1}{4}\begin{bmatrix} 1 & 2 & 1 & & & & & \\ & & 1 & 2 & 1 & & & \\ & & & 1 & 2 & 1 & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & 1 & 2 & 1 \end{bmatrix}.$$

Die Matrizen lassen sich ähnlich wie bei einem Diskretisierungs-Stern (vgl. Abbildungen 2.1, 2.3) abkürzend als $\frac{1}{4}\,[\,1\ 2\ 1\,]$ aufschreiben. Für ein analoges 2D Problem erhält man den Stern

$$\frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}.$$

Damit wurde das zu lösende Problem auf ein gröberes Gitter restringiert. Auf diesem Gitter muss das Problem schließlich gelöst bzw. die Lösung approximiert und anschließend das Ergebnis auf das feine Gitter interpoliert werden. Dieser Operator wird mit $P_H^h : \Omega_H \to \Omega_h$ bezeichnet. Auch hier gibt es verschiedene Möglichkeiten, diese Interpolation durchzuführen. Es soll im Eindimensionalen zunächst für $H = 2h$ gelten $e_{2j}^h = e_j^{2h}$. Für die hinzukommenden Punkte muss dann gewichtet interpoliert werden, z.B.

$$e_{2j+1}^h = \tfrac{e_j^{2h} + e_{j+1}^{2h}}{2}.$$

Dies ergibt $e^h = P_{2h}^h e^{2h}$ mit

$$P_{2h}^h = \frac{1}{2}\begin{bmatrix} 1 & & & & & & \\ 2 & & & & & & \\ 1 & 1 & & & & & \\ & 2 & & & & & \\ & 1 & 1 & & & & \\ & & \ddots & \ddots & & & \\ & & & 1 & 1 & & \\ & & & & 2 & & \\ & & & & 1 & & \end{bmatrix} \text{ und in der Stern-Schreibweise } \frac{1}{2}\Big]\ 1 \quad 2 \quad 1\ \Big[.$$

Die offenen Klammern bezeichnen die spaltenweise Schreibweise. So erhält man analog für den 2D Fall den Stern

$$\frac{1}{4}\Big]\begin{matrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{matrix}\Big[.$$

Die Verwendung von lediglich zwei Gittern ist für praktische Probleme im Allgemeinen jedoch ungeeignet, da das Grobgitter-Problem mit dem Gitterabstand H immer noch zu groß ist, um es exakt zu lösen. Überträgt man die vorgestellten Ideen auf mehrere Gitter, so erhält man ein Mehrgitterverfahren. Für die theoretische Analyse von Mehrgitterverfahren reicht es jedoch häufig aus, das zugehörige Zweigitterverfahren zu betrachten, da sich die Ergebnisse durch eine rekursive Anwendung auf weitere Gitter übertragen lassen.

Ein Mehrgitterverfahren basiert in der Regel auf zwei Phasen:

1. Es müssen geeignete Gitter bzw. Diskretisierungen auf diesen Gittern angegeben werden. Des Weiteren benötigt man Interpolations- und Restriktionsoperatoren. Diese erste Phase nennt man *Setup- Phase* und sie kann wie folgt beschrieben werden.

Algorithmus 2: Setup-Phase eines Mehrgitterverfahrens

Input : Ein Problem, z.B. eine partielle Differentialgleichung; Anzahl der Gitterebenen L, eine Gitterhierarchie $\Omega_{h_1}, \ldots, \Omega_{h_L}$.

Output : Systemmatrizen $A^{(k)}$, $k = 1, \ldots, L$; Interpolationen P_{k+1}^k, Restriktionen R_k^{k+1}, $k = 1, \ldots, L-1$; Glätter $M^{(k)}$, $k = 1, \ldots, L-1$.

```
for k = 1, ..., L do
    Finde durch Diskretisierung auf jedem Gitter k eine Systemmatrix A^(k). Hierbei
    ist A^(L) die Matrix auf dem gröbsten und A^(1) die auf dem feinsten Gitter.
for k = 1, ..., L-1 do
    Definiere geeignete Interpolationen P_{k+1}^k : Ω_{h_{k+1}} → Ω_{h_k} und Restriktionen
    R_k^{k+1} : Ω_{h_k} → Ω_{h_{k+1}} sowie Glätter M^(k).
```

2. Der zweite Teil eines Mehrgitterverfahrens ist die Lösungsphase. Diese wird durch den folgenden Algorithmus realisiert.

Algorithmus 3: $x^{(k)} = \text{Mehrgitter}(A^{(k)}, x_0^{(k)}, b^{(k)})$

1 $x^{(k)} = \text{Glätter}^{\nu_1}(A^{(k)}, M^{(k)}, x_0^{(k)}, b^{(k)})$, // *Vorglättung*
2 $r^{(k)} = b - A^{(1)}x^{(k)}$, // *Berechne das Residuum*
3 $r^{(k+1)} = R_k^{k+1} r^{(k)}$, //*Restringiere auf das gröbere Gitter*
4 **if** $k = L - 1$ **then**
5 Löse: $A^{(L)}x^{(L)} = r^{(L)}$, // *exakte Lösung auf dem gröbsten Gitter*
6 **else**
7 $x^{(k+1)} = \text{Mehrgitter}(A^{(k+1)}, 0, r^{(k+1)})$ // *Rekursion*
8 $x^{(k)} = x^{(k)} + P_{k+1}^k x^{(k+1)}$, //*Korrektur durch prolongierte Grobgitterlösung*
9 $x^{(k)} = \text{Glätter}^{\nu_2}(A^{(k)}, M^{(k)}, x^{(k)}, b)$, // *Nachglättung*
10 **return** $x^{(l)}$.

Mit Algorithmus 3 erhält man für $l = 1$ eine Approximation an $x = {A^{(1)}}^{-1} b^{(1)}$. Dazu werden ν_1 Vorglättungsschritte und ν_2 Nachglättungsschritte verwendet, welche durch Algorithmus 1 gegeben sind.

Wird $L = 2$ gesetzt, so erhält man den Algorithmus für das Zweigitterverfahren. Anstatt in 5 exakt zu lösen, ist es jedoch für den Zweigitterfall empfehlenswert, die Lösung $x^{(L)}$ zu approximieren.

Die resultierende Iterationsmatrix des Zweigitterverfahrens ist gegeben durch

$$T_{2g} := I - C_{2g}A := T_s^{\nu_2} \cdot T_c \cdot T_s^{\nu_1},$$

wobei hier T_s der Glätter (smoother) und T_c die sogenannte Grobgitterkorrektur (coarse grid correction) ist. Die beiden Matrizen haben folgende Gestalt

$$T_s := I - M^{-1}A, \tag{3.4}$$

$$T_c := I - PA_c^{-1}RA. \tag{3.5}$$

Die Matrix M wird normalerweise als $M = \text{diag}(A)$ (Jacobi), $M = \frac{1}{\omega} \cdot \text{diag}(A)$ (gewichtetes Jacobi) oder als obere/ untere Dreiecksmatrix von A (Gauß-Seidel) gewählt. Die Matrix $R = R_1^2 : \Omega_h \to \Omega_H$ beschreibt die Restriktion. Die Matrix $P = P_2^1$ stellt die Interpolation dar, also $P : \Omega_H \to \Omega_h$. Wenn aus dem Kontext der Bild- und Urbildbereich von P und R deutlich wird, wird auf die Indexierung verzichtet. Da in dieser Arbeit ausschließlich symmetrische bzw. hermitesche Probleme betrachten werden, wird $R = P^T$ bzw. $R = P^H$ gefordert.

Die Iterationsmatrix des Mehrgitterverfahrens kann am einfachsten rekursiv angegeben werden:

$$T_{\text{mg}}^{(1)} := I - C_{\text{mg}}^{(1)} A^{(1)}$$

wobei $C_{\text{mg}}^{(k)}$ für $k = 1, \ldots, L-1$ gegeben ist durch

$$I - C_{\text{mg}}^{(k)} A^{(k)} := T_s^{(k)\nu_2} (I - P_{k+1}^k C_{\text{mg}}^{(k+1)} P_{k+1}^{k^T} A^{(l)}) \cdot T_s^{(k)\nu_1}.$$

Die Matrix $C_{\text{mg}}^{(L)}$ ist dabei eine Approximation an ${A^{(L)}}^{-1}$ bzw. wenn exakt gelöst wird $C_{\text{mg}}^{(L)} = {A^{(L)}}^{-1}$ gesetzt. Außerdem sind

$$T_s^{(k)} := I - {M^{(k)}}^{-1} A^{(k)} \tag{3.6}$$

die Glätter auf dem k-ten Gitter.

Eine Abwandlung des Mehrgitterverfahrens wird durch eine mehrmalige Anwendung der Rekursion in Schritt 7 erreicht. Dies führt zur Lösungsphase, gegeben durch Algorithmus 4.

In diesem Algorithmus wird in Schritt 7 derselbige γ-mal durchgeführt. Für $\gamma = 1$ erhält man den Algorithmus 3 und nennt das resultierende Verfahren V-Zyklus, siehe Abbildung 3.3. Wählt man $\gamma = 2$, so erhält man den W-Zyklus, dargestellt in Abbildung 3.4. Im Allgemeinen nennt man diese Art von Rekursion γ-Zyklus.

Algorithmus 4: $x^{(k)} = \mathrm{MG}(A^{(k)}, x_0^{(k)}, b^{(k)})$

1 $x^{(k)} = \text{Glätter}^{\nu_1}(A^{(k)}, M^{(k)}, x_0^{(k)}, b^{(k)})$, *// Vorglättung*
2 $r^{(k)} = b - A^{(k)}x^{(k)}$, *// Berechne das Residuum*
3 $r^{(k+1)} = R_k^{k+1} r^{(k)}$, *//Restringiere auf das gröbere Gitter*
4 **if** $k = L - 1$ **then**
5 Löse: $A^{(L)}x^{(L)} = r^{(L)}$, *// exakte Lösung auf dem gröbsten Gitter*
6 **else**
7 $x^{(k+1)} = \mathrm{MG}^{\gamma}(A^{(k+1)}, 0, r^{(k+1)})$ *// Rekursion*
8 $x^{(k)} = x^{(k)} + P_{k+1}^{k} x^{(k+1)}$, *//Korrektur durch prolongierte Grobgitterlösung*
9 $x^{(k)} = \text{Glätter}^{\nu_2}(A^{(k)}, M^{(k)}, x^{(k)}, b)$, *//Nachglättung*
10 **return** $x^{(l)}$.

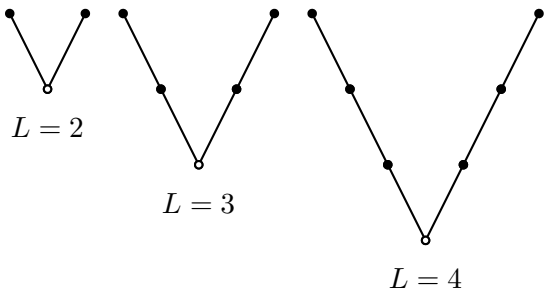


Abbildung 3.3.: Veranschaulichung des V-Zyklus' für verschiedene Gitterhierarchien.

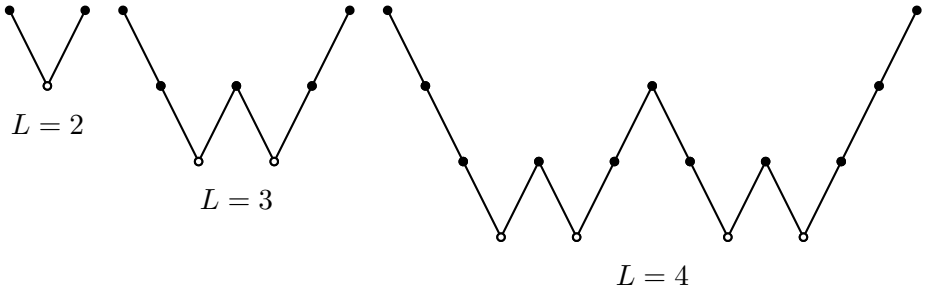


Abbildung 3.4.: Veranschaulichung des W-Zyklus' für verschiedene Gitterhierarchien.

Die Iterationsmatrizen für die verschiedenen Zyklen lassen sich sehr elegant über die Verwendung eines Polynoms $P_\gamma(t) = (1-t)^\gamma$ definieren. Für den allgemeinen γ-Zyklus erhält man

$$T^{(1)}_{\mathrm{mg},\gamma} := I - C^{(1)}_{\mathrm{mg},\gamma} A^{(1)}$$

wobei für $k = 1, \ldots, L-1$ die Matrizen $C^{(k)}_{\mathrm{mg},\gamma}$ rekursiv durch

$$I - C^{(k)}_{\mathrm{mg},\gamma} A^{(k)} := T_s^{(k)^{\nu_2}} (I - P^k_{k+1} S^{(k+1)^{-1}}_{\mathrm{mg}_\gamma} P^{k^T}_{k+1} A^{(k)}) \cdot T_s^{(k)^{\nu_1}} \quad \text{mit} \tag{3.7}$$

$$S^{(k+1)}_{\mathrm{mg},\gamma} := A^{(k+1)} \Big[I - (I - C^{(k+1)}_{\mathrm{mg},\gamma} A^{(k+1)})^\gamma \Big]^{-1} \tag{3.8}$$

gegeben sind. Die Matrix $C^{(L)}_{\mathrm{mg},\gamma}$ ist erneut eine Approximation an $A^{(L)^{-1}}$ bzw. durch $C^{(L)}_{\mathrm{mg}} = A^{(L)^{-1}}$ definiert. Die Glätter $T_s^{(k)}$ auf dem k-ten Level sind gegeben durch (3.6). Für $\gamma = 1$ ergibt sich $T^{(1)}_{\mathrm{mg},\gamma} = T^{(1)}_{\mathrm{mg}}$.

In den 80er und 90er Jahren wurde daraus ein Ansatz entwickelt, der erst wieder in den letzten Jahren aufgegriffen wurde, siehe [60, 61]. Die Idee stammt von Axelsson und Vassilevski [11] und beinhaltet die Verwendung von Tschebysheff Polynomen für die Konstruktion der Grobgittermatrix. Man betrachtet in (3.8) anstatt des Polynoms $(1-t)^\gamma$ ein beliebiges Polynom $P_\gamma(t)$ vom Grad γ und den Eigenschaften $P(0) = 1$ sowie $P_\gamma(t) \neq 1$ auf $\sigma(C^{(k+1)}_{\mathrm{mg},\gamma} A^{(k+1)})$ und erhält

$$S^{(k+1)}_{\mathrm{mg},\gamma} = A^{(k+1)} \Big[I - P_\gamma(C^{(k+1)}_{\mathrm{mg},\gamma} A^{(k+1)}) \Big]^{-1}. \tag{3.9}$$

Der nächste Abschnitt behandelt bestimmte Polynome, die den Ausdruck $1 - P(t)$ möglichst auf 1 skalieren sollen bzw. die auf den Eigenwerten von $C^{(k+1)}_{\mathrm{mg},\gamma} A^{(k+1)}$ minimal sind.

3.4 Tschebysheff Polynome.

Die Tschebysheff Polynome spielen in vielen Bereichen der numerischen linearen Algebra eine wichtige Rolle. Ihre Darstellung in diesem Abschnitt basiert auf [86].

Definition 3.2. Das *reelle Tschebysheff Polynom erster Art vom Grad* γ wird für $t \in [-1, 1]$ definiert durch

$$T_\gamma(t) := \cos\big(\gamma \cos^{-1}(t)\big).$$

Eine äquivalente Darstellung erhält man durch die trigonometrische Identität

$$\cos\big((\gamma+1)\theta\big) + \cos\big((\gamma-1)\theta\big) = 2\cos(\theta)\cos(\gamma\theta) \quad \text{für alle}\, \gamma \geq 1,\ \theta \in [0, 2\pi)$$

und mittels der Substitution $\theta = \cos^{-1}(t)$. Es ergibt sich die rekursive Darstellung

$$T_{\gamma+1}(t) = 2tT_\gamma(t) - T_{\gamma-1}(t) \quad \text{mit} \quad T_0(t) = 1,\ T_1(t) = t. \tag{3.10}$$

Damit kann der Definitionsbereich von T_γ für $|\,t\,| > 1$ erweitert werden. Eine besondere Eigenschaft der Tschebysheff Polynome ist eine gewisse Minimierungseigenschaft auf dem Intervall $[-1, 1]$, die durch eine Translation auf ein beliebiges Intervall $[a, b]$ überführt werden kann. Die Darstellung dieses Satz findet sich in [86, Theorem 6.25] wieder.

Satz 3.3. *Sei $c \in \mathbb{R}$ mit $c \notin [a,b]$ mit $a < b$, dann wird das Minimum*

$$\min_{p \in \mathbb{R}^c_{\leq \gamma}[t]} \max_{t \in [a,b]} |\, p(t) \,|$$

durch das Polynom

$$\Pi^c_\gamma(t) := \frac{T_\gamma(\frac{b+a-2t}{b-a})}{T_\gamma(\frac{b+a}{b-a})}$$

angenommen. Hierbei ist $\mathbb{R}^c_{\leq \gamma}[t] := \{p \in \mathbb{R}_{\leq \gamma}[t]$ mit $p(c) = 1\}$ und $\mathbb{R}_{\leq \gamma}[t]$ bezeichnet die Menge aller Polynome mit reellen Koeffizienten vom Grad kleiner oder gleich γ.

In [42, Seite 61] wird diese Aussage für das Intervall $[-1,1]$ bewiesen. Die allgemeinere Aussage für das Intervall $[a,b]$ erfolgt durch eine Variablentransformation $\widetilde{t} = \frac{b+a-2t}{b-a}$, siehe [6, Theorem B.1].

Durch ihre Minimierungseigenschaft nehmen die Tschebysheff Polynome eine wichtige Rolle im späteren Verlauf dieser Arbeit ein. Für den Spezialfall $c = 0$ und $b > a > 0$ ergibt sich

$$\Pi^0_\gamma(t) = \frac{T_\gamma(\frac{b+a-2t}{b-a})}{T_\gamma(\frac{b+a}{b-a})}. \tag{3.11}$$

Beispiel 3.4. Betrachtet man die Matrix $S^{(k+1)}_{\mathrm{mg},\gamma}$ aus (3.9) und nimmt an, dass $C^{(k+1)}_{\mathrm{mg},\gamma}$ und $A^{(k+1)}$ spd sind, dann gilt für das Spektrum von $C^{(k+1)}_{\mathrm{mg},\gamma} A^{(k+1)}$, dass $\sigma(C^{(k+1)}_{\mathrm{mg},\gamma} A^{(k+1)}) \subset \mathbb{R}$. Das Polynom P_γ sollte dann möglichst auf einem Intervall $[a,b]$ mit $\sigma(C^{(k+1)}_{\mathrm{mg},\gamma} A^{(k+1)}) \subset [a,b]$ minimal sein. Daher ist die optimale Wahl des Polynoms durch (3.11) gegeben.

Es stellt sich die Frage, ob auch ein geeignetes Polynom gewählt werden kann, wenn $C^{(k+1)}_{\mathrm{mg},\gamma}$ unsymmetrisch ist. Damit befinden sich die Eigenwerte von $C^{(k+1)}_{\mathrm{mg},\gamma} A^{(k+1)}$ nicht mehr auf der reellen Achse. Mit der Annahme, dass die beiden Matrizen ein konvergentes Iterationsverfahren induzieren, d.h. es gilt

$$\left\| I - C^{(k+1)}_{\mathrm{mg},\gamma} A^{(k+1)} \right\|_{A^{(k+1)}} = \rho < 1, \tag{3.12}$$

liegen die Eigenwerte von $C^{(k+1)}_{\mathrm{mg},\gamma} A^{(k+1)}$ in $B_\rho(1) := \{t \in \mathbb{C} \text{ mit } |\, t - 1 \,| \leq \rho\}$.

Da die Minimierungseigenschaft der Tschebysheff Polyonome die für diese Arbeit zentrale Eigenschaft dieser Polynome ist, wird eine verallgemeinerte und in der komplexen Analysis übliche Definition der Tschebysheff Polynome eingeführt, siehe z.B. [88, Seite 46] oder [95, Seite 89].

Definition 3.5. Sei $Q \subseteq \mathbb{C}$. Ein Polynom $\Pi^c_\gamma\{Q\} \in \mathbb{C}^c_{\leq \gamma}[t]$ heißt *in $c \in \mathbb{R}$ normiertes Tschebysheff Polynom auf Q vom Grad höchstens γ*, wenn es für $t \in \mathbb{C}$ die Minimierungseigenschaft

$$\max_{t \in Q} |\, \Pi^c_\gamma\{Q\}(t) \,| = \min_{p \in \mathbb{C}^c_{\leq \gamma}[t]} \max_{t \in Q} |\, p(t) \,|,$$

erfüllt. Hierbei werden $\mathbb{C}^c_{\leq \gamma}[t]$ und $\mathbb{C}_{\leq \gamma}[t]$ analog zu den reellen Pandons definiert.

Für $Q = [-1, 1]$ stimmt Definition 3.5 durch Anwendung von Satz 3.3 mit der Definition 3.2 überein. Außerdem ist für $Q = [a, b]$

$$\Pi_\gamma^0\{[a,b]\}(t) = \frac{T_\gamma\left(\frac{b+a-2t}{b-a}\right)}{T_\gamma\left(\frac{b+a}{b-a}\right)}. \tag{3.13}$$

Die Gleichung (3.12) verlangt eine Untersuchung, wie das auf c normierte Tschebysheff Polynom für $Q = B_\rho(1)$ mit $\rho < 1$ aussieht. Dazu ist das folgende Resultat hilfreich, dass in dieser Form in [86, Lemma 6.26] wiedergefunden werden kann.

Satz 3.6 (Zarantonello)**.** *Sei $B_\rho(0) \subset \mathbb{C}$ ein Kreis um den Ursprung mit Radius ρ und $c \notin B_\rho(0)$, dann gilt*

$$\min_{p \in \mathbb{C}^c_{\leq \gamma}[t]} \max_{t \in B_\rho(0)} |\, p(t) \,| = \max_{t \in B_\rho(0)} \left(\frac{t}{c}\right)^\gamma = \left(\frac{\rho}{|c|}\right)^\gamma.$$

Ein Beweis kann in [88, Theorem 3, Seite 367] gefunden werden. Durch Verwendung einer Variablentransformation erhält man das gewünschte Resultat für die Minimierung auf $B_\rho(1)$, also

$$\min_{p \in \mathbb{C}^c_{\leq \gamma}[t]} \max_{t \in B_\rho(1)} |\, p(t) \,| = \max_{t \in B_\rho(1)} \left(\frac{t-1}{c-1}\right)^\gamma.$$

Daher ist

$$\Pi_\gamma^0\{B_\rho(1)\}(t) = (1-t)^\gamma.$$

Kapitel 4

Algebraische Mehrgitterverfahren

Im vorangegangenen Kapitel wurde bei der Einführung der Mehrgitterverfahren die Bedeutung der Gitterhierarchie sowie der zugrunde liegenden partiellen Differentialgleichung (partial differential equation, kurz PDE) deutlich. Jedoch kann nicht immer ein geeignetes Gitter und damit verbundene Operatoren angegeben werden. Sei es wegen eines schwer zu diskretisierenden Gebietes Ω oder, dass für das gegebene Problem keine partielle Differentialgleichung zu Grunde liegt bzw. diese unbekannt ist. Abhilfe schaffen algebraische Varianten von Mehrgitterverfahren, die erstmals in einer Arbeit von Brandt, McCormick und Ruge [30] eingeführt wurden und ein Jahr später durch Ruge und Stüben [85] Bekanntheit erlangten. In diesem Kapitel werden zunächst die Hauptideen algebraischer Mehrgitterverfahren (AMG) anhand grundlegender Konzepte von Ruge und Stüben erläutert. Einer der wesentlichen Unterschiede zu den Mehrgitterverfahren aus Abschnitt 2 ist, dass die Matrizen $A^{(k)}$ nicht mehr über eine Diskretisierung gröberer Gitter erklärt werden können. Des Weiteren können auch Interpolationen und Restriktionen nicht über die Gitterhierarchie definiert werden und der Begriff der glatten Fehler, so wie er definiert wurde, macht ohne eine Geometrie nur noch wenig Sinn.

Das AMG von Ruge und Stüben lieferte die ersten Ideen zur Entwicklung eines Mehrgitterverfahrens ohne Nutzung der Geometrien. Eine detaillierte Darstellung dieses Verfahrens kann unter anderem in [51] gefunden werden. Dieses AMG liefert auch heute noch die Grundlagen für immer weitere Abwandlungen. Im Anschluss an diese Schilderung werden einige wichtige Varianten verschiedener AMG's vorgestellt, wie z.B. *Multilevel-Block-Faktorisierungs-Verfahren* oder auch das sogenannte *reduction-based AMG (AMGr)*. Diese Verfahren nehmen im weiteren Verlauf dieser Arbeit eine zentrale Rolle ein.

4.1

Grobgittermatrix, Interpolationsmatrix und Coarsening. Im Folgenden wird erläutert, wie ein algebraisches Mehrgitterverfahren konstruiert werden kann. Ein wesentlicher Unterschied zu den bisher betrachteten Mehrgitterverfahren ist, dass die Matrizen $A^{(k)}$ auf den verschiedenen Gittern nicht mehr durch eine geeignete Diskretisierung entstehen. Geht man davon aus, dass eine Interpolation P_{k+1}^k von Gitter $k+1$ zu Gitter k bereits gefunden wurde, so ist ein häufig gewählter Ansatz für die Konstruktion der Matrix auf dem nächst gröberen Gitter der sogenannte *Galerkin-Ansatz*[1]. Hier wird vorausgesetzt, dass A symmetrisch und damit die entsprechende Restriktion durch $R_k^{k+1} = \left(P_{k+1}^k\right)^T$ definiert ist. Setzt man auf dem $(k+1)$-ten Gitter die Matrix als

$$A^{(k+1)} = (P_{k+1}^k)^T A^{(k)} P_{k+1}^k,$$

[1] Benannt nach Boris Grigorjewitsch Galjorkin (1871-1945), einem sowjetischen Ingenieur und Mathematiker.

so nennt man $A^{(k+1)}$ die *Galerkin-Matrix*. Im Laufe dieser Arbeit werden noch weitere Möglichkeiten dargestellt, eine Gitterhierarchie aufzubauen; jedoch werden in diesem Abschnitt AMG's anhand des Galerkin-Ansatzes eingeführt. Der Einfachheit halber werden ausschließlich Zweigitterverfahren betrachtet. Die Verallgemeinerung auf Mehrgitterverfahren erfolgt rekursiv, wie im letzten Kapitel beschrieben.

Im nächsten Satz wird dargelegt, dass der Galerkin-Ansatz für spd Matrizen als optimal interpretiert werden kann. Dazu wird eine bestimmte Norm, die A-Norm oder auch Energie-Norm, benötigt, die sich wie folgt definiert.

Definition 4.1. Sei $A \in \mathbb{C}^{n,n}$ hermitesch positiv definit (hpd), dann ist die *A-Norm eines Vektors* $x \in \mathbb{C}^n$ definiert durch

$$\|x\|_A := \sqrt{x^H A x}.$$

Die zugehörige *Matrix-Norm* ist für $B \in \mathbb{C}^{n,n}$ gegeben durch

$$\|B\|_A := \sup_{\|x\|_A=1} \|Bx\|_A.$$

Damit erhält man das folgende Resultat, was allgemein bekannt ist und in einer äquivalenten Form in [67] wiedergegeben wird. Zur Vollständigkeit wird der Beweis vorgeführt.

Satz 4.2. *Sei $A \in \mathbb{R}^{n,n}$ spd und $P \in \mathbb{R}^{n,n_1}$ eine Interpolationsmatrix mit* $\mathrm{Rang}(P) = n_1$, *$n_1 < n$. Dann gilt für alle $x \in \mathbb{R}^n$*

$$\|x + Py\|_A = \min \quad \Leftrightarrow \quad y = -(P^T A P)^{-1} P^T A x.$$

Beweis. Zuerst sei angemerkt, dass $P^T A P$ spd und damit nichtsingulär ist, da P vollen Rang hat. Weiter ist

$$\|x + Py\|_A = x^T A x + 2y^T P^T A x + y^T P^T A P y =: f(y).$$

Die Funktion f wird extremal, wenn

$$0 = f'(y) = 2P^T A x + 2P^T A P y,$$

also für $y_0 = -(P^T A P)^{-1} P^T A x$. Da $f''(y_0) = 2P^T A P$ positiv definit ist, befindet sich bei y_0 ein Minimum. □

Satz 4.2 besagt insbesondere, dass zur Minimierung der A-Norm die Iterationsmatrix $T_c = I - P(P^T A P)^{-1} P^T A$ zu wählen ist. Vergleicht man dies mit (3.5), so erhält man als optimale Wahl der Grobgittermatrix $A_c = P^T A P$.

Anschließend wird untersucht, wie eine geeignete Interpolationsmatrix P gewählt wird. In Kapitel 3 wurde gezeigt, dass es gewisse Anteile im Fehler $e^{[k]} = x - x^{[k]}$ gibt, die durch den Glätter nur langsam reduziert werden. Es konnte beobachtet werden, dass diese Anteile zu entsprechend glatten Eigenvektoren von A gehören. Diese Glattheit hängt von der gewählten Diskretisierung und insbesondere vom Gitterabstand h ab. Daher kann diese Beobachtung nicht auf algebraische Mehrgitterverfahren übertragen werden. Trotzdem gibt es weiterhin entsprechende Anteile im Fehler $e^{[k]}$, die nur langsam reduziert werden. Diese hängen jedoch von dem gewählten Glätter $T_s := I - M^{-1} A$ ab. Mithilfe dieses Glätters lassen sich algebraisch glatte Vektoren definieren.

Definition 4.3. Sei $A \in \mathbb{R}^{n,n}$ spd und $T_s = I - M^{-1}A$ ein Glätter. Dann heißt $e \in \mathbb{R}^n$ ein *glatter Fehler bzgl.* T_s, wenn gilt

$$\|T_s e\|_A \approx \|e\|_A .$$

Es gibt weitere Varianten, glatte Fehler zu definieren, jedoch ist Definition 4.3 die geläufigste und kann z.B. in [85] und [86] wiedergefunden werden. Dabei ist zu beachten, dass die Glattheit von Vektoren im Gegensatz zu geometrisch glatten Vektoren nicht mehr von geometrischen Gegebenheiten eines Gitters abhängt, sondern einzig von der Wahl des Glätters.

Betrachtet man einen in seinen glatten und seinen nicht glatten, genannt oszillierenden, Anteil aufgeteilten Fehler, also $e = e_{glatt} + e_{osz}$, und nimmt an, dass der Glätter T_s den Anteil e_{osz} weitestgehend eliminiert, also $T_s e \in \{e_{glatt} \in \mathbb{R}^n \mid \|T_s e_{glatt}\|_A \approx \|e_{glatt}\|_A\}$, dann resultiert daraus eine Eigenschaft, die von der Interpolationsmatrix P erfüllt werden sollte: Betrachtet man die Iterationsmatrix $T_{2g} = T_c \cdot T_s$, so ist

$$T_{2g} e = 0,$$

falls

$$\{e_{glatt} \in \mathbb{R}^n \mid \|T_s e_{glatt}\|_A \approx \|e_{glatt}\|_A\} \in \text{Bild}(P). \tag{4.1}$$

Denn unter dieser Voraussetzung existiert ein x mit $Px = T_s e$ und daher gilt

$$T_{2g} e = (I - P(P^T AP)^{-1} P^T A)Px = Px - Px = 0.$$

Für die Interpolation P wird folglich gefordert, dass P vollen Rang hat und die algebraisch glatten Vektoren möglichst im Bild von P liegen.

Bisher ist ungeklärt, auf welchen Raum P abbildet. Den Prozess, diesen Raum zu finden, nennt man *Coarsening* oder *Coarsing-Prozess*. Dazu werden die Unbekannten in sogenannte *Fein* (F)- und *Grobgitterunbekannte* ($\mathcal{C}$) aufgeteilt, was auf verschiedene Weisen geschehen kann. Eine der ersten Varianten war die von Ruge und Stüben [85].

Aufgrund algebraischer Umformungen und der Annahme, dass A eine symmetrische M-Matrix ist, d.h. $A = A^T$ nichtsingulär, A^{-1} hat nur nicht negative Einträge und die Nichtdiagonaleinträge von A sind nicht positiv, kann gezeigt werden, dass für glatte Fehler $e = [e_j]$ gilt

$$\sum_{k \neq j} \frac{|a_{kj}|}{a_{jj}} \frac{(e_j - e_k)^2}{e_j^2} \ll 1. \tag{4.2}$$

Eine Herleitung von 4.2 findet sich in [86, Seite 457 (13.70)].

Dies bedeutet, dass Verbindungen zwischen j und k charakterisiert werden können. Man sagt, dass zwischen j und k eine starke Verbindung besteht, wenn der Faktor

$$\frac{|a_{jk}|}{a_{jj}}$$

hinreichend groß ist. Mit Hilfe eines zu wählenden Parameters θ, kann eine Einteilung wie folgt vorgenommen werden: k hat eine starke Verbindung zu j, wenn

$$\frac{|a_{jk}|}{a_{jj}} \geq \theta.$$

Für diese starken Verbindungen erhält man durch (4.2), dass

$$\frac{(e_j - e_k)^2}{e_j^2}$$

klein ist, also die j-te und k-te Komponente des Fehlers nur gering variiert. Daher lässt sich der Eintrag e_k gut durch e_j interpolieren. Ziel ist es, eine Einteilung von $\mathcal{U} := \{1, \ldots, n\}$ in feine und grobe Unbekannte zu finden, so dass für jedes $j \in F$ mindestens ein $k \in \mathcal{C}$ existiert, welches sich eine starke Verbindung mit j teilt. Im Laufe der letzten 30 Jahre haben sich viele Varianten entwickelt, die eine Einteilung von $\mathcal{U}$ mithilfe starker Verbindungen vornehmen. Der Fokus dieser Arbeit soll aber unter anderem darauf liegen, auch komplexwertige Matrizen zu behandeln. Dies fällt bei der Charakterisierung mit starken Verbindungen im Allgemeinen schwer, siehe dazu [69].

Ungeachtet dessen wird ein Coarsing-Prozess benötigt. Die Methode der Wahl ist der sogenannte *Greedy-Coarser* aus [71], der in Kapitel 6 eingeführt wird. Ist ein Coarsing-Prozess durchgeführt worden, erhält man eine Permutation der Matrix A, gegeben durch

$$\pi A \pi^T = \begin{bmatrix} A_{FF} & A_{FC} \\ A_{CF} & A_{CC} \end{bmatrix},$$

wobei die Permutation $\pi \in \mathbb{R}^{n,n}$ so konstruiert wird, dass die Unbekannten x in Fein-und Grobgitter-Unbekannte aufgeteilt sind, also $\pi^T x = \left[\begin{smallmatrix} x_F \\ x_C \end{smallmatrix}\right]$ erfüllt ist.

Abgesehen von den hier vorgestellten Abweichungen entsprechen algebraische Mehrgitterverfahren den im letzten Kapitel vorgestellten Mehrgitterverfahren. Insbesondere kann weiterhin der Algorithmus 3 für die Lösungsphase verwendet werden. Die Setup-Phase wird jedoch durch eine algebraische Variante ersetzt, in der auch der zugrunde liegende Coarser verwendet wird. Dies führt zu einem höheren Rechenaufwand der algebraischen Mehrgitterverfahren, im Vergleich zu Mehrgitterverfahren, die auf einer Geometrie basieren.

Im Verlauf dieses Abschnittes werden ein paar Vereinbarungen getroffen. In Kapitel 3 und auch in diesem Abschnitt wurden ausschließlich reellwertige Matrizen betrachtet, jedoch lassen sich die vorgestellten Zusammenhänge problemlos auf komplexwertige Matrizen verallgemeinern. Da zum Ende der Arbeit das Modellproblem die Gauge-Laplace-Matrix aus Abschnitt **??** untersucht werden soll, ist A in Zukunft eine hermitesch positiv definite (hpd) Matrix, die bereits eine Aufteilung

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \in \mathbb{C}^{n,n} \tag{4.3}$$

mit $A_{11} \in \mathbb{C}^{n_1,n_1}$, $A_{22} \in \mathbb{C}^{n_2,n_2}$ und $n = n_1 + n_2$ besitzt. Die häufig verwendete Darstellung $A = \left[\begin{smallmatrix} A_{FF} & A_{FC} \\ A_{CF} & A_{CC} \end{smallmatrix}\right]$ wird vermieden, um deutlich zu machen, dass der verwendete Ansatz keine starken Verbindungen benötigt, es daher keine „typische“ Aufteilung in Fein-und Grobgitterpunkte gibt. Des Weiteren wird der Begriff „Gitter“ durch „Level“ ersetzt, um dem Leser vor Augen zu führen, dass keine Gitter mehr im herkömmlichen Sinn für die Konstruktion der hier zu untersuchenden Verfahren benötigt werden. Einzig in der Bezeichnung „algebraisches Mehrgitterverfahren“, kurz AMG, wird weiterhin dieses Wort verwendet, da die Bezeichnung AML eher unüblich und eine Verwechslung mit dem sogenannten *AMLI-Verfahren*, welches in Kapitel 5 eingeführt wird, möglich ist.

Zum Ende dieses Kapitels werden zwei Varianten von AMG's besprochen, die ihre Gemeinsamkeit in der Nutzung eines sogenannten *F-Glätters* haben. Dies ist ein Glätter, der ausschließlich die Fehler glättet, die nicht mit den Grobgitterpunkten assoziiert werden können, was nicht mit einem Glätter zu verwechseln ist, der auf einem feinen Gitter operiert, um die oszillierenden Anteile zu eliminieren. Vielmehr arbeitet dieser auf Fehleranteilen, die durch die Grobgittermatrix unberührt bleiben. Das hat zur Folge, dass der numerische Aufwand verkleinert wird, da die glatten Anteile des Fehlers vom Glätter nicht erfasst werden. Für eine Aufteilung wie in (4.3) hat solch ein Glätter die Gestalt

$$T_{s_F} = I - \begin{bmatrix} M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A, \tag{4.4}$$

wobei M_s eine leicht zu invertierende Approximation an A_{11} sein soll.

4.2

F-Glättung. Eine F-Glättung ist eine besondere Art der Glättung, welche unter anderem in [33, 68, 77, 78] betrachtet wird. Für den Rest der Arbeit sei $A \in \mathbb{C}^{n,n}$ hpd und abkürzend wird dies wie folgt beschrieben: Wenn

$$x^H B x \geq x^H C x$$

für alle $x \in \mathbb{C}^{n,}$ und zwei hpd Matrizen $B, C \in \mathbb{C}^{n,n}$ erfüllt ist, wird die Notation $B \succeq C$ verwendet. In ähnlicher Weise werden $B \succ C, B \prec C, B \preceq C$ definiert. Für A gilt somit $A \succ 0$. Weiter sei A wie in (4.3) partitioniert.

Wie im letzten Abschnitt besprochen (siehe Definition 4.3), hängen die algebraisch glatten Vektoren von der Wahl des Glätters ab. Hier wird ein Glätter der Form (4.4), der nur auf einem Teil der Unbekannten operiert, betrachtet. Für die algebraisch glatten Vektoren $v = \left[\begin{smallmatrix} v_1 \\ v_2 \end{smallmatrix}\right] \in \mathbb{C}^n$ soll gelten $T_{s_F} v = v$, also dass

$$\left(I - \begin{bmatrix} M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A \right) v = v \Leftrightarrow \begin{bmatrix} M_s^{-1} A_{11} v_1 + M_s^{-1} A_{12} v_2 \\ 0 \end{bmatrix} = 0$$

bzw. $v_1 = -A_{11}^{-1} A_{12} v_2$ erfüllt ist. Diese Vektoren werden vom Glätter nicht verändert und spielen daher bei der Konstruktion der Grobgitterkorrektur eine zentrale Rolle. Wie zuvor besprochen (siehe (4.1)), sollen glatte Vektoren möglichst im Bild der Interpolationsmatrix P liegen, d.h.

$$P_{\text{opt}} = \begin{bmatrix} -A_{11}^{-1} A_{12} \\ I \end{bmatrix} \tag{4.5}$$

ist die bestmögliche Wahl. Daher wird diese Interpolationsmatrix als *optimal* bezeichnet.

Betrachtet man eine Zerlegung eines Vektors $e \in \mathbb{C}^n$ in seinen glatten und oszillierenden Anteil

$$e = \underbrace{\begin{bmatrix} -A_{11}^{-1} A_{12} \\ I \end{bmatrix} v}_{=e_{\text{glatt}}} + \underbrace{\begin{bmatrix} I \\ 0 \end{bmatrix} w}_{=e_{\text{osz}}}$$

mit eindeutig bestimmten $v \in \mathbb{C}^{n_2}$, $w \in \mathbb{C}^{n_1}$, wird deutlich, dass ein algebraisches Mehrgitterverfahren mit F-Glättung eine Grobgitterkorrektur benötigt, die e_{glatt} ausreichend reduziert,

da der Glätter nur auf den oszillierenden Anteil wirkt, den glatten Anteil aber weitestgehend unberührt lässt. Der Glätter benötigt zur Reduktion des oszillierenden Anteils eine leicht zu invertierende Matrix M_s, so dass $\rho(I - M_s^{-1}A_{11})$ möglichst klein ist.

Da im Allgemeinen die Berechnung der Inversen von A_{11} sehr aufwändig ist, wird diese in (4.5) durch eine weitere, leicht zu invertierende Matrix $D_p \in \mathbb{C}^n$ ersetzt, wobei D_p hpd sein soll. Für die Anschauung kann D_p als diagonal vorausgesetzt werden. Dann ist die Grobgitterkorrektur durch

$$T_c = I - \begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix} A_c^{-1} \begin{bmatrix} -A_{21}D_p^{-1} & I \end{bmatrix} A$$

gegeben. Hier ist A_c die Matrix auf dem groben Level. Bezüglich der Minimierung der A-Norm wurde in Satz 4.2 festgestellt, dass die beste Wahl $A_c = \left[\begin{smallmatrix} -A_{21}D_p^{-1} & I \end{smallmatrix}\right] A \left[\begin{smallmatrix} -D_p^{-1}A_{12} \\ I \end{smallmatrix}\right]$ ist. Allerdings berücksichtigt diese Optimalität nicht, dass die Grobgitterkorrektur nur gewisse Vektoren reduzieren muss. Durch das Ersetzen von A_{11} durch D_p in P_{opt} entsteht nicht nur eine Bedingung für D_p an A_{11}, sondern auch eine für A_c. Betrachtet man die Wirkung der Grobgitterkorrektur auf die glatten Vektoren, so erhält man

$$\begin{aligned}
&\left(I - \begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix} A_c^{-1} \begin{bmatrix} -A_{21}D_p^{-1} & I \end{bmatrix} A\right) \begin{bmatrix} -A_{11}^{-1}A_{12} \\ I \end{bmatrix} v \\
&= \begin{bmatrix} -A_{11}^{-1}A_{12} \\ I \end{bmatrix} v - \begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix} A_c^{-1}(A_{22} - A_{21}A_{11}^{-1}A_{12})v \\
&= \left(\begin{bmatrix} -(A_{11}^{-1} - D_p^{-1})A_{12} \\ I \end{bmatrix} + \begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix} (I - A_c^{-1}(A_{22} - A_{21}A_{11}^{-1}A_{12}))\right) v. \qquad (4.6)
\end{aligned}$$

Um glatte Fehler zu reduzieren, wird zum einen eine gute Approximation D_p an A_{11}, als auch eine Grobgittermatrix A_c, die das Schurkomplement von A bzgl. A_{11} approximiert, benötigt. Für $D_p = A_{11}$ sind trivialerweise beide Bedingungen erfüllt. Jedoch bedeutet ein gut durch D_p approximiertes A_{11} nicht automatisch, dass

$$A_c = \begin{bmatrix} -A_{21}D_p^{-1} & I \end{bmatrix} A \begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix}$$

auch $A_{22} - A_{21}A_{11}^{-1}A_{12}$ ausreichend gut approximiert. Im späteren Stadium dieser Arbeit wird dieses Problem näher erläutert und die Güte einer solchen Approximation charakterisiert. Da das Schurkomplement im Weiteren eine zentrale Rolle einnimmt, wird abkürzend

$$S(A, B_{11}) := A_{22} - A_{21}B_{11}^{-1}A_{12}$$

für eine Matrix A partitioniert wie in (4.3) und ein invertierbares $B_{11} \in \mathbb{C}^{n_1,n_1}$ definiert.

4.3 Spezielle algebraische Mehrgitterverfahren.

In diesem Abschnitt werden die wichtigsten Multilevel-Verfahren eingeführt, die eine F-Glättung verwenden. Das erste dieser Verfahren ist das sogenannte reduction-based AMG oder kurz AMGr-Verfahren.

4.3.1 Reduction-based AMG.

Dieses Verfahren wurde erstmals von MacLachlan, Manteuffel und McCormick im Jahr 2006 in [68] vorgestellt.

Das AMGr-Verfahren nutzt in seiner ursprünglichen Form einen F-Glätter. In [36] wurde dieses Verfahren modifiziert und ein Gauß-Seidel-Glätter verwendet, der auf allen Unbekannten arbeitet. Aus denselben Gründen, die im letzten Abschnitt formuliert wurden, wird die Interpolationsmatrix als $P = \begin{bmatrix} -D^{-1}A_{12} \\ I \end{bmatrix}$ gewählt. Des Weiteren nutzt das AMGr-Verfahren in seiner Ursprungsform für die Interpolation und den F-Glätter dieselbe Approximation D von A_{11}, wobei D als diagonal vorausgesetzt wird.

Definition 4.4. Sei $A \in \mathbb{C}^{n,n}$ hpd und partitioniert wie in (4.3). Dann ist *AMGr-Verfahren* gegeben durch

$$T_{\mathrm{amg}_r} := I - B_{\mathrm{amg}_r}^{-1} A := T_c^{\mathrm{amg}_r} \cdot T_s^{\mathrm{amg}_r}, \tag{4.7}$$

wobei

$$\begin{aligned} T_c^{\mathrm{amg}_r} &:= I - P(P^H A P)^{-1} P^H A, \\ T_s^{\mathrm{amg}_r} &:= I - \sigma \begin{bmatrix} D^{-1} & 0 \\ 0 & 0 \end{bmatrix} A. \end{aligned}$$

Hierbei ist $\sigma \in \mathbb{R}$ noch geeignet zu bestimmen.

Das Besondere bei der Definition des AMGr-Verfahrens ist die Gewichtung des Glätters mit der Konstante σ. Diese soll so gewählt werden, dass

$$\rho(I - A_{11}^{\frac{1}{2}} D^{-1} A_{11}^{\frac{1}{2}})$$

möglichst klein ist. Für die Approximation D wird außerdem gefordert, dass $\sigma(D^{-1}A_{11}) \subseteq [1, 1+\varepsilon]$ für ein $\varepsilon > 0$ ist. Hierbei bezeichnet $\sigma(\cdot)$ das Spektrum einer Matrix.

Mit dieser Eigenschaft gilt

$$\rho(I - \sigma A_{11}^{\frac{1}{2}} D^{-1} A_{11}^{\frac{1}{2}}) \leq \max(|\, 1 - \sigma \,|, |\, (1+\varepsilon)\sigma - 1 \,|).$$

Dieser Term wird minimal, wenn

$$|\, 1 - \sigma \,| = 1 - \sigma = (1+\varepsilon)\sigma - 1 = |\, (1+\varepsilon)\sigma - 1 \,|.$$

Daraus folgt die Wahl $\sigma = \frac{2}{2+\varepsilon}$.

Um die Konvergenz des AMGr-Verfahrens, also $\rho(I - B_{\mathrm{amg}_r}^{-1} A) < 1$, zu zeigen, benötigt es noch eine weitere Eigenschaft, die in dieser Arbeit eine zentrale Rolle einnimmt. Die Matrix $\begin{bmatrix} D & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ soll hermitesch positiv semidefinit (hpsd) sein. Nimmt man alle Voraussetzungen an D zusammen, erhält man zum einen $\frac{1}{1+\varepsilon} A_{11} \preceq D \preceq A_{11}$, d.h. D soll in der positiv definiten Relation kleiner als A_{11} sein, doch zum anderen soll die Matrix A trotz des Ersetzens von A_{11} durch D weiterhin hpsd bleiben. Folgendes Konvergenzresultat wird in [68] für den reellen Fall, also A spd, gezeigt.

Satz 4.5 ([68]). *Sei A hpd und partitioniert wie in* (4.3). *Des Weiteren sei D eine Approximation an A_{11} und es gelte $\frac{1}{1+\varepsilon}A_{11} \preceq D \preceq A_{11}$, sowie $\begin{bmatrix} D & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ sei hpsd. Dann ist*

$$\left\| I - B_{\mathrm{amg}_r}^{-1} A \right\|_A \leq \left(\frac{\varepsilon}{1+\varepsilon} \left(1 + \left(\frac{\varepsilon}{(2+\varepsilon)^2} \right) \right) \right)^{\frac{1}{2}} < 1, \tag{4.8}$$

wobei B_{amg_r} durch (4.7) *mit $\sigma = \frac{2}{2+\varepsilon}$ definiert ist.*

Üblicherweise wird in derartigen Beweisen ausgenutzt, dass $x^T Ax \in \mathbb{R}$ für alle x abgeschätzt werden kann. Da für hermitesche Matrizen $A \in \mathbb{C}^{n,n}$ ebenfalls $x^H Ax \in \mathbb{R}$ für alle $x \in \mathbb{C}^n$ erfüllt ist, stellt das Verallgemeinern auf komplexwertige Matrizen keine Schwierigkeit dar. Die gleiche Argumentation kann bei den Sätzen 5.1, 6.2, 6.3, 6.6, 7.1, 8.1 angewendet werden.

Betrachtet man das AMGr-Verfahren mit mehr als nur einem Glätter, erhält man folgendes Resultat.

Satz 4.6 ([68]). *Sei A hpd und partitioniert wie in* (4.3). *Des Weiteren sei D eine Approximation an A_{11} und es gelte $\frac{1}{1+\varepsilon}A_{11} \preceq D \preceq A_{11}$, sowie $\begin{bmatrix} D & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ sei hpsd. Dann ist*

$$\left\| I - B_{\mathrm{amg}_{r,\nu}}^{-1} A \right\|_A \leq \left(\frac{\varepsilon}{1+\varepsilon} \left(1 + \left(\frac{\varepsilon}{2+\varepsilon} \right)^{2(\nu-1)} \left(\frac{\varepsilon}{(2+\varepsilon)^2} \right) \right) \right)^{\frac{1}{2}} < 1,$$

wobei $B_{\mathrm{amg}_{r,\nu}}$ für $\sigma = \frac{2}{2+\varepsilon}$ definiert ist durch

$$I - B_{\mathrm{amg}_{r,\nu}}^{-1} A := T_c^{\mathrm{amg}_r} \cdot \left(T_s^{\mathrm{amg}_r} \right)^{\nu}.$$

Die zweite Klasse von AMG's mit F-Glättung, die betrachtet werden, basieren auf einer Block-Faktorisierung der Matrix A.

4.3.2 Multilevel-Block-Faktorisierungs-Verfahren.

Basierend auf einer Block-Faktorisierung der Systemmatrix A lassen sich AMG-ähnliche Verfahren entwickeln.

Betrachtet man eine Block-Faktorisierung der Matrix A

$$A = \begin{bmatrix} I & 0 \\ A_{21}A_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & 0 \\ 0 & S(A, A_{11}) \end{bmatrix} \begin{bmatrix} I & A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix}, \tag{4.9}$$

mit $S(A, A_{11}) = A_{22} - A_{21}A_{11}^{-1}A_{12}$, dann ist die Inverse von A durch

$$A^{-1} = \begin{bmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & S(A, A_{11})^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{bmatrix}$$

gegeben. Man reduziert somit das zu lösende System mit A in zwei kleinere Probleme mit A_{11} bzw. $S(A, A_{11})$. Anstatt diese Systeme exakt zu lösen, bietet es sich an, diese zu approximieren, was auf die folgende Definition führt.

Definition 4.7. Sei $A \in \mathbb{C}^{n,n}$ hpd und partitioniert wie in (4.3). Weiter seien E und S hermitesch positiv definite Approximationen an A_{11} bzw. $S(A,E) = A_{22} - A_{21}E^{-1}A_{12}$. Dann wird das Zweilevel-Block-Faktorisierungs-Verfahren (2BF) durch

$$B_{\text{2bf}} := \begin{bmatrix} E & A_{12} \\ A_{21} & S + A_{21}E^{-1}A_{12} \end{bmatrix} \tag{4.10}$$

definiert.

In dieser Definition wird auf die Darstellung $I - B_{\text{2bf}}^{-1}A$ als Iterationsmatrix verzichtet. Im weiteren Verlauf der Arbeit soll nicht nur die Konvergenz, sondern auch eine mögliche Verwendung algebraischer Mehrgitterverfahren als Vorkonditionierer untersucht werden. Durch eine Analyse des Spektrums von $B_{\text{2bf}}^{-1}A$ kann sowohl eine Aussage über den Spektralradius von $\rho(I - B_{\text{2bf}}^{-1}A)$ gemacht, als auch eine Abschätzung von

$$\kappa(B_{\text{2bf}}^{-1}A) := \frac{\lambda_{\max}(B_{\text{2bf}}^{-1}A)}{\lambda_{\min}(B_{\text{2bf}}^{-1}A)}$$

erreicht werden. Hierbei und im weiteren Verlauf ist $\kappa(\cdot)$ die Konditionszahl einer hpd Matrix bzgl. der 2-Norm und $\lambda_{\max}(\cdot)$ bzw. $\lambda_{\min}(\cdot)$ der größte bzw. der kleinste Eigenwert einer hpd Matrix.

Die durch (4.10) definierte Matrix ist erstmals durch Bank and Dupont [13] als Vorkonditionierer publiziert und von Axelsson und Gustafsson [7] sowie von Vassilevski [92] studiert worden. Ein auf Probleme der QCD, wie z.B. der Gauge-Laplace-Matrix aus Abschnitt **??**, angepasstes Block-Faktorisierungsverfahren wurde in [74] entwickelt, auf das nicht näher eingegangen wird, da im weiteren Verlauf Block-Faktorisierungs-Verfahren sowie algebraische Mehrgitterverfahren im Allgemeinen untersucht und nicht auf ein bestimmtes Problem angepasst werden sollen.

Bevor dieses Verfahren rekursiv auf ein Mehrlevelverfahren überführt wird, wird der Zusammenhang solcher Verfahren mit „typischen“ AMG's, die in Abschnitt 4.2 besprochen wurden, erläutert. Dieser basiert auf den Arbeiten Mense und Nabben [76–78], in denen verschiedene Varianten von Multilevel-Verfahren auf Konvergenz untersucht werden. Zwei dieser Verfahren sind gegeben durch die Iterationsmatrizen

$$I - B_{\text{mamli}}^{-1}A := I - B_{\text{mamli}}^{-1}A := T_c^{\text{mamli}} \cdot T_s^{\text{mamli}} \quad \text{und} \tag{4.11}$$

$$I - B_{\text{smamli}}^{-1}A := I - B_{\text{smamli}}^{-1}A := T_s^{\text{mamli}} \cdot T_c^{\text{mamli}} \cdot T_s^{\text{mamli}}, \tag{4.12}$$

wobei

$$T_c^{\text{mamli}} := I - \begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix} S^{-1} \begin{bmatrix} -A_{21}D_p^{-1} & I \end{bmatrix} A, \tag{4.13}$$

$$T_s^{\text{mamli}} := I - \begin{bmatrix} M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A \tag{4.14}$$

sind. In den Arbeiten von Mense und Nabben werden diese Iterationsmatrizen in allgemeiner Form betrachtet, da die Konvergenz für nichtsymmetrische M-Matrizen untersucht wird. Hier werden ausschließlich hermitesche Matrizen untersucht, daher wird die hier aufgezeigte

Einschränkung betrachtet und insbesondere gefordert, dass Nachglätter und Vorglätter bei dem SMAMLI-Verfahren identisch sind. Die Matrix S in (4.13) soll eine Approximation an die Grobgittermatrix A_c sein, was einen wesentlichen Unterschied zum AMGr-Verfahren darstellt, indem für die Konvergenzanalyse eine exakte Lösung auf dem groben Gitter gefordert wird. Als Grobgittermatrix werden in den Arbeiten von Mense und Nabben drei Fälle besprochen:

$$A_c = A_{22},\ A_c = S(A, D_p) \text{ und } A_c = \left[\begin{smallmatrix} -A_{21}D_p^{-1} & I \end{smallmatrix}\right] A \left[\begin{smallmatrix} -D_p^{-1}A_{12} \\ I \end{smallmatrix}\right].$$

Die Konvergenz dieser Verfahren wird in [76] über die Konstruktion eines *allgemeinen Mehrgitterverfahrens* gezeigt. Die wesentliche Aussage beinhaltet das folgende Lemma.

Lemma 4.8 ([76] Lemma 5.3)**.** *Sei $A \in \mathbb{C}^{n,n}$ partitioniert wie in* (4.3) *and A_{11} nichtsingulär. Desweiteren sei $\nu \in \mathbb{N}$ mit $\nu \geq 1$ und es sei die folgende Iterationsmatrix eines algebraischen Mehrgitterverfahrens gegeben*

$$\begin{aligned} I - B_{\mathrm{amg}_\nu}^{-1} A := &\left(I - \begin{bmatrix} M_s^{-H} & 0 \\ 0 & 0 \end{bmatrix} A\right)^\nu \cdot \left(I - \begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix} S^{-1} \begin{bmatrix} -A_{21}D_p^{-1} & I \end{bmatrix} A\right) \\ &\cdot \left(I - \begin{bmatrix} M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A\right)^\nu . \end{aligned} \tag{4.15}$$

Hier seien M_s, D_p Approximationen an A_{11} und die Matrix S approximiere die Galerkin-Matrix $A_c := \left[\begin{smallmatrix} -A_{21}D_P^{-1} & I \end{smallmatrix}\right] A \left[\begin{smallmatrix} -D_P^{-1}A_{12} \\ I \end{smallmatrix}\right]$. Dann erfüllt B_{amg_ν} die Identität

$$B_{\mathrm{amg}_\nu}^{-1} = \begin{bmatrix} \widetilde{M}_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -E_p^{-1}A_{12} \\ I \end{bmatrix} S^{-1} \begin{bmatrix} -A_{21}E_p^{-H} & I \end{bmatrix}, \tag{4.16}$$

wobei

$$\widetilde{M}_s := A_{11}\Big[I - (I - M_s^{-H}A_{11})^\nu (I - M_s^{-1}A_{11})^\nu\Big]^{-1}, \tag{4.17}$$

$$E_p := A_{11}\Big[I - (I - M_s^{-H}A_{11})^\nu (I - D_p^{-1}A_{11})\Big]^{-1} \tag{4.18}$$

sind.

Dieses Resultat wird in [76], wo AMG's für nicht symmetrische M-Matrizen untersucht werden, in einer allgemeineren Form dargestellt. Lemma 4.8 ist eine, für diese Arbeit ausreichende, Simplifikation für symmetrische bzw. hermitesche Matrizen. Des Weiteren wird gefordert, dass S die Galerkin-Matrix approximiert, was, wie in Satz 4.2 gesehen, die natürliche Wahl darstellt. In den Arbeiten [76–78] wird das obige Lemma verwendet, um Konvergenzaussagen für allgemeine Mehrgitterverfahren bei Verwendung eines F-Glätters für M-Matrizen zu zeigen. Dazu werden gewisse Splitting-Eigenschaften auf einem groben Level gefordert, um anschließend zu zeigen, dass diese auf das nächste Level transportiert werden. In Kapitel 5 wird eine ähnliche Untersuchung für den hpd Fall durchgeführt.

Doch zuvor wird die Matrix in (4.16) etwas genauer betrachtet. Diese erfüllt die Gleichung

$$B_{\mathrm{amg}_\nu}^{-1} = \begin{bmatrix} I & -E_p^{-1}A_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} \widetilde{M}_s^{-1} & 0 \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -A_{21}E_p^{-H} & I \end{bmatrix}, \tag{4.19}$$

stellt also eine Block-Faktorisierung dar. Wird $M_s = D_p$ sowie $\nu = 1$ gewählt und definiert man $E := A_{11}\Big[I - (I - M_s^{-1}A_{11})^2\Big]^{-1}$, so ist

$$B_{\mathrm{amg}_\nu} = \begin{bmatrix} E & A_{12} \\ A_{12} & S + A_{21}E^{-1}A_{12} \end{bmatrix}.$$

Für diese spezielle Wahl der Approximation sowie Anzahl der Glättungsschritte entspricht B_{amg_ν} dem Zweilevel-Block-Faktorisierungs-Verfahren aus Definition 4.7. Daher ist eine weitere Betrachtung solcher Faktorisierungen auch für das Verständnis sowie der Analyse von algebraischen Mehrgitterverfahren sinnvoll. Aufbauend auf dem 2BF-Verfahren kann ein Multilevel-Block-Faktorisierungs-Verfahren definiert werden.

Definition 4.9. Sei $A^{(1)} \in \mathbb{R}^{n,n}$ hpd und partitioniert wie in (4.3). Für $k = 1, \ldots, L-1$ werden hpd Approximationen $E^{(k)}$ für $A_{11}^{(k)}$ und aufbauend auf $E^{(k)}$ die Grobgittermatrix $A^{(k+1)} := S(A^k, E^k)$ definiert, welche wiederum wie in (4.3) mit entsprechenden Abwandlungen der Räume partitioniert werden. Auf dem gröbsten Level L sei $B_{\mathrm{mbf}}^{(L)}$ eine hpd Approximation an $A^{(L)}$. Das *Multilevel-Block-Faktorisierungsverfahren* wird rekursiv durch

$$B_{\mathrm{mbf}}^{(k)} := \begin{bmatrix} E^{(k)} & A_{12}^{(k)} \\ A_{21}^{(k)} & B_{\mathrm{mbf}}^{(k+1)} + A_{21}^{(k)} E^{(k)^{-1}} A_{12}^{(k)} \end{bmatrix}. \tag{4.20}$$

definiert.

Kapitel 5

Multilevel-Block-Faktorisierungs-Verfahren: Der Transport von Informationen

Hier wird das Multilevel-Block-Faktorisierungsverfahren, das die Grundlage für viele algebraische Mehrgitterverfahren mit F-Glättung bildet, betrachtet. In [76–78] wurde dieses Verfahren für unsymmetrische M-Matrizen analysiert, indem gezeigt wird, wie Konvergenzeigenschaften von einem groben zu einem feinen Level transportiert werden. Dass eine ähnliche Analyse im hpd Fall möglich ist, zeigt das folgende Resultat, welches in [11] für reellwertige Matrizen A bewiesen wurde, hier aber für komplexwertige Matrizen wiedergegeben wird.

Satz 5.1 ([11])**.** *Sei A hpd und partitioniert wie in (4.3) sowie B_{2bf} geben durch* (4.10)*, wobei E und S hpd Approximationen an A_{11} and A_{22} sind, so dass*

$$A_{11} \preceq E \preceq \beta A_{11}, \tag{5.1}$$

$$A_{22} \preceq S \preceq \theta A_{22} \tag{5.2}$$

erfüllt sind. Dann gilt

$$A \preceq B_{\mathrm{2bf}} \preceq \theta_{[11]} A$$

mit

$$\theta_{[11]} := 1 + \tfrac{1}{1-\gamma(A)^2} \left\{ \tfrac{1}{2}(b+d) + \left[\tfrac{1}{4}(b-d)^2 + bd\gamma(A)^2 \right]^{\frac{1}{2}} \right\} \tag{5.3}$$

und $b = \beta - 1$, $d = \theta + \gamma(A)^2 - 1$. Hierbei ist $\gamma(A)$ die kleinste Konstante, die die Ungleichung $x_1^H A_{12} x_2 \leq \gamma(A)(x_1^H A_{11} x_1 x_2^H A_{22} x_2)^{\frac{1}{2}}$ für alle $x_1 \in \mathbb{C}^{n_1}, x_2 \in \mathbb{C}^{n_2}$ erfüllt. Diese Konstante wird Cauchy-Bunyakovski-Schwarz (C.B.S.) Konstante von A genannt.

Satz 5.1 gibt den weiteren Ablauf vor. Es wird gefordert, dass die Approximationen in der positiv definiten Halbordnung größer als die Originale sind. Im letzten Abschnitt wurde gezeigt, dass bei dem AMGr-Verfahren genau die andere Richtung verlangt wird. Es stellt sich somit die Frage, ob die Relationen (5.1) und (5.2) aus Satz 5.1 in der dargestellten Halbordnung benötigt werden. Den zweiten Ansatzpunkt für eine detaillierte Analyse bietet die Wahl der Grobgittermatrix $A_c = A_{22}$. Obwohl für das numerische Verfahren nur die Matrix S, also die Approximation von A_c, benötigt wird und nicht A_c selbst, können numerische Berechnungen nur dann erfolgreich sein, wenn S eine Approximation einer geeigneten Matrix darstellt. Beachtet man, dass die hier betrachteten Verfahren aus einer Block-Faktorisierung entstanden sind, ist die beste Wahl $A_c = S(A, A_{11})$ und daher bietet sich $A_c = S(A, E)$ an, wobei $S(A, E) = A_{22} - A_{21}E^{-1}A_{12}$ ist. Mit der eben angesprochenen Variation zur Originalarbeit [11] wird im Weiteren gezeigt, dass gewisse spektrale Eigenschaften von einem groben Level zum nächst feineren „transportiert“ werden.

Vorher wird jedoch auf die wichtigsten Eigenschaften der im Satz 5.1 erwähnten C.B.S. Konstante eingegangen, die später in Kapitel 6 detailliert behandelt wird.

Lemma 5.2 ([3, 4, 7, 11])**.** *Sei A hpd und partitioniert wie in (4.3). Weiter sei $\gamma(A)$ die zu A assoziierte C.B.S. Konstante. Dann gilt*

$$\gamma(A)^2 = \max_{x_2 \neq 0} \frac{x_2^H A_{21} A_{11}^{-1} A_{12} x_2}{x_2^H A_{22} x_2} = \max_{x_1 \neq 0} \frac{x_1^H A_{12} A_{22}^{-1} A_{21} x_1}{x_1^H A_{11} x_1}, \tag{5.4}$$

$$1 - \gamma(A)^2 = \min_{x_2 \neq 0} \frac{x_2^H S(A, A_{11}) x_2}{x_2^H A_{22} x_2} = \min_{x_1 \neq 0} \frac{x_1^H S(A, A_{22}) x_1}{x_1^H A_{11} x_1}, \tag{5.5}$$

$$\max_{x_2 \neq 0} \frac{x_2^H S(A, A_{11}) x_2}{x_2^H A_{22} x_2} \leq 1, \max_{x_1 \neq 0} \frac{x_1^H S(A, A_{22}) x_1}{x_1^H A_{11} x_1} \leq 1. \tag{5.6}$$

Außerdem erhält man für alle zur Partition von A assoziierten $x = \left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right] \in \mathbb{C}^n$ und für alle μ, $\gamma \leq \mu \leq \gamma^{-1}$ die Ungleichung

$$x^H A x \geq (1 - \mu\gamma) x_1^H A_{11} x_1 + (1 - \mu^{-1}\gamma) x_2^H A_{22} x_2, \tag{5.7}$$

$$x^H A x \geq x_2^H S(A, A_{11}) x_2. \tag{5.8}$$

Die Resultate dieses Kapitels können in [52] wiedergefunden werden.

5.1 Spektrale Äquivalenz.

In der Arbeit [11] werden nur Approximationen betrachtet, die in der positiv definiten Halbordnung größer als die Originale sind. Dies wird verallgemeinert, wofür das Konzept der spektralen Äquivalenz benötigt wird, das in vielen Arbeiten eine zentrale Rolle einnimmt, siehe z.B. [11, 68, 93].

Definition 5.3. Seien $A, B \in \mathbb{C}^{n,n}$ hpd. B heißt *spektral äquivalent zu A mit* $0 < \alpha \leq 1 \leq \beta$, wenn die Relation

$$\alpha A \preceq B \preceq \beta A$$

erfüllt ist.

Wie bereits festgestellt, spielt es oft eine wichtige Rolle, ob eine Approximation größer oder kleiner als das Original ist. Daher lohnt sich eine Unterteilung der spektralen Äquivalenz.

Definition 5.4. Seien $A, B \in \mathbb{C}^{n,n}$ hpd.

1. B heißt *spektral-kleiner (σ-kleiner) als A mit* $0 < \alpha \leq 1$, wenn

   $$\alpha A \preceq B \preceq A$$

   erfüllt ist.

2. B heißt *spektral-größer (σ-größer) als A mit* $\beta \geq 1$, wenn

   $$A \preceq B \preceq \beta A$$

   erfüllt ist.

3. B heißt *spektral-unsortiert (σ-unsortiert) zu A mit* $0 < \alpha \leq 1 \leq \beta$, wenn

$$\alpha A \preceq B \preceq \beta A,\ A \not\preceq B,\ B \not\preceq A$$

erfüllt ist.

Diese Unterteilung hat verschiedene Vorteile. Zum einen werden die verschiedenen Fälle in der gängigen Literatur bereits genutzt, ohne die obige Terminologie zu verwenden, siehe [8, 11, 79]. Definition 5.4 hilft, zwischen den verschiedenen Fällen zu unterscheiden.

Zum anderen resultiert aus dieser Unterteilung das folgende, in der Literatur häufig verwendete Lemma, das bei der Abschätzung des Spektrums und damit der Konditionszahl verwendet werden kann.

Lemma 5.5. *Seien $A, B \in \mathbb{R}^{n,n}$ hpd.*

1. *Wenn B σ-kleiner als A mit α ist, dann gilt*

$$\sigma(B^{-1}A) \subset [1, \tfrac{1}{\alpha}] \text{ und } \kappa_2(B^{-1}A) \leq \tfrac{1}{\alpha}.$$

2. *Wenn B σ-größer als A mit β ist, dann gilt*

$$\sigma(B^{-1}A) \subset [\tfrac{1}{\beta}, 1] \text{ und } \kappa_2(B^{-1}A) \leq \beta.$$

3. *Wenn B σ-unsortiert zu A mit α und β ist, dann gilt*

$$\sigma(B^{-1}A) \subset [\tfrac{1}{\beta}, \tfrac{1}{\alpha}] \text{ und } \kappa_2(B^{-1}A) \leq \tfrac{\beta}{\alpha}.$$

Beweis. Dieses Lemma folgt durch elementare Umformungen. □

5.2 Analyse der 2BF- und MBF-Verfahren.

Dieser Abschnitt behandelt die Fragestellung: Wenn E bzgl. A_{11} und S bzgl. $S(A, E)$ dieselbe spektrale Äquivalenz vorweisen, hat dann $B_{2\text{bf}}$ ebenfalls diese Eigenschaft bzgl. A? Im folgenden Resultat wird der Transport spektraler Eigenschaften vom groben zum feinen Level erläutert.

Satz 5.6. *Sei A hpd und partitioniert wie in (4.3). Weiter sei $B_{2\text{bf}}$ definiert durch* (4.10) *und es seien $\gamma(A)$ und $\gamma(B)$ die C.B.S. Konstanten A bzw. $B_{2\text{bf}}$. Außerdem wird gefordert, dass die Matrix $S(A, E)$ hpd ist.*

1. *Wenn E σ-kleiner als A_{11} mit α und S σ-kleiner als $S(A, E)$ mit ξ ist, dann ist $B_{2\text{bf}}$ σ-kleiner als A mit der Konstante*

$$\xi_{2\text{bf}} = \left(\frac{\frac{1}{\alpha} - 1}{1 - \gamma(B)^2} + \frac{1}{\xi} \right)^{-1}. \tag{5.9}$$

2. *Wenn E σ-größer als A_{11} mit β und S σ-größer als $S(A, E)$ mit θ ist, dann ist $B_{2\text{bf}}$ σ-größer als A mit der Konstante*

$$\theta_{2\text{bf}} = 1 + \tfrac{1}{\beta} \left\{ \theta - 1 + \tfrac{1}{2} \tfrac{\beta - 1}{1 - \gamma(A)^2} \cdot \left(\beta + \theta - 1 + \left[(\beta + \theta - 1)^2 + \beta(\theta - 1)\gamma(A)^2 \right]^{\frac{1}{2}} \right) \right\}. \tag{5.10}$$

3. *Wenn E σ-unsortiert zu A_{11} mit α und β sowie S σ-unsortiert zu $S(A,E)$ mit ξ und θ ist, dann ist $B_{2\text{bf}}$ σ-unsortiert zu A mit den Konstanten $\xi_{2\text{bf}}$ und $\theta_{2\text{bf}}$, die durch* (5.9) *bzw.* (5.10) *gegeben sind.*

Beweis. Zunächst werden die Schranken $\theta_{2\text{bf}}$ und $\xi_{2\text{bf}}$ mit der Annahme, dass E spektral äquivalent zu A_{11} mit $\alpha \leq 1$ und $\beta \geq 1$ ist sowie, dass S spektral äquivalent zu $S(A,E)$ mit $\xi \leq 1$ und $\theta \geq 1$ ist, hergeleitet. Im Anschluss werden die speziellen spektralen Eigenschaften gezeigt.

Sei zunächst $x = \left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right] \in \mathbb{C}^n$ in der zu A gehörenden Partition gegeben. Um die obere Schranke herzuleiten, betrachte

$$\begin{aligned} x^H(B_{2\text{bf}} - A)x &= x^H \left(\begin{bmatrix} E & A_{12} \\ A_{21} & S + A_{21}E^{-1}A_{12} \end{bmatrix} - \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \right) x \\ &= x^H \left(\begin{bmatrix} E - A_{11} & 0 \\ 0 & S - (A_{22} - A_{21}E^{-1}A_{12}) \end{bmatrix} \right) x \\ &= x_1^H(E - A_{11})x_1 + x_2^H(S - S(A,E))x_2. \end{aligned} \tag{5.11}$$

Durch die Voraussetzungen lassen sich die Approximationen durch ihre Originale abschätzen und es gilt

$$\begin{aligned} x^H(B_{2\text{bf}} - A)x &= x_1^H(E - A_{11})x_1 + x_2^H(S - S(A,E))x_2 \\ &\leq (\beta - 1)x_1^H A_{11}x_1 + (\theta - 1)x_2^H S(A,E)x_2 \\ &\leq (\beta - 1)x_1^H A_{11}x_1 + (\theta - 1)x_2^H(A_{22} - A_{21}E^{-1}A_{12})x_2. \end{aligned} \tag{5.12}$$

Mit der spektralen Äquivalenz von E zu A_{11} gilt $x_1^H E x_1 \leq \beta x_1^H A_{11} x_1$ und daher $-x_2^H A_{21}E^{-1}A_{12}x_2 \leq -\frac{1}{\beta}x_2^H A_{21}A_{11}^{-1}A_{12}x_2$. Wird dies in (5.12) eingesetzt, so erhält man

$$\begin{aligned} x^H(B_{2\text{bf}} - A)x \leq& (\beta - 1)x_1^H A_{11}x_1 + (\theta - 1)x_2^H(A_{22} - \tfrac{1}{\beta}A_{21}A_{11}^{-1}A_{12})x_2 \\ =& (\beta - 1)x_1^H A_{11}x_1 + \tfrac{\theta-1}{\beta}x_2^H S(A, A_{11})x_2 + (\theta - 1)\left(1 - \tfrac{1}{\beta}\right)x_2^H A_{22}x_2 \\ =& (\beta - 1)\Big(\underbrace{x_1^H A_{11}x_1 + \tfrac{\theta-1}{\beta}x_2^H A_{22}x_2}_{(\star)}\Big) + \frac{\theta - 1}{\beta}x_2^H S(A, A_{11})x_2. \end{aligned} \tag{5.13}$$

Es bleibt der Term $(\star)$ abzuschätzen. Mit (5.7) folgt, dass

$$x_1^H A_{11}x_1 + \tfrac{1-\mu^{-1}\gamma(A)}{1-\mu\gamma(A)}x_2^H A_{22}x_2 \leq \tfrac{1}{1-\mu\gamma(A)}x^H Ax$$

für alle $x = \left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]$ und $\mu \in [\gamma(A), \gamma(A)^{-1}]$ erfüllt ist.

Die bestmögliche Abschätzung ergibt sich für ein μ mit $\frac{1-\mu^{-1}\gamma(A)}{1-\mu\gamma(A)} = \frac{\theta-1}{\beta}$. Für $\theta > 1$ ist die zu lösende Gleichung gegeben durch

$$\begin{aligned} &\tfrac{\theta-1}{\beta} = \tfrac{\mu-\gamma(A)}{\mu(1-\mu\gamma(A))} \Leftrightarrow \mu - \gamma(A) = \left(\mu - \mu^2\gamma(A)\right) \cdot \tfrac{\theta-1}{\beta} \\ \Leftrightarrow\quad & \left(\mu\gamma(A)\right)^2 + \tfrac{\beta-(\theta-1)}{\theta-1} \cdot \left(\mu\gamma(A)\right) - \tfrac{\beta\gamma(A)^2}{\theta-1} = 0. \end{aligned}$$

Diese Gleichung hat die positive Nullstelle

$$\mu\gamma(A) = \tfrac{1}{2}\left(1 - \tfrac{\beta}{\theta-1}\right) + \left[\tfrac{1}{4}\left(1 - \tfrac{\beta}{\theta-1}\right)^2 + \tfrac{\beta}{\theta-1}\gamma(A)^2\right]^{\frac{1}{2}}.$$

Um den Term $\frac{1}{1-\mu\gamma(A)}$ zu vereinfachen, betrachte

$$\begin{aligned}
& 1 - \mu\gamma(A) = \tfrac{1}{2}\left(1 + \tfrac{\beta}{\theta-1}\right) - \left[\tfrac{1}{4}\left(1 - \tfrac{\beta}{\theta-1}\right)^2 + \tfrac{\beta}{\theta-1}\gamma(A)^2\right]^{\frac{1}{2}} \\
\Leftrightarrow \quad & (1 - \mu\gamma(A)) \cdot \left\{\tfrac{1}{2}\left(1 + \tfrac{\beta}{\theta-1}\right) + \left[\tfrac{1}{4}\left(1 - \tfrac{\beta}{\theta-1}\right)^2 + \tfrac{\beta}{\theta-1}\gamma(A)^2\right]^{\frac{1}{2}}\right\} \\
& = \tfrac{1}{4}\left(1 + \tfrac{\beta}{\theta-1}\right)^2 - \tfrac{1}{4}\left(1 - \tfrac{\beta}{\theta-1}\right)^2 - \tfrac{\beta}{\theta-1}\gamma(A)^2 \\
& = \tfrac{\beta}{\theta-1} \cdot (1 - \gamma(A)^2).
\end{aligned}$$

Dies führt auf die Identität

$$\tfrac{1}{1-\mu\gamma(A)} = \tfrac{1}{1-\gamma(A)^2}\left\{\tfrac{1}{2}\left(1 + \tfrac{\theta-1}{\beta}\right) + \left[\tfrac{1}{4}\left(1 - \tfrac{\theta-1}{\beta}\right)^2 + \tfrac{\theta-1}{\beta}\gamma(A)^2\right]^{\frac{1}{2}}\right\}.$$

Damit folgt aus (5.13)

$$\begin{aligned}
x^H(B_{2\mathrm{bf}} - A)x \leq & \tfrac{\beta-1}{1-\gamma(A)^2}\left\{\tfrac{1}{2}\left(1 + \tfrac{\theta-1}{\beta}\right) + \left[\tfrac{1}{4}\left(1 - \tfrac{\theta-1}{\beta}\right)^2 + \tfrac{\theta-1}{\beta}\gamma(A)^2\right]^{\frac{1}{2}}\right\} x^H A x \\
& + \tfrac{\theta-1}{\beta} x^H A x.
\end{aligned}$$

Zusammenfassend wurde eine obere Schranke unabhängig von α und ξ durch

$$x^H B_{2\mathrm{bf}} x \leq \theta_{2\mathrm{bf}} x^H A x$$

mit

$$\theta_{2\mathrm{bf}} = 1 + \tfrac{1}{\beta}\left\{\theta - 1 + \tfrac{1}{2}\tfrac{\beta-1}{1-\gamma(A)^2}\left(\beta + \theta - 1 + \left[(\beta+\theta-1)^2 + \beta(\theta-1)\gamma(A)^2\right]^{\frac{1}{2}}\right)\right\}$$

gefunden.

Die untere Schranke kann vergleichsweise einfach gezeigt werden. Mit der Voraussetzung $\alpha A_{11} \preceq E$ und $\xi S(A,E) \preceq S$, was gleichbedeutend zu $A_{11} - E \preceq (\frac{1}{\alpha} - 1)E$ und $S(A,E) - S \preceq (\frac{1}{\xi} - 1)S$ ist, und mit (5.11) folgt für alle $x = \left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]$

$$\begin{aligned}
x^H(A - B_{2\mathrm{bf}})x &= x_1^H(A_{11} - E)x_1 + x_2^H(S(A,E) - S)x_2 \\
&= (\tfrac{1}{\alpha} - 1)x_1^H E x_1 + (\tfrac{1}{\xi} - 1)x_2^H S x_2. \qquad (5.14)
\end{aligned}$$

Da

$$B_{2\mathrm{bf}}^{-1} = \begin{bmatrix} I & 0 \end{bmatrix}^H E^{-1} \begin{bmatrix} I & 0 \end{bmatrix} + \begin{bmatrix} A_{21}E^{-1} & I \end{bmatrix}^H S^{-1} \begin{bmatrix} A_{21}E^{-1} & I \end{bmatrix} \succ 0,$$

ist $B_{2\text{bf}}$ hpd und $\gamma(B)$ wohldefiniert. Die Gleichung (5.7) induziert weiter, dass $x_1^H E x_1 \leq \frac{1}{1-\gamma(B)^2} x^H B_{2\text{bf}} x$ gilt. Außerdem ist

$$S(B_{2\text{bf}}, E) = S + A_{21}E^{-1}A_{12} - A_{21}E^{-1}A_{12} = S,$$

d.h. S ist das Schurkomplement von $B_{2\text{bf}}$ bzgl. E. Zusammen mit (5.8) ergibt das die Ungleichung $x^H B_{2\text{bf}} x \geq x_2^H S x_2$.

Aus (5.14) folgt

$$x^H(A - B_{2\text{bf}})x \leq \left(\frac{\frac{1}{\alpha} - 1}{1 - \gamma(B)^2} + \frac{1}{\xi} - 1 \right) x^H B_{2\text{bf}} x$$

und damit

$$x^H B_{2\text{bf}} x \geq \left(\frac{\frac{1}{\alpha} - 1}{1 - \gamma(B)^2} + \frac{1}{\xi} \right)^{-1} x^H A x.$$

Hiermit erhält man, dass $B_{2\text{bf}}$ spektral äquivalent zu A mit $\xi_{2\text{bf}} = \left(\frac{\frac{1}{\alpha}-1}{1-\gamma(B)^2} + \frac{1}{\xi} \right)^{-1}$ und $\theta_{2\text{bf}} = \frac{\beta-\gamma^2}{1-\gamma^2} \cdot \frac{\beta+\theta-1}{\beta}$ ist.

Um zu zeigen, dass $B_{2\text{bf}}$ σ-unsortiert zu A ist, muss dargelegt werden, dass $B_{2\text{bf}}^{-1}A$ mindestens einen Eigenwert besitzt, der größer oder gleich Eins ist, und mindestens einen, der kleiner oder gleich Eins ist. Da E und S σ-unsortiert zu A_{11} bzw. $S(A, E)$ sind, erhält man mit Lemma 5.5, dass Vektoren $x_1 \in \mathbb{C}^{n_1}, x_2 \in \mathbb{C}^{n_2}$ mit $x_1 \neq 0$ und $x_2 \neq 0$ zu Eigenwerten $\lambda \leq 1$ und $\mu \leq 1$ existieren, so dass

$$\frac{x_1^H A_{11} x_1}{x_1^H E x_1} = \lambda \leq 1 \text{ und } \frac{x_2^H S(A, E) x_2}{x_2^H S x_2} = \mu \leq 1$$

erfüllt ist. Damit sind die Ungleichungen

$$x_1^H A_{11} x_1 \leq x_1^H E x_1 \text{ und } x_2^H S(A, E) x_2 \leq x_2^H S x_2$$

gezeigt. Setzt man $x = \left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right] \in \mathbb{C}^n$ und $\widetilde{A} = \left[\begin{smallmatrix} E & A_{12} \\ A_{21} & A_{22} \end{smallmatrix}\right] \in \mathbb{R}^{n,n}$, dann induziert die Voraussetzung $S(A, E) \succ 0$, dass $\widetilde{A} \succ 0$ gilt und damit

$$\begin{aligned}
\frac{x^H A x}{x^H B_{2\text{bf}} x} &= \frac{x^H A x}{x^H \widetilde{A} x} \cdot \frac{x^H \widetilde{A} x}{x^H B_{2\text{bf}} x} \\
&= \left(1 + \frac{x_1^H (A_{11} - E) x_1}{x^H \widetilde{A} x} \right) \cdot \left(1 + \frac{x_2^H (S(A, E) - S) x_2}{x^H B_{2\text{bf}} x} \right) \\
&= \left(1 - \frac{x_1^H E x_1}{x^H \widetilde{A} x} + \frac{x_1^H A_{11} x_1}{x^H \widetilde{A} x} \right) \cdot \left(1 - \frac{x_2^H S x_2}{x^H B_{2\text{bf}} x} + \frac{x_2^H S(A, E) x_2}{x^H B_{2\text{bf}} x} \right) \\
&\leq \left(1 - \frac{x_1^H E x_1}{x^H \widetilde{A} x} + \frac{x_1^H E x_1}{x^H \widetilde{A} x} \right) \cdot \left(1 - \frac{x_2^H S x_2}{x^H B_{2\text{bf}} x} + \frac{x_2^H S x_2}{x^H B_{2\text{bf}} x} \right) = 1.
\end{aligned}$$

Die obige Rechnung zeigt, dass der kleinste Eigenwert von $C_{2\text{bf}}A$ kleiner oder gleich Eins ist. Mit denselben Argumenten erhält man einen Eigenwert, der größer oder gleich Eins ist.

Ferner bleibt zu zeigen, dass die anderen beiden spektralen Eigenschaften auf dem groben Level jeweils dieselben auf dem feinen Level implizieren. Dazu setzt man $\beta = 1$, um die untere Schranke zu beweisen. Es bleibt zu überprüfen, ob $A \succeq B_{2\text{bf}}$ erfüllt ist. Da $A_{11} \succeq E$ und $S(A, E) \succeq S$ gilt

$$A - B_{2\text{bf}} = \begin{bmatrix} A_{11} - E & 0 \\ 0 & S(A,E) - S \end{bmatrix} \succeq 0.$$

In ähnlicher Weise kann Teil 2 des Satzes gezeigt werden. □

Die untere Schranke $\xi_{2\text{bf}}$ aus Satz 5.6 hängt von der C.B.S. Konstante von $B_{2\text{bf}}$ ab. Dies macht es schwierig, qualitative Aussagen über die Schranke zu treffen. Jedoch kann diese C.B.S. Konstante unter gewissen Voraussetzungen an die Approximationen abgeschätzt werden.

Satz 5.7. *Sei A hpd und partitioniert wie in (4.3). Weiter sei $B_{2\text{bf}}$ definiert durch* (4.10) *und $\gamma(A), \gamma(B)$ die C.B.S. Konstanten von A bzw. $B_{2\text{bf}}$. Zusätzlich wird gefordert, dass*

$$\begin{aligned} &E \preceq \alpha A_{11} \text{ mit } \alpha > \gamma(A)^2, \\ &S \preceq \xi S(A,E). \end{aligned}$$

Dann lässt sich die C.B.S. Konstante von $B_{2\text{bf}}$ durch

$$\gamma(B)^2 \leq \frac{\gamma(A)^2}{\gamma(A)^2 + (\alpha - \gamma(A)^2)\xi} < 1$$

abschätzen.

Beweis. Für diesen Beweis wird die Gleichung (5.4) und die Darstellung der C.B.S. Konstanten

$$\gamma(B)^2 = \max_{x_2 \neq 0} \frac{x_2^H A_{21} E^{-1} A_{12} x_2}{x_2^H (S + A_{21} E^{-1} A_{12}) x_2}.$$

sowie

$$\gamma(A)^2 = \max_{x_2 \neq 0} \frac{x_2^H A_{21} A_{11}^{-1} A_{12} x_2}{x_2^H A_{22} x_2}$$

benötigt. Da $S \preceq \xi S(A,E)$ sowie $E \preceq \alpha A_{11}$ vorausgesetzt werden, erhält man

$$\begin{aligned} \gamma(B)^{-2} &= \min_{x_2 \neq 0} \frac{x_2^H (S + A_{21} E^{-1} A_{12}) x_2}{x_2^H A_{21} E^{-1} A_{12} x_2} \geq \min_{x_2 \neq 0} \frac{x_2^H (\xi A_{22} + (1-\xi) A_{21} E^{-1} A_{12}) x_2}{x_2^H A_{21} E^{-1} A_{12} x_2} \\ &= \min_{x_2 \neq 0} \frac{x_2^H \xi A_{22} x_2}{\alpha^{-1} x_2^H A_{21} A_{11}^{-1} A_{12} x_2} + 1 - \xi = \alpha \xi \gamma(A)^{-2} + 1 - \xi \\ &= \frac{\alpha\xi + (1-\xi)\gamma(A)^2}{\gamma(A)^2}. \end{aligned}$$

Daher lässt sich die C.B.S. Konstante von $B_{2\text{bf}}$ durch

$$\gamma(B)^2 \leq \frac{\gamma(A)^2}{\alpha\xi + (1-\xi)\gamma(A)^2} = \frac{\gamma(A)^2}{\gamma(A)^2 + (\alpha - \gamma(A)^2)\xi}$$

abschätzen. Mit der Voraussetzung $\alpha > \gamma(A)^2$, ergibt sich

$$\gamma(B)^2 \leq \frac{\gamma(A)^2}{\gamma(A)^2 + (\alpha - \gamma(A)^2)\xi} < 1,$$

was zu zeigen war. □

Mit dem bisher Gezeigten und der rekursiven Konstruktion des Multilevel-Block-Faktorisierungs-Verfahrens kann dieses analysiert werden.

Satz 5.8. *Sei $A^{(1)} \in \mathbb{C}^{n,n}$ hpd und partitioniert wie in (4.3). Weiter sei $B^{(k)}_{\text{mbf}}$ für $k = 1, \ldots, L-1$ gegeben durch* (4.20).

1. *Wird gefordert, dass für $j \in \{1, \ldots, L-1\}$, $E^{(k)}$ σ-kleiner als $A^{(k)}_{11}$ mit $\alpha^{(k)}$ für jedes $k = j, \ldots, L-1$ und $B^{(L)}_{\text{mbf}}$ σ-kleiner als $A^{(L)}$ mit $\xi^{(L)}_{\text{mbf}}$ ist, dann ist $B^{(j)}_{\text{mbf}}$ σ-kleiner als $A^{(j)}$ mit*

$$\xi^{(j)}_{\text{mbf}} = \left(\sum_{k=j}^{L-1} \frac{\alpha^{(k)^{-1}} - 1}{1 - \gamma(B^{(k)})} + \xi^{(L)^{-1}}_{\text{mbf}} \right)^{-1}. \tag{5.15}$$

2. *Wird gefordert, dass für $j \in \{1, \ldots, L-1\}$, $E^{(k)}$ σ-größer als $A^{(k)}_{11}$ mit $\beta^{(k)}$ für jedes $k = j, \ldots, L-1$ und $B^{(L)}_{\text{mbf}}$ σ-größer als $A^{(L)}$ mit $\theta^{(L)}_{\text{mbf}}$ ist, dann ist $B^{(j)}_{\text{mbf}}$ σ-größer als $A^{(j)}$ mit*

$$\begin{aligned} \theta^{(j)}_{\text{mbf}} = 1 + \tfrac{1}{\beta^{(j+1)}} \Big\{ \theta^{(j+1)}_{\text{mbf}} - 1 + \tfrac{1}{2} \tfrac{\beta^{(j+1)} - 1}{1 - \gamma(A^{(j+1)})^2} \cdot \Big(\beta^{(j+1)} + \theta^{(j+1)}_{\text{mbf}} - 1 \\ + \big[(\beta^{(j+1)} + \theta^{(j+1)}_{\text{mbf}} - 1)^2 + \beta^{(j+1)} (\theta^{(j+1)}_{\text{mbf}} - 1) \gamma(A^{(j+1)})^2 \big]^{\frac{1}{2}} \Big) \Big\}. \end{aligned} \tag{5.16}$$

3. *Wird gefordert, dass für $j \in \{1, \ldots, L-1\}$, $E^{(k)}$ σ-unsortiert zu $A^{(k)}_{11}$ mit $\alpha^{(k)}$ und $\beta^{(k)}$ für jedes $k = j, \ldots, L-1$ sowie $B^{(L)}_{\text{mbf}}$ σ-unsortiert zu $A^{(L)}$ mit $\xi^{(L)}_{\text{mbf}}$ und $\theta^{(L)}_{\text{mbf}}$ ist, dann ist $B^{(j)}_{\text{mbf}}$ σ-unsortiert zu $A^{(j)}$ mit $\xi^{(j)}_{\text{mbf}}$ and $\theta^{(j)}_{\text{mbf}}$, wobei die Konstanten durch* (5.15) *und* (5.16) *gegeben sind.*

Beweis. Es genügt, die Konstante (5.15) herzuleiten, da (5.16) direkt durch die rekursive Anwendung von Satz 5.6 folgt.

Aus Satz 5.6 folgt

$$\xi^{(L-1)}_{\text{mbf}} = \left(\frac{\alpha^{(L-1)^{-1}} - 1}{1 - \gamma(B^{(L-1)})} + \xi^{(L)^{-1}}_{\text{mbf}} \right)^{-1}.$$

Sei $\xi_{\text{mbf}}^{(j)} = \left(\sum_{k=j}^{L-1} \frac{\alpha^{(k)^{-1}}-1}{1-\gamma(B^{(k)})} + \xi_{\text{mbf}}^{(L)^{-1}} \right)^{-1}$ für $j \in \{L-1, L-2, \ldots, m\}$. Durch Anwendung von Satz 5.6 ergibt sich

$$\begin{aligned}
\xi_{\text{mbf}}^{(j)} &= \left(\frac{\alpha^{(j-1)^{-1}} - 1}{1-\gamma(B^{(j-1)})} + \xi_{\text{mbf}}^{(j)^{-1}} \right)^{-1} \\
&= \left(\frac{\alpha^{(j-1)^{-1}} - 1}{1-\gamma(B^{(j-1)})} + \sum_{k=j}^{L-1} \frac{\alpha^{(k)^{-1}} - 1}{1-\gamma(B^{(k)})} + \xi_{\text{mbf}}^{(L)^{-1}} \right)^{-1} \\
&= \left(\sum_{k=h-1}^{L-1} \frac{\alpha^{(k)^{-1}} - 1}{1-\gamma(B^{(k)})} + \xi_{\text{mbf}}^{(L)^{-1}} \right)^{-1}.
\end{aligned}$$

Dies bedeutet, es ist $\xi_{\text{mbf}}^{(j)} = \left(\sum_{k=j}^{L-1} \frac{\alpha^{(k)^{-1}}-1}{1-\gamma(B^{(k)})} + \xi_{\text{mbf}}^{(L)^{-1}} \right)^{-1}$ für $j = \{L-1, L-2, \ldots, m-1\}$. □

Satz 5.8 zeigt, wie durch die Mehrlevelstruktur und ihre rekursive Konstruktion Abschätzungen der Eigenwerte auf jedem Level erreicht werden. Außerdem ist mit (5.15) sogar eine explizite Formel gegeben, aus der abgelesen werden kann, wie stark die jeweiligen, durch die Approximation gegeben, Größen die Eigenwerte beeinflussen. Des Weiteren stellt Satz 5.8 eine vollständige Analyse der MBF-Verfahren für alle möglichen Varianten von Approximationen dar.

5.3 Polynombasierte Approximationen.

Polynome können auf verschiedene Weise benutzt werden, um Mehrgitter- oder auch Mehrlevelverfahren zu beschleunigen. In Kapitel 3 und insbesondere in (3.8) wurden Polynome verwendet, um verschiedene Zyklen, wie z.B. den V-Zyklus und den W-Zyklus von Mehrgitterverfahren darzustellen. Diese Idee wurde von Axelsson and Vassilevski [11] aufgegriffen, um das MBF-Verfahren zu „stabilisieren“. Jedoch wurde die Wahl des Polynoms nicht auf $P_\gamma(t) := (1-t)^\gamma$ eingeschränkt. Dies führt auf das sogenannte *AMLI-Verfahren.*

Definition 5.9. Sei $A^{(1)} \in \mathbb{R}^{n,n}$ hpd und partitioniert wie in (4.3). Für $k = 1, \ldots, L-1$ werden hpd Approximationen $E^{(k)}$ für $A_{11}^{(k)}$ und Grobgittermatrizen durch $A^{(k+1)} = S(A^{(k)}, E^{(k)})$ definiert, welche wiederum wie in (4.3) partitioniert seien. Auf dem gröbsten Level L sei $B_{\text{amli}}^{(L)}$ eine hpd Approximation für $A^{(L)}$. Das *AMLI Verfahren* $B_{\text{amli}}^{(1)}$ ist dann rekursiv gegeben durch

$$B_{\text{amli}}^{(k)} = \begin{bmatrix} E^{(k)} & A_{12}^{(k)} \\ A_{21}^{(k)} & S_{\text{amli}}^{(k+1)} + A_{21}^{(k)} E^{(k)^{-1}} A_{12}^{(k)} \end{bmatrix},$$

wobei

$$S_{\text{amli}}^{(k+1)} = A^{(k+1)} \Big[I - P_{\gamma_{k+1}}^{(k+1)} \big(B_{\text{amli}}^{(k+1)^{-1}} A^{(k+1)} \big) \Big]^{-1}. \tag{5.17}$$

Hierbei seien $P_{\gamma_{k+1}}^{(k+1)} \in \mathbb{R}_{\leq \gamma_{k+1}}[t]$ Polynome jeweils vom Grad γ_{k+1}, die die Bedingungen $P_{\gamma_{k+1}}^{(k+1)}(0) = 1$ und $|P_{\gamma_{k+1}}^{(k+1)}(t)| < 1$ für $t \in [\underline{t}^{(k+1)}, \overline{t}^{(k+1)}]$ erfüllen, wobei die Intervalle $[\underline{t}^{(k+1)}, \overline{t}^{(k+1)}]$ jeweils die Eigenwerte von $B_{\text{amli}}^{(k+1)^{-1}} A^{(k+1)}$ einschließen.

Im Gegensatz zur Originalarbeit [11], in der Polynome ausschließlich für die Konstruktion einer Grobgittermatrix benutzt werden, siehe (5.17), wird im Folgenden ein allgemeinerer Ansatz verfolgt. So wird, ähnlich wie in [11], analysiert, wie die spektralen Eigenschaften durch verschiedene Möglichkeiten der Wahl des Polynoms transportiert werden. Außerdem wird die Verwendung eines Polynoms zur Verbesserung der Approximation E an A_{11} besprochen.

Um die folgende Theorie aus den eben genannten Gründen allgemein zu halten, wird eine Approximation definiert, die auf einem Polynom basiert.

Definition 5.10. Seien $m \in \mathbb{N}$ und die Matrizen $C, \widetilde{C}_0 \in \mathbb{C}^{m,m}$ hpd, wobei $\widetilde{C}_0$ eine Approximation an C darstellt. Weiter sei $P_\gamma \in \mathbb{R}_{\leq\gamma}[t]$ ein Polynom mit reellen Koeffizienten und $P_\gamma(0) = 1$ sowie $\left| P_\gamma(t) \right| < 1$ für $t \in \sigma(\widetilde{C}_0^{-1}C)$. Dann heißt die Matrix

$$\widetilde{C} := C \left[I - P_\gamma \left(\widetilde{C}_0^{-1} C \right) \right]^{-1} \tag{5.18}$$

die *P_γ-Approximation an C von $\widetilde{C}_0$*. Wenn aus dem Zusammenhang hervorgeht, dass sich das Polynom auf C und $\widetilde{C}_0$ bezieht, wird sie auch mit *P_γ-Approximation* bezeichnet.

Bemerkung 5.11. Anstatt der Voraussetzung $\left| P_\gamma(t) \right| < 1$ genügt es in diesem Kontext $P_\gamma(t) < 1$ für $t \in \sigma(\widetilde{C}_0^{-1}C)$ zu fordern. Für die Invertierbarkeit von $I - P_\gamma(\widetilde{C}_0^{-1}C)$ muss $P_\gamma(t) \neq 1$ auf $\sigma(\widetilde{C}_0^{-1}C)$ verlangt werden. Um sicherzustellen, dass die neue Approximation $\widetilde{C}$ hpd ist, benötigt man $P_\gamma(t) < 1$ für $t \in \sigma(\widetilde{C}_0^{-1}C)$. Da im späteren Verlauf der Arbeit Matrizen mit komplexen Eigenwerten als Argument des Polynoms betrachtet werden, wird zusätzlich in der ganzen Arbeit $\left| P_\gamma(t) \right| < 1$ gefordert. Dies ist keine große Einschränkung und es wird sichergestellt, dass $\sigma(\widetilde{C}^{-1}C) \subset (0, 2)$ ist.

Zu beachten ist außerdem bei der Definition 5.10, dass bei der Verwendung der Approximation (5.18) der Term $I - P_\gamma(\widetilde{C}_0^{-1}C)$ in der Praxis nicht invertiert werden muss bzw. darf. Normalerweise ist man nur an der Wirkung von $\widetilde{C}^{-1}$ auf einen Vektor $r \in \mathbb{C}^n$ interessiert, also

$$\widetilde{C}^{-1} r = \left[I - P_\gamma \left(\widetilde{C}_0^{-1} C \right) \right] C^{-1} r.$$

Durch die Einführung eines zweiten Polynoms

$$Q_{\gamma-1}(t) := \frac{1 - P_\gamma(t)}{t}$$

erhält man

$$\widetilde{C}^{-1} r = Q_{\gamma-1} \left(\widetilde{C}_0^{-1} C \right) \widetilde{C}_0 r,$$

eine Berechnung der Wirkung von $\widetilde{C}^{-1}$, wobei nur die Inverse von $\widetilde{C}_0$ benötigt wird. Diese Möglichkeit der Polynomauswertung wird *Horner's Methode* genannt.

Das nächste Resultat beantwortet die Frage, welche spektralen Eigenschaften eine P_γ-Approximation in Abhängigkeit von P_γ und dem Spektrum von $\widetilde{C}_0^{-1}C$, auf dem das Polynom operiert, erhält.

Satz 5.12. *Seien $m \in \mathbb{N}$ sowie $C, \widetilde{C}_0 \in \mathbb{C}^{m,m}$ hpd. Weiter sei $P_\gamma \in \mathbb{R}_{\leq\gamma}[t]$ und $\widetilde{C}$ eine P_γ-Approximation wie in* (5.18)*. Dann gilt*

1. *$\widetilde{C}$ ist σ-kleiner als C mit α genau dann, wenn $P_\gamma(t) \in [1-\frac{1}{\alpha}, 0]$ für $t \in \sigma\left(\widetilde{C}_0^{-1}C\right)$.*
2. *$\widetilde{C}$ ist σ-größer to C mit β genau dann, wenn $P_\gamma(t) \in [0, 1-\frac{1}{\beta}]$ für $t \in \sigma\left(\widetilde{C}_0^{-1}C\right)$.*
3. *$\widetilde{C}$ ist σ-unsortiert zu C mit α und β genau dann, wenn $P_\gamma(t) \in [1-\frac{1}{\alpha}, 1-\frac{1}{\beta}]$ für $t \in \sigma\left(\widetilde{C}_0^{-1}C\right)$ und Eigenwerte $\overline{\lambda}$ sowie $\underline{\lambda}$ von $\widetilde{C}_0^{-1}C$ existieren, so dass $P_\gamma(\overline{\lambda}) > 0$ und $P_\gamma(\underline{\lambda}) < 0$.*

Beweis. Die P_γ-Approximation lässt sich zu

$$\begin{aligned}\widetilde{C} &= C\Big[I - P_\gamma(\widetilde{C}_0^{-1}C)\Big]^{-1} \\ &= C^{\frac{1}{2}}\Big[I - P_\gamma(C^{\frac{1}{2}}\widetilde{C}_0^{-1}C^{\frac{1}{2}})\Big]^{-1}C^{\frac{1}{2}}\end{aligned}$$

umschreiben. Damit können die spektralen Eigenschaften mit den folgenden Umformungen beschrieben werden.

$$\begin{aligned}
& 1-\tfrac{1}{\alpha} \leq P_\gamma(t) \leq 1-\tfrac{1}{\beta} \quad \forall t \in \sigma\left(\widetilde{C}_0^{-1}C\right) \\
\Leftrightarrow\quad & \left(1-\tfrac{1}{\alpha}\right) I \preceq P_\gamma\left(C^{\frac{1}{2}}\widetilde{C}_0^{-1}C^{\frac{1}{2}}\right) \preceq \left(1-\tfrac{1}{\beta}\right) I \\
\Leftrightarrow\quad & \alpha\Big[I - P_\gamma(C^{\frac{1}{2}}\widetilde{C}_0^{-1}C^{\frac{1}{2}})\Big] \preceq I \preceq \beta\left[I - P_\gamma\left(C^{\frac{1}{2}}\widetilde{C}_0^{-1}C^{\frac{1}{2}}\right)\right] \\
\Leftrightarrow\quad & \tfrac{1}{\beta}\Big[I - P_\gamma(C^{\frac{1}{2}}\widetilde{C}_0^{-1}C^{\frac{1}{2}})\Big]^{-1} \preceq I \preceq \tfrac{1}{\alpha}\left[I - P_\gamma\left(C^{\frac{1}{2}}\widetilde{C}_0^{-1}C^{\frac{1}{2}}\right)\right]^{-1} \\
\Leftrightarrow\quad & \tfrac{1}{\beta}C^{-\frac{1}{2}}\widetilde{C}C^{-\frac{1}{2}} \preceq I \preceq \tfrac{1}{\alpha}C^{-\frac{1}{2}}\widetilde{C}C^{-\frac{1}{2}} \\
\Leftrightarrow\quad & \tfrac{1}{\beta}\widetilde{C} \preceq C \preceq \tfrac{1}{\alpha}\widetilde{C} \qquad \Leftrightarrow \qquad \alpha C \preceq \widetilde{C} \preceq \beta C.
\end{aligned}$$

Damit ist gezeigt, dass

$$1-\frac{1}{\alpha} \leq P_\gamma(t) \leq 1-\frac{1}{\beta} \text{ für alle } t \in \sigma\left(\widetilde{C}_0^{-1}C\right)$$

genau dann gilt, wenn $\widetilde{C}$ spektral äquivalent zu C mit α und β ist. Mit der Voraussetzung, dass $P_\gamma(\overline{\lambda}) > 0$ und $P_\gamma(\underline{\lambda}) < 0$ für gewisse Eigenwerte $\overline{\lambda}$ und $\underline{\lambda}$ von $\widetilde{C}_0^{-1}C$ erfüllt ist, folgt, dass $\widetilde{C}$ σ-unsortiert zu C ist. Für den Beweis der anderen beiden spektralen Eigenschaften setzt man $\alpha = 1$ bzw. $\beta = 1$. □

Satz 5.12 beschreibt, wie mithilfe eines Polynoms gewisse spektrale Eigenschaften erreicht werden können. Es wurde gezeigt, dass dafür eine Kenntnis des Spektrums von $\widetilde{C}_0^{-1}C$ notwendig ist. In dem folgenden Resultat wird dies ausgenutzt, um die Approximation E von A_{11} im Vergleich zu einer initialen Approximation E_0 zu verbessern. Sei E_0 mit einer der spektralen Eigenschaften gegeben, dann soll mithilfe eines Polynoms eine Matrix E, wie in (5.18), erzeugt werden, so dass das Spektrum von $E^{-1}A_{11}$ nahe Eins liegt.

Satz 5.13. *Seien $n_1 \in \mathbb{N}$ sowie $A_{11}, E_0 \in \mathbb{C}^{n_1,n_1}$ hpd.*

1. *Mit der Annahme, dass E_0 σ-kleiner als A_{11} mit α_0 ist und $P_\gamma \in \mathbb{R}_{\leq\gamma}[t]$ mit $P_\gamma(0) = 1$ und $P_\gamma(\frac{1}{\alpha_0}) = \min\{P_\gamma(t) \,|\, t \in [1, \frac{1}{\alpha_0}]\}$ sowie $P_\gamma(t) \leq 0$ auf $[1, \frac{1}{\alpha_0}]$, erhält man, dass die P_γ-Approximation E aus (5.18) σ-kleiner als A_{11} mit*

$$\alpha = \frac{\alpha_0}{Q_{\gamma-1}\left(\frac{1}{\alpha_0}\right)} \tag{5.19}$$

ist, wobei $Q_{\gamma-1}$ durch (5.3) gegeben ist.

2. *Mit der Annahme, dass E_0 σ-größer als A_{11} mit β_0 ist und $P_\gamma \in \mathbb{R}_{\leq\gamma}[t]$ mit $P_\gamma(0) = 1$ und $P_\gamma(\frac{1}{\beta_0}) = \max\{P_\gamma(t) \,|\, t \in [\frac{1}{\beta_0}, 1]\}$ sowie $0 \leq P_\gamma(t) < 1$ auf $[\frac{1}{\beta_0}, 1]$ ist, erhält man, dass die P_γ-Approximation E aus (5.18) σ-größer als A_{11} mit*

$$\beta = \frac{\beta_0}{Q_{\gamma-1}\left(\frac{1}{\beta_0}\right)} \tag{5.20}$$

ist, wobei $Q_{\gamma-1}$ durch (5.3) gegeben ist.

3. *Mit der Annahme, dass E_0 σ-unsortiert zu A_{11} mit α_0 und β_0 und $P_\gamma \in \mathbb{R}_{\leq\gamma}[t]$ mit $P_\gamma(0) = 1$, $P_\gamma(t) < 1$ auf $[\frac{1}{\beta_0}, \frac{1}{\alpha_0}]$, $P_\gamma(\frac{1}{\alpha_0}) = \min\{P_\gamma(t) \,|\, t \in [\frac{1}{\beta_0}, \frac{1}{\alpha_0}]\}$ und $P_\gamma(\frac{1}{\beta_0}) = \max\{P_\gamma(t) \,|\, t \in [\frac{1}{\beta_0}, \frac{1}{\alpha_0}]\}$ sowie es $\overline{\lambda}, \underline{\lambda} \in \sigma(\widetilde{C}_0^{-1}C)$ gibt mit $P(\overline{\lambda}) > 0$ und $P(\underline{\lambda}) < 0$, erhält man, dass die P_γ-Approximation E aus (5.18) σ-unsortiert zu A_{11} mit α und β gegeben durch (5.19) bzw. (5.20) ist.*

Beweis. Hier reicht es aus, die dritte Aussage zu beweisen. Die ersten beiden folgen dann durch die jeweiligen Voraussetzungen und Satz 5.12.

Da für $\sigma(E_0^{-1}A_{11}) \subset [\beta_0^{-1}, \alpha_0^{-1}]$ die Ungleichung

$$\min_{t\in[\frac{1}{\beta_0},\frac{1}{\alpha_0}]} P_\gamma(t) \leq P_\gamma(t) \leq \max_{t\in[\frac{1}{\beta_0},\frac{1}{\alpha_0}]} P_\gamma(t)$$

gilt, erhält man mit Satz 5.12, dass E σ-unsortiert zu

$$\alpha = \left(1 - \min_{t\in I} P_\gamma^{(k)}(t)\right)^{-1} \text{ und } \beta = \left(1 - \max_{t\in I} P_\gamma^{(k)}(t)\right)^{-1}$$

ist. Durch die Definition von α_0 sowie β_0 folgt

$$\alpha = \left(1 - P_\gamma(\alpha_0^{-1})\right)^{-1} = \left(\alpha_0^{-1} Q_{\gamma-1}(\alpha_0^{-1})\right)^{-1} = \alpha_0 \left(Q_{\gamma-1}(\alpha_0^{-1})\right)^{-1}$$

und

$$\beta = \left(1 - P_\gamma(\beta_0^{-1})\right)^{-1} = \left(\beta_0^{-1} Q_{\gamma-1}(\beta_0^{-1})\right)^{-1} = \beta_0 \left(Q_{\gamma-1}(\beta^{-1})\right)^{-1}.$$

Damit sind die Schranken (5.19) und (5.20) gezeigt. □

Im nächsten Beispiel werden zwei Möglichkeiten angesprochen, ein Polynom zu wählen, welches die Approximation E im Vergleich zu E_0 verbessert.

Beispiel 5.14. Es sei E_0 σ-größer als A_{11} mit $\beta_0 \geq 1$, was der verwendeten Voraussetzung aus [11] entspricht.

1. Eine mögliche Wahl des Polynoms ist gegeben durch $P_2(t) = (1-t)^2$. Dies führt auf $Q_1(t) = -t + 2$. Da das Polynom P_2 für $t > 0$ streng monoton wachsend ist, ist die Identität $P_2(\frac{1}{\beta_0}) = \max\{P_2(t) \,|\, t \in [\frac{1}{\beta_0}, 1]\}$ erfüllt. Mit dem Satz 5.13 erhält man für
$$E = A_{11}\Big[I - \big(I - E_0^{-1}A_{11}\big)^2\Big]^{-1},$$
dass E σ-größer als A_{11} mit
$$\beta = \frac{\beta_0}{-\frac{1}{\beta_0} + 2} \leq \beta_0$$
ist. Hierbei ging die Ungleichung $-\frac{1}{\beta_0} + 2 \geq 1$ ein.
2. Im zweiten Beispiel wird ein Tschebysheff Polynom, das in Abschnitt 3.4 definiert wurde, betrachtet. Damit die neue Approximation
$$E = A_{11}\Big[I - \widetilde{P}_2(E_0^{-1}A_{11})\Big]^{-1}$$
weiterhin σ-größer als A_{11} ist, wird das Polynom verschoben und es ergibt sich
$$\widetilde{P}_2(t) = \frac{T_2\left(\frac{\beta_0+1-2t}{\beta_0-1}\right) + 1}{T_2\left(\frac{\beta_0+1}{\beta_0-1}\right) + 1}$$
mit $T_2(t) = 2t^2 - 1$ und $\widetilde{P}_2(\frac{1}{\beta_0}) = \max\{\widetilde{P}_2(t) \,|\, t \in [\frac{1}{\beta_0}, 1]\}$ sowie
$$\widetilde{P}_2(t) = 1 - \frac{4}{\beta+1}t + \frac{4}{(\beta+1)^2}t^2.$$
Daraus resultiert
$$\widetilde{Q}_1(t) = \frac{4}{\beta+1} - \frac{4}{(\beta+1)^2}t,$$
also $\widetilde{Q}_1(\frac{1}{\beta_0}) > 1$. Dies führt auf
$$\beta = \frac{\beta_0}{\widetilde{Q}_1\left(\frac{1}{\beta_0}\right)} \leq \beta_0.$$

Abschließend wird in diesem Kapitel Satz 5.13 verwendet, um das AMLI-Verfahren, siehe Definition 5.9, auf den Transport der spektralen Eigenschaften zu untersuchen.

Korollar 5.15. *Sei $A^{(k)}$ hpd, $S^{(k)}_{\text{amli}}$ definiert durch (5.17) und $Q^{(k)}_{\gamma-1} \in \mathbb{R}_{\leq\gamma-1}[t]$ gegeben durch*
$$Q^{(k)}_{\gamma-1}(t) = \frac{1 - P^{(k)}_{\gamma}(t)}{t}.$$
Dann erhält man die folgenden Zusammenhänge:

1. *Sei $B^{(k)}_{\text{amli}}$ σ-kleiner als $A^{(k)}$ mit $\xi^{(k)}$. Weiter erfülle $P^{(k)}_{\gamma}$ die Bedingungen $P^{(k)}_{\gamma}(t) \leq 0$ sowie $P_\gamma(\xi^{(k)^{-1}}) = \min\{P_\gamma(t) \,|\, t \in [1, \xi^{(k)^{-1}}]\}$. Dann ist $S^{(k)}_{\text{amli}}$ σ-kleiner als $A^{(k)}$ mit*
$$\xi^{(k)}_S = \xi^{(k)} \cdot \Big(Q^{(k)}_{\gamma-1}(\xi^{(k)^{-1}})\Big)^{-1}.$$

2. *Sei* $B_{\text{amli}}^{(k)}$ *σ-größer als* $A^{(k)}$ *mit* $\theta^{(k)}$ *ist. Weiter erfülle* $P_\gamma^{(k)}$ *die Bedingungen* $P_\gamma^{(k)}(t) \geq 0$ *und* $P_\gamma(\theta^{(k)^{-1}}) = \max\{P_\gamma(t) \,|\, t \in [\theta^{(k)^{-1}}, 1]\}$. *Dann ist* $S_{\text{amli}}^{(k)}$ *σ-größer als* $A^{(k)}$ *mit*

$$\theta_S^{(k)} = \theta^{(k)} \cdot \left(Q_{\gamma-1}^{(k)}(\theta^{(k)^{-1}})\right)^{-1}.$$

3. *Sei* $B_{\text{amli}}^{(k)}$ *σ-unsortiert zu* $A^{(k)}$ *mit* $\xi^{(k)}$ *und* $\theta^{(k)}$ *ist. Weiter existieren* $\overline{\lambda}, \underline{\lambda} \in \sigma(B_{\text{amli}}^{(k)^{-1}} A^{(k)})$ *mit* $P(\overline{\lambda}) > 0$ *und* $P(\underline{\lambda}) < 0$ *und* P_γ *erfülle die Bedingungen*

$$\begin{aligned} P_\gamma(\xi^{(k)^{-1}}) &= \min\{P_\gamma(t) \,|\, t \in \sigma(B_{\text{amli}}^{(k)^{-1}} A^{(k)})\} \quad \textit{sowie} \\ P_\gamma(\theta^{(k)^{-1}}) &= \max\{P_\gamma(t) \,|\, t \in \sigma(B_{\text{amli}}^{(k)^{-1}} A^{(k)})\}. \end{aligned}$$

Dann ist $S_{\text{amli}}^{(k)}$ *σ-unsortiert zu* $A^{(k)}$ *mit*

$$\xi_S^{(k)} = \xi^{(k)} \cdot \left(Q_{\gamma-1}^{(k)}(\xi^{(k)^{-1}})\right)^{-1} \quad \textit{and} \quad \theta_S^{(k)} = \theta^{(k)} \cdot \left(Q_{\gamma-1}^{(k)}(\theta^{(k)^{-1}})\right)^{-1}.$$

Zusammen mit Satz 5.8 führt Korollar 5.15 zu einer vollständigen Analyse des AMLI-Verfahrens in Bezug auf den Transport von spektralen Eigenschaften.

In diesem Abschnitt wurde eine Analyse von spektralen Eigenschaften durchgeführt. Dies führt zu Eigenwert- bzw. Konditionszahl-Abschätzungen, wie in Lemma 5.5 beschrieben, wobei eine Diskussion und Analyse der Güte dieser Abschätzungen vernachlässigt wurde. In Theorem 5.6 wurde gezeigt, dass diese Güte stark von der C.B.S. Konstante von A abhängt, welche in Satz 5.1 eingeführt wurde. Insbesondere sind die Schranken nicht aussagekräftig, wenn diese Konstante $\gamma(A)$ nahe Eins ist.

Im folgenden Kapitel wird detaillierter auf die Konstante eingegangen und Möglichkeiten angegeben, auch bei Matrizen A mit $\gamma(A) \approx 1$ Block-Faktorisierungs-Verfahren zu entwickeln, die eine gute Performance versprechen.

Kapitel 6

Die C.B.S. Konstante

Die sogenannte Cauchy-Bunyakovski-Schwarz (C.B.S.)-Ungleichung und ihre assoziierte Konstante nehmen in der Analyse von Zwei- und somit auch von Mehrlevelverfahren eine zentrale Rolle ein, [4, 6, 46]. In Satz 5.1 wurde diese Konstante ohne detaillierte Betrachtung eingeführt. Die Konstante darf aber bei der Konstruktion einer Grobgittermatrix nicht vernachlässigt werden, da durch sie Aussagen über die Approximation der optimalen Grobgittermatrix, also das Schurkomplement, getroffen werden können. Der Vollständigkeit halber wird die Definition der C.B.S. Konstante erneut angegeben.

Definition 6.1. Sei A hpd und partitioniert wie in (4.3). Die zu dieser Matrix und ihrer Partition assoziierte *C.B.S. Konstante* wird durch

$$\gamma(A) = \inf\{\widehat{\gamma} \,|\, x_1^H A_{12} x_2 \leq \widehat{\gamma}(x_1^H A_{11} x_1 x_2^H A_{22} x_2)^{\frac{1}{2}} \text{ für alle } x_1 \in \mathbb{C}^{n_1}, x_2 \in \mathbb{C}^{n_2}\}$$

definiert.

6.1

Hierarchische Basis. Die Abschätzungen der Eigenwerte aus den Sätzen 5.1 sowie 5.6 und deren Folgerungen sind nur dann hilfreich, wenn die C.B.S. Konstante $\gamma(A)$ von A, die durch das jeweilige Problem gegeben ist, von Eins weg beschränkt ist. Daher sind Variationen von den in Kapitel 5 vorgestellten Verfahren zu finden, die trotz einer C.B.S. Konstante $\gamma(A) \approx 1$ gut arbeiten.

Die Methoden, solche Verfahren zu entwickeln, basieren auf der folgenden Idee: Man führt eine Transformation der Matrix A durch, so dass die C.B.S. Konstante der transformierten Matrix möglichst klein wird, also

$$A_{\text{trafo}} = J^H A J \tag{6.1}$$

mit $\gamma(A_{\text{trafo}}) \ll 1$ und einer geeigneten Matrix J, siehe [6, 14, 81]. Doch anstatt ein AMG, definiert durch $I - B_{\text{amg}}^{-1} A$, auf die neue Matrix anzuwenden, nutzt man die Identität

$$\sigma(B_{\text{amg}}^{-1} A_{\text{trafo}}) = \sigma\left(B_{\text{amg}}^{-1} J^H A J\right) = \sigma\left((J^{-H} B_{\text{amg}} J^{-1})^{-1} A\right)$$

und definiert ein neues AMG durch $B_{\text{amg,mod}} := J^{-H} B_{\text{amg}} J^{-1}$.

In diesem Abschnitt werden Möglichkeiten aufgezeigt, wie die Matrix J gewählt werden kann. Die erste Variante wurde von Bank, Dupont und Yserentant [14] entwickelt. Diese Arbeit nutzt jedoch die Geometrie der gegebenen PDE und ist daher auf algebraische Mehrgitterverfahren nicht anwendbar. Um eine Idee zu erhalten, wie die Matrix J auf einem algebraischen Weg konstruiert werden kann, ist die Theorie aus [14] jedoch hilfreich.

Ausgehend von einer Diskretisierung einer partiellen Differentialgleichung mit finiten Elementen ist es möglich, durch einen Wechsel auf eine geeignete finite Elemente Basis, die C.B.S. Konstante der neuen Matrix zu reduzieren. Wird zunächst das Problem auf einem groben Gitter diskretisiert und werden anschließend die finiten Elemente vom groben Gitter durch eine geeignete Ergänzung auch auf dem feinen Gitter verwendet, erhält man eine sogenannte *hierarchische Basis*, siehe Abbildung 6.1. In [14] wird gezeigt, dass sich die Transformation von der sogenannten Knoten-Basis auf eine hierarchische Basis wie eine Kongruenztransformation der Form (6.1) mit

$$J = \begin{bmatrix} I & J_{12} \\ 0 & I \end{bmatrix}$$

verhält und für die C.B.S. Konstante der neuen Matrix die Relation $\gamma(A_{\text{trafo}}) \leq \frac{1}{1+c} < 1$ mit $c > 0$ unabhängig von der Anzahl der Gitterpunkte erfüllt und die C.B.S. Konstante damit von der Eins weg beschränkt ist, siehe [6]. Die Matrix J_{12} stellt die Transformation von der Knoten- zur hierarchischen Basis dar.

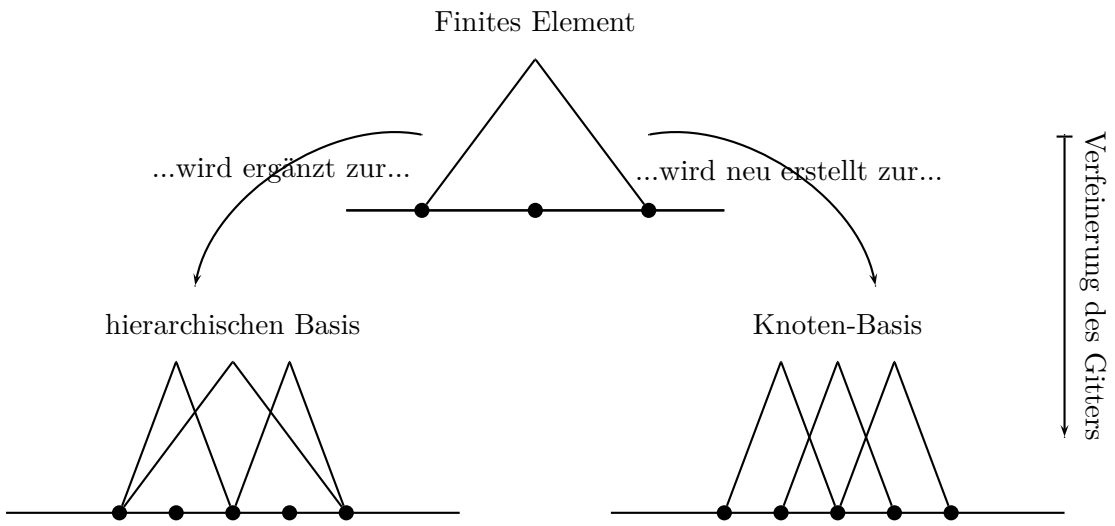


Abbildung 6.1.: Entwicklung der hierarchischen Basis im Vergleich zur Knoten-Basis.

Auf dieser Transformation basiert auch die Analyse in [10] sowie [11], was die Qualität der Abschätzung der Eigenwerte in Satz 5.1 begründet. Es wird also angenommen, dass die gegebene Matrix in der hierarchischen Basis vorliegt und daher eine C.B.S. Konstante besitzt, die von Eins weg beschränkt ist.

Aufbauend auf diesem Ansatz führte Axelsson [6] im Jahre 2003 eine algebraische Variante ein. Der Vorteil des algebraischen Ansatzes ist, dass im Gegensatz zu der hierarchischen Basis eine Verwendung auf AMG's möglich ist. Ausgehend von der Block-Faktorisierung

$$\begin{bmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{bmatrix} A \begin{bmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix} = \begin{bmatrix} A_{11} & 0 \\ 0 & S(A, A_{11}) \end{bmatrix} =: A_D$$

wird ersichtlich, dass die Transformation aus (6.1) mit

$$J = \begin{bmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix}$$

zu einer Matrix A_D mit $\gamma(A_D) = 0$ führt.

Ähnlich zum Kapitel 5 muss die Matrix A_{11} durch eine leicht zu invertierende Matrix ersetzt werden, um den numerischen Aufwand zu verringern. In dem bisherigen Verlauf wurden die hpd Approximationen E und D an A_{11} eingeführt. D soll dabei verdeutlichen, dass die Matrix möglichst dünn besetzt bzw. sogar diagonal sein soll. In der Regel wird D verwendet, um eine Grobgittermatrix A_c zu konstruieren. Im Gegensatz dazu stellt Matrix E eine beliebige, leicht zu invertierende Matrix, die nicht notwendigerweise dünn besetzt ist, dar. Um die Theorie allgemein wie möglich zu gestalten, wird für die hierarchischen Transformationen eine weitere invertierbare Approximation $H \in \mathbb{C}^{n_1,n_1}$ eingeführt. Im Folgenden werden Transformationen der Form

$$\widehat{A} := J^H A J \quad \text{mit} \quad J := \begin{bmatrix} I & -H^{-1}A_{12} \\ 0 & I \end{bmatrix} \tag{6.2}$$

untersucht.

In [6] wird der folgende Satz für reelle Martizen bewiesen. In diesem Kontext wird das Resultat direkt für den komplexen Fall dargestellt.

Satz 6.2 ([6])**.** *Sei A hpd und partitioniert wie in* (4.3)*. Außerdem sei $\widehat{A}$ definiert durch* (6.2) *mit $H \in \mathbb{C}^{n_1,n_1}$ nichtsingulär, so dass $\rho = \|I - H^{-1}A_{11}\|_{A_{11}} < 1$. Des Weiteren sei $\gamma(A)$die C.B.S. Konstante von A. Dann erfüllt die C.B.S. Konstante von $\widehat{A}$ die Ungleichung*

$$\gamma(\widehat{A}) \leq \sqrt{1 - \frac{1-\gamma(A)^2}{\rho^2\gamma(A)^2 + 1 - \gamma(A)^2}}. \tag{6.3}$$

Wendet man diese Art der Transformation auf die Block-Faktorisierungs-Verfahren aus Kapitel 5 an, so erhält man für den Zweilevelfall und den Spezialfall $H = E$ (vergleiche Definition 4.7) ein transformiertes Verfahren, gegeben durch

$$\widehat{B}_{2\text{bf}}^{-1} = \begin{bmatrix} E^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -(2E^{-1}-E^{-1}A_{11}E^{-1})A_{12} \\ I \end{bmatrix} S^{-1} \begin{bmatrix} -A_{21}(2E^{-1}-E^{-1}A_{11}E^{-1}), \; I \end{bmatrix}. \tag{6.4}$$

Ebenso ergibt sich ein modifiziertes Mehrlevelverfahren, siehe Algorithmus 5.

Aus Theorem 6.2 wird jedoch nicht ersichtlich, ob die Transformation hilfreich ist. Die Schranke der C.B.S. Konstante von $\widehat{A}$ hängt ebenfalls stark von der C.B.S. Konstante von A ab. Es gilt $\gamma(\widehat{A}) \approx 1$, falls $\gamma(A) \approx 1$.

Für einen Spezialfall fand Notay 2005 eine Abhilfe, siehe [81]. Dort wird gezeigt, dass unter starken Voraussetzungen an die Matrix A die C.B.S. Konstante $\gamma(\widehat{A})$ unabhängig von $\gamma(A)$ abgeschätzt werden kann. Dieses Resultat wird ebenfalls für den komplexen Fall verallgemeinert.

Satz 6.3 ([81]). *Sei A hpd und partitioniert wie in* (4.3). *Weiter sei $\widehat{A}$ definiert wie in* (6.2) *mit H hpd, so dass*

- $H \preceq A_{11}$,
- $\begin{bmatrix} H & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ *hermitesch positiv semidefinit ist.*

Dann erfüllt die C.B.S. Konstante von $\widehat{A}$ die Ungleichung

$$\gamma(\widehat{A}) \leq \sqrt{1 - \frac{1}{\lambda_{\max}(H^{-1}A_{11})}}. \tag{6.5}$$

Algorithmus 5: Berechnung von $d = \widehat{B}_{\text{mult}}^{-1} r$

Input : $r = \left[\begin{smallmatrix} r_1 \\ r_2 \end{smallmatrix}\right]$
Output : $\widehat{B}_{\text{mult}}^{-1} r$

```
//Restriktion:
for k = 1, ..., L − 1 do
    löse E^(k) w_E^(k) = r_1,
    löse E^(k) f = A_11^(k) w_E^(k),
    setze r = r_2 − A_21^(k) (2 w_E^(k) − f),        //Restringiere auf das nächste Level
    partitioniere r = [r_1; r_2].
//Lösen auf dem gröbsten Gitter:
Löse A^(L) d = r.
//Interpolation:
for k = L − 1, ..., 1 do
    Setze e = A_12^(k) d,
    löse E^(k) f_E = e,
    löse E^(k) f = A_11^(k) f_E,
    setze d_1 = w_E^(k) − 2 f_E + f                  //Update durch den additiven Glätter
    setze d_2 = d und d = [d_1; d_2].                //Interpoliere auf das nächste Level
```

Die vorgestellten algebraischen Kongruenztransformationen der Matrix A werden in Anlehnung an die Transformation auf eine hierarchische Basis aus [14] in der Literatur als Transformationen auf generalisierte hierarchische Basen bezeichnet. Ein Überblick einiger Variationen kann in [46, 48, 92] gefunden werden. Satz 6.3 gibt wahrscheinlich die einzige bekannte Abschätzung der C.B.S. Konstante von $\widehat{A}$ an, die unabhängig von der C.B.S. Konstante von A ist. In Abschnitt 6.5 wird eine weitere Abschätzung unter schwächeren Voraussetzungen angegeben.

Im nächsten Abschnitt wird gezeigt, dass die Voraussetzungen von Satz 6.3 garantieren, dass die Galerkin-Matrix $P^H AP$ eine gute Approximation an das Schurkomplement $S(A, A_{11})$ darstellt. Dieser Zusammenhang wird bei dem AMGr-Verfahren benötigt.

6.2 Zusammenhang zur Grobgittermatrix.

Eine der wichtigsten Fragen bei algebraischen Mehrgitterverfahren ist, wie eine Grobgittermatrix erzeugt werden kann, die die notwendigen Informationen auf dem groben Level ausreichend enthält. Wie in Kapitel 4 beschrieben, ist für eine gegebene Interpolationsmatrix P eine häufige Wahl

$$A_c = P^H AP. \tag{6.6}$$

Mit Satz 4.2 folgt, dass diese Wahl für die Minimierung der A-Norm in Bezug auf einen Rekursionsschritt als optimal aufgefasst werden kann.

Dies allein garantiert nicht, dass die Wahl der Grobgittermatrix zu einer schnellen Konvergenz führt. Für die hier betrachteten F-Glätter wurde in (4.6) gezeigt, dass die Grobgittermatrix A_c eine gute Approximation von $S(A, A_{11})$ darstellen soll, wobei die Identität

$$S(A, A_{11}) = P_{\text{opt}}^H A P_{\text{opt}} \quad \text{für} \quad P_{\text{opt}} = \begin{bmatrix} -A_{11}^{-1}A_{12} \\ I \end{bmatrix}$$

gilt. Um eine schnelle Konvergenz zu erhalten, muss folglich eine Interpolationsmatrix P gefunden werden, so dass $A_c = P^H AP$ die Relation

$$\alpha S(A, A_{11}) \preceq A_c \preceq \beta S(A, A_{11})$$

für Konstanten α und β, die möglichst nahe der Eins liegen, erfüllt ist.

Um diese Konstanten zu erhalten, kann die C.B.S. Konstante verwendet werden. Dazu betrachtet man die Matrix $\widehat{A}$ aus (6.2) und ihre 2×2-Blockstruktur. Insbesondere ist

$$\widehat{A}_{22} = \begin{bmatrix} -H^{-1}A_{12} \\ I \end{bmatrix}^H A \begin{bmatrix} -H^{-1}A_{12} \\ I \end{bmatrix}. \tag{6.7}$$

Definiert man die Grobgittermatrix als

$$A_c := \widehat{A}_{22}, \tag{6.8}$$

so folgt aus Lemma 5.2 das folgende Hilfsresultat.

Lemma 6.4. *Sei A hpd und partitioniert wie in (4.3). Des Weiteren seien A_c und $\widehat{A}$ gegeben durch* (6.7) *bzw.* (6.2)*. Außerdem sei $\gamma(\widehat{A})$ die C.B.S. Konstante von $\widehat{A}$. Dann ist*

$$S(A, A_{11}) \preceq A_c \preceq \frac{1}{1 - \gamma(\widehat{A})^2} S(A, A_{11}). \tag{6.9}$$

Es ist folglich notwendig, die C.B.S. Konstante von $\widehat{A}$ für die Konstruktion einer Interpolationsmatrix zu berücksichtigen. Mit Lemma 5.2 kann festgestellt werden, dass die Abschätzung nach oben scharf ist, also

$$\min_{x \neq 0} \frac{x^H S(A, A_{11}) x}{x^H A_c x} = 1 - \gamma(\widehat{A})^2.$$

Die Voraussetzungen des Konvergenzresultats des AMGr-Verfahrens (Satz 4.5) implizieren mit Anwendung des Satzes 6.3, dass die optimale Interpolation P_{opt} durch die Matrix

$$P = \begin{bmatrix} -H^{-1}A_{12} \\ I \end{bmatrix}$$

ausreichend gut approximiert wird, wenn H hpd ist und $\lambda_{\max}(H^{-1}A_{11})$ nicht zu groß wird.

Im nächsten Abschnitt wird untersucht, für welche Art von Matrizen die hierfür notwendigen Bedingungen

$$0 \preceq H \preceq A_{11} \quad \text{und} \quad \begin{bmatrix} H & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \succeq 0 \tag{6.10}$$

stets erfüllt werden können.

6.3 Der Greedy-Coarser und diagonaldominante Matrizen.

Wie in [71] gezeigt wird, ist es für diagonaldominante hpd Matrizen möglich, durch eine geeignete Diagonalmatrix H die Voraussetzung 6.10 zu erfüllen. Außerdem stellen MacLachlan und Saad in [71] einen Coarser vor, den sogenannten Greedy-Coarser, der eine Abschätzung der Eigenwerte von $H^{-1}A_{11}$, basierend auf dem Satz von Gershgorin, liefert, was für die Konstruktion von AMG's, wie zum Beispiel dem AMGr-Verfahren und solchen, die in Kapitel 7 vorgestellt werden, hilfreich ist. Der folgende Satz kann z.B. in [86, Theorem 4.6.] gefunden werden.

Satz 6.5 (Gershgorin). *Sei $A \in \mathbb{C}^{n,n}$ mit den Einträgen a_{ij}. Dann ist jeder Eigenwert $\lambda(A)$ von A in einer geschlossenen Kreisscheibe mit dem Mittelpunkt a_{ii}, $i \in \{1, \ldots, n\}$ und dem Radius*

$$\rho_i := \sum_{i \neq j} |\, a_{ij} \,|$$

lokalisiert, d.h. für alle $\lambda(A) \in \sigma(A)$ existiert ein $i \in \{1, \ldots, n\}$, so dass

$$|\, \lambda - a_{ii} \,| \leq \rho_i.$$

Der Greedy-Coarser verfolgt zwei Ziele:

1. Die Kondition der Untermatrizen auf den F-Punkten soll möglichst klein sein. Nach der Anwendung des Greedy-Coarsers auf die Matrix A ist diese in einer Blockstruktur wie in (4.3) gegeben. Dann kann eine Diagonalmatrix H angegeben werden, so dass $\kappa_2(H^{-1}A_{11})$ abhängig von einem Parameter $\varphi \in (\frac{1}{2}, 1)$ abgeschätzt werden kann. Insbesondere ist die Konditionszahl durch eine geeignete Wahl von φ klein. Dies ist für algebraische Mehrgitterverfahren und die Verwendung eines F-Glätters notwendig, wie in [81] gezeigt wird. Es gilt

$$\kappa_2(B_{2\mathrm{bf}}^{-1}A) \geq \kappa_2(H^{-1}A_{11}).$$

2. Die zweite Eigenschaft des Greedy-Coarsers ist eine gewisse Diagonaldominanz bezüglich der F-Punkte, d.h. für $\varphi \in (\frac{1}{2}, 1)$ und $i \in F$ wird die Eigenschaft

$$|\, a_{ii} \,| \geq \varphi \sum_{j \in F} |\, a_{ij} \,|$$

erfüllt. Für $\varphi = \frac{1}{2}$ heißt dies, dass A auf den F-Punkten diagonaldominant ist. Umso größer φ gewählt wird, desto größer wird die Anzahl der Grobgitter-Punkte.

Der Coarser ist durch den folgenden Algorithmus gegeben.

Algorithmus 6: Der Greedy-Coarser

Input : $A \in \mathbb{C}^{n,n}, \varphi \in (\frac{1}{2}, 1)$.
Output : F, C, $\varphi_i, i = 1, 2 \ldots, n$.

Setze $U = \{1, 2, \ldots, n\}$, $F = C = \emptyset$.
Berechne $\varphi_i = \dfrac{|a_{ii}|}{\sum_{j \in F \cup U} |a_{ij}|}$.
for $i = 1, \ldots, n$ **do**
  if $\varphi_i \geq \varphi$ **then**
    $F = F \cup \{i\}$, $U = U \backslash \{i\}$.
while $U \neq \emptyset$ **do**
  Finde $j = \operatorname{argmin}_{i \in U}\{\varphi_i\}$.
  $C = C \cup \{j\}$, $U = U \backslash \{j\}$.
  for $i \in U \cap \operatorname{Adj}(j)$, *wobei* $\operatorname{Adj}(j) := \{k : a_{jk} \neq 0\}$ **do**
    Aktualisiere $\varphi_i = \dfrac{|a_{ii}|}{\sum_{j \in F \cup U} |a_{ij}|}$.
    if $\varphi_i \geq \varphi$ **then**
      $F = F \cup \{i\}$, $U = U \backslash \{i\}$.

Im folgenden Resultat werden die wichtigsten Eigenschaften der Matrix A nach dem Coarsing-Prozess angegeben.

Satz 6.6 ([71])**.** *Sei A hpd. Setzt man $\varphi \in (\frac{1}{2}, 1)$ und wendet den Greedy-Coarser auf A an, dann erhält man eine zu A permutierte Matrix*

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

1. *Wählt man $H = \operatorname{diag}(A_{11})$, so ist*

$$\lambda_{\min}(H^{-1}A_{11}) \geq 2 - \tfrac{1}{\varphi}, \quad \lambda_{\max}(H^{-1}A_{11}) \leq \tfrac{1}{\varphi}.$$

2. *Ist $A \in \mathbb{C}^{n,n}$ diagonaldominant sowie $A_{11} = [a_{jk}^{(11)}] \in \mathbb{C}^{n_1,n_1}$ und wählt man $H = [h_{jk}]$ mit*

$$h_{jj} = a_{jj}^{(11)} - \sum_{j \neq k} |\, a_{jk}^{(11)} \,| \quad \textit{und} \quad h_{jk} = 0 \textit{ für } j \neq k, \tag{6.11}$$

dann gilt die Abschätzung

$$\lambda_{\min}(H^{-1}A_{11}) \geq 1, \quad \lambda_{\max}(H^{-1}A_{11}) \leq \tfrac{1}{2\varphi - 1}.$$

Außerdem ist die Matrix

$$\begin{bmatrix} H & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

hpsd.

Durch die Kombination des Greedy-Coarsers und der Diagonaldominanz von A (die nicht strikt sein muss) erhält man die Erfüllbarkeit der Voraussetzung 6.10, die für die Sätze 4.5 sowie 6.3 notwendig ist, sowie eine Abschätzung der Eigenwerte von $H^{-1}A_{11}$. In den folgenden Kapiteln nehmen die Eigenwerte von $H^{-1}A_{11}$ eine zentrale Rolle für die Entwicklung weiterer AMG's ein. Aus diesem Grund wird der Greedy-Coarser die Grundlage für alle numerischen Berechnungen in Kapitel 9 sein.

Damit wurde eine Klasse von Matrizen (diagonaldominante hermitesche Matrizen) vorgestellt, für die effektive Block-Faktorisierungs-Verfahren konstruiert werden können.

6.4 Adaptive Konstruktion der Interpolationsmatrix und M-Matrizen.

Eine weitere Möglichkeit, Interpolationsmatrizen für algebraische Mehrgitterverfahren zu erzeugen, bieten sogenannte *adaptive Verfahren*. Die Idee der Adaption wurde in den Arbeiten [35–37] entwickelt. Dabei ist für $A \in \mathbb{R}^{n,n}$ eine Interpolationsmatrix zu konstruieren, die auf glatten Vektoren $w \in \mathbb{R}^n$ mit $Aw \approx 0$ arbeitet. In diesem Abschnitt werden Varianten vorgestellt, die für die Verwendung eines F-Glätters entwickelt wurden. Details dazu können in [33, 67] sowie [68] gefunden werden. Als Grundlage wird in diesem Abschnitt der Körper der reellen Zahlen betrachtet.

Das Hauptaugenmerk bei adaptiven Varianten von Mehrgitterverfahren liegt auf der Bestimmung von Vektoren, die nahe dem Kern von A liegen. Eine Möglichkeit, dies zu erreichen, ist, parallel zu dem System $Ax = b$ mit $b \in \mathbb{R}^n$, das gelöst werden soll, das homogene System $Ax = 0$ zu betrachten. Dadurch erhält man eine Folge von Lösungen $\{x_{\text{hom}}^{[k]}\}$ des homogenen Systems. Stellt man im k-ten Schritt fest, z.B. durch Berechnung und Vergleich der Norm von $x_{\text{hom}}^{[k]}$, dass die Konvergenzgeschwindigkeit abnimmt, so nutzt man $x_{\text{hom}}^{[k]}$ für die Konstruktion einer neuen Interpolationsmatrix. Dabei geht man davon aus, dass $Ax_{\text{hom}}^{[k]} \approx 0$ erfüllt ist. Bei AMG's mit F-Glättung kann ein anderer Ansatz aus [68] verfolgt werden, der im Folgenden beschrieben wird.

In (4.5) wurde festgestellt, dass

$$P_{\text{opt}} = \begin{bmatrix} -A_{11}^{-1}A_{12} \\ I \end{bmatrix}$$

die optimale Interpolationsmatrix ist. Ausgehend von dieser Matrix wird eine neue Interpolationsmatrix konstruiert. Dazu wird zunächst ein sogenannter *Prototyp* $w^{[1]} \in \mathbb{R}^n$ „geraten“, der nahe des Kerns von A liegt. Glatte Vektoren sollen möglichst im Bild der Interpolationsmatrix liegen, d.h. für P_{opt} gilt

$$P_{\text{opt}} w_2^{[1]} = \begin{bmatrix} -A_{11}^{-1}A_{12}w_2^{[1]} \\ w_2^{[1]} \end{bmatrix} \quad \text{mit} \quad w^{[1]} = \begin{bmatrix} w_1^{[1]} \\ w_2^{[1]} \end{bmatrix}. \tag{6.12}$$

Mit dem Ansatz

$$P = \begin{bmatrix} -H^{-1}A_{12} \\ I \end{bmatrix}$$

als Interpolationsmatrix, wobei $H \in \mathbb{R}^{n_1,n_1}$ eine Diagonalmatrix ist, erhält man mit (6.12) die Gleichung

$$-A_{11}^{-1}A_{12}w_2^{[1]} = -H^{-1}A_{12}w_2^{[1]} \tag{6.13}$$

und den folgenden Algorithmus aus [68].

Algorithmus 7: Adaptive Bestimmung der Interpolation

Input : $A \in \mathbb{R}^{n,n}$, einen Prototypen $w^{[1]} \in \mathbb{R}^n$, Anzahl der Iterationen $k_{\max}$.
Output : Interpolationsmatrix P.

1 **for** $k = 1, \ldots, k_{\max}$ **do**
2 Es ist $w^{[k]} = \begin{bmatrix} {w_1^{[k]}}^T & {w_2^{[k]}}^T \end{bmatrix}^T$.
3 Löse $A_{11} w_1^{[k]} = -A_{12} w_2^{[k]}$.
4 Definiere $P = \begin{bmatrix} -H^{-1}A_{12} \\ I \end{bmatrix}$ für die Diagonalmatrix H, die $H^{-1}A_{12}w_2^{[k]} = -w_1^{[k]}$ erfüllt.
5 Setze $w^{[k+1]} = P\left(\operatorname*{argmin}_{w \in \mathbb{R}^{n_2}} \frac{w^T P^T A P w}{w^T P^T P w} \right)$.

In Schritt 5 muss das verallgemeinerte Eigenwertproblem

$$P^T A P w = \lambda_{\min} P^T P w$$

gelöst werden, siehe z.B. [67]. Wenn $k_{\max}$ groß genug gewählt wird, ist $Aw^{[k_{\max}]} \approx 0$. Die bisher dargestellten Ideen der adaptiven Verfahren basieren auf Beobachtungen und numerischen Berechnungen. In [68] wurde gezeigt, dass der Algorithmus 7 das AMGr-Verfahren beschleunigt und die beste Performance durch die Interpolationsmatrix erzielt wird, die die Gleichung (6.13) erfüllt, wenn $w^{[1]}$ als Eigenvektor zum kleinsten Eigenwert von A gewählt wird.

Im zweiten Teil dieses Abschnittes wird gezeigt, dass die Wahl der Interpolationsmatrix durch den Algorithmus (7) für eine bestimmte Klasse von Matrizen auch in der Theorie zu einer guten Performance im Sinne von Satz 6.3 bzw. Lemma 6.4 führt.

Die angesprochene Klasse von Matrizen, die ähnliche Eigenschaften wie diagonaldominante Matrizen besitzen, ist die Menge der M-Matrizen.

Definition 6.7. Man definiert die Menge

$$\mathcal{Z}^{n,n} := \{A = [a_{ij}] \in \mathbb{R}^{n,n} \,|\, a_{ij} \leq 0,\ i \neq j\}.$$

Eine Matrix $A \in \mathcal{Z}^{n,n}$ heißt *(nichtsinguläre) M-Matrix*, wenn ein Vektor $w \in \mathbb{R}^n$, $w > 0$ mit

$$Aw > 0$$

existiert. Gilt $Aw = 0$ für ein $w > 0$, so heißt A *singuläre M-Matrix*. Hierbei wird die Relation „$>$ “ durch $\mathbb{R}^n \ni x = [x_k] > 0 \Leftrightarrow x_k > 0$ für alle k definiert.

Dies ist eine von vielen möglichen Charakterisierungen von M-Matrizen, die in [17] gefunden werden können. Insbesondere gibt es einen engen Zusammenhang zu nicht negativen Matrizen, denn es gilt, dass A genau dann eine M-Matrix ist, wenn eine Matrix $B \geq 0$ mit

$$A = sI - B \quad \text{und} \quad s > \rho(B)$$

existiert.

Mithilfe des Satzes von Perron[1]-Frobenius[2] können nicht negative Matrizen und daher auch M-Matrizen analysiert werden. Ein Beweis dieses Satzes kann u.a. in [17] gefunden werden.

Satz 6.8 (Perron-Frobenius)**.** *Sei $A \in \mathbb{R}^{n,n}$ nicht negativ, d.h. $A \geq 0$, und irreduzibel, das bedeutet, es existiert keine Permutationsmatrix $\pi \in \mathbb{R}^{n,n}$, so dass*

$$\pi A \pi^T = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}.$$

Dann gilt:

1. *Der Spektralradius $\rho(A)$ ist einfacher Eigenwert von A.*
2. *Zu $\rho(A)$ existiert ein positiver Eigenvektor $w \in \mathbb{R}^n$, d.h. $Aw = \rho(A)w$, $w > 0$. w heißt dann* Perron-Vektor *und ist - bis auf Skalierung - eindeutig.*
3. *Jeder Eigenvektor $v \in \mathbb{R}^{n,n}$ mit $v > 0$ von A ist Eigenvektor zum Eigenwert $\rho(A)$, ist also ein Perron-Vektor.*

Wendet man dieses Resultat auf M-Matrizen an, erhält man das folgende Korollar.

Korollar 6.9. *Sei $A \in \mathbb{R}^{n,n}$ eine (singuläre) M-Matrix, dann existiert ein Eigenvektor $w \in \mathbb{R}^n$ von A mit $w > 0$ und $Aw \geq 0$. Ist A zusätzlich symmetrisch, dann ist A symmetrisch positiv semidefinit (spsd) und es gilt*

$$Aw = \lambda_{\min}(A)w.$$

Beweis. Da A eine (singuläre) M-Matrix ist, existieren $B \geq 0$ und $s > \rho(B)$ mit $A = sI - B$. Mit Satz 6.8 existiert $w > 0$, der ein Eigenvektor von B zum Eigenwert $\rho(B)$ ist. Damit gilt

$$Aw = sw - Bw = (s - \rho(B))w,$$

also ist w Eigenvektor von A zum Eigenwert $s - \rho(B) \geq 0$. Ist A symmetrisch, so ist $\sigma(A) \subset \mathbb{R}$ und der kleinste Eigenwert von A ist gegeben durch $s - \rho(B)$. Daher ist A spsd. □

Dieses Korollar vereinfacht die Überprüfung einer Matrix auf ihre positive Definitheit. Anstatt zu überprüfen, ob für eine symmetrische Matrix $A \in \mathbb{R}^{n,n}$ und allen $x \in \mathbb{R}^n$ die Ungleichung $x^T Ax > 0$ gilt, reicht es für eine symmetrische Matrix $A \in \mathcal{Z}^{n,n}$ zu zeigen, dass ein Vektor $w \in \mathbb{R}^n$ mit $w > 0$ existiert, so dass $Aw \geq 0$ ist.

Sei im Folgenden $A \in \mathbb{R}^{n,n}$ eine symmetrische M-Matrix mit $Aw > 0$ und partitioniert wie in (4.3). Mit dem eben gezeigten Zusammenhang ist die Voraussetzung von Satz 6.3 erfüllt, wenn eine Matrix $H \in \mathcal{Z}^{n_1,n_1}$ mit

$$H \preceq A_{11} \quad \text{und} \tag{6.14}$$

$$\begin{bmatrix} H & A_{12} \\ A_{21} & A_{22} \end{bmatrix} w \geq 0 \tag{6.15}$$

[1] Oskar Perron (1880-1975) war ein deutscher Mathematiker. Unter anderem beschäftige er sich mit mehrdimensionale Kettenbrüche, himmelsmechanischen Problemen und Matrizentheorie.

[2] Ferdinand Georg Frobenius (1849-1917) war ein deutscher Mathematiker. Er beschäftigte sich hauptsächlich mit Gruppentheorie.

existiert. Dabei soll H leicht zu invertieren, z.B. eine Diagonalmatrix sein.

Da A eine M-Matrix ist, existiert ein positiver Vektor $w = [\, w_1^T \; w_2^T \,]^T$ mit $Aw > 0$. Sei H so konstruiert, dass H eine Diagonalmatrix ist und die Gleichung

$$w_1 = -H^{-1} A_{12} w_2 \tag{6.16}$$

erfüllt, dann gilt

$$\begin{bmatrix} H & A_{12} \\ A_{21} & A_{22} \end{bmatrix} w \geq 0, \tag{6.17}$$

also ist die Bedingung (6.15) gezeigt.

Um die Bedingung (6.14) nachzuweisen, setzt man $\widetilde{w}_1 := H^{\frac{1}{2}} w_1 > 0$. (6.16) impliziert

$$H w_1 \leq A_{11} w_1$$

und damit

$$H^{-\frac{1}{2}} A_{11} H^{-\frac{1}{2}} \widetilde{w}_1 \geq \widetilde{w}_1. \tag{6.18}$$

Seien a_{ii}, h_{ii} die Diagonaleinträge von A_{11} bzw. H, dann erhält man

$$h_{ii} \leq a_{ii} \quad \text{und damit} \quad H^{-\frac{1}{2}} A_{11} H^{-\frac{1}{2}} - I \in \mathcal{Z}^{n,n}.$$

Da mit (6.18) $(H^{-\frac{1}{2}} A_{11} H^{-\frac{1}{2}} - I)\widetilde{w}_1 \geq 0$ ist, ist $H^{-\frac{1}{2}} A_{11} H^{-\frac{1}{2}} - I$ eine (singuläre) M-Matrix und die Eigenwerte von $H^{-\frac{1}{2}} A_{11} H^{-\frac{1}{2}} - I$ liegen in der rechten Halbebene. Außerdem ist $H^{-\frac{1}{2}} A_{11} H^{-\frac{1}{2}} - I$ symmetrisch, daher auch spd und es folgt, dass $H^{-\frac{1}{2}} A_{11} H^{-\frac{1}{2}} \succeq I$ und damit (6.14) erfüllt ist. Der folgende Zusammenhang wurde gezeigt:

Satz 6.10. *Sei $A \in \mathbb{R}^{n,n}$ eine symmetrische M-Matrix und partitioniert wie in* (4.3)*. Weiter sei $H \in \mathbb{R}^{n_1,n_1}$ hpd und eine Diagonalmatrix, die die Gleichung*

$$w_1 = -H^{-1} A_{12} w_2$$

erfüllt, wobei $w = [\, w_1^T \; w_2^T \,]^T \in \mathbb{R}^n$ mit $w > 0$ und $Aw > 0$. Dann lässt sich die C.B.S. Konstante von $\widehat{A}$, definiert wie in (6.2)*, durch* (6.5) *abschätzen.*

Der benötigte Vektor w kann dabei durch Algorithmus 7 approximiert werden.

6.5 Verallgemeinerung der Abschätzung der C.B.S. Konstante von Notay.

In diesem Abschnitt wird eine weitere Abschätzung der C.B.S. Konstante, ähnlich der aus Satz 6.3, vorgestellt und damit eine Alternative zu der Voraussetzung $H \preceq A_{11}$ besprochen.

Satz 6.11. *Sei A hpd und partitioniert wie in* (4.3)*. Weiter sei $\widehat{A}$ definiert durch* (6.2) *mit einer Approximation H, die die Bedingungen*

- *$H \succeq A_{11}$ und*
- $\begin{bmatrix} H(H + HA_{11}^{-1}H - A_{11})^{-1}H & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ *sei hpsd*

erfüllt. Dann gilt für die C.B.S. Konstante von $\widehat{A}$ die Ungleichung

$$\gamma(\widehat{A}) \leq \sqrt{1 - \lambda_{\min}(H^{-1}A_{11})}. \tag{6.19}$$

Beweis. Um dies zu beweisen, ist eine obere Schranke von

$$\frac{x_2^H \widehat{A}_{22} x_2}{x_2^H S(\widehat{A}, \widehat{A}_{11}) x_2} \text{ für alle } x_2 \in \mathbb{C}^{n_2}$$

zu finden. Mithilfe der Gleichung (5.5) aus Lemma 5.2 erhält man eine Abschätzung der C.B.S. Konstante.

Da die Matrix

$$\begin{bmatrix} H(H+HA_{11}^{-1}H-A_{11})^{-1}H & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

nach Voraussetzung hpsd ist, folgt

$$x_2^H A_{22} x_2 \geq x_2^H A_{21}(H^{-1} + A_{11}^{-1} - H^{-1}A_{11}H^{-1})A_{12}x_2. \tag{6.20}$$

Dies führt für alle $x_2 \in \mathbb{C}^{n_2}$ und $c = \beta - 1$, $\beta = \big(\lambda_{\min}(H^{-1}A_{11})\big)^{-1}$ auf

$$\begin{aligned}
&(1+c)x_2^H S(\widehat{A}, \widehat{A}_{11})x_2 - x_2^H \widehat{A}_{22} x_2 \\
&= (1+c)x_2^H(A_{22} - A_{21}A_{11}^{-1}A_{12})x_2 - x_2^H(A_{22} - A_{21}(2H^{-1} - H^{-1}A_{11}H^{-1})A_{12})x_2 \\
&\overset{(6.20)}{\geq} c x_2^H A_{21}(H^{-1} + A_{11}^{-1} - H^{-1}A_{11}H^{-1})A_{12}x_2 \\
&\quad + x_2^H A_{21}(2H^{-1} - (1+c)A_{11}^{-1} - H^{-1}A_{11}H^{-1})A_{12}x_2 \\
&= x_2^H A_{21} A_{11}^{-\frac{1}{2}} \left((2+c)A_{11}^{\frac{1}{2}} H^{-1} A_{11}^{\frac{1}{2}} - I - (1+c)A_{11}^{\frac{1}{2}} H^{-1} A_{11} H^{-1} A_{11}^{\frac{1}{2}} \right) A_{11}^{-\frac{1}{2}} A_{12} x_2.
\end{aligned}$$

Der letzte Term ist genau dann nicht negativ, wenn das Polynom

$$p(t) = (2+c)t - 1 - (1+c)t^2$$

auf $\sigma(H^{-1}A_{11}) \subseteq [\frac{1}{\beta}, 1]$ nicht negativ ist. Durch

$$p(t) = -\Big(1+c\Big)\Big(t^2 - \frac{2+c}{1+c}t + \frac{1}{1+c}\Big) = -\Big(1+c\Big)\Big(t-1\Big)\Big(t - \frac{1}{1+c}\Big),$$

folgt, dass p auf $\left[\frac{1}{1+c}, 1\right] = \left[\frac{1}{\beta}, 1\right]$ nicht negativ ist.

Damit wurde gezeigt, dass

$$\frac{x_2^H \widehat{A}_{22} x_2}{x_2^H S(\widehat{A}, \widehat{A}_{11}) x_2} \leq 1 + c = \beta \text{ für alle } x_2 \in \mathbb{C}^{n_2}.$$

□

Betrachtet man die Voraussetzungen des obigen Satzes, erkennt man die Ähnlichkeit zu Satz 6.3. Auch hier wird in der Matrix A der (1×1)-Block durch eine weitere Matrix ersetzt, die im hpd Sinn kleiner als A_{11} ist.

Nachfolgend wird gezeigt, dass Satz 6.11 eine Verallgemeinerung von Satz 6.3 ist. Satz 6.3 benötigt eine Approximation H mit $H \preceq A_{11}$ und bei Satz 6.11 soll $H \succeq A_{11}$ erfüllt sein. Sei zunächst H so konstruiert, dass die Voraussetzung von Satz 6.11 erfüllt ist. Dann gilt für diese Matrix die Bedingung $\sigma(H^{-1}A_{11}) \subseteq [a, 1]$ mit $a = \lambda_{\min}(H^{-1}A)$. Nach einer Skalierung der Approximation H mit a und anschließender Umbenennung zu $\widetilde{H}$ erhält man $\sigma(\widetilde{H}^{-1}A_{11}) \subseteq [1, \frac{1}{a}]$. Die Approximation $\widetilde{H}$ erfüllt die erste Voraussetzung von Satz 6.3. Es gilt sogar

$$\frac{1}{\lambda_{\max}(\widetilde{H}^{-1}A_{11})} = \frac{1}{\lambda_{\max}((aH)^{-1}A_{11})} = \lambda_{\min}(H^{-1}A_{11}).$$

Das heißt, beide Sätze erzielen dieselbe Abschätzung für die C.B.S. Konstante von $\widehat{A}$.

Wie verhält es sich mit der jeweiligen zweiten Voraussetzung für H bzw. $\widetilde{H}$? Dazu betrachtet man den jeweiligen (1×1)-Block von der 2×2 Matrix, die nach Voraussetzung hpsd sein soll. Für Satz 6.11 ist dies $H(H + HA_{11}^{-1}H - A_{11})^{-1}H$ und für Satz 6.3 $\widetilde{H} = aH$. Damit erhält man für alle $x_1 \in \mathbb{C}^{n_1}$

$$\begin{aligned} & x_1^H H(H + HA_{11}^{-1}H - A_{11})^{-1}Hx_1 \geq x_1^H aHx_1 \\ \Leftrightarrow \quad & x_1^H Hx_1 \geq x_1^H a(H + HA_{11}^{-1}H - A_{11})x_1 \\ \Leftrightarrow \quad & x_1^H(1-a)Hx_1 + ax_1^H H^{\frac{1}{2}}(H^{\frac{1}{2}}A_{11}^{-1}H^{\frac{1}{2}} - H^{-\frac{1}{2}}A_{11}H^{-\frac{1}{2}})H^{\frac{1}{2}}x_1 \geq 0. \end{aligned} \tag{6.21}$$

Da $a \leq 1$ ist, ist der erste Term $x_1^H(1-a)Hx_1$ nicht negativ. Der zweite Term ist ebenfalls nicht negativ, da die Abbildung $t \mapsto \frac{1}{t} - t$ für $0 \leq t \leq 1$ nicht negativ ist.

Zusammenfassend wurde gezeigt, dass die Voraussetzungen von Satz 6.11 schwächer, als die von Satz 6.3 sind. Deshalb stellt Satz 6.11 eine Verallgemeinerung von Satz 6.3 dar und es können die für Satz 6.11 benötigten Voraussetzungen durch eine Skalierung bei diagonaldominanten Matrizen immer erfüllt werden, siehe Satz 6.6. Durch Anwendung des Greedy-Coarsers kann außerdem eine weitere Approximation für diagonaldominante hpd Matrizen angegeben werden, mit der die Voraussetzung von Satz 6.11 erfüllt ist.

Satz 6.12. *Sei $A = [a_{jk}]$ hpd und diagonaldominant. Weiter seien $\varphi \in (\frac{3}{4}, 1)$ und*

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

die permutierte Matrix nach Anwendung des Greedy-Coarsers auf A. Außerdem sei $H = \operatorname{diag}(h_{11}, \ldots, h_{n_1,n_1}) \in \mathbb{C}^{n_1,n_1}$ mit $h_{jj} := \sum_{k=1}^{n_1} |\, a_{jk} \,|$, dann gilt

$$A_{11} \preceq H \preceq \tfrac{1}{2\varphi - 1} A_{11}$$

und

$$\begin{bmatrix} H(H + HA_{11}^{-1}H - A_{11})^{-1}H & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \text{ ist hpsd.}$$

Beweis. Definiert man $\varphi_j := \frac{|\, a_{jj} \,|}{\sum_{k=1}^{n_1} |\, a_{jk} \,|}$, so ist $\varphi_j \geq \varphi$ und $h_{jj} = \frac{|\, a_{jj} \,|}{\varphi_j}$. Daher folgt

$$P = [p_{jk}] := H^{-1}A_{11} \quad \text{mit} \quad p_{jk} = \frac{a_{jk}}{|\, a_{jj} \,|}\varphi_j,$$

wobei $|\,a_{jj}\,| = a_{jj}$. Mit dem Satz von Gershgorin (Satz 6.5) ergibt sich für die Eigenwerte $\lambda(P)$ von P

$$\begin{aligned}|\,\lambda(P) - \varphi_j\,| = |\,\lambda(P) - p_{jj}\,| &\leq \sum_{\substack{k=1\\k\neq j}}^{n_1} p_{jk} = \sum_{\substack{k=1\\k\neq j}}^{n_1} \frac{|\,a_{jk}\,|}{|\,a_{jj}\,|}\varphi_j \\ &= \sum_{k=1}^{n_1} \Big(\frac{|\,a_{jk}\,|}{|\,a_{jj}\,|} - 1\Big)\varphi_j = \Big(\tfrac{1}{\varphi_j} - 1\Big)\varphi_j = 1 - \varphi_j.\end{aligned}$$

Das impliziert

$$\sigma(H^{-1}A_{11}) \subset [2\varphi - 1, 1].$$

Damit folgt außerdem, dass $\lambda_{\min}(H^{-1}A_{11}) \geq 2 \cdot \frac{3}{4} - 1 = \frac{1}{2}$.

Um die zweite Bedingung zu zeigen, sei $\widehat{H}$ definiert durch Gleichung (6.11). Mit Satz 6.6 folgt, dass

$$\begin{bmatrix} \widehat{H} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

hpsd ist und $\sigma(\widehat{H}^{-1}A_{11}) \subset [1, \frac{1}{2\varphi-1}]$. Mit (6.21) folgt für $\widetilde{H} := \frac{1}{2\varphi-1}\widehat{H}$, dass

$$A_{11} \preceq \widetilde{H} \preceq \tfrac{1}{2\varphi-1}A_{11} \quad \text{und} \quad \begin{bmatrix} \widetilde{H}(\widetilde{H} + \widetilde{H}A_{11}^{-1}\widetilde{H} - A_{11})^{-1}\widetilde{H} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \succeq 0.$$

Im weiteren Verlauf stellen $h_{jj}, \widehat{h}_{jj}$ und $\widetilde{h}_{jj}$ die Diagonaleinträge der Matrizen $H, \widehat{H}$ bzw. $\widetilde{H}$ dar. Da $\widehat{h}_{jj} = (2 - \frac{1}{\varphi_j})a_{jj}$ ist, erhält man

$$\widetilde{h}_{jj} = \tfrac{1}{2\varphi-1}\widehat{h}_{jj} = \tfrac{1}{2\varphi-1}(2 - \tfrac{1}{\varphi_j})a_{jj} \geq \tfrac{1}{\varphi_j}a_{jj} = h_{jj},$$

also $H \preceq \widetilde{H}$. Da das Polynom $P(t) = t + 1 - t^2$ auf $[\frac{1}{2}, 1]$ monoton wachsend ist, folgt

$$A^{-\frac{1}{2}}\Big[A^{\frac{1}{2}}H^{-1}A^{\frac{1}{2}} + I - A^{\frac{1}{2}}H^{-1}AH^{-1}A^{\frac{1}{2}}\Big]A^{-\frac{1}{2}} \preceq A^{-\frac{1}{2}}\Big[A^{\frac{1}{2}}\widetilde{H}^{-1}A^{\frac{1}{2}} + I - A^{\frac{1}{2}}\widetilde{H}^{-1}A\widetilde{H}^{-1}A^{\frac{1}{2}}\Big]A^{-\frac{1}{2}}$$

und daher

$$H(H + HA_{11}^{-1}H - A_{11})^{-1}H \succeq \widetilde{H}(\widetilde{H} + \widetilde{H}A_{11}^{-1}\widetilde{H} - A_{11})^{-1}\widetilde{H}.$$

□

Außerdem gibt es weitere Klassen von Matrizen, für die die Erfüllbarkeit der Voraussetzungen von Satz 6.11 nachgewiesen werden kann.

Satz 6.13. *Sei A hpd und partitioniert wie in* (4.3)*. Weiter sei $\gamma(A)$ die C.B.S. Konstante von A und H eine Approximation von A_{11} mit $H \succeq A_{11}$. Falls eine der folgenden Bedingungen*

1. $\gamma(A) \leq \sqrt{\frac{4}{5}} \approx 0.8944$ *oder*
2. $\lambda_{\min}(H^{-1}A_{11}) \geq \frac{\sqrt{5\gamma(A)^2-4}+\gamma(A)}{2\gamma(A)}$ *für* $\gamma(A) > \sqrt{\frac{4}{5}}$

erfüllt ist, erfüllt H die Voraussetzung von Satz 6.11.

Beweis. Die 2×2-Block Matrix ist genau dann hpsd, wenn der (1×1)-Block sowie das zugehörige Schurkomplement hpsd sind. Durch (5.4) folgt $x_2^H A_{22} x_2 \geq \frac{1}{\gamma(A)^2} x_2^H A_{21} A_{11}^{-1} A_{12} x_2$ für alle $x_2 \in \mathbb{C}^{n_2}$. Mit

$$H(H + HA_{11}^{-1}H - A_{11})^{-1}H = (H^{-1} + A_{11}^{-1} - H^{-1}A_{11}H^{-1})^{-1}$$

und $x_1^H H x_1 \geq x_1^H A_{11} x_1$ ist die Voraussetzung von Satz 6.11 erfüllt, wenn für alle $x_2 \in \mathbb{C}^{n_2}$ gilt

$$x_2^H \Big(A_{22} - A_{21}(H^{-1} + A_{11}^{-1} - H^{-1}A_{11}H^{-1})A_{12} \Big) x_2 \geq 0.$$

Damit ist

$$\begin{aligned} & x_2^H \Big(A_{22} - A_{21}(H^{-1} + A_{11}^{-1} - H^{-1}A_{11}H^{-1})A_{12} \Big) x_2 \\ \geq & x_2^H \left(\frac{1}{\gamma(A)^2} A_{21} A_{11}^{-1} A_{12} - A_{21}(H^{-1} + A_{11}^{-1} - H^{-1}A_{11}H^{-1})A_{12} \right) x_2 \\ = & x_2^H A_{21} A_{11}^{-\frac{1}{2}} \left((\frac{1}{\gamma(A)^2} - 1)I - X + X^2 \right) A_{11}^{-\frac{1}{2}} A_{12} x_2 \end{aligned} \tag{6.22}$$

mit $X = A_{11}^{\frac{1}{2}} H^{-1} A_{11}^{\frac{1}{2}}$. Hierbei ist (6.22) nicht negativ, wenn $p(t) := t^2 - t + (\frac{1}{\gamma(A)^2} - 1) \geq 0$ auf $\sigma(H^{-1}A_{11})$. Für das Polynom p gilt

$$\begin{aligned} & p = t^2 - t + (\frac{1}{\gamma(A)^2} - 1) = (t - \frac{1}{2})^2 + \frac{1}{\gamma(A)^2} - \frac{5}{4} \geq 0 \\ \Leftrightarrow \quad & (t - \frac{1}{2})^2 \geq \frac{5\gamma(A)^2 - 4}{4\gamma(A)^2} \end{aligned} \tag{6.23}$$

für $t \in \sigma(H^{-1}A_{11})$.

Daher ist für $\gamma(A) \leq \sqrt{\frac{4}{5}}$ die Ungleichung (6.23) erfüllt. Falls $\gamma(A) \geq \sqrt{\frac{4}{5}}$ ist, folgt die Bedingungen für die Eigenwerte von $H^{-1}A_{11}$. □

Insbesondere das erste Kriterium ist erwähnenswert. In [6] nutzt Axelsson solche Approximationen H für die Reduktion der C.B.S. Konstante einer Matrix A, die in einer hierarchischen Basis gegeben ist, vergleiche Abschnitt 6.1. In solch einer Basis ist die C.B.S. Konstante gewöhnlich nach oben durch $\sqrt{\frac{4}{5}}$ beschränkt, siehe [18], [63] und [83]. Des Weiteren kann für generelle elliptische 2D Probleme mit stückweisen linearen Basisfunktionen gezeigt werden, dass $\gamma(A)^2 < \frac{3}{4}$, [5, 72]. In [73] wurde dieselbe Schranke für 2D Elastizitäts-Probleme gefunden, siehe dazu [6].

Mit dem Gezeigten lässt sich das 2BF-Verfahren nach einem Basiswechsel analysieren.

Korollar 6.14. *Sei A hpd und partitioniert wie in* (4.3) *mit der C.B.S. Konstante $\gamma(A)$. Weiter sei $\widehat{B}_{2\mathrm{bf}}$ definiert durch* (6.4)*, so dass E σ-größer als A_{11} mit β ist und eine der folgenden Bedingungen erfüllt ist*

1. $\gamma(A) \leq \sqrt{\frac{4}{5}} \approx 0.8944$,
2. $\lambda_{\min}(E^{-1}A_{11}) \geq \frac{\sqrt{5\gamma(A)^2 - 4} + \gamma(A)}{2\gamma(A)}$, *für* $\gamma(A) > \sqrt{\frac{4}{5}}$.

Außerdem sei S σ-größer als

$$A_c := A_{22} - A_{21}\Big[I - \big(I - E^{-1}A_{11}\big)^3\Big]A_{11}^{-1}A_{12} \tag{6.24}$$

mit θ. Dann ist $\widehat{B}_{2\mathrm{bf}}$ σ-größer als A mit

$$\theta_{two} = 1 + \frac{1}{\beta}\left\{\theta - 1 + \frac{1}{2}\frac{\beta-1}{\beta}\left(\beta + \theta - 1 + \Big[(\beta+\theta-1)^2 + (\theta-1)(\beta-1)\Big]^{\frac{1}{2}}\right)\right\}. \tag{6.25}$$

Beweis. Mit Satz 5.6 folgt, dass

$$\theta_{two} = 1+\frac{1}{\beta}\left\{\theta - 1 + \frac{1}{2}\frac{\beta-1}{1-\gamma(\widehat{A})^2}\left(\beta + \theta - 1 + \Big[(\beta+\theta-1)^2 + \beta(\theta-1)\gamma(A)^2\Big]^{\frac{1}{2}}\right)\right\}, \tag{6.26}$$

wobei $\gamma(\widehat{A})$ die C.B.S Konstante von der durch (6.2) definierten Matrix $\widehat{A}$ ist. Durch die Anwendung von Satz 6.13 kann diese durch $\sqrt{1-\lambda_{\min}(E^{-1}A_{11})}$ abgeschätzt werden. Dies führt auf

$$\frac{1}{1-\gamma(\widehat{A})^2} \leq \frac{1}{\lambda_{\min}(E^{-1}A_{11})} \leq \beta$$

und

$$\gamma(\widehat{A})^2 \leq 1 - \frac{1}{\lambda_{\min}(E^{-1}A_{11})} \leq \frac{\beta-1}{\beta}.$$

Diese Ungleichung eingesetzt in (6.26) schließt den Beweis ab. □

Die durch (6.25) erhaltene obere Schranke der Konditionszahl von $\widehat{B}_{2\mathrm{bf}}^{-1}A$ ist unabhängig von der C.B.S. Konstante von A. Anhand dieses Korollars kann beobachtet werden, dass insbesondere eine geeignete Wahl der Grobgittermatrix A_c, gegeben durch (6.24), zu einer Konditionsabschätzung unabhängig von $\gamma(A)$ führen kann.

6.6

Konstruktion eines polynombasierten Vorkonditionierers. Bisher wurde der Vorkonditionierer $B_{2\mathrm{bf}}$ und einige Abwandlungen analysiert. Dieser Vorkonditionierer basiert auf einer Block-Faktorisierung der Systemmatrix A, siehe Abschnitt 4.3.2. Durch eine Kongruenztransformation auf verallgemeinerte hierarchische Basen kann das Konvergenz- bzw. Vorkonditionierungsverhalten verbessert werden, siehe Abschnitte 6.1 und 6.5. In diesem Abschnitt wird eine Möglichkeit hergeleitet, um dieses Verhalten durch rekursive Anwendung geeigneter Kongruenztransformationen weiter zu verbessern. Dazu wird erneut die Block-Faktorisierung (4.9) und der daraus entstandene Vorkonditionierer

$$B_{2\mathrm{bf}} = \begin{bmatrix} I & 0 \\ A_{21}E^{-1} & I \end{bmatrix}\begin{bmatrix} E & 0 \\ 0 & S \end{bmatrix}\begin{bmatrix} I & E^{-1}A_{12} \\ 0 & I \end{bmatrix} \tag{6.27}$$

mit E und S (beide hpd) betrachtet. Setzt man

$$J^{-1} = \begin{bmatrix} I & E^{-1}A_{12} \\ 0 & I \end{bmatrix} \quad \text{bzw.} \quad J = \begin{bmatrix} I & -E^{-1}A_{12} \\ 0 & I \end{bmatrix},$$

so ergibt (6.27) die Gleichung $B_{2\mathrm{bf}} = J^{-H} B_{\mathrm{diag}} J^{-1}$ bzw. $B_{2\mathrm{bf}}^{-1} = J B_{\mathrm{diag}}^{-1} J^H$ mit

$$B_{\mathrm{diag}} := \begin{bmatrix} E & 0 \\ 0 & S \end{bmatrix}.$$

Durch die spektrale Identität

$$\sigma\Big(B_{2\mathrm{bf}}^{-1} A\Big) = \sigma(J B_{\mathrm{diag}}^{-1} J^H A) = \sigma(B_{\mathrm{diag}}^{-1} J^H A J)$$

wird deutlich, dass der Vorkonditionierer $B_{2\mathrm{bf}}$ als eine, durch hierarchische Basen transfomierte, Abwandlung des Block-Jacobi-Vorkonditionierers B_{diag} gesehen werden kann. Verfolgt man diesen Gedanken weiter, so entsteht $\widehat{B}_{2\mathrm{bf}}$ aus der Matrix B_{diag} durch zweimalige Anwendung der Kongruenztransformation. Im weiteren Verlauf werden Vorkonditionierer untersucht, die durch eine k-fache rekursive Anwendung solcher Transformationen hervorgehen. Als Grundlage bzw. 0-te Iteration wird die Matrix B_{diag} verwendet, die unter anderem in [4, 5, 9] und [81] untersucht wird.

Sei $A \in \mathbb{C}^{n,n}$ gegeben und partitioniert wie in (4.3). Diese Matrix stellt die „Ausgangsmatrix“ dar und wird im Folgenden auch als $A_{[0]}$ bezeichnet. Durch rekursiv definierte generalisierte hierarchische Basen kann eine ganze Klasse von Matrizen konstruiert werden.

Lemma 6.15. *Sei $A = A_{[0]}$ partitioniert wie in* (4.3)*. Weiter seien für $k = 0, 1, 2, \ldots$ Matrizen auf* rekursiv definierten generalisierten hierarchischen Basen *durch*

$$A_{[k+1]} := J_{[k]}^H A_{[k]} J_{[k]} \tag{6.28}$$

definiert, wobei die Transformationsmatrizen durch

$$J_{[k]} := \begin{bmatrix} I & -H^{-1} A_{[k]_{12}} \\ 0 & I \end{bmatrix},$$

mit $H \approx A_{11}$ hpd gegeben sind. Dann erfüllt $A_{[k]}$ die Gleichung

$$A_{[k]} = \begin{bmatrix} A_{11} & (I - A_{11} H^{-1})^k A_{12} \\ A_{21}(I - H^{-1} A_{11})^k & A_{22} + A_{21}\Big[I - (I - H^{-1} A_{11})^{2k}\Big] A_{11}^{-1} A_{12} \end{bmatrix}. \tag{6.29}$$

Beweis. Der Beweis wird per vollständiger Induktion durchgeführt. Für den Induktionsanfang ist

$$\begin{aligned} A_{[1]} = J_{[0]}^H A J_{[0]} &= \begin{bmatrix} I & 0 \\ -A_{21} H^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & (I - A_{11} H^{-1}) A_{12} \\ A_{21} & A_{22} - A_{21} H^{-1} A_{12} \end{bmatrix} \\ &= \begin{bmatrix} A_{11} & (I - A_{11} H^{-1}) A_{12} \\ A_{21}(I - H^{-1} A_{11}) & A_{22} - A_{21}(2H^{-1} - H^{-1} A_{11} H^{-1}) A_{12} \end{bmatrix}. \end{aligned}$$

Mit

$$A_{22} - A_{21}(2H^{-1} - H^{-1} A_{11} H^{-1}) A_{12} = A_{22} + A_{21}\Big[I - (I - H^{-1} A_{11})^2\Big] A_{11}^{-1} A_{12}$$

folgt die Behauptung für $k = 1$. Nun sei die Gleichung

$$A_{[k]} = \begin{bmatrix} A_{11} & (I - A_{11}H^{-1})^k A_{12} \\ A_{21}(I - H^{-1}A_{11})^k & A_{22} + A_{21}\Big[I - (I - H^{-1}A_{11})^{2k}\Big]A_{11}^{-1}A_{12} \end{bmatrix}$$

für ein $k \in \mathbb{N}$ erfüllt. Dann folgt für $k + 1$

$$\begin{aligned}
A_{[k+1]} = J_{[k]}^H A_{[k]} J_{[k]} &= \begin{bmatrix} I & 0 \\ -A_{[k]_{21}}H^{-1} & I \end{bmatrix} A_{[k]} \begin{bmatrix} I & -H^{-1}A_{[k]_{12}} \\ 0 & I \end{bmatrix} \\
&= \begin{bmatrix} I & 0 \\ -A_{21}(I - H^{-1}A_{11})^k H^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & (I - A_{11}H^{-1})^k A_{12} \\ A_{21}(I - H^{-1}A_{11})^k & \star \end{bmatrix} \\
&\quad \cdot \begin{bmatrix} I & -H^{-1}(I - A_{11}H^{-1})^k A_{12} \\ 0 & I \end{bmatrix} \\
&= \begin{bmatrix} A_{11} & (I - A_{11}H^{-1})^k A_{12} \\ A_{21}(I - H^{-1}A_{11})^k (I - H^{-1}A_{11}) & \star \end{bmatrix} \\
&\quad \cdot \begin{bmatrix} I & -H^{-1}(I - A_{11}H^{-1})^k A_{12} \\ 0 & I \end{bmatrix} \\
&= \begin{bmatrix} A_{11} & (I - A_{11}H^{-1})^{k+1} A_{12} \\ A_{21}(I - H^{-1}A_{11})^{k+1} & A_{[k+1]_{22}} \end{bmatrix}.
\end{aligned}$$

Es bleibt zu zeigen, dass

$$A_{[k+1]_{22}} = A_{22} + A_{21}(I - (I - H^{-1}A_{11})^{2(k+1)})A_{11}^{-1}A_{12}$$

ist, was mithilfe des Schurkomplements gezeigt werden kann. Durch Transformationen der Art (6.28) wird sowohl der (1×1)-Block sowie das Schurkomplement nicht verändert. Es gilt daher für alle $k = 1, 2, 3, \ldots$, dass $A_{11} = A_{[k]_{11}}$ und

$$S(A_{[k]}, A_{11}) = S(A, A_{11}) = A_{22} - A_{21}A_{11}^{-1}A_{12}.$$

Daher ist

$$S(A_{[k+1]}, A_{11}) = A_{22} - A_{21}A_{11}^{-1}A_{12} = A_{[k+1]_{22}} - A_{[k+1]_{21}}A_{11}^{-1}A_{[k+1]_{12}}$$

und das impliziert

$$\begin{aligned}
A_{[k+1]_{22}} &= A_{22} - A_{21}A_{11}^{-1}A_{12} + A_{[k+1]_{21}}A_{11}^{-1}A_{[k+1]_{12}} \\
&= A_{22} - A_{21}A_{11}^{-1}A_{12} + A_{21}(I - H^{-1}A_{11})^{(k+1)}A_{11}^{-1}(I - A_{11}H^{-1})^{(k+1)}A_{12} \\
&= A_{22} - A_{21}A_{11}^{-1}A_{12} + A_{21}(I - H^{-1}A_{11})^{(k+1)}(I - H^{-1}A_{11})^{(k+1)}A_{11}^{-1}A_{12} \\
&= A_{22} - A_{21}\left[I - (I - H^{-1}A_{11})^{2(k+1)}\right]A_{11}^{-1}A_{12}.
\end{aligned}$$

Damit ist die Induktion abgeschlossen. □

Durch Lemma 6.15 ist es gelungen, die Matrix A durch eine rekursiv definierte Basistransformation auf eine generalisierte hierarchische Basis zu bringen. Das nächste Resultat gibt eine explizite Darstellung der Transformationsmatrizen an.

Lemma 6.16. *Sei $A = A_{[0]}$ partitioniert wie in* (4.3) *und die Voraussetzungen von Lemma 6.15 erfüllt . Weiter seien für $k = 0, 1, 2, \ldots$ Matrizen $A_{[k]}$ rekursiv definiert durch* (6.28)*. Dann ist*

$$A_{[k]} = J_{[\to k]}^H A J_{[\to k]}$$

mit

$$J_{[\to k]} := \begin{bmatrix} I & -\sum_{j=0}^{k-1} H^{-1}(I - A_{11}H^{-1})^j A_{12} \\ 0 & I \end{bmatrix}. \tag{6.30}$$

Beweis. Mit (6.28) ist

$$A_{[k]} = J_{[k-1]}^H A_{[k-1]} J_{[k-1]} = J_{[k-1]}^H \big(J_{[k-2]}^H A_{[k-2]} J_{[k-2]}\big) J_{[k-1]} = \big(\prod_{j=0}^{k-1} J_{[j]}\big)^H A \big(\prod_{j=0}^{k-1} J_{[j]}\big),$$

mit

$$J_{[j]} = \begin{bmatrix} I & -H^{-1}A_{[j]_{12}} \\ 0 & I \end{bmatrix} = \begin{bmatrix} I & -H^{-1}(I - A_{11}H^{-1})^j A_{12} \\ 0 & I \end{bmatrix}.$$

Weiter ist

$$\prod_{j=0}^{k-1} J_{[j]} = \prod_{j=0}^{k-1} \begin{bmatrix} I & -H^{-1}(I - A_{11}H^{-1})^j A_{12} \\ 0 & I \end{bmatrix} = \begin{bmatrix} I & -\sum_{j=0}^{k-1} H^{-1}(I - A_{11}H^{-1})^j A_{12} \\ 0 & I \end{bmatrix}.$$

Mit der Definition von $J_{[\to k]}$ folgt daher

$$A_{[k]} = J_{[\to k]}^H A J_{[\to k]},$$

was zu zeigen war. □

Die Konstruktion der rekursiv definierten hierarchischen Basen ist eine elegante Möglichkeit, die Ideen der generalisierten hierarchischen Basis aus [6, 81] zu verallgemeinern. Dass diese Verallgemeinerung einen Nutzen im Sinne eines besseren Konvergenzverhaltens mit sich bringt, wird in Kapitel 7 erläutert. Doch zuvor motiviert die folgende Betrachtung die Verwendung dieser Transformationen.

Mit der Annahme, dass A_{11} und H ein konvergentes Splitting erzeugen, also $\rho(H^{-1}A_{11}) < 1$ ist, folgt mit der Neumann'schen Reihe

$$\begin{aligned} \lim_{k\to\infty} \prod_{j=0}^{k} J_{[j]} &= \lim_{k\to\infty} \begin{bmatrix} I & -\sum_{j=0}^{k} H^{-1}(I - A_{11}H^{-1})^j A_{12} \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} I & -H^{-1}\sum_{j=0}^{\infty}(I - A_{11}H^{-1})^j A_{12} \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} I & -H^{-1}\big[I - (I - A_{11}H^{-1})\big]^{-1} A_{12} \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix}, \end{aligned}$$

also

$$\gamma\Big(\lim_{k\to\infty} A_{[k]}\Big) = 0.$$

Der Sinn der rekursiv definierten hierarchischen Basis ist es, Vorkonditionierer für die Matrix A zu konstruieren, die wie folgt definiert werden.

Definition 6.17. Sei $A \in \mathbb{C}^{n,n}$ hpd und partitioniert wie in (4.3). Dann wird der *k-fache h-transformierte Vorkonditionierer von A* durch

$$B_{[k]} := J_{[\to k]}^{-H} B_{\text{diag}} J_{[\to k]}^{-1}$$

definiert. Hierbei ist

$$B_{\text{diag}} = \begin{bmatrix} E & 0 \\ 0 & S \end{bmatrix} \tag{6.31}$$

mit E und S hpd und $J_{[\to k]}$ definiert durch (6.30).

Für diesen Vorkonditionierer gilt

$$\sigma\Big(B_{[k]}^{-1}A\Big) = \sigma\Big(J_{[\to k]}B_{\text{diag}}^{-1}J_{[\to k]}^{H}A\Big) = \sigma\Big(B_{\text{diag}}^{-1}J_{[\to k]}^{H}AJ_{[\to k]}\Big) = \sigma\Big(B_{\text{diag}}^{-1}A_{[k]}\Big) \tag{6.32}$$

mit $A_{[k]}$ gegeben durch (6.29). Des Weiteren kann eine explizite Darstellung gefunden werden.

Lemma 6.18. *Sei $B_{[k]}$ gegeben wie in Definition 6.17, dann gilt*

$$B_{[k]} = \begin{bmatrix} E & EH^{-1}\sum_{j=0}^{k-1}(I-A_{11}H^{-1})^j A_{12} \\ A_{21}\sum_{j=0}^{k-1}(I-H^{-1}A_{11})^j H^{-1}E, & S+A_{21}\sum_{j=0}^{k-1}(I-H^{-1}A_{11})^j H^{-1}EH^{-1}\sum_{j=0}^{k-1}(I-A_{11}H^{-1})^j A_{12} \end{bmatrix}.$$

Beweis. Durch die Definition von $B_{[k]}$ folgt

$$\begin{aligned}
B_{[k]} &= J_{[\to k]}^{-H} B_{\text{diag}} J_{[\to k]}^{-1} \\
&= \begin{bmatrix} I & -\sum_{j=0}^{k-1} H^{-1}(I-H^{-1}A_{11})^j A_{12} \\ 0 & I \end{bmatrix}^{-H} B_{\text{diag}} \begin{bmatrix} I & -\sum_{j=0}^{k-1} H^{-1}(I-A_{11}H^{-1})^j A_{12} \\ 0 & I \end{bmatrix}^{-1} \\
&= \begin{bmatrix} I & 0 \\ A_{21}\sum_{j=0}^{k-1}(I-H^{-1}A_{11})^j H^{-1} & I \end{bmatrix} B_{\text{diag}} \begin{bmatrix} I & H^{-1}\sum_{j=0}^{k-1}(I-A_{11}H^{-1})^j A_{12} \\ 0 & I \end{bmatrix} \\
&= \begin{bmatrix} E & 0 \\ A_{21}\sum_{j=0}^{k-1}(I-H^{-1}A_{11})^j H^{-1}E & S \end{bmatrix} \begin{bmatrix} I & H^{-1}\sum_{j=0}^{k-1}(I-A_{11}H^{-1})^j A_{12} \\ 0 & I \end{bmatrix} \\
&= \begin{bmatrix} E & EH^{-1}\sum_{j=0}^{k-1}(I-A_{11}H^{-1})^j A_{12} \\ A_{21}\sum_{j=0}^{k-1}(I-H^{-1}A_{11})^j H^{-1}E, & S+A_{21}\sum_{j=0}^{k-1}(I-H^{-1}A_{11})^j H^{-1}EH^{-1}\sum_{j=0}^{k-1}(I-A_{11}H^{-1})^j A_{12} \end{bmatrix}.
\end{aligned}$$

Dies war zu zeigen. □

Dass der Vorkonditionierer $B_{[k]}$ eine echte Verallgemeinerung bekannter Vorkonditionierer darstellt, zeigt das folgende Beispiel.

Beispiel 6.19. Für den Spezialfall $H = E$ ergibt sich:

1. Wäht man $k = 1$, so ist

$$B_{[1]} = \begin{bmatrix} E & A_{12} \\ A_{21} & S + A_{21}E^{-1}A_{12} \end{bmatrix},$$

   was identisch zu dem Zweilevel-Block-Faktorisierungs-Vorkonditionierer aus Definition 4.7 ist.

2. Für $k = 2$ ist $B_{[k]}$ gegeben durch

$$B_{[2]} = \begin{bmatrix} E & (2I - A_{11}E^{-1})A_{12} \\ A_{21}(2I - E^{-1}A_{11}) & S + A_{21}(2I - E^{-1}A_{11})E^{-1}(2I - A_{11}E^{-1})A_{12} \end{bmatrix}.$$

   Diese Matrix entspricht $\widehat{B}_{2\mathrm{bf}}$, definiert in (6.4).

Im nächsten Kapitel wird der k-fache h-transformierte Vorkonditionierer analysiert, indem dieser ein weiteres Mal verallgemeinert wird.

Dazu wird die Inverse von $B_{[k]}$, die durch

$$B_{[k]}^{-1} = \begin{bmatrix} E^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -H^{-1}\sum_{j=0}^{k-1}(I-A_{11}H^{-1})^j A_{12} \\ I \end{bmatrix} S^{-1} \begin{bmatrix} -A_{21}\sum_{j=0}^{k-1}(I-H^{-1}A_{11})^j H^{-1}, & I \end{bmatrix} \tag{6.33}$$

gegeben ist, betrachtet. Die Terme

$$H^{-1}\sum_{j=0}^{k-1}(I - A_{11}H^{-1})^j \text{ und } \sum_{j=0}^{k-1}(I - H^{-1}A_{11})^j H^{-1}$$

sollen durch ein geeignetes Polynom ersetzt werden, was die Darstellung von (6.33) vereinfacht. Zu diesem Zweck wird das Polynom Q_{k-1} aus (5.3) für ein Polynom P_k mit $P_k(0) = 1$ definiert. Für den hier gegebenen Spezialfall bietet sich die Wahl

$$P_k(t) := (1 - t)^k$$

und damit

$$Q_{k-1}(t) = \tfrac{1-P_k(t)}{t} = \tfrac{1-(1-t)^k}{1-(1-t)} = \sum_{j=1}^{k-1}(1 - t)^j$$

an. Damit erhält man

$$\begin{aligned}\sum_{j=0}^{k-1}(I - H^{-1}A_{11})^j H^{-1} = Q_{k-1}(H^{-1}A_{11})H^{-1} &= \Big[I - P_k(H^{-1}A_{11})\Big]A_{11}^{-1}HH^{-1} \\ &= \Big[I - \big(I - H^{-1}A_{11}\big)^k\Big]A_{11}^{-1},\end{aligned}$$

was auf

$$B_{[k]}^{-1} = \begin{bmatrix} E^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -A_{11}^{-1}\big[I-P_k(A_{11}H^{-1})\big]A_{12} \\ I \end{bmatrix} S^{-1} \begin{bmatrix} -A_{21}\big[I-P_k(H^{-1}A_{11})\big]A_{11}^{-1} & I \end{bmatrix} \tag{6.34}$$

mit $P_k(H^{-1}A_{11}) = (I - H^{-1}A_{11})^k$ führt. Anstatt der Spezialisierung mit $P_k(t) = (1-t)^k$ wird der Vorkonditionierer für ein beliebiges Polynom $P_k \in \mathbb{R}_{\leq k}[t]$ mit $P_k(0) = 1$ untersucht. Abschließend sei bemerkt, dass $B_{[k]}^{-1}$ aus (6.34) die Gleichung

$$B_{[k]}^{-1} = \begin{bmatrix} E^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\left[I - P_k(H^{-1}A_{11})\right] A_{11}^{-1} A_{12} \\ I \end{bmatrix} S^{-1} \left[-A_{21} A_{11}^{-1} \left[I - P_k(A_{11} H^{-1}) \right],\, I \right] \tag{6.35}$$

erfüllt. Die Darstellung (6.35) wird im nächsten Kapitel verwendet.

Kapitel 7

Polynombasierte algebraische Mehrgitterverfahren

In Abschnitt 6.6 wurde gezeigt, dass sich die Analyse von Verfahren, gegeben durch (6.35), auf die Untersuchung des Block-Diagonal-Vorkonditionierers der Form (6.31) zurückführen lässt. Mithilfe dieses Resultats werden in diesem Kapitel polynombasierte Block-Faktorisierungs-Verfahren sowie polynombasierte algebraische Mehrgitterverfahren untersucht. Dazu bedarf es der Analyse des Block-Diagonal-Vorkonditionierers.

Satz 7.1 ([4], Theorem 9.3)**.** *Sei A hpd und partitioniert wie in* (4.3) *sowie $\gamma(A)$ ihre assoziierte C.B.S. Konstante. Weiter sei B_{diag} durch* (6.31) *mit*

$$\begin{aligned} \alpha A_{11} \preceq E \preceq \beta A_{11}, \\ \xi A_{22} \preceq S \preceq \theta A_{22} \end{aligned}$$

gegeben. Dann erfüllen die extremen Eigenwerte von $B_{\text{diag}}^{-1}A$ die Abschätzungen

$$\begin{aligned} \lambda_{\max}(B_{\text{diag}}^{-1}A) &\leq \tfrac{1}{\alpha}\left\{\tfrac{1}{2}(1+\tfrac{\alpha}{\xi}) + \left[\tfrac{1}{4}(1-\tfrac{\alpha}{\xi})^2 + \tfrac{\alpha}{\xi}\gamma(A)^2\right]^{\frac{1}{2}}\right\}, \\ \lambda_{\min}(B_{\text{diag}}^{-1}A) &\geq \tfrac{1-\gamma(A)^2}{\beta}\left\{\tfrac{1}{2}(1+\tfrac{\theta}{\beta}) + \left[\tfrac{1}{4}(1-\tfrac{\theta}{\beta})^2 + \tfrac{\theta}{\beta}\gamma(A)^2\right]^{\frac{1}{2}}\right\}^{-1}. \end{aligned}$$

Mithilfe dieses Resultats wird zunächst der Vorkonditionierer aus (6.35) analysiert.

7.1

Analyse der Block-Faktorisierungs-Verfahren. Ziel dieses Abschnittes ist es, Abschätzungen der Eigenwerte von $B_{[k]}^{-1}A$ zu erhalten, wobei $B_{[k]}$ durch (6.35) gegeben ist. Zunächst wird jedoch das Polynom in dieser Darstellung nicht festgelegt. Dies führt zu der folgenden Verallgemeinerung des k-fachen h-transformierten Vorkonditionieres aus Definition 6.17.

Definition 7.2. Sei $A \in \mathbb{C}^{n,n}$ hpd und partitioniert wie in (4.3). Weiterhin seien E und H hpd Approximationen von A_{11}. Außerdem seien $S \in \mathbb{C}^{n_2,n_2}$ hpd und $P_k \in \mathbb{R}_{\leq k}[t]$ mit $P_k(0) = 1$. Dann wird der *polynombasierte Block-Faktorisierungs-Vorkonditionierer* durch

$$B_{\text{pbf}}^{-1} := \begin{bmatrix} E^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\left[I - P_k(H^{-1}A_{11})\right]A_{11}^{-1}A_{12} \\ I \end{bmatrix} S^{-1} \left[\, -A_{21}A_{11}^{-1}\left[I - P_k(A_{11}H^{-1})\right],\, I \,\right] \tag{7.1}$$

definiert.

Zur Analyse dieses Vorkonditionierers kann die spektrale Identität aus (6.32) und der Satz 7.1 verwendet werden, was erneut auf die Notwendigkeit der Abschätzung einer C.B.S. Konstante führt. In den Sätzen 6.3 bzw. 6.11 wird die Voraussetzung $H \preceq A_{11}$ bzw. $H \succeq A_{11}$ gefordert, wobei in Abschnitt 6.5 gezeigt wurde, dass die Voraussetzungen von Satz 6.11 einfacher, als die von Satz 6.3 erfüllt werden können. Daher wird ferner vorausgesetzt, dass $H \succeq A_{11}$ erfüllt ist. Im Anhang A werden Abschätzungen der C.B.S. Konstante vorgestellt, die weder $H \preceq A_{11}$ noch $H \succeq A_{11}$ benötigen, aber von starker technischer Natur sind.

Satz 7.3. *Sei $A \in \mathbb{C}^{n,n}$ hpd und partitioniert wie in* (4.3)*. Weiter sei B_{pbf} definiert durch* (7.1) *mit*

$$\begin{aligned} \alpha_e A_{11} \preceq E \preceq \beta_e A_{11}, \\ A_{11} \preceq H \preceq \beta_h A_{11}, \\ \xi A_c \preceq S \preceq \theta A_c, \end{aligned}$$

wobei

$$A_c := A_{22} - A_{21}\Big[I - \Big(P_k(H^{-1}A_{11})\Big)^2\Big]A_{11}^{-1}A_{12}. \tag{7.2}$$

Das für die Definition von B_{pbf} notwendige Polynom $P_k \in \mathbb{R}_{\leq k}[t]$ erfülle die Bedingungen $P_k(0) = 1$, $P_k(1) = 0$ und $\big|\, P_k(t) \,\big| < 1$ für $t \in [\frac{1}{\beta_e}, 1]$. Außerdem sei

$$\Pi_{2k}(t) := P_k^2(t) \cdot \tfrac{t}{1-t} \in \mathbb{R}_{\leq 2k}[t]$$

ein weiteres Polynom und es gelte

$$A_{22} - A_{21}\Big[I + \Pi_{2k}(H^{-1}A_{11})\Big]A_{11}^{-1}A_{12} \text{ ist hpsd.}$$

Dann ist

$$\begin{aligned} \lambda_{\max}(B_{\text{pbf}}^{-1}A) &\leq \tfrac{1}{\alpha_e}\Big\{\tfrac{1}{2}(1+\tfrac{\alpha_e}{\xi}) + \Big[\tfrac{1}{4}(1-\tfrac{\alpha_e}{\xi})^2 + \tfrac{\alpha_e}{\xi}(1-\tfrac{1}{\beta_h})\Big]^{\frac{1}{2}}\Big\} \quad \text{und} \\ \lambda_{\min}(B_{\text{pbf}}^{-1}A) &\geq \tfrac{1}{\beta_e\beta_h}\Big\{\tfrac{1}{2}(1+\tfrac{\theta}{\beta_e}) + \Big[\tfrac{1}{4}(1-\tfrac{\theta}{\beta_e})^2 + \tfrac{\theta}{\beta_e}(1-\tfrac{1}{\beta_h})\Big]^{\frac{1}{2}}\Big\}^{-1}. \end{aligned}$$

Beweis. Der Beweis gliedert sich in mehrere Teile.

1. *Spektrale Identität*

   Der polynombasierte Block-Faktorisierungs-Vorkonditionierer aus (7.1) ist durch

$$B_{\text{pbf}}^{-1} = \begin{bmatrix} I & -H_p^{-1}A_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} E^{-1} & 0 \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -A_{21}H_p^{-H} & I \end{bmatrix} \tag{7.3}$$

   mit $H_p := A_{11}\big[I - P_k(H^{-1}A_{11})\big]^{-1}$ darstellbar. Da $P_k(t) < 1$ auf $\sigma(H^{-1}A_{11})$, ist H_p invertierbar und es gilt $H_p = H_p^H$. Die Hermitizität von H_p wird für diesen Beweis jedoch nicht benötigt, was ein analoges Vorgehen für den Beweis von Satz 7.9 aus dem nächsten Abschnitt erlaubt, bei dem diese Eigenschaft nicht vorliegt.

Mit der Faktorisierung (7.3) erhält man die Identität

$$\sigma\left(B_{\text{pbf}}^{-1}A\right) = \sigma\left(\begin{bmatrix} E^{-1} & 0 \\ 0 & S^{-1} \end{bmatrix}\begin{bmatrix} I & 0 \\ -A_{21}H_p^{-H} & I \end{bmatrix} A \begin{bmatrix} I & -H_p^{-1}A_{12} \\ 0 & I \end{bmatrix}\right). \tag{7.4}$$

Definiert man die Matrix

$$\widehat{A} := \begin{bmatrix} I & 0 \\ -A_{21}H_p^{-H} & I \end{bmatrix} A \begin{bmatrix} I & -H_p^{-1}A_{12} \\ 0 & I \end{bmatrix}, \tag{7.5}$$

so ist die vorkonditionierte Matrix $B_{\text{diag}}^{-1}\widehat{A}$ mithilfe von Satz 7.1 zu untersuchen.

2. *Abschätzung der C.B.S. Konstante*

Um die C.B.S. Konstante von $\widehat{A}$ abzuschätzen, wird Satz 6.11 verwendet. Dazu benötigt es eine Matrix $A^{\text{vor}} \in \mathbb{C}^{n,n}$, die die Gleichung

$$\widehat{A} = \begin{bmatrix} I & 0 \\ -A_{21}^{\text{vor}}H^{-H} & I \end{bmatrix} A^{\text{vor}} \begin{bmatrix} I & -H^{-1}A_{12}^{\text{vor}} \\ 0 & I \end{bmatrix} \tag{7.6}$$

erfüllt. Da $P_k(1) = 0$ ist, existiert ein $\widetilde{P}_{k-1} \in \mathbb{R}_{\leq k-1}[t]$ mit $P_k(t) = (1-t)\widetilde{P}_{k-1}(t)$. Daher erfüllt A^{vor} die Gleichung

$$A^{\text{vor}} = \begin{bmatrix} A_{11} & \widetilde{P}_{k-1}(A_{11}H^{-1})A_{12} \\ A_{21}\widetilde{P}_{k-1}(H^{-H}A_{11}) & A_{22}^{\text{vor}} \end{bmatrix} \tag{7.7}$$

mit $A_{22}^{\text{vor}} := A_{22} - A_{21}\Big[I - \widetilde{P}_{k-1}(H^{-H}A_{11})\widetilde{P}_{k-1}(H^{-1}A_{11})\Big]A_{11}^{-1}A_{12}$.

Um zu bestätigen, dass durch diese Wahl (7.6) erfüllt ist, betrachtet man $\widehat{A}$ aus (7.5). Es gilt

$$\begin{aligned}
\widehat{A} &= \begin{bmatrix} I & 0 \\ -A_{21}H_p^{-H} & I \end{bmatrix}\begin{bmatrix} A_{11} & (I - A_{11}H_p^{-1})A_{12} \\ A_{21} & A_{22} - A_{21}H_p^{-1}A_{12} \end{bmatrix} \\
&= \begin{bmatrix} I & 0 \\ -A_{21}H_p^{-1} & I \end{bmatrix}\begin{bmatrix} A_{11} & P_k(A_{11}H^{-1})A_{12} \\ A_{21} & A_{22} - A_{21}\left[I - P_k(H^{-1}A_{11})\right]A_{11}^{-H}A_{12} \end{bmatrix} \\
&= \begin{bmatrix} A_{11} & P_k(A_{11}H^{-1})A_{12} \\ A_{21}(I - H_p^{-H}A_{11}) & \widehat{A}_{22} \end{bmatrix} \\
&= \begin{bmatrix} A_{11} & P_k(A_{11}H^{-1})A_{12} \\ A_{21}P_k(H^{-H}A_{11}) & \widehat{A}_{22} \end{bmatrix}
\end{aligned}$$

mit

$$\widehat{A}_{22} := A_{22} - A_{21}\Big[I - P_k(H^{-H}A_{11})P_k(H^{-1}A_{11})\Big]A_{11}^{-1}A_{12} (= A_c).$$

Andererseits erhält man mit (7.6)

$$\begin{aligned}
\widehat{A} &= \begin{bmatrix} I & 0 \\ -A_{21}\widetilde{P}_{k-1}(H^{-H}A_{11})H^{-H} & I \end{bmatrix} \\
&\quad \cdot \begin{bmatrix} A_{11} & (I - A_{11}H^{-1})\widetilde{P}_{k-1}(A_{11}H^{-1})A_{12} \\ A_{21}\widetilde{P}_{k-1}(H^{-H}A_{11}) & \star \end{bmatrix} \\
&= \begin{bmatrix} A_{11} & P_k(A_{11}H^{-1})A_{12} \\ A_{21}P_k(H^{-H}A_{11}) & \star \end{bmatrix}.
\end{aligned}$$

Da sich das Schurkomplement durch die gegebenen Transformationen nicht ändert, muss $\star = \widehat{A}_{22}$ sein und damit folgt die Gleichheit. Wendet man Satz (6.11) auf A^{vor} an, erhält man die folgende Abschätzung für die C.B.S. Konstante von $\widehat{A}$

$$\gamma(\widehat{A})^2 \leq 1 - \lambda_{\min}(H^{-1}A_{11}) \leq 1 - \tfrac{1}{\beta_h}.$$

Dabei muss H die beiden Voraussetzungen

- $H \succeq A_{11}$,
- $\begin{bmatrix} H(H+HA_{11}^{-1}H-A_{11})^{-1}H & A_{12}^{\text{vor}} \\ A_{21}^{\text{vor}} & A_{22}^{\text{vor}} \end{bmatrix} \succeq 0$

erfüllen. Die zweite Bedingung ist äquivalent zu

$$\begin{aligned}
&A_{22}^{\text{vor}} - A_{21}^{\text{vor}} H^{-1}(H + HA_{11}^{-1}H - A_{11})H^{-1}A_{12}^{\text{vor}} \\
&= A_{22} - A_{21}\Big[I - \big(\widetilde{P}_{k-1}(H^{-1}A_{11})\big)^2\Big]A_{11}^{-1}A_{12} \\
&\quad - A_{21}\widetilde{P}_{k-1}(H^{-1}A_{11})(H^{-1} + A_{11}^{-1} - H^{-1}A_{11}H^{-1})\widetilde{P}_{k-1}(A_{11}H^{-1})A_{12} \\
&= A_{22} - A_{21}\Big[I + \widetilde{P}_{k-1}(H^{-1}A_{11})H^{-1}A_{11}(I - H^{-1}A_{11})\widetilde{P}_{k-1}(H^{-1}A_{11})\Big]A_{11}^{-1}A_{12} \\
&= A_{22} - A_{21}\Big[I + \Pi_{2k}(H^{-1}A_{11})\Big]A_{11}^{-1}A_{12}
\end{aligned}$$

mit $\Pi_{2k}(t) = t(1-t)\widetilde{P}_{k-1}^2(t) \in \mathbb{R}_{\leq 2k}[t]$. Damit folgt

$$\begin{bmatrix} H(H+HA_{11}^{-1}H-A_{11})^{-1}H & A_{12}^{\text{vor}} \\ A_{21}^{\text{vor}} & A_{22}^{\text{vor}} \end{bmatrix} \succeq 0 \Leftrightarrow A_{22} - A_{21}\Big[I + \Pi_{2k}(H^{-1}A_{11})\Big]A_{11}^{-1}A_{12} \succeq 0.$$

3. *Abschätzung der Eigenwerte*

Mithilfe von Satz 7.1 können die extremen Eigenwerte von $B_{\text{pbf}}^{-1}A$ abgeschätzt werden. Es gilt

$$\begin{aligned}
\lambda_{\max}(B_{\text{pbf}}^{-1}A) &= \lambda_{\max}(B_{\text{diag}}^{-1}\widehat{A}) \\
&\leq \tfrac{1}{\alpha_e}\left\{\tfrac{1}{2}(1 + \tfrac{\alpha_e}{\xi}) + \left[\tfrac{1}{4}(1 - \tfrac{\alpha_e}{\xi})^2 + \tfrac{\alpha_e}{\xi}\gamma(\widehat{A})^2\right]^{\frac{1}{2}}\right\} \\
&\leq \tfrac{1}{\alpha_e}\left\{\tfrac{1}{2}(1 + \tfrac{\alpha_e}{\xi}) + \left[\tfrac{1}{4}(1 - \tfrac{\alpha_e}{\xi})^2 + \tfrac{\alpha_e}{\xi}(1 - \tfrac{1}{\beta_h})\right]^{\frac{1}{2}}\right\}
\end{aligned}$$

und

$$\begin{aligned}
\lambda_{\min}(B_{\text{pbf}}^{-1}A) &= \lambda_{\min}(B_{\text{diag}}^{-1}\widehat{A}) \\
&\geq \tfrac{1-\gamma(\widehat{A})^2}{\beta_e}\left\{\tfrac{1}{2}(1 + \tfrac{\theta}{\beta_e}) + \left[\tfrac{1}{4}(1 - \tfrac{\theta}{\beta_e})^2 + \tfrac{\theta}{\beta_e}\gamma(\widehat{A})^2\right]^{\frac{1}{2}}\right\}^{-1} \\
&\geq \tfrac{1}{\beta_e\beta_h}\left\{\tfrac{1}{2}(1 + \tfrac{\theta}{\beta_e}) + \left[\tfrac{1}{4}(1 - \tfrac{\theta}{\beta_e})^2 + \tfrac{\theta}{\beta_e}(1 - \tfrac{1}{\beta_h})\right]^{\frac{1}{2}}\right\}^{-1}.
\end{aligned}$$

□

Die Aussage von Satz 7.3 erlaubt die folgende Interpretation. Die schwer erfüllbaren Voraussetzungen des Satzes 6.11 lassen sich auf Kosten einer aufwändig zu berechnenden Grobgittermatrix A_c, definiert wie in (7.2), abschwächen. Es muss allerdings bei der Suche nach einem geeigneten Vorkonditonierer der Mehraufwand beachtet werden.

Beispiel 7.4. Das Polynom $P_k(t) = (1-t)^k$, $k \geq 2$ erfüllt die Voraussetzungen des Satzes 7.3. Wählt man H, so dass $H \succeq A_{11}$, so muss k so groß gewählt werden, dass für $\Pi_{2k}(t) = t(1-t)^{2k-1}$ die Matrix

$$A_{22} - A_{21}\Big[I + \Pi_{2k}(H^{-1}A_{11})\Big]A_{11}^{-1}A_{12} \text{ hpsd ist.}$$

Dies kann in jedem Fall erreicht werden; jedoch unter Umständen nur durch eine aufwändig zu berechnende und voll besetzte Grobgittermatrix A_c, die durch (7.2) gegeben ist.

Bemerkung 7.5. Obwohl die Voraussetzungen des Satzes zu mehr Aufwand des Mehrlevelverfahrens führen, können diese, im Gegensatz zu der Voraussetzung 6.10, stets erfüllt werden. Dazu bedarf es keiner weiteren Bedingung an die Matrix A.

Zusammenfassend wurde gezeigt, dass Block-Faktorisierungs-Verfahren durch ihre einfache Darstellung umfangreich analysiert werden können. Für das Lösen von linearen Gleichungssystemen finden diese allerdings nur bei einfachen Problemen Anwendung, da sie im Vergleich zum numerischen Aufwand keine gute Performance erreichen. Stattdessen werden algebraische Mehrgitterverfahren für die verschiedensten Probleme verwendet. Mithilfe von Lemma 4.8 kann die Analyse der MBF-Verfahren verwendet werden, um AMG's zu untersuchen.

7.2 Analyse des polynombasierten AMGs.

Mit Lemma 4.8 steht ein Werkzeug zur Verfügung, welches eine Umwandlung von algebraischen Mehrgitterverfahren zu Block-Faktorisierungs-Verfahren ermöglicht. In Abschnitt 7.1 wurden diese Faktorisierungs-Verfahren untersucht. Dabei wurde gezeigt, dass die Verwendung eines Polynoms die Abschätzungen der Eigenwerte verbessert. Hier werden diese Ergebnisse verwendet, um polynombasierte algebraische Mehrgitterverfahren für eine hpd Matrix A zu analysieren.

Definition 7.6. Sei $A \in \mathbb{C}^{n,n}$ hpd und partitioniert wie in (4.3). Dann wird das *polynombasierte algebraische Mehrgitterverfahren* bzw. das *symmetrische polynombasierte algebraische Mehrgitterverfahren* durch

$$I - B_{\mathrm{amg}_p}^{-1}A := \mathrm{C}\cdot\mathrm{R}_{\mathrm{pre}}, \tag{7.8}$$

$$I - B_{\mathrm{amg}_{p,s}}^{-1}A := \mathrm{R}_{\mathrm{post}}\cdot\mathrm{C}\cdot\mathrm{R}_{\mathrm{pre}} \tag{7.9}$$

definiert, wobei

$$\mathrm{C} := \left(I - \begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix} S^{-1} \begin{bmatrix} -A_{21}D_p^{-1} & I \end{bmatrix} A\right), \tag{7.10}$$

$$\mathrm{R}_{\mathrm{pre}} := \left(I - w_1 \begin{bmatrix} M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A\right)\cdots\left(I - w_\nu \begin{bmatrix} M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A\right), \tag{7.11}$$

$$\mathrm{R}_{\mathrm{post}} := \left(I - w_\nu \begin{bmatrix} M_s^{-H} & 0 \\ 0 & 0 \end{bmatrix} A\right)\cdots\left(I - w_1 \begin{bmatrix} M_s^{-H} & 0 \\ 0 & 0 \end{bmatrix} A\right). \tag{7.12}$$

Hierbei sind $w_1, \ldots, w_\nu \in \mathbb{R}$ zu bestimmende Gewichte. Des Weiteren seien D_p, M_s invertierbare Approximationen von A_{11}, wobei D_p zusätzlich hpd sein soll. Außerdem sei S ebenfalls hpd.

Die Definition 7.6 enthält einige Besonderheiten. Zunächst werden Gewichte $w_1, \ldots, w_\nu$ benötigt, was an die Definition des AMGr-Verfahrens erinnert (Definition 4.4). Außerdem wurde nicht spezifiziert, welche Matrix durch S approximiert werden soll. Dies wird die Grobgittermatrix A_c sein, auf die zunächst nicht näher eingegangen wird. Ferner sei darauf hingewiesen, dass die Matrix M_s nicht als hpd vorausgesetzt wird.

Um diese Art von algebraischen Mehrgitterverfahren zu untersuchen, werden sie auf Block-Faktorisierungs-Verfahren transformiert. Hierbei ist der erste Schritt das Zusammenfassen aller Vor- bzw. Nachglätter.

Satz 7.7. *Sei A hpd und partitioniert wie in* (4.3)*. Weiter seien* $\mathrm{R}_{\mathrm{pre}}$ *und* $\mathrm{R}_{\mathrm{post}}$ *gegeben durch* (7.11) *bzw.* (7.12) *mit* $w_1, w_2, \ldots, w_\nu \in \mathbb{R}$*. Dann ist*

$$\mathrm{R}_{\mathrm{pre}} = I - \begin{bmatrix} [I - \prod_{i=\nu}^{1}(I - w_i M_s^{-1} A_{11})]A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} A,$$

$$\mathrm{R}_{\mathrm{post}} = I - \begin{bmatrix} [I - \prod_{i=1}^{\nu}(I - w_i M_s^{-H} A_{11})]A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} A.$$

Beweis. Zunächst werden zwei Glätter $\left(I - \left[\begin{smallmatrix} w_\nu M_s^{-1} & 0 \\ 0 & 0 \end{smallmatrix}\right] A\right) \cdot \left(I - \left[\begin{smallmatrix} w_{\nu-1} M_s^{-1} & 0 \\ 0 & 0 \end{smallmatrix}\right] A\right)$ betrachtet. Durch eine kurze Rechnung erhält man

$$\begin{aligned}
&\left(I - \begin{bmatrix} w_\nu M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A\right) \cdot \left(I - \begin{bmatrix} w_{\nu-1} M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A\right) \\
=& I - \left(\begin{bmatrix} w_\nu M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} w_{\nu-1} M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} w_\nu M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A \begin{bmatrix} w_{\nu-1} M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix}\right) A \\
=& I - \begin{bmatrix} (w_\nu + w_{\nu-1}) M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} w_\nu M_s^{-1} A_{11} & w_\nu M_s^{-1} A_{12} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w_{\nu-1} M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A \\
=& I - \left(\begin{bmatrix} (w_\nu + w_{\nu-1}) M_s^{-1} - w_\nu w_{\nu-1} M_s^{-1} A_{11} M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix}\right) A \\
=& I - \begin{bmatrix} [I - (I - (w_\nu + w_{\nu-1}) M_s^{-1} A_{11} - w_\nu w_{\nu-1} M_s^{-1} A_{11} M_s^{-1} A_{11}] A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} A \\
=& I - \begin{bmatrix} [I - (I - w_\nu M_s^{-1} A_{11})(I - w_{\nu-1} M_s^{-1} A_{11})] A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} A.
\end{aligned}$$

Induktiv folgt schließlich für $j = \nu, \nu - 1, \ldots, 1$

$$\begin{aligned}
&\left(I - \begin{bmatrix} [I - \prod_{i=\nu}^{j-1}(I - w_i M_s^{-1} A_{11})]A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} A\right) \cdot \left(I - \begin{bmatrix} w_j M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A\right) \\
&= I - \begin{bmatrix} [I - \prod_{i=\nu}^{j}(I - w_i M_s^{-1} A_{11})]A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} A.
\end{aligned}$$

Dies folgt durch

$$
\begin{aligned}
&\left(I - \begin{bmatrix} [I - \prod_{i=\nu}^{j-1}(I - w_i M_s^{-1} A_{11})]A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} A\right) \cdot \left(I - \begin{bmatrix} w_j M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A\right) \\
&= I - \left(\begin{bmatrix} [I - \prod_{i=\nu}^{j-1}(I - w_i M_s^{-1} A_{11})]A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} w_j M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} \right) A \\
&\quad + \begin{bmatrix} [I - \prod_{i=\nu}^{j-1}(I - w_i M_s^{-1} A_{11})]A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} A \begin{bmatrix} w_j M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A \\
&= I - \begin{bmatrix} [I - \prod_{i=\nu}^{j-1}(I - w_i M_s^{-1} A_{11})]A_{11}^{-1} + w_j M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A \\
&\quad + \begin{bmatrix} [I - \prod_{i=\nu}^{j-1}(I - w_i M_s^{-1} A_{11})]w_j M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A \\
&= I - \begin{bmatrix} A_{11}^{-1} - \prod_{i=\nu}^{j-1}(I - w_i M_s^{-1} A_{11})A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} A \\
&\quad - \begin{bmatrix} \prod_{i=\nu}^{j-1}(I - w_i M_s^{-1} A_{11}) w_j M_s^{-1} A_{11} A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} A \\
&= I - \begin{bmatrix} A_{11}^{-1} - \prod_{i=\nu}^{j-1}(I - w_i M_s^{-1} A_{11})(I - w_j M_s^{-1} A_{11})A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} A \\
&= I - \begin{bmatrix} [I - \prod_{i=\nu}^{j}(I - w_i M_s^{-1} A_{11})]A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} A.
\end{aligned}
$$

Damit ist gezeigt

$$
\begin{aligned}
&\left(I - \begin{bmatrix} w_\nu M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A\right) \cdot \left(I - \begin{bmatrix} w_{\nu-1} M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A\right) \dots \left(I - \begin{bmatrix} w_1 M_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A\right) \\
=& I - \begin{bmatrix} [I - \prod_{i=\nu}^{1}(I - w_i M_s^{-1} A_{11})]A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} A.
\end{aligned}
$$

Analog funktioniert die Berechnung für $\mathrm{R}_{\mathrm{post}}$. □

Im nächsten Schritt wird der Zusammenhang zwischen den Gewichten $w_1, \dots, w_\nu$ und den vorher angesprochenen Polynomen verdeutlicht.

Korollar 7.8. *Sei A hpd und partitioniert wie in* (4.3)*. Betrachtet man das polynombasierte AMG aus Definition 7.6 mit Gewichten $w_i = \frac{1}{r_i}$, $i = 1, \ldots, \nu$, wobei $r_1, r_2, \ldots, r_\nu \in \mathbb{R}$ ν reelle Nullstellen eines Polynoms $P_\nu \in \mathbb{R}_{\leq \nu}[t]$ vom Grad ν mit $P_\nu(0) = 1$ sowie $\left| P_\nu(t) \right| < 1$ für $t \in \sigma(M_s^{-1} A_{11})$ seien, dann erfüllen $\mathrm{R}_{\mathrm{pre}}$ und $\mathrm{R}_{\mathrm{post}}$ die Gleichungen*

$$\mathrm{R}_{\mathrm{pre}} = I - \begin{bmatrix} \left[I - P_\nu(M_s^{-1} A_{11})\right] A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} A,$$

$$\mathrm{R}_{\mathrm{post}} = I - \begin{bmatrix} \left[I - P_\nu(M_s^{-H} A_{11})\right] A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} A$$

und die Matrizen $\left[I - P_\nu(M_s^{-1} A_{11})\right] A_{11}^{-1}$ sowie $\left[I - P_\nu(M_s^{-H} A_{11})\right] A_{11}^{-1}$ sind invertierbar.

Beweis.

Das Polynom P_ν sei gegeben. Zu zeigen ist

$$P_\nu(t) \stackrel{!}{=} \prod_{i=1}^{\nu} (1 - w_i t) = \prod_{i=1}^{\nu} (1 - \tfrac{t}{r_i}) =: \widetilde{P}(t).$$

Da $\widetilde{P}(r_i) = 0 = P_\nu(r_i)$ für alle $j = 1, \ldots, \nu$ und $\widetilde{P}(0) = 1 = P_\nu(0)$ ist, stimmen die beiden Polynome an $\nu + 1$ Interpolationspunkten überein und sind identisch.

Da $\left| P_\nu(t) \right| < 1$ auf $\sigma(M_s^{-1} A_{11})$ gefordert ist, folgt die geforderte Invertierbarkeit. □

Mithilfe von Lemma 4.8 und Korollar 7.8 kann die Analyse des symmetrischen polynombasierten AMGs auf die Untersuchung von Block-Faktorisierungs-Verfahren und damit Satz 7.3 zurückgeführt werden.

Satz 7.9. *Sei A hpd und partitioniert wie in* (4.3)*. Weiter seien $D_p, M_s \in \mathbb{R}^{n_1, n_1}$ Approximationen von A_{11}. Hierbei soll M_s invertierbar sein und D_p hpd mit*

$$A_{11} \preceq D_p \preceq \beta_d A_{11}.$$

Weiter sei $P_\nu \in \mathbb{R}_{\leq k}[t]$ ein Polynom mit ν reellen Nullstellen $r_1, \ldots, r_\nu$ und den Eigenschaften $P_\nu(0) = 1$ sowie $\left| P_\nu(t) \right| < 1$ für $t \in \sigma(M_s^{-1} A_{11})$. Außerdem approximiere S die Grobgittermatrix A_c, definiert durch

$$A_{22} - A_{21} \Big[I - P_\nu(M_s^{-H} A_{11}) (I - D_p^{-1} A_{11})^2 P_\nu(M_s^{-1} A_{11}) \Big] A_{11}^{-1} A_{12}. \tag{7.13}$$

Hier soll die spektrale Äquivalenz

$$\xi A_c \preceq S \preceq \theta A_c$$

gelten. Zusätzlich sei

$$A_{22} - A_{21} \Big[I + P_\nu(M_s^{-H} A_{11}) D_p^{-1} A_{11} (I - D_p^{-1} A_{11}) P_\nu(M_s^{-1} A_{11}) \Big] A_{11}^{-1} A_{12} \succeq 0. \tag{7.14}$$

Betrachtet man mit diesen Voraussetzungen die Matrix $B_{amg_{p,s}}$ definiert durch (7.9) mit den Gewichten $w_i = \frac{1}{r_i}$, dann gilt

$$\lambda_{\max}(B_{amg_{p,s}}^{-1}A) \leq \left\{ \tfrac{1}{2}(1+\tfrac{1}{\xi}) + \left[\tfrac{1}{4}(1-\tfrac{1}{\xi})^2 + \tfrac{1}{\xi}(1-\tfrac{1}{\beta_d})\right]^{\frac{1}{2}} \right\} \quad \textit{und} \tag{7.15}$$

$$\begin{aligned}\lambda_{\min}(B_{amg_{p,s}}^{-1}A) \geq \tfrac{1-\sigma_{\max}}{\beta_d} \Big\{ & \tfrac{1}{2}\Big(1+\theta(1-\sigma_{\max})\Big) \\ & + \left[\tfrac{1}{4}(1-\theta(1-\sigma_{\max}))^2 + \theta(1-\sigma_{\max})(1-\tfrac{1}{\beta_d})\right]^{\frac{1}{2}} \Big\}^{-1},\end{aligned} \tag{7.16}$$

wobei $\sigma_{\max} := \sigma_{\max}\Big(P_\nu(A_{11}^{\frac{1}{2}}M_s^{-1}A_{11}^{\frac{1}{2}})\Big)$ der größte Singulärwert von $P_\nu(A_{11}^{\frac{1}{2}}M_s^{-1}A_{11}^{\frac{1}{2}})$ ist.

Beweis. Der Beweis gliedert sich erneut in mehrere Teile.

1. *Transformation auf ein PBF-Verfahren*

   Mit Lemma 4.8 erhält man

$$\begin{aligned} B_{amg_{p,s}}^{-1} &= \begin{bmatrix} \widetilde{M}_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -H_p^{-1}A_{12} \\ I \end{bmatrix} S^{-1} \begin{bmatrix} -A_{21}H_p^{-H} & I \end{bmatrix} \\ &= \begin{bmatrix} I & -H_p^{-1}A_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} \widetilde{M}_s^{-1} & 0 \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -A_{21}H_p^{-H} & I \end{bmatrix} \end{aligned} \tag{7.17}$$

   mit

$$\widetilde{M}_s := A_{11}\left[I - P_\nu(M_s^{-H}A_{11})P_\nu(M_s^{-1}A_{11})\right]^{-1}, \tag{7.18}$$

$$H_p := A_{11}\left[I - P_\nu(M_s^{-1}A_{11})(I - D_p^{-1}A_{11})\right]^{-1}. \tag{7.19}$$

   Hierbei lässt sich feststellen, dass (7.17) die gleiche Struktur wie (7.3) besitzt. Einzig die Matrix E aus (7.3) muss durch $\widetilde{M}_s$ ersetzt und H_p modifiziert werden. Daher kann die Abschätzung der Eigenwerte in gleicher Weise wie im Beweis von Satz 7.3 erfolgen. Dort wurde die Eigenschaft $P_\nu(1) = 0$ vorausgesetzt. Das Polynom in (7.19) hat die Gestalt $P_\nu(t) \cdot (1-\widetilde{t})$, also einen um Eins größeren Grad und besitzt bei Eins eine Nullstelle. Weiterhin muss beachtet werden, dass dieses, im Gegensatz zum Beweis von Satz 7.3, kein Polynom in einer Variable ist. Vielmehr bezieht sich P_ν auf $M_s^{-1}A_{11}$ und das Polynom $(1-\widetilde{t})$ auf $D_p^{-1}A_{11}$. In der Transformation (7.6) und den darauf folgenden Umformungen wurde nicht berücksichtigt, dass die Matrizen in den Polynomen dieselben sein müssen.

2. *Nutzung des Beweises von Satz 7.3*

   Die Matrix D_p und das Polynom $(1-\widetilde{t})$ benötigt man für die Transformation auf eine generalisierte hierarchische Basis, siehe (7.6). Im Gegensatz dazu benötigt man das Polynom $P_\nu(t)$ und die Matrix M_s für die „Voraussetzungsmatrix “

$$A^{\text{vor}} := \begin{bmatrix} A_{11} & P_\nu(A_{11}M_s^{-1})A_{12} \\ A_{21}P_\nu(M_s^{-H}A_{11}) & A_{22} - A_{21}\Big[I - P_\nu(M_s^{-H}A_{11})P_\nu(M_s^{-1}A_{11})\Big]A_{11}^{-1}A_{12} \end{bmatrix},$$

analog zu (7.7). Damit ergibt sich die natürliche Grobgittermatrix durch eine analoge Definition von $\widehat{A}$ wie in (7.6) durch

$$\begin{aligned} A_c = \widehat{A}_{22} &:= \begin{bmatrix} -A_{21}^{\text{vor}} D_p^{-1} & I \end{bmatrix} A^{\text{vor}} \begin{bmatrix} -D_p^{-1} A_{12}^{\text{vor}} \\ I \end{bmatrix} \\ &= A_{22} - A_{21}\Big[I - P_\nu(M_s^{-H} A_{11})(I - D_p^{-1} A_{11})^2 P_\nu(M_s^{-1} A_{11})\Big] A_{11}^{-1} A_{12}. \end{aligned}$$

Die Voraussetzung zur Abschätzung der C.B.S. Konstante verändert sich zu

$$\begin{aligned} &D_p \succeq A_{11} \quad \text{und} \\ &\begin{bmatrix} D_p(D_p + D_p A_{11}^{-1} D_p - A_{11})^{-1} D_p & P_\nu(A_{11} M_s^{-1}) A_{12} \\ A_{21} P_\nu(M_s^{-H} A_{11}) & A_{22} - A_{21}\Big[I - P_\nu(M_s^{-H} A_{11}) P_\nu(M_s^{-1} A_{11}))\Big] A_{11}^{-1} A_{12} \end{bmatrix} \succeq 0. \end{aligned} \tag{7.20}$$

Die Bedingung (7.20) ist dabei äquivalent zu

$$A_{22} - A_{21}\Big[I + P_\nu(M_s^{-H} A_{11}) D_p^{-1} A_{11} (I - D_p^{-1} A_{11}) P_\nu(M_s^{-1} A_{11})\Big] A_{11}^{-1} A_{12} \succeq 0.$$

Dies führt zu den Abschätzungen der extremen Eigenwerte

$$\lambda_{\max}(B_{amg_{p,s}}^{-1} A) \leq \frac{1}{\alpha_e}\left\{\frac{1}{2}(1 + \frac{\alpha_e}{\xi}) + \Big[\frac{1}{4}(1 - \frac{\alpha_e}{\xi})^2 + \frac{\alpha_e}{\xi}(1 - \frac{1}{\beta_d})\Big]^{\frac{1}{2}}\right\} \tag{7.21}$$

und

$$\lambda_{\min}(B_{amg_{p,s}}^{-1} A) \geq \frac{1}{\beta_e \beta_d}\left\{\frac{1}{2}(1 + \frac{\theta}{\beta_e}) + \Big[\frac{1}{4}(1 - \frac{\theta}{\beta_e})^2 + \frac{\theta}{\beta_e}(1 - \frac{1}{\beta_d})\Big]^{\frac{1}{2}}\right\}^{-1}. \tag{7.22}$$

Hierbei sind $\alpha_e := (\lambda_{\max}(\widetilde{M}_s^{-1} A_{11}))^{-1}$ und $\beta_e := (\lambda_{\min}(\widetilde{M}_s^{-1} A_{11}))^{-1}$.

3. *Abschätzung der Konstanten α_e und β_e*

   Um die Eigenwerte von $\widetilde{M}_s^{-1} A_{11}$ abzuschätzen, betrachtet man

$$A_{11}^{\frac{1}{2}} \widetilde{M}_s^{-1} A_{11}^{\frac{1}{2}} = I - P_\nu(A_{11}^{\frac{1}{2}} M_s^{-H} A_{11}^{\frac{1}{2}}) P_\nu(A_{11}^{\frac{1}{2}} M_s^{-1} A_{11}^{\frac{1}{2}}). \tag{7.23}$$

   Damit erhält man

$$\begin{aligned} \alpha_e^{-1} &= \lambda_{\max}(\widetilde{M}_s^{-1} A_{11}) = 1 - \min_{\|x\|_2 = 1} \left\| P_\nu(A_{11}^{\frac{1}{2}} M_s^{-1} A_{11}^{\frac{1}{2}}) \right\|_2 \leq 1, \\ \beta_e^{-1} &= \lambda_{\min}(\widetilde{M}_s^{-1} A_{11}) = 1 - \max_{\|x\|_2 = 1} \left\| P_\nu(A_{11}^{\frac{1}{2}} M_s^{-1} A_{11}^{\frac{1}{2}}) \right\|_2 \geq 1 - \sigma_{\max}, \end{aligned}$$

   wobei $\sigma_{\max} := \sigma_{\max}(P_\nu(A_{11}^{\frac{1}{2}} M_s^{-1} A_{11}^{\frac{1}{2}}))$ der größte Singulärwert von $P_\nu(A_{11}^{\frac{1}{2}} M_s^{-1} A_{11}^{\frac{1}{2}})$ ist. Eingesetzt in (7.21) bzw. (7.22) ergibt dies

$$\begin{aligned} \lambda_{\max}(B_{amg_{p,s}}^{-1} A) &\leq \left\{\frac{1}{2}(1 + \frac{1}{\xi}) + \Big[\frac{1}{4}(1 - \frac{1}{\xi})^2 + \frac{1}{\xi}(1 - \frac{1}{\beta_d})\Big]^{\frac{1}{2}}\right\} \quad \text{bzw.} \\ \lambda_{\min}(B_{amg_{p,s}}^{-1} A) &\geq \frac{1 - \sigma_{\max}}{\beta_d}\Bigg\{\frac{1}{2}\Big(1 + \theta(1 - \sigma_{\max})\Big) \\ &\qquad + \Big[\frac{1}{4}(1 - \theta(1 - \sigma_{\max}))^2 + \theta(1 - \sigma_{\max})(1 - \frac{1}{\beta_d})\Big]^{\frac{1}{2}}\Bigg\}^{-1}. \end{aligned}$$

Dies beendet den Beweis. □

Mithilfe von Satz 7.9 kann für jede hpd Matrix A ein algebraisches Mehrgitterverfahren entwickelt werden, das eine gute Performance sichert. Ungeachtet dessen muss auf einige Details eingegangen werden.

Das Polynom $P_\nu \in \mathbb{R}_{\leq\nu}[t]$ stellt den Hauptbestandteil für dieses AMG dar und garantiert durch eine eventuelle Vergrößerung des Grades ν die Realisierbarkeit der Voraussetzung (7.14). Anders als bei dem AMGr-Verfahren, siehe Satz 4.5, wird bei dem polynombasierten AMG keine bestimmte Struktur, wie z.B. Diagonaldominanz, der Matrix A benötigt. Trotz dieses enormen Vorteils darf dabei nicht außer Acht gelassen werden, dass eine Erhöhung des Grades den numerischen Aufwand bei der Bildung der Grobgittermatrix (7.13) vergrößert. Des Weiteren scheint diese Wahl der Grobgittermatrix widersprüchlich zu der Optimalität der Galerkin-Matrix zu sein, siehe Satz 4.2. Führt man einen Iterationsschritt der Grobgitterkorrektur durch, verspricht in der Tat der Galerkin-Ansatz die schnellste Reduktion des Fehlers in der A-Norm, wobei dabei vernachlässigt wird, dass ein AMG nicht nur aus einer Grobgitterkorrektur besteht, sondern für die Konvergenz mindestens ein Glättungsschritt benötigt wird. Bei der Verwendung von F-Glättern ist es möglich, diesen Zusammenhang auch bei der Frage der optimalen Grobgittermatrix zu verwenden.

Satz 7.10. *Sei $A \in \mathbb{C}^{n,n}$ hpd und partitioniert wie in* (4.3)*. Weiter sei (oB.d.A für $b = 0$) ein Iterationsschritt gegeben durch*

$$x^{[k+1]} = x^{[k]} - \begin{bmatrix} \widetilde{M}_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} Ax^{[k]} + Py$$

mit $x^{[k]} \in \mathbb{C}^n$, $\widetilde{M} \in \mathbb{C}^{n_1,n_1}$ hpd und $P = \begin{bmatrix} -H_p^{-1}A_{12} \\ I \end{bmatrix}$ mit $H_p \in \mathbb{C}^{n_1,n_1}$ invertierbar. Dann gilt

$$\left\|x^{[k+1]}\right\|_A = \min \quad \Leftrightarrow \quad y = -A_c^{-1}RAx^{[k]}$$

mit $A_c := P^H AP$ und $R = \begin{bmatrix} -A_{21}\Big[I - (I - H_p^{-H}A_{11})(I - \widetilde{M}_s^{-1}A_{11})\Big]A_{11}^{-1} & I \end{bmatrix}$.

Beweis. Für die A-Norm von $x^{[k+1]}$ gilt

$$\begin{aligned}\left\|x^{[k+1]}\right\|_A^2 = \Bigg\{ &\begin{bmatrix} x_1^{[k]^H}(I - A_{11}\widetilde{M}_s^{-1}) - x_2^{[k]^H}A_{21}\widetilde{M}_s^{-1} - y^H A_{21}H_p^{-H}, & x_2^{[k]^H} + y^H \end{bmatrix} \cdot A \\ &\cdot \begin{bmatrix} (I - \widetilde{M}_s^{-1}A_{11})x_1^{[k]} - \widetilde{M}_s^{-1}A_{11}x_2^{[k]} - H_p^{-1}A_{12}y \\ x_2^{[k]} + y \end{bmatrix} \Bigg\}.\end{aligned}$$

Leitet man diesen Term nach y ab und vereinfacht ihn, erhält man

$$\begin{aligned}\nabla_y \left\|x^{[k+1]}\right\|_A^2 &= A_{21}H_p^{-H}A_{11}(I - \widetilde{M}_s^{-1}A_{11})x_1^{[k]} + A_{21}H_p^{-H}A_{11}\widetilde{M}_s^{-1}A_{12}x_2^{[k]} \\ &\quad + A_{21}H_p^{-H}A_{11}H_p^{-1}A_{12}y - A_{21}H_p^{-1}A_{12}x_2^{[k]} + A_{21}(I - \widetilde{M}_s^{-1}A_{11})x_1^{[k]} \\ &\quad - A_{21}\widetilde{M}_s^{-1}A_{12}x_2^{[k]} - 2A_{21}H_p^{-H}A_{12}y + A_{21}\widetilde{M}_s^{-1}b_1 + A_{22}x_2^{[k]} + A_{22}y \\ &= A_c y + RAx^{[k]}\end{aligned}$$

mit $A_c = P^H AP$, $P = \begin{bmatrix} -H_p^{-1}A_{12} \\ I \end{bmatrix}$ und

$$R = \begin{bmatrix} -A_{21}[I - (I - H_p^{-H}A_{11})(I - \widetilde{M}_s^{-1}A_{11})]A_{11}^{-1} & I \end{bmatrix}.$$

Daher wird die A-Norm minimal für

$$x^{[k+1]} = \left(I - \left(\begin{bmatrix} \widetilde{M}_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} + PA_c^{-1}R\right)A\right)x^{[k]}$$

mit $A_c = P^H AP$. □

Dieses Resultat lässt interessante Interpretationen zu. Die Interpolationsmatrix P wurde durch $P = \begin{bmatrix} -H_P^{-1}A_{12} \\ I \end{bmatrix}$ vorgegeben. Diese ist ausschließlich für die Konstruktion der Grobgittermatrix $A_c = P^H AP$ verantwortlich, was dem Satz 4.2 entspricht. Wird dieser Zusammenhang auf das AMG aus (7.9) angewendet, stecken in der Matrix H_p bereits Informationen aus dem Glätter, denn H_p ist gegeben durch (7.19). Die Grobgittermatrix ist daher gegeben durch

$$\begin{aligned} A_c = P^H AP &= \begin{bmatrix} -H_p^{-1}A_{12} \\ I \end{bmatrix}^H A \begin{bmatrix} -H_p^{-1}A_{12} \\ I \end{bmatrix} \\ &= A_{22} - A_{21}\left(H_p^{-1} + H_p^{-H} - H_p^{-H}A_{11}H_p^{-1}\right)A_{12} \\ &= A_{22} - A_{21}\left[I - (I - H_p^{-H}A_{11})(I - H_p^{-1}A_{11})\right]A_{11}^{-1}A_{12} \\ &= A_{22} - A_{21}\left[I - P_\nu(M_s^{-H}A_{11})(I - D_p^{-1}A_{11})^2 P_\nu(M_s^{-1}A_{11})\right]A_{11}^{-1}A_{12}, \end{aligned}$$

was der Matrix aus (7.13) entspricht.

Ebenfalls auffallend ist die Wahl der Restriktion R, die nicht $R = P^H$ entspricht, sondern

$$R = \begin{bmatrix} -A_{21}[I - (I - H_p^{-H}A_{11})(I - \widetilde{M}_s^{-1}A_{11})]A_{11}^{-1} & I \end{bmatrix}.$$

Zur Konstruktion dieser Restriktion wird folglich der Nachglätter benötigt. Es ergibt sich die Gleichung

$$R = \left(\left(I - \begin{bmatrix} \widetilde{M}_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A\right)P\right)^H. \tag{7.24}$$

Zusammengefasst wird die Grobgittermatrix durch die Interpolation konstruiert, aber auf die Restriktion nimmt zusätzlich der Nachglätter Einfluss. Bei dieser Wahl von R erhält man einen nicht hermiteschen Vorkonditionierer $B_{(\cdot)}$. Für den Beweis des Satzes 7.9 wird jedoch vorausgesetzt, dass $B_{\mathrm{amg}_{p,s}} = B^H_{\mathrm{amg}_{p,s}}$ erfüllt ist, daher ist $R = P^H$ die geeignete Wahl. Dennoch entspricht A_c in diesem Satz der optimalen Wahl aus Satz 7.10.

Im Satz 7.9 ist außerdem erwähnenswert, dass unsymmetrische Approximationen M_s im Glätter zugelassen werden. Abhängig von dieser Approximation muss ein geeignetes Polynom P_ν bestimmt werden. Im nächsten Abschnitt werden die folgenden Möglichkeiten betrachtet:

- M_s ist hpd, dann existieren $a, b > 0$ mit $\sigma(M_s^{-1}A_{11}) \subseteq [a, b]$.
- M_s ist unsymmetrisch und invertierbar, sowie es existiert ein $\rho < 1$ mit $\|I - M_s^{-1}A_{11}\|_{A_{11}} \leq \rho$.

7.3 Optimierung durch Tschebysheff Polynome.

Das Ziel ist es, den größten Singulärwert aus dem Satz 7.9 für die obigen Varianten abzuschätzen. Dies wird für ganz bestimmte Polynome, die Tschebysheff Polynome aus Abschnitt 3.4, geschehen.

7.3.1 Der hermitesche Fall.

Die in Satz 7.9 vorgestellten Abschätzungen der Eigenwerte hängen vom größten Singulärwert der Matrix

$$P_\nu(A_{11}^{\frac{1}{2}} M_s^{-1} A_{11}^{\frac{1}{2}})$$

mit $P_\nu \in \mathbb{R}_{\leq \nu}[t]$ ab. Für eine hpd Matrix $M_s \in \mathbb{C}^{n_1,n_1}$ mit $\sigma(M_s^{-1}A_{11}) \subseteq [a,b]$ kann dieser direkt angegeben werden, denn es gilt

$$\sigma_{\max} = \sigma_{\max}\Big(P_\nu(A_{11}^{\frac{1}{2}} M_s^{-1} A_{11}^{\frac{1}{2}})\Big) = \max_k |\, \lambda_k(P_\nu(A_{11}^{\frac{1}{2}} M_s^{-1} A_{11}^{\frac{1}{2}}))\,| \leq \max_{t\in[a,b]} |\, P_\nu(t)\,|, \tag{7.25}$$

da $P_\nu(A_{11}^{\frac{1}{2}} M_s^{-1} A_{11}^{\frac{1}{2}})$ hermitesch ist. Um den Singulärwert zu minimieren, muss ein Polynom P_ν gewählt werden, welches auf dem Intervall $[a,b]$ im Betrag minimal wird. Polynome mit dieser Eigenschaft wurden in Abschnitt 3.4 eingeführt. Betrachtet man das in 1 normierte Tschebysheff Polynom auf $[a,b]$ aus (3.13), ergibt sich die Abschätzung

$$\sigma_{\max} \leq \max_{t\in[a,b]} \left| \frac{T_\nu(\frac{b+a-2t}{b-a})}{T_\nu(\frac{b+a}{b-a})} \right| = \frac{1}{T_\nu(\frac{b+a}{b-a})}. \tag{7.26}$$

Mit den Eigenschaften der Tschebysheff Polynome, siehe dazu [6], kann der Nenner umgeformt werden und man erhält

$$\begin{aligned}
T_\nu(\tfrac{b+a}{b-a}) &= \tfrac{1}{2}\Big[(\tfrac{b+a}{b-a} + \sqrt{(\tfrac{b+a}{b-a})^2 - 1})^\nu + (\tfrac{b+a}{b-a} - \sqrt{(\tfrac{b+a}{b-a})^2 - 1})^\nu\Big] \\
&= \tfrac{1}{2}\Big[(\tfrac{b+a+\sqrt{(b+a)^2-(b-a)^2}}{b-a})^\nu + (\tfrac{b+a-\sqrt{(b+a)^2-(b-a)^2}}{b-a})^\nu\Big] \\
&= \tfrac{1}{2}\Big[(\tfrac{b+a+2\sqrt{ab}}{b-a})^\nu + (\tfrac{b+a-2\sqrt{ab}}{b-a})^\nu\Big] = \tfrac{1}{2}\Big[(\tfrac{(\sqrt{b}+\sqrt{a})^2}{b-a})^\nu + (\tfrac{(\sqrt{b}-\sqrt{a})^2}{b-a})^\nu\Big] \\
&= \tfrac{1}{2}\Big[(\tfrac{\sqrt{b}+\sqrt{a}}{\sqrt{b}-\sqrt{a}})^\nu + (\tfrac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}})^\nu\Big].
\end{aligned} \tag{7.27}$$

Damit ergibt sich

$$\begin{aligned}
\sigma_{\max} \leq \frac{1}{T_\nu(\frac{b+a}{b-a})} &= \left\{\tfrac{1}{2}\Big[(\tfrac{\sqrt{b}+\sqrt{a}}{\sqrt{b}-\sqrt{a}})^\nu + (\tfrac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}})^\nu\Big]\right\}^{-1} = \left\{\tfrac{(\sqrt{b}+\sqrt{a})^{2\nu}+(\sqrt{b}-\sqrt{a})^{2\nu}}{2(\sqrt{b}+\sqrt{a})^\nu(\sqrt{b}-\sqrt{a})^\nu}\right\}^{-1} \\
&= 2\tfrac{(\sqrt{b}+\sqrt{a})^\nu(\sqrt{b}-\sqrt{a})^\nu}{(\sqrt{b}+\sqrt{a})^{2\nu}+(\sqrt{b}-\sqrt{a})^{2\nu}},
\end{aligned} \tag{7.28}$$

was Folgendes impliziert

$$\begin{aligned}
1 - \sigma_{\max} &\geq 1 - 2\tfrac{(\sqrt{b}+\sqrt{a})^\nu(\sqrt{b}-\sqrt{a})^\nu}{(\sqrt{b}+\sqrt{a})^{2\nu}+(\sqrt{b}-\sqrt{a})^{2\nu}} \\
&= \tfrac{\big((\sqrt{b}+\sqrt{a})^\nu-(\sqrt{b}-\sqrt{a})^\nu\big)^2}{(\sqrt{b}+\sqrt{a})^{2\nu}+(\sqrt{b}-\sqrt{a})^{2\nu}}.
\end{aligned}$$

Somit lässt sich der Term $1-\sigma_{\max}$ in (7.16) aus Satz 7.9 durch

$$1-\sigma_{\max} \geq \frac{\left((\sqrt{b}+\sqrt{a})^{\nu}-(\sqrt{b}-\sqrt{a})^{\nu}\right)^2}{(\sqrt{b}+\sqrt{a})^{2\nu}+(\sqrt{b}-\sqrt{a})^{2\nu}} \tag{7.29}$$

abschätzen.

Für Satz 7.9 wird vorausgesetzt, dass das Polynom ν reelle Nullstellen besitzt. Die Nullstellen des hier betrachteten Polynoms sind

$$r_i = \tfrac{1}{2}\Big(b+a-t_i(b-a)\Big) \quad \text{für} \quad t_i = \cos\left(\tfrac{\pi}{2}\cdot\tfrac{2i-1}{\nu}\right), \quad i=1,2,\ldots,\nu. \tag{7.30}$$

Im Satz 7.9 werden hpd Approximationen M_s und auch allgemeinere und insbesondere nicht hermitesche Matrizen betrachtet.

7.3.2 Der nicht hermitesche Fall. Mit der Annahme, dass M_s eine nicht hermitesche, invertierbare Matrix mit $\|I-M_s^{-1}A_{11}\|_{A_{11}} \leq \rho$ für ein $\rho<1$ ist, ist das zugehörige in 1 normierte Tschebysheff Polynom auf $B_\rho(1)$ gegeben durch $P_\nu(t)=(1-t)^\nu$, siehe dazu Abschnitt 3.4. Damit ergibt sich für den größten Singulärwert von

$$P_\nu(A_{11}^{\frac{1}{2}}M_s^{-1}A_{11}^{\frac{1}{2}})$$

die Abschätzung

$$\begin{aligned}\Big(\sigma_{\max}(P_\nu(A_{11}^{\frac{1}{2}}M_s^{-1}A_{11}^{\frac{1}{2}}))\Big)^2 &= \lambda_{\max}\Big(P_\nu(A_{11}^{\frac{1}{2}}M_s^{-1}A_{11}^{\frac{1}{2}})^H P_\nu(A_{11}^{\frac{1}{2}}M_s^{-1}A_{11}^{\frac{1}{2}})\Big)\\ &= \max_{x_1\neq 0}\frac{x_1^H P_\nu(A_{11}M_s^{-H})A_{11}P_\nu(M_s^{-1}A_{11})x_1}{x_1^H A_{11}x_1}\\ &= \left\|P_\nu(M_s^{-1}A_{11})\right\|_{A_{11}}^2 = \left\|(I-M_s^{-1}A_{11})^\nu\right\|_{A_{11}}^2 \leq (\rho^\nu)^2.\end{aligned} \tag{7.31}$$

Daher ist für (7.16) die Identität

$$1-\sigma_{\max} \geq 1-\rho^\nu$$

gegeben. Hier sind die Nullstellen des Polynoms $r_1=\cdots=r_\nu=1$ und damit die Gewichte ebenfalls $w_1=\cdots=w_\nu=1$.

Für hermitesche Approximationen M_s wurden in Satz 6.6 Abschätzungen der Eigenwerte von $M_s^{-1}A_{11}$ vorgestellt, die es ermöglichen, die auf 1 normierten Tschebysheff Polynome auf $[a,b]$ zu konstruieren. Für die in 1 normierten Tschebysheff Polynome auf $B_\rho(1)$ ist die Kenntnis von ρ nicht nötig, da die Polynome stets durch $(1-t)^\nu$ gegeben sind. Ist man aber an der Abschätzung $1-\sigma_{\max}=1-\rho^\nu$ interessiert, so ist es hilfreich, ρ angeben zu können. Für den Spezialfall, dass M_s als obere Dreiecksmatrix von A_{11} gewählt wird, bezeichnet mit $\operatorname{triu}(A_{11})$, kann ρ bei Verwendung des Greedy-Coarsers angegeben werden:

In [82, Theorem 10.24] wird gezeigt, dass die Ungleichung

$$\left\|I-\operatorname{triu}(A_{11})^{-1}A_{11}\right\|_\infty \leq \left\|I-\operatorname{diag}(A_{11})^{-1}A_{11}\right\|_\infty \tag{7.32}$$

für diagonaldominante Matrizen A_{11} erfüllt ist. Hierbei ist $\|\cdot\|_\infty : \mathbb{C}^{n,n} \to \mathbb{R}^+$ die Matrix-Norm bzgl. der Vektor-Norm

$$\|x\|_\infty := \max_j |\, x_j \,|, \quad \text{wobei} \quad x = [x_j].$$

Es kann gezeigt werden, dass

$$\|B\|_\infty = \max_j \sum_{k=1}^{n} |\, b_{jk} \,|, \quad \text{wobei} \quad B = [b_{jk}].$$

Für eine bessere Übersicht wird $C_{\text{jac}} := [c_{jk}^{\text{jac}}] := \text{diag}(A_{11})^{-1} A_{11}$ und $C_{\text{gs}} := [c_{jk}^{\text{gs}}] := \text{triu}(A_{11})^{-1} A_{11}$ definiert. Um die Ungleichung (7.32) zu beweisen, wird in [82] für den Beweis von Theorem 10.24 die Beziehung

$$|\, I - C_{\text{gs}} \,| e \leq \|I - C_{\text{jac}}\|_\infty e \tag{7.33}$$

mit $e = [\,1 \ldots 1\,]^H$ verwendet.

(7.33) impliziert für alle j

$$|\, 1 - c_{jj}^{\text{gs}} \,| + \sum_{k \neq j} |\, c_{jk}^{\text{gs}} \,| \leq \max_j \Big(\underbrace{|\, 1 - c_{jj}^{\text{jac}} \,|}_{=0} + \sum_{k \neq j} |\, c_{jk}^{\text{jac}} \,| \Big). \tag{7.34}$$

Für die Einträge c_{jk}^{jac} gilt die Identität $c_{jk}^{\text{jac}} = \frac{a_{jk}}{a_{jj}}$, wobei a_{jk} die Einträge von A_{11} sind. Durch die Anwendung des Satzes von Gershgorin (Satz 6.5) existiert für jeden Eigenwert von C_{gs} ein j, so dass die Ungleichung

$$\begin{aligned} |\, \lambda(C_{\text{gs}}) - c_{jj}^{\text{gs}} \,| &\leq \sum_{k \neq j} |\, c_{jk}^{\text{gs}} \,| \overset{(7.34)}{\leq} \Big(\max_j \sum_{k \neq j} |\, c_{jk}^{\text{jac}} \,| \Big) - |\, 1 - c_{jj}^{\text{gs}} \,| \\ &= \Big(\max_j \sum_{k \neq j} \frac{|\, a_{jk} \,|}{|\, a_{jj} \,|} \Big) - |\, 1 - c_{jj}^{\text{gs}} \,| \end{aligned} \tag{7.35}$$

erfüllt ist. Betrachtet man φ und φ_j aus Algorithmus 6, dann gilt $\varphi \leq \varphi_j$ und die Abschätzung

$$\varphi_j^{-1} = \frac{\sum_k |\, a_{jk} \,|}{|\, a_{jj} \,|} = 1 + \sum_{k \neq j} \frac{|\, a_{jk} \,|}{|\, a_{jj} \,|}. \tag{7.36}$$

Aus (7.35) und (7.36) folgt

$$\begin{aligned} |\, \lambda(C_{\text{gs}}) - 1 \,| &\leq |\, \lambda(C_{\text{gs}}) - c_{jj}^{\text{gs}} \,| + |\, 1 - c_{jj}^{\text{gs}} \,| \leq \max_j \sum_{k \neq j} \frac{|\, a_{jk} \,|}{|\, a_{jj} \,|} \\ &\leq \max_j \varphi_j^{-1} - 1 \leq \tfrac{1}{\varphi} - 1. \end{aligned}$$

Damit lassen sich die Eigenwerte von $\text{triu}(A_{11})^{-1} A_{11}$ abschätzen und es gilt

$$\sigma(\text{triu}(A_{11})^{-1} A_{11}) \subseteq B_{(\frac{1}{\varphi} - 1)}(1),$$

also $1 - \sigma_{\max} \leq 1 - \rho^\nu \leq 1 - (\frac{1}{\varphi} - 1)^\nu$.

7.4

Konvergenz des polynombasierten AMGs. In diesem Abschnitt wird das polynombasierte AMG aus (7.8) betrachtet. Die hier aufgeführten Resultate können in [53] wiedergefunden werden. Da B_{amg_p} keine hermitesche Matrix ist, benötigt es andere Hilfsmittel, als die bisher betrachteten. Transformiert man dieses Verfahren auf ein Block-Faktorisierungs-Verfahren, so erhält man

$$B_{\mathrm{amg}_p}^{-1} = \left(\begin{bmatrix} \overline{M}_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix} A_c^{-1} \begin{bmatrix} -A_{21}H_p^{-H} & I \end{bmatrix} \right) \tag{7.37}$$

mit H_p definiert wie in (7.19) und

$$\overline{M}_s := A_{11}\Big[I - P_\nu(M_s^{-1}A_{11})\Big]^{-1}. \tag{7.38}$$

Die Interpolationsmatrix ist bei diesem Verfahren gegeben durch $P = \left[\begin{smallmatrix} -D_p^{-1}A_{12} \\ I \end{smallmatrix}\right]$, daher ist nach Satz 7.10 die optimale Grobgittermatrix gegeben durch

$$A_c = \left[\begin{smallmatrix} -D_p^{-1}A_{12} \\ I \end{smallmatrix}\right]^H A \left[\begin{smallmatrix} -D_p^{-1}A_{12} \\ I \end{smallmatrix}\right]. \tag{7.39}$$

Mit Satz 7.10 und der Gleichung (7.24) erhält man als Restriktion die optimale Matrix

$$\begin{aligned} R^H &= \Big(I - \begin{bmatrix} \overline{M}_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A\Big)P \\ &= \begin{bmatrix} I - \Big[I - P_\nu(M_s^{-1}A_{11})\Big] & -\Big[I - P_\nu(M_s^{-1}A_{11})\Big]A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix} \\ &= \begin{bmatrix} -P_\nu(M_s^{-1}A_{11})D_p^{-1}A_{12} - \Big[I - P_\nu(M_s^{-1}A_{11})\Big]A_{11}^{-1}A_{12} \\ I \end{bmatrix} \\ &= \begin{bmatrix} -\Big[I - P_\nu(M_s^{-1}A_{11})(I - D_p^{-1}A_{11})\Big]A_{11}^{-1}A_{12} \\ I \end{bmatrix} \\ &= \begin{bmatrix} -H_p^{-1}A_{12} \\ I \end{bmatrix}. \end{aligned}$$

Folglich ist mit (7.37) und Satz 7.10 die optimale Kombination bzgl. der Interpolation P gewählt. Es lässt sich zeigen, dass dieses Verfahren konvergent ist. Der Beweis orientiert sich zum Teil an der Arbeit [68]. Im Unterschied zu Satz 7.9 wird vorausgesetzt, dass $D_p \preceq A_{11}$ ist. Dies ermöglicht einen Vergleich mit dem AMGr-Verfahren aus Abschnitt 4.3.1.

Satz 7.11. *Sei A hpd und partitioniert wie in* (4.3)*. Weiter sei D_p eine hpd Approximation von A_{11} mit*

$$\alpha A_{11} \preceq D_p \preceq A_{11} \tag{7.40}$$

für $\alpha \leq 1$, so dass

$$\left[\begin{smallmatrix} D_p & A_{12} \\ A_{21} & A_{22} \end{smallmatrix}\right] \quad \textit{hpsd ist.} \tag{7.41}$$

Außerdem sei M_s eine weitere (möglicherweise nicht hermitesche) Approximation von A_{11} und $P_\nu \in \mathbb{R}_{\leq\nu}[t]$ ein Polynom mit ν reellen Nullstellen $r_1, r_2, \ldots, r_\nu$, $P_\nu(0) = 1$ sowie

$|\,P_\nu(t)\,| < 1$ *für* $t \in \sigma(M_s^{-1}A_{11})$, *so dass der größte Singulärwert von* $P_\nu(A_{11}^{\frac{1}{2}}M_s^{-1}A_{11}^{\frac{1}{2}})$, *also* $\sigma_{\max} := \sigma_{\max}\left(P_\nu(A_{11}^{\frac{1}{2}}M_s^{-1}A_{11}^{\frac{1}{2}})\right)$, *die Ungleichung* $\sigma_{\max} < 1$ *erfüllt. Dann gilt*

$$\left\|I - B_{\mathrm{amg}_p}^{-1}A\right\|_A \leq \left(1 - \alpha(1-\sigma_{\max}^2)\right)^{\frac{1}{2}} < 1, \tag{7.42}$$

wobei B_{amg_p} *durch* (7.8) *mit den Gewichten* $w_i = \frac{1}{r_i}$, $i = 1, \ldots, \nu$ *gegeben ist und* $S = A_c$ *gilt, wobei* A_c *durch die Galerkin-Matrix aus* (7.39) *definiert ist.*

Beweis. Jeder Vektor $\overline{e} \in \mathbb{C}^n$ kann als eine A-orthogonale Summe

$$\overline{e} = \eta \begin{bmatrix} -A_{11}^{-1}A_{12} \\ I \end{bmatrix} v + \mu \begin{bmatrix} I \\ 0 \end{bmatrix} w, \tag{7.43}$$

mit $v \in \mathbb{C}^{n_2}, w \in \mathbb{C}^{n_1}$ und $\|v\|_{S(A,A_{11})} = \|w\|_{A_{11}} = 1$ geschrieben werden. Hierbei sind $\eta, \mu \in \mathbb{C}$ die entsprechenden Koordinaten. Um die A-Norm der Iterationsmatrix abzuschätzen, wird ein normierter Vektor, also $e \in \mathbb{C}^n$ mit $\|e\|_A = 1$, betrachtet. Für die Koordinaten η und μ in (7.43) impliziert dies $\eta^2 + \mu^2 = 1$.

Mit (7.37) erhält man

$$I - B_{\mathrm{amg}_p}^{-1}A = I - \left(\begin{bmatrix} \overline{M}_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix} A_c^{-1} \begin{bmatrix} -A_{21}H_p^{-H} & I \end{bmatrix}\right) A$$

mit $\overline{M}_s$ und H_p wie in (7.38) bzw. (7.19).

Durch die Anwendung von Satz 7.10 gilt die Identität

$$\left\|(I - B_{\mathrm{amg}_p}^{-1}A)e\right\|_A = \min_y \left(\left\|(I - \begin{bmatrix} \overline{M}_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A)e + \begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix} y\right\|_A\right). \tag{7.44}$$

Der erste Summand kann zu

$$\begin{aligned} \left(I - \begin{bmatrix} \overline{M}_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A\right) e &= e - \mu \begin{bmatrix} [I - P_\nu(M_s^{-1}A_{11})]\, w \\ 0 \end{bmatrix} \\ &= \eta \begin{bmatrix} -A_{11}^{-1}A_{12} \\ I \end{bmatrix} v + \mu \begin{bmatrix} P_\nu(M_s^{-1}A_{11}) \\ 0 \end{bmatrix} w \end{aligned} \tag{7.45}$$

umgeformt werden.

Folgt man den Ideen aus [68] und schätzt den zweiten Summanden aus (7.45) bzgl. der A-Norm ab, so erhält man

$$\begin{aligned} \left\|\begin{bmatrix} P_\nu(M_s^{-1}A_{11}) \\ 0 \end{bmatrix} w\right\|_A^2 &= \left\|P_\nu(M_s^{-1}A_{11})w\right\|_{A_{11}}^2 \\ &\leq \max_{x_1 \neq 0} \frac{x_1^H P_\nu(A_{11}^{\frac{1}{2}}M_s^{-H}A_{11}^{\frac{1}{2}})P_\nu(A_{11}^{\frac{1}{2}}M_s^{-1}A_{11}^{\frac{1}{2}})x_1}{x_1^H x_1} \\ &= \left(\sigma_{\max}(P_\nu(A_{11}^{\frac{1}{2}}M_s^{-1}A_{11}^{\frac{1}{2}}))\right)^2 = \sigma_{\max}^2. \end{aligned}$$

Es existiert folglich ein $\widehat{w} \in \mathbb{C}^{n_1}$ mit $\|\widehat{w}\|_{A_{11}} = 1$ und

$$\left(I - \begin{bmatrix} \overline{M}_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} A\right) e = \eta \begin{bmatrix} -A_{11}^{-1}A_{12} \\ I \end{bmatrix} v + \widehat{\mu} \begin{bmatrix} I \\ 0 \end{bmatrix} \widehat{w},$$

wobei die Ungleichung $|\,\widehat{\mu}\,| \leq |\,\mu\sigma_{\max}\,|$ erfüllt ist.

Durch Einsetzen von $y = \eta\theta v$ in (7.44) mit $\theta \in \mathbb{C}$ und der Gleichung

$$\begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix} v = \begin{bmatrix} -A_{11}^{-1}A_{12} \\ I \end{bmatrix} v + \begin{bmatrix} -(D_p^{-1} - A_{11}^{-1})A_{12} \\ 0 \end{bmatrix} v$$

erhält man

$$\begin{aligned}
&\left\|(I - B_{\mathrm{amg}_p}^{-1}A)e\right\|_A^2 \\
&= \min_\theta \left\|\eta(1-\theta) \begin{bmatrix} -A_{11}^{-1}A_{12} \\ I \end{bmatrix} v + \widehat{\mu} \begin{bmatrix} I \\ 0 \end{bmatrix} \widehat{w} - \eta\theta \begin{bmatrix} -(D_p^{-1} - A_{11}^{-1})A_{12} \\ 0 \end{bmatrix} v\right\|_A^2 \\
&\leq \min_\theta \left[\eta^2(1-\theta)^2 \left\| \begin{bmatrix} -A_{11}^{-1}A_{12} \\ I \end{bmatrix} v\right\|_A^2 + \left\|\widehat{\mu}\widehat{w} + \eta\theta(D_p^{-1} - A_{11}^{-1})A_{12}v\right\|_{A_{11}}^2\right] \\
&\leq \min_\theta \left[\eta^2(1-\theta)^2 + \left(\widehat{\mu} + \eta\theta \left\|(D_p^{-1} - A_{11}^{-1})A_{12}v\right\|_{A_{11}}\right)^2\right]. \qquad (7.46)
\end{aligned}$$

Da die Matrix $\left[\begin{smallmatrix} D_p & A_{12} \\ A_{21} & A_{22} \end{smallmatrix}\right]$ als hpsd vorausgesetzt ist, ist $A_{22} \succeq A_{21}D_p^{-1}A_{12}$. Des Weiteren wurde gefordert, dass $\lambda_{\max}(D_p^{-1}A_{11}) \leq \frac{1}{\alpha}$. Definiert man $\varepsilon := \frac{1}{\alpha} - 1$, so ergibt sich

$$\begin{aligned}
&\left\|(D_p^{-1} - A_{11}^{-1})A_{12}v\right\|_{A_{11}}^2 = v^H A_{21}A_{11}^{-\frac{1}{2}}(A_{11}^{\frac{1}{2}}D_p^{-1}A_{11}^{\frac{1}{2}} - I)^2 A_{11}^{-\frac{1}{2}}A_{12}v \qquad (7.47) \\
&\leq \varepsilon v^H A_{21}A_{11}^{-\frac{1}{2}}(A_{11}^{\frac{1}{2}}D_p^{-1}A_{11}^{\frac{1}{2}} - I)A_{11}^{-\frac{1}{2}}A_{12}v \\
&\leq \varepsilon v^H A_{21}(D_p^{-1} - A_{11}^{-1})A_{12}v \leq \varepsilon v^H(A_{22} - A_{21}A_{11}^{-1}A_{12})v \\
&\leq \varepsilon \|v\|_{S(A,A_{11})}^2 = \varepsilon.
\end{aligned}$$

Dies führt auf

$$\left\|(I - B_{\mathrm{amg}_p}^{-1}A)e\right\|_A^2 \leq \min_\theta \left[\eta^2(1-\theta)^2 + (\widehat{\mu} + \eta\theta\sqrt{\varepsilon})^2\right].$$

Zur Minimierung dieses Ausdrucks wird die Ableitung bestimmt und diese gleich Null gesetzt. Mit dem Ergebnis

$$\theta = \frac{\eta - \sqrt{\varepsilon}\widehat{\mu}}{(1+\varepsilon)\eta}$$

ergibt sich

$$\begin{aligned}\left\|(I - B_{\mathrm{amg}_p}^{-1}A)e\right\|_A^2 &\leq \eta^2 \left[\frac{(1+\varepsilon)\eta - \eta + \sqrt{\varepsilon}\widehat{\mu}}{(1+\varepsilon)\eta}\right]^2 + \left[\widehat{\mu} + \frac{\eta - \sqrt{\varepsilon}\widehat{\mu}}{1+\varepsilon}\sqrt{\varepsilon}\right]^2 \\ &= \frac{\varepsilon}{(1+\varepsilon)^2}\left(\left[\sqrt{\varepsilon}\eta + \widehat{\mu}\right]^2 + \left[\frac{\widehat{\mu}}{\sqrt{\varepsilon}} + \eta\right]^2\right) \\ &= \frac{\varepsilon}{(1+\varepsilon)^2}\left((1+\varepsilon)\eta^2 + (1+\varepsilon)(\frac{\widehat{\mu}}{\sqrt{\varepsilon}})^2 + (1+\varepsilon)2\eta\frac{\widehat{\mu}}{\sqrt{\varepsilon}}\right) \\ &= \frac{\varepsilon}{1+\varepsilon}\left(\eta + \frac{\widehat{\mu}}{\sqrt{\varepsilon}}\right)^2.\end{aligned}$$

Zum Abschluss wird die rechte Seite maximiert, wobei zu beachten ist, dass $\mu = \sqrt{1-\eta^2}$ mit $\eta \in [0,1]$ und $|\,\widehat{\mu}\,| \leq |\,\mu\sigma_{\max}\,|$ gilt. Das Maximum erhält man bei

$$\eta = \sqrt{\frac{\varepsilon}{\sigma_{\max}^2 + \varepsilon}} \text{ und } \mu = \sqrt{\frac{\sigma_{\max}^2}{\sigma_{\max}^2 + \varepsilon}}.$$

Dies führt auf

$$\begin{aligned}\left\|(I - B_{\mathrm{amg}_p}^{-1}A)e\right\|_A^2 &\leq \frac{\varepsilon}{1+\varepsilon}\left(\sqrt{\frac{\varepsilon}{\sigma_{\max}^2+\varepsilon}} + \sqrt{\frac{\sigma_{\max}^2}{\varepsilon}\cdot\frac{\sigma_{\max}^2}{\sigma_{\max}^2+\varepsilon}}\right)^2 \\ &= \frac{\varepsilon}{1+\varepsilon}\left(\frac{\varepsilon + \sigma_{\max}^2}{\sqrt{\sigma_{\max}^2+\varepsilon}\cdot\sqrt{\varepsilon}}\right)^2 = \frac{1}{1+\varepsilon}\cdot\frac{1}{\sigma_{\max}^2+\varepsilon}\left(\varepsilon + \sigma_{\max}^2\right)^2 \\ &= \frac{\varepsilon}{1+\varepsilon}\left(1 + \frac{\sigma_{\max}^2}{\varepsilon}\right).\end{aligned}$$

Durch die Resubstitution $\varepsilon = \frac{1}{\alpha} - 1$ ergibt sich

$$\begin{aligned}\left\|(I - B_{\mathrm{amg}_p}^{-1}A)e\right\|_A^2 &\leq \frac{\frac{1}{\alpha}-1}{\frac{1}{\alpha}}\left(1 + \frac{\sigma_{\max}^2}{\frac{1}{\alpha}-1}\right) = (1-\alpha)\left(1 + \frac{\sigma_{\max}^2\alpha}{1-\alpha}\right) \\ &= 1 - \alpha + \sigma_{\max}^2\alpha = 1 - \alpha(1-\sigma_{\max}^2).\end{aligned}$$

Dies schließt den Beweis ab. □

Die Abschätzung der A-Norm in (7.42) hängt ebenfalls von dem größten Singulärwert von $P_\nu(A^{\frac{1}{2}}M_s^{-1}A^{\frac{1}{2}})$ ab, wie dies auch bei Satz 7.9 der Fall war. Der Singulärwert kann, wie in Abschnitt 7.3, durch die Verwendung von Tschebysheff Polynomen optimal abgeschätzt werden.

Das polynombasierte AMG ähnelt dem AMGr-Verfahren aus Definition 4.7, wobei die jeweiligen Gewichte verschieden hergeleitet wurden. Das folgende Beispiel zeigt, dass die beiden Verfahren in einem Spezialfall identisch sind.

Beispiel 7.12. Setze $M_s = D_P$ hpd und sei $\nu = 1$ sowie $P_1 \in \mathbb{R}_{\leq 1}[t]$ das Tschebysheff Polynom, welches auf dem Intervall $[1, 1+\varepsilon]$ für ein $\varepsilon > 0$ minimal ist. Dann hat das polynombasierte AMG die Form

$$I - B_{\mathrm{amg}_p}^{-1}A = \left(I - \begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix} A_c^{-1} \begin{bmatrix} -A_{21}D_p^{-1} & I \end{bmatrix} A\right)\left(I - w_1 \begin{bmatrix} D_p^{-1} & 0 \\ 0 & 0 \end{bmatrix} A\right)$$

mit $w_1 = \frac{1}{r_1}$ und r_1 gegeben durch (7.30), also

$$w_1 = \left[\tfrac{1}{2}(1+\varepsilon+1)\right]^{-1} = \tfrac{2}{2+\varepsilon}.$$

Dies bedeutet, dass für den Spezialfall $k = 1$ sowie $a = 1$ und $b = 1+\varepsilon$ das polynombasierte AMG und das AMGr-Verfahren übereinstimmen, vergleiche mit Definition 4.7. Außerdem sind die Konvergenzschranken ebenfalls identisch, denn mit (7.28) folgt

$$\sigma_{\max} \leq 2\tfrac{(\sqrt{b}+\sqrt{a})^1(\sqrt{b}-\sqrt{a})^1}{(\sqrt{b}+\sqrt{a})^2+(\sqrt{b}-\sqrt{a})^2} = 2\tfrac{(\sqrt{1+\varepsilon}+1)(\sqrt{1+\varepsilon}-1)}{(\sqrt{1+\varepsilon}+1)^2+(\sqrt{1+\varepsilon}-1)^2} = \tfrac{2\varepsilon}{2\varepsilon+4}.$$

Setzt man $\alpha = \frac{1}{1+\varepsilon}$, so ergibt sich für die Normabschätzung (7.42)

$$\begin{aligned}
&\left\|I - B_{\mathrm{amg}_r}A\right\|_A^2 \leq 1-\alpha\Big(1-\sigma_{\max}^2\Big) \leq 1-\tfrac{1}{1+\varepsilon}\Big(1-\tfrac{4\varepsilon^2}{4\varepsilon^2+16\varepsilon+16}\Big)\\
&= \tfrac{\varepsilon}{1+\varepsilon}\Big(\tfrac{1+\varepsilon}{\varepsilon} - \tfrac{1}{\varepsilon}\cdot\tfrac{4\varepsilon+4}{\varepsilon^2+4\varepsilon+4}\Big) = \tfrac{\varepsilon}{1+\varepsilon}\Big(1+\tfrac{(1+\varepsilon)(2+\varepsilon)^2-(4\varepsilon+4)-\varepsilon(2+\varepsilon)^2}{\varepsilon(2+\varepsilon)^2}\Big)\\
&= \tfrac{\varepsilon}{1+\varepsilon}\Big(1+\tfrac{(2+\varepsilon)^2-4\varepsilon-4)}{\varepsilon(2+\varepsilon)^2}\Big) = \tfrac{\varepsilon}{1+\varepsilon}\Big(1+\tfrac{\varepsilon}{(2+\varepsilon)^2}\Big),
\end{aligned}$$

was der Abschätzung aus Satz 4.5 entspricht.

Obiges Beispiel verdeutlicht, dass das polynombasierte AMG als eine Verallgemeinerung des AMGr-Verfahrens betrachtet werden kann. Es gibt zwei wesentliche Verbesserungen.

1. In dem Glätter werden auch nicht hermitesche Approximationen M_s zugelassen. In diesem Fall ist $w_j = 1$, $j = 1, \ldots, n$. Bei dieser Wahl der Gewichte entspricht das polynombasierte AMG dem *MAMLI-Verfahren* aus (4.11) und das symmetrische polynombasierte AMG dem *SMAMLI-Verfahren* aus (4.12), siehe [76–78]. In diesen Arbeiten werden nichtsymmetrische M-Matrizen betrachtet, daher stimmen die Beobachtungen überein.

2. Für hpd Approximationen M_s erhält man das AMGr-Verfahren. Jedoch nutzt dies bei der Verwendung mehrerer Glätter stets das Gewicht $w_j = \sigma = \frac{2}{2+\varepsilon}$, siehe Satz 4.6. Für die Analyse mit Satz 7.11 führt dies auf das Polynom

$$P_\nu(t) = \Big(1-\tfrac{2t}{2+\varepsilon}\Big)^\nu.$$

   Das Maximum dieses Polynoms auf $[1, 1+\varepsilon]$ ist größer, als bei der Verwendung der Tschebysheff Polynome. Daher erhält man eine Verbesserung der Konvergenzschranke durch die polynombasierten AMG's. Wie stark diese Verbesserung in der Praxis ausfällt, wird in Kapitel 9 untersucht.

Sowohl beim AMGr-Verfahren und dem polynombasierten AMG (AMGp-Verfahren), als auch bei der Abschätzung der C.B.S Konstante von $\widehat{A}$ in Satz 6.3 wurde dieselbe Voraussetzung

$$D_p \preceq A_{11}, \quad \begin{bmatrix} D_p & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \succeq 0 \tag{7.48}$$

verwendet. Das folgende Lemma zeigt, dass bei den AMGr- und AMGp-Verfahren implizit die C.B.S. Konstante abgeschätzt wurde. Dafür sei daran erinnert, dass die Voraussetzung (7.48) bei der Abschätzung (7.47), also bei

$$\left\|(D_p^{-1} - A_{11}^{-1})A_{12}v\right\|_{A_{11}}^2 \leq \varepsilon \left\|v\right\|_{S(A,A_{11})}^2,$$

verwendet wurde.

Lemma 7.13. *Sei $A \in \mathbb{C}^{n,n}$ hpd und partitioniert wie in (4.3) und $\gamma(\widehat{A})$ sei die C.B.S. Konstante von*

$$\widehat{A} = \begin{bmatrix} I & -D_p^{-1}A_{12} \\ 0 & I \end{bmatrix}^H A \begin{bmatrix} I & -D_p^{-1}A_{12} \\ 0 & I \end{bmatrix}.$$

Dann gilt

$$\left\|(D_p^{-1} - A_{11}^{-1})A_{12}v\right\|_{A_{11}} \leq \sqrt{\varepsilon}\,\|v\|_{S(A,A_{11})} \tag{7.49}$$

für $\varepsilon > 0$ und alle $v \in \mathbb{C}^{n_2}$ genau dann, wenn $\gamma(\widehat{A}) \leq \sqrt{1 - \frac{1}{1+\varepsilon}}$.

Beweis. Sei (7.49) erfüllt, dann gilt für

$$\widehat{A}_{22} = \begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix}^H A \begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix}$$

die Ungleichung

$$\begin{aligned} v^H(\widehat{A}_{22} - S(A,A_{11}))v &= v^H A_{21}\Big[A_{11}^{-1} - (D_p^{-1} + D_p^{-H} - D_p^{-H}A_{11}D_p^{-1})\Big]A_{12}v \\ = \left\|(D_p^{-1} - A_{11}^{-1})A_{12}v\right\|_{A_{11}}^2 &\leq \varepsilon\,\|v\|_{S(A,A_{11})}^2 = \varepsilon v^H S(A,A_{11})v \end{aligned}$$

Daher ist (7.49) äquivalent zu

$$v^H\widehat{A}_{22}v \leq (1+\varepsilon)v^H S(A,A_{11})v$$

bzw. zu

$$\frac{v^H S(A,A_{11})v}{v^H\widehat{A}_{22}v} \geq \frac{1}{1+\varepsilon}$$

für alle $v \in \mathbb{C}^{n_2}$. Mit (5.5) erhält man

$$1 - \gamma(\widehat{A})^2 = \min_{v\neq 0} \frac{v^H S(A,A_{11})v}{v^H\widehat{A}_{22}v} \geq \frac{1}{1+\varepsilon}$$

und damit die Behauptung. □

Durch Anwendung von Lemma 7.13 kann die Voraussetzung von Satz 7.11 modifiziert werden. Z.B. kann

$$D_p \preceq A_{11}, \quad \begin{bmatrix} D_p & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \succeq 0$$

durch

$$D_p \succeq A_{11}, \quad \begin{bmatrix} D_p(D_p + D_pA_{11}^{-1}D_p - A_{11})^{-1}D_p & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \succeq 0$$

ersetzt werden, siehe Satz (6.11). Weitere Varianten werden im Anhang A vorgestellt.

In Kapitel 6 wurde gezeigt, dass die Voraussetzung von Satz 7.11 sowohl für diagonal dominante, als auch für M-Matrizen erfüllt werden kann. Im Fall der M-Matrizen muss dafür ein Eigenvektor zum kleinsten Eigenwert bekannt sein bzw. gut approximiert werden können. Der

Ansatz, diesen Eigenvektor zur Konstruktion der Interpolationsmatrix zu verwenden, ist auch für Matrizen mit weniger Struktur bzw. komplexwertige Matrizen ein erfolgreicher Ansatz, wie in den numerischen Berechnungen in [33] dargestellt wird, führt aber zu einer Diagonalmatrix D_p mit nicht reellen Einträgen, also insbesondere $D_p^H \neq D_p$. In dem folgenden Lemma wird dargelegt, wie Satz 7.11 trotzdem angewendet werden kann.

Lemma 7.14. *Sei $A \in \mathbb{C}^{n,n}$ hpd und partitioniert wie in (4.3). Weiter sei $D_p \in \mathbb{C}^{n_1,n_1}$ nichtsingulär und $\varepsilon > 0$, so dass*

$$D_p^{-1} + D_p^{-H} - A_{11}^{-1} \succeq \varepsilon A_{11}^{-1} \quad \textit{und} \tag{7.50}$$

$$\varepsilon A_{22} \succeq A_{21} D_p^{-H} A_{11} D_p^{-1} A_{12}, \tag{7.51}$$

dann gilt für alle $v \in \mathbb{C}^{n_2}$

$$\left\| (D_p^{-1} - A_{11}^{-1}) A_{12} v \right\|_{A_{11}} \leq \sqrt{\varepsilon}\, \|v\|_{S(A,A_{11})}.$$

Beweis. Es ist für alle $v \in \mathbb{C}^{n_2}$

$$\begin{aligned}
\left\| (D_p^{-1} - A_{11}^{-1}) A_{12} v \right\|_{A_{11}}^2 &= v^H A_{21} \left[A_{11}^{-1} - (D_p^{-H} + D_p^{-1}) + D_p^{-H} A_{11} D_p^{-1}) \right] A_{12} v \\
&\overset{(7.50)}{\leq} v^H A_{21} D_p^{-H} A_{11} D_p^{-1} A_{12} v - \varepsilon v^H A_{21} A_{11}^{-1} A_{12} v \\
&\overset{(7.51)}{\leq} \varepsilon v^H \Big(A_{22} - A_{21} A_{11}^{-1} A_{12} \Big) v = \varepsilon \|v\|_{S(A,A_{11})}^2 .
\end{aligned}$$

Damit ist die Behauptung gezeigt. □

Dieses Lemma erlaubt es, nicht hermitesche Approximationen D_p für das AMGp-Verfahren zu betrachten.

Abschließend wird ein weiteres Mal das symmetrische polynombasierte AMG betrachtet, nun aber mit derselben Voraussetzung, die für Satz 7.11 verwendet wurde und unter anderem für diagonaldominante hpd Matrizen erfüllbar ist, siehe Abschnitt 6.3.

Dafür wird benötigt, dass C definiert durch (7.10) eine Projektion ist, also $C^2 = C$ gilt. Für die A-Norm der Iterationsmatrix bedeutet das

$$\begin{aligned}
\|I - B_{\mathrm{amg}_{p,s}}^{-1}\|_A = \|R_{\mathrm{post}} C R_{\mathrm{pre}}\|_A = \|R_{\mathrm{post}} C C R_{\mathrm{pre}}\|_A &\leq \|R_{\mathrm{post}} C\|_A \, \|C R_{\mathrm{pre}}\|_A \\
= \|C R_{\mathrm{pre}}\|_A^2 = \|I - B_{\mathrm{amg}_p}^{-1} A\|_A^2 &\leq 1 - \alpha(1 - \sigma_{\max}^2).
\end{aligned}$$

Damit folgt schließlich

$$\lambda_{\max}\Big(B_{\mathrm{amg}_{p,s}}^{-1} A \Big) \leq 2 + \alpha(1 - \sigma_{\max}^2) \quad \text{und} \tag{7.52}$$

$$\lambda_{\min}\Big(B_{\mathrm{amg}_{p,s}}^{-1} A \Big) \geq \alpha(1 - \sigma_{\max}^2). \tag{7.53}$$

Dies führt auf das folgende Korollar.

Korollar 7.15. *Sei A hpd und partitioniert wie in* (4.3)*. Weiter seien die Voraussetzungen von Satz 7.11 erfüllt. Dann gilt*

$$\kappa_2\Big(I - B_{\mathrm{amg}_{p,s}}^{-1} A\Big) \leq 1 + \tfrac{2}{\alpha(1-\sigma_{\max}^2)}, \tag{7.54}$$

wobei $B_{\mathrm{amg}_{p,s}}$ durch (7.9) *mit den Gewichten $w_i = \frac{1}{r_i}$, $i = 1, \ldots, \nu$ gegeben ist und $S = A_c$ gilt. Die Matrix A_c sei wie in* (7.39) *definiert.*

Kapitel 8

AMG's mit voller Glättung basierend auf Tschebysheff Polynomen

In dieser Arbeit werden AMG's mit F-Glättung untersucht und dafür polynombasierte AMG's betrachtet. In der ersten Hälfte dieses Kapitels soll gezeigt werden, dass einige der vorgestellten Ideen auch bei AMG's mit voller Glättung Anwendung finden. Insbesondere durch die Theorie aus [48] und [93] können AMG's mit voller Glättung und der Verwendung von Tschebysheff Polynomen untersucht werden.

In der Arbeit [48] wird ein zu $\mathbb{C}$ isomorpher Vektorraum V in eine Summe

$$V = SV_1 + PV_2, \tag{8.1}$$

für gewisse Räume V_1 und V_2 aufgeteilt. Ein *allgemeines Zweilevelverfahren* wird definiert durch

$$T_{\mathrm{tl}} := I - B_{\mathrm{tl}}^{-1}A := (I - SM_s^{-H}S^HA)(I - PA_c^{-1}P^HA)(I - SM_s^{-1}S^HA). \tag{8.2}$$

Hierbei sind $M_s \approx A_s := S^HAS$ and $A_c := P^HAP$. Wird $S = \left[\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right]$ gewählt, so entspricht (8.2) den in dieser Arbeit betrachteten AMG's mit F-Glättung. Für $S = I$ ergibt sich ein *Zweigitterverfahren mit Glättung auf allen Unbekannten*

$$T_{\mathrm{tg}} = I - B_{\mathrm{tg}}^{-1}A = (I - M^{-H}A)(I - PA_c^{-1}P^HA)(I - M^{-1}A). \tag{8.3}$$

In [48, 93] werden sowohl T_{tl}, als auch T_{tg} analysiert und insbesondere wird eine scharfe Abschätzung der Konditionszahl angegeben.

Für die in [48] bzw. [93] vorgestellte Analyse wird ein sogenannter *symmetrisierter Glätter*

$$\widetilde{M} := M^H(M^H + M - A)^{-1}M = A\left[I - (I - M^{-1}A)(I - M^{-H}A)\right]^{-1} \tag{8.4}$$

eingeführt und damit der folgende Satz gezeigt. Dieses Resultat wird in [48] für reellwertige Matrizen bewiesen, eine Verallgemeinerung auf komplexwertige Matrizen folgt denselben Ideen.

Satz 8.1 ([48], Theorem 5.1)**.** *Seien $A \in \mathbb{C}^{n,n}$ hpd, $\widetilde{M} \approx A$ mit*

$$A \preceq \widetilde{M} \preceq \beta A \quad \textit{und} \quad A_c \preceq \theta S_A,$$

wobei S_A das Schurkomplement ist, definiert durch

$$x^HS_Ax = \inf_y (Sy + Px)^HA(Sy + Px), \tag{8.5}$$

dann gilt

$$A \preceq B_{\mathrm{tg}} \preceq \beta\theta A.$$

Bemerkung 8.2. Das Schurkomplement aus (8.5) entspricht dem in dieser Arbeit verwendeten Schurkomplement. Um das einzusehen, betrachtet man den zu minimierenden Term aus (8.5)

$$(Sy + Px)^H A(Sy + Px) = \begin{bmatrix} x \\ y \end{bmatrix}^H \begin{bmatrix} S^H AS & S^H AP \\ P^H AS & P^H AP \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

Definiert man die übliche Blockdarstellung

$$\begin{bmatrix} S^H AS & S^H AP \\ P^H AS & P^H AP \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

so erhält man

$$\inf_y (Sy + Px)^H A(Sy + Px) = x^H A_{22} x + \inf_y \Big(y^H A_{11} y + 2y^H A_{12} x \Big).$$

Dabei wird das Minimum bei $y = -A_{11}^{-1} A_{12} x$ angenommen und es ergibt sich mit (8.5)

$$x^H S_A x = \inf_y (Sy + Px)^H A(Sy + Px) = x^H \Big(A_{22} - A_{21} A_{11}^{-1} A_{12} \Big) x,$$

also $S_A = A_{22} - A_{21} A_{11}^{-1} A_{12}$.

Im Folgenden wird zunächst das durch (8.3) gegebene Zweigitterverfahren mit voller Glättung betrachtet und untersucht. In der Darstellung (8.3) ist dies ein Zweilevelverfahren, basierend auf einem $V(1,1)$-Zyklus, also einem Vor- und einem Nachglätter. Analog dazu definiert man ein polynombasiertes AMG auf einem $V(\nu,\nu)$-Zyklus durch

$$I - B_{\mathrm{tg}_p}^{-1} A := (I - M_p^{-H} A)(I - PA_c^{-1} P^H A)(I - M_p^{-1} A),$$

vergleiche Kapitel 7. Dabei ist der Glätter gegeben durch

$$I - M_p^{-1} A := \Big(I - \tfrac{1}{r_1} M^{-1} A \Big) \dots \Big(I - \tfrac{1}{r_\nu} M^{-1} A \Big)$$

mit $M \approx A$ und r_j definiert durch die Nullstellen der auf 1 normierten Tschebysheff Polynome. Analog zum Beweis von Korollar 7.8 ergibt sich die Darstellung

$$M_p = A \Big[I - P_\nu(M^{-1} A) \Big]^{-1}$$

mit P_ν gegeben durch (3.11) für hermitesch positiv definites M bzw. (3.4) für nichtsinguläres M mit $\|I - M^{-1} A\|_A < 1$. Eingesetzt in (8.4) erhält man

$$\widetilde{M}_p = A \Big[I - P_\nu(M^{-H} A) P_\nu(M^{-1} A) \Big]^{-1},$$

und damit (vgl. (7.23))

$$A^{\frac{1}{2}} \widetilde{M}_p^{-1} A^{\frac{1}{2}} = I - P_\nu(A^{\frac{1}{2}} M^{-H} A^{\frac{1}{2}}) P_\nu(A^{\frac{1}{2}} M^{-1} A^{\frac{1}{2}}).$$

Mit der Abschätzung

$$\lambda_{\min}(\widetilde{M}_p^{-1} A) = 1 - \max_{\|x\|_2 = 1} \Big\| P_\nu(A^{\frac{1}{2}} M^{-1} A^{\frac{1}{2}}) x \Big\|_2 \geq 1 - \sigma_{\max}, \tag{8.6}$$

wobei $\sigma_{\max} = \sigma_{\max}(P_\nu(A^{\frac{1}{2}} M^{-1} A^{\frac{1}{2}}))$ der größte Singulärwert von $P_\nu(A^{\frac{1}{2}} M^{-1} A^{\frac{1}{2}})$ ist, ergeben sich die folgenden Aussagen.

Satz 8.3. *Seien $A \in \mathbb{C}^{n,n}$ hpd und*

$$A_c \preceq \theta S_A,$$

wobei S_A das Schurkomplement ist, was durch

$$x^H S_A x = \inf_y (y + Px)^H A (y + Px)$$

definiert ist. Weiter sei

$$\begin{aligned} I - B_{\mathrm{tg}_p} A = &(I - \omega_\nu M^{-H} A) \cdots (I - \omega_1 M^{-H} A) \\ &\cdot (I - P A_c^{-1} P^H A) \\ &\cdot (I - \omega_1 M^{-1} A) \cdots (I - \omega_\nu M^{-1} A). \end{aligned}$$

1. *Falls $M \approx A$ hpd mit $\sigma(M^{-1}A) \subseteq [a, b]$ ist, sowie die Gewichte durch $\omega_j = \frac{1}{r_j}$ mit r_j aus* (7.30) *gegeben sind, dann gilt*

   $$A \preceq B_{\mathrm{tg}_p} \preceq \tfrac{\theta}{1-\sigma_{\max}} A,$$

   wobei $\sigma_{\max}$ durch (7.28) *gegeben und damit die Ungleichung*

   $$1 - \sigma_{\max} \geq \frac{\left((\sqrt{b}+\sqrt{a})^\nu - (\sqrt{b}-\sqrt{a})^\nu\right)^2}{(\sqrt{b}+\sqrt{a})^{2\nu} + (\sqrt{b}-\sqrt{a})^{2\nu}}$$

   erfüllt ist.

2. *Falls $M \approx A$ nichtsingulär mit $\|I - M^{-1}A\|_A \leq \rho$, $\rho < 1$ ist und $\omega_j = 1$, $j = 1, \ldots, \nu$, so gilt*

   $$A \preceq B_{\mathrm{tg}_p} \preceq \tfrac{\theta}{1-\rho^\nu} A.$$

Beweis. Es ist Satz 8.1 anzuwenden. Für diesen Satz wird gefordert, dass $A \preceq \widetilde{M}_p \preceq \beta A$ und $A_c \preceq \theta S_A$ erfüllt ist. Damit ist ein geeignetes β zu bestimmen, dass diese Ungleichung erfüllt. Mit (8.6) ergibt sich

$$\beta^{-1} := \lambda_{\min}(\widetilde{M}_p^{-1} A) \geq 1 - \sigma_{\max}.$$

Damit folgt durch Satz 8.1

$$A \preceq B_{\mathrm{tg}_p} \preceq \tfrac{\theta}{1-\sigma_{\max}} A.$$

Für den Fall, dass M hpd ist, ist P_ν gegeben durch (3.11) und damit kann der Singulärwert durch (7.28) abgeschätzt werden. Im zweiten Fall ist P_ν definiert durch $P_\nu(t) = (1-t)^\nu$ und daher $\sigma_{\max} \leq \rho^\nu$, vergleiche (7.31). □

Folglich bietet die Konstruktion eines polynombasierten AMGs auch bei voller Glättung eine Möglichkeit zur Beschleunigung. Es bleibt aber die Schwierigkeit bestehen, eine geeignete Interpolationsmatrix P zu bestimmen, so dass θ nicht zu groß wird, was erneut eine Abschätzung der C.B.S. Konstante einer Matrix $\widehat{A}$ mit sich bringt, siehe Lemma 6.4.

Die vorgestellte Arbeit von Falgout, Vassilevski, Zikatanov [48] verallgemeinert klassische Resultate für AMG's und gibt eine scharfe Schranke der Eigenwerte von $B_{\mathrm{tl}}^{-1}A$ und $B_{\mathrm{tg}}^{-1}A$ an;

hierbei sind $B_{\mathrm{tl}}^{-1}, B_{\mathrm{tg}}^{-1}$ gegeben wie in (8.2) bzw. (8.3). Außerdem werden Kriterien hergeleitet, wie die Interpolationsmatrix P und damit auch die Grobgittermatrix $A_c = P^H AP$ im optimalen Fall gewählt werden muss.

Im zweiten Teil dieses Kapitels wird untersucht, ob diese allgemeinen Resultate hilfreich sind, AMG's mit F-Glättung zu analysieren und gegebenenfalls die Ergebnisse dieser Arbeit verallgemeinert werden können. Dazu werden einige Resultate aus Abschnitt 4 von [48] zusammengefasst. Sei A partitioniert wie in (4.3). Durch die Wahl von $S = \left[\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right]$ bei dem Zweilevelverfahren aus (8.2) erhält man ein AMG, dass sich, mit den in dieser Arbeit besprochenen, vergleichen lässt.

Um einen effizienten Vorkonditionerer zu konstruieren, müssen zwei Bedingungen erfüllt sein.

1. Der symmetrisierte Glätter $\widetilde{M_s} := M_s^H(M_s^H + M - A)^{-1}M$ erfülle die Bedingung

$$x^H A_s x \leq x^H \widetilde{M_s} x \leq \beta x^H A_s x, \tag{8.7}$$

   für ein möglichst kleines $\beta \geq 1$ und alle $x \in V_1$ mit V_1 wie in (8.1). Mit der Wahl $S = \left[\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right]$ entspricht $A_s = A_{11}$ und die Bedingung (8.7) wird durch die Wahl der Tschebysheff Polynome erfüllt.

2. Die zweite Bedingung charakterisiert die Wahl der Interpolation und der Grobgittermatrix. Es soll gelten, dass

$$x^H A_c x \leq \theta \inf_y (Sy + Px)^H A(Sy + Px) \tag{8.8}$$

   für alle $x \in V_2$ erfüllt ist. Hierbei sei $\theta \geq 1$ möglichst klein. Setzt man $S = \left[\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right]$ ein und betrachtet die Interpolationsmatrix $P = \left[\begin{smallmatrix} -H^{-1}A_{12} \\ I \end{smallmatrix}\right]$, so erhält man

$$\begin{bmatrix} S & P \end{bmatrix}^H A \begin{bmatrix} S & P \end{bmatrix} = \begin{bmatrix} I & -H^{-1}A_{12} \\ 0 & I \end{bmatrix}^H A \begin{bmatrix} I & -H^{-1}A_{12} \\ 0 & I \end{bmatrix}$$

   und damit

$$\begin{bmatrix} S & P \end{bmatrix}^H A \begin{bmatrix} S & P \end{bmatrix} = \widehat{A}$$

   mit $\widehat{A}$ definiert wie in (6.2). Wie in Bemerkung 8.2 festgestellt wurde, gilt

$$\inf_y (Sy + Px)^H A(Sy + Px) = S(\widehat{A}, \widehat{A}_{11}).$$

   Da $\widehat{A}_{11} = A_{11}$ und $S(\widehat{A}, \widehat{A}_{11}) = S(A, A_{11})$ ist, wie im Beweis von Lemma 6.15 bemerkt wurde, ist (8.8) äquivalent zu

$$x^H A_c x \leq \theta x^H S(A, A_{11}) x.$$

   Dies entspricht der Aussage aus Lemma 7.13. Daher muss auch hier die C.B.S. Konstante abgeschätzt werden (siehe Sätze 6.3 und 6.11).

Folglich scheint durch die Theorie aus [48] und [93] für den Spezialfall der F-Glättung und $P = \left[\begin{smallmatrix} -H^{-1}A_{12} \\ I \end{smallmatrix}\right]$ keine Verbesserung möglich zu sein.

Kapitel 9

Numerische Resultate

In diesem Kapitel werden die theoretischen Resultate anhand numerischer Berechnungen verifiziert. Als Coarser wird für alle Experimente der Greedy-Coarser, Algorithmus 6, verwendet. Daher wird vorausgesetzt, dass A wie in (4.3) partitioniert ist. Durch die Wahl von $\varphi \in (\frac{1}{2}, 1)$ ermöglicht dieser Coarser, das Verhältnis n_f/n_c mit $n_f = \#$Feingitterpunkte und $n_c = \#$Grobgitterpunkte zu variieren, was bei der Analyse der Verfahren hilfreich ist. Hierbei bezeichnet $\#M$ die Anzahl der Elemente in einer Menge M.

9.1

Poisson-Gleichung. Die ersten numerischen Berechnungen werden anhand der 2D Poisson-Gleichung aus Abschnitt 2.1 durchgeführt. Dieses Modellproblem wurde unter anderem auch in [68] zur Verifizierung des AMGr-Verfahrens betrachtet. Da in den Arbeiten [33, 68] bereits gezeigt wurde, dass das AMGr-Verfahren für verschiedene Probleme gute Ergebnisse liefert, wird in diesem Abschnitt lediglich ein Vergleich auf einem kleinen Testproblem zwischen dem AMGr- und dem AMGp-Verfahren durchgeführt, wobei folgende Varianten betrachtet werden:

1. Ein Vergleich anhand des Spektralradius' zwischen dem AMGr-Verfahren und dem polynombasierten AMG, das mit AMGp bezeichnet wird.
2. Ein Vergleich verschiedener hpd Approximationen M_s an A_{11}, wobei M_s die Matrix im Glätter darstellt.
3. Ein Vergleich verschiedener hpd Approximationen D_p an A_{11}, wobei D_p die Matrix in der Interpolation darstellt.
4. Ein Vergleich einer hpd Approximation M_s an A_{11} mit einer nicht hermiteschen Approximation M_s.
5. Die Verwendung eines adaptiv-ähnlichen Prozesses für die Konstruktion der Interpolationsmatrix.

Zunächst wird das AMGr-Verfahren mit seiner Verallgemeinerung, dem polynombasierten AMG, verglichen. Wie bei dem klassischen AMGr-Verfahren aus [68] wird $D = D_p = M_s$ so gewählt, dass $D \preceq A_{11}$ und die Matrix

$$\begin{bmatrix} D & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

hpsd ist, vergleiche Definition 4.4 und Satz 4.5. Dafür wird $D = \mathrm{diag}(d_{11}, \ldots, d_{n_1,n_1})$ mit

$$d_{jj} = 2a_{jj} - \sum_{j=1}^{n_f} |\, a_{jk} \,|$$

gesetzt. Durch die M-Matrix Struktur der Matrix A entspricht dies

$$D = \mathrm{diag}(A_{11}e)$$

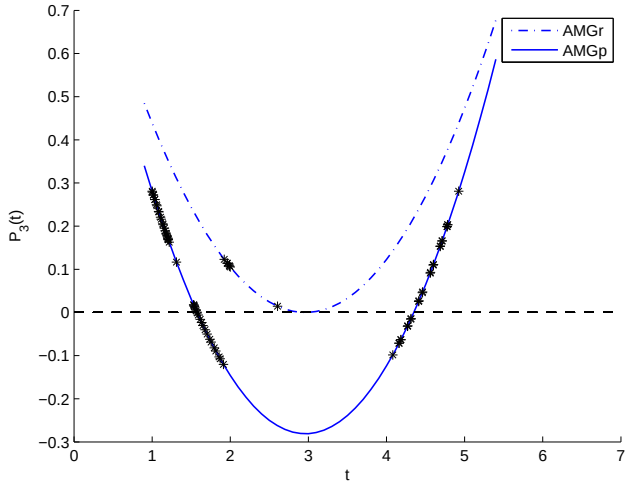
mit $e = [\,1 \ldots 1\,]^H$. Mit Satz 6.6 folgt, dass D die benötigten Voraussetzungen erfüllt.

Für den Vergleich mit dem polynombasierten AMG werden die Gewichte als Kehrwerte der Nullstellen von dem in 1 normierten Tschebysheff Polynom auf $[1, 1+\varepsilon]$ definiert, wobei ε mithilfe von Satz 6.6 abgeschätzt oder durch ein verallgemeinertes Eigenwertproblem approximiert werden kann. Das abgeschätzte ε wird mit ε_t und das approximierte bzw. exakte mit ε_e bezeichnet. Des Weiteren wird das Zweilevelverfahren betrachtet und $S = A_c = P^H AP$ als Galerkin-Matrix gesetzt, siehe dazu Definition 7.6.

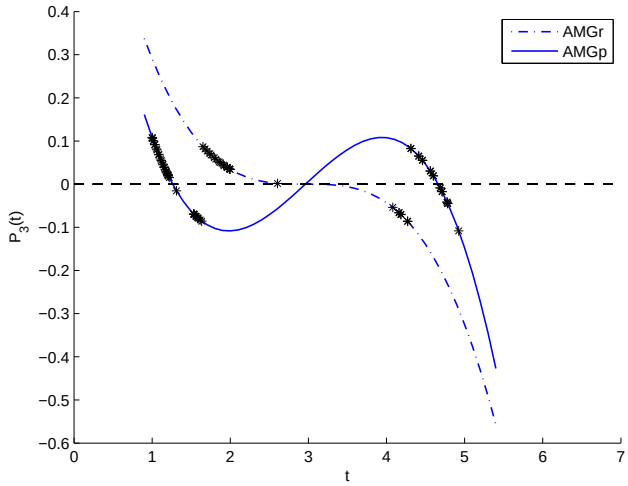
	φ	n_c	Verfahren	ε_t	$\rho(\varepsilon_t)$ $\nu=1$	$\nu=2$	$\nu=3$	$\nu=4$	ε_e	$\rho(\varepsilon_e)$ $\nu=1$	$\nu=2$	$\nu=3$	$\nu=4$
16×16	0.65	126	AMGr	2.33	.54	.29	.16	.10	1.33	.40	.24	.06	.11
			AMGp		.54	.17	.11	.10		.40	.17	.08	.10
	0.60	126	AMGr	4	.67	.44	.30	.20	1.33	.40	.24	.06	.11
			AMGp		.67	.29	.11	.10		.40	.17	.08	.10
	4/7	98	AMGr	6	.75	.59	.55	.54	3.92	.66	.71	.45	.60
			AMGp		.75	.42	.59	.54		.66	.63	.50	.54
	.55	98	AMGr	9	.82	.67	.57	.55	3.92	.66	.71	.45	.60
			AMGp		.82	.50	.48	.58		.66	.63	.50	.54
32×32	0.65	510	AMGr	2.33	.54	.29	.16	.10	1.33	.40	.24	.06	.12
			AMGp		.54	.17	.11	.10		.40	.17	.08	.10
	0.60	510	AMGr	4	.67	.44	.30	.20	1.33	.40	.24	.06	.12
			AMGp		.67	.29	.11	.10		.40	.17	.08	.10
	4/7	450	AMGr	6	.75	.59	.55	.54	3.92	.66	.71	.45	.60
			AMGp		.75	.42	.59	.54		.66	.63	.51	.54
	.55	450	AMGr	9	.82	.67	.57	.55	3.92	.66	.71	.45	.60
			AMGp		.82	.50	.48	.58		.66	.63	.50	.54

Tabelle 9.1.: Vergleich des AMGr Verfahrens und des polynombasierten AMGs (AMGp), angewandt auf die 2D Poisson-Gleichung auf einem 16×16 bzw. 32×32 Gitter.

Aus der Tabelle 9.1 wird die bessere Performance von AMGp im Vergleich zu AMGr deutlich. Insbesondere, wenn das Intervall $[1, 1+\varepsilon]$ nicht zu groß ist, geht die Wirkung des Tschebysheff Polynoms in das AMGp-Verfahren ein. Man erkennt aber auch, dass in Spezialfällen der Spektralradius des AMGr-Verfahrens kleiner, als der des AMGp-Verfahrens ist. Um diese Beobachtung zu erklären, werden in Abbildung 9.1 die verwendeten Polynome der beiden Verfahren betrachtet.



(a) $\nu = 2$



(b) $\nu = 3$

Abbildung 9.1.: Die Polynome der AMGr- und AMGp-Verfahren vom Grad ν bei $\varphi = 0.55$, wobei die Intervallgrenzen exakt bestimmt wurden. Mit $*$ sind die Eigenwerte von $P_\nu(D^{-1}A_{11})$ für das jeweilige Polynom mit dem jeweils kleinen Betrag markiert: $t = \min\left(\lambda_j(P_\nu^{\text{amg}_r}(D^{-1}A_{11})), \lambda_j(P_\nu^{\text{amg}_p}(D^{-1}A_{11}))\right)$.

In Abbildung 9.1b ist einer der Fälle skizziert, in dem der Spekralradius des AMGr-Verfahrens kleiner als der des AMGp-Verfahrens ist. In der Konstruktion des AMGp-Verfahrens geht ausschließlich der kleinste und der größte Eigenwert von $D^{-1}A_{11}$ ein. Es wird lediglich auf dem, durch diese beiden Eigenwerte gegebenen, Intervall minimiert. Dabei werden die anderen Eigenwerte außer Acht gelassen. Daher ist es in Spezialfällen möglich, dass das AMGr-Verfahren asymptotisch schneller, als das AMGp-Verfahren konvergiert. In Tabelle 9.1 wird aber deutlich, dass dies tatsächlich nur in Ausnahmen der Fall ist und das AMGp-Verfahren im Allgemeinem einen kleineren Konvergenzradius als das AMGr-Verfahren besitzt.

Ein weiterer Vorteil des AMGp-Verfahrens ist, dass eine gewisse Freiheit bei der Wahl der Approximation M_s besteht. Folgende drei Varianten werden untersucht.

1. $M_s^{[1]} := \mathrm{diag}(m_1^{[1]}, \ldots, m_{n_1}^{[1]})$ mit $m_j^{[1]} := 2a_{jj} - \sum_{k=1}^{n_f} |\, a_{jk} \,|$,
2. $M_s^{[2]} := \mathrm{diag}(m_1^{[2]}, \ldots, m_{n_1}^{[2]})$ mit $m_j^{[2]} := a_{jj}$,
3. $M_s^{[3]} := \mathrm{diag}(m_1^{[3]}, \ldots, m_{n_1}^{[3]})$ mit $m_j^{[3]} := \sum_{k=1}^{n_f} |\, a_{jk} \,|$.

Die Matrix D_p wird zunächst durch $D_p = M_s$ festgelegt.

φ	Glätter	$[a_t, b_t]$	$\rho(a_t, b_t)$ $\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 4$	$[a_e, b_e]$	$\rho(a_e, b_e)$ $\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 4$
0.65	$M_s^{[1]}$	$[1, 3.33]$	.54	.17	.11	.11	$[1, 2.33]$	.40	.17	.08	.10
	$M_s^{[2]}$	$[0.46, 1.54]$	.37	.17	.07	.04	$[0.65, 1.35]$	.37	.19	.06	.06
	$M_s^{[3]}$	$[0.3, 1]$	.54	.17	.05	.06	$[0.47, 1]$	.41	.14	.08	.08
0.6	$M_s^{[1]}$	$[1, 5]$	.67	.29	.11	.10	$[1, 2.33]$	.40	.17	.08	.10
	$M_s^{[2]}$	$[0.33, 1.67]$	.38	.29	.13	.04	$[0.65, 1.35]$	.37	.19	.06	.06
	$M_s^{[3]}$	$[0.2, 1]$	.67	.29	.11	.08	$[0.47, 1]$	.41	.14	.08	.08
0.55	$M_s^{[1]}$	$[1, 10]$	.81	.50	.48	.58	$[1, 4.9]$	.66	.63	.50	.54
	$M_s^{[2]}$	$[0.18, 1.82]$	.70	.50	.37	.27	$[0.41, 1.59]$	.70	.47	.38	.37
	$M_s^{[3]}$	$[0.1, 1]$	.82	.50	.37	.34	$[0.25, 1]$	.74	.50	.43	.41

Tabelle 9.2.: Vergleich verschiedener Approximationen M_s im Glätter, wobei $D_p = M_s$ gesetzt wird. Hier wird das AMGp-Verfahren auf die Poisson-Gleichung auf einem 16×16 Gitter angewendet.

In Tabelle 9.2 ist ersichtlich, dass für dieses Beispiel die Diagonale von A_{11} die beste Approximation von den drei betrachteten darstellt. In der nächsten Berechnung wird erneut $M_s = \mathrm{diag}(A_{11})$ gesetzt, doch dieses Mal wird D_p variiert. Die gewählten Approximationen werden analog zu M_s definiert und mit $D_p^{[1]}, D_p^{[2]}, D_p^{[3]}$ bezeichnet.

φ	Interpolation	ρ_t $\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 4$	ρ_e $\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 4$
0.65	$D_p^{[1]}$	.36	.17	.13	.08	.36	.15	.08	.10
	$D_p^{[2]}$	.38	.17	.07	.04	.37	.12	.06	.06
	$D_p^{[3]}$	.40	.17	.05	.06	.40	.13	.08	.07
0.6	$D_p^{[1]}$	.36	.29	.19	.07	.36	.15	.08	.10
	$D_p^{[2]}$	.38	.29	.12	.04	.37	.12	.06	.06
	$D_p^{[3]}$	.40	.29	.11	.05	.40	.13	.08	.07
0.55	$D_p^{[1]}$	.63	.51	.61	.49	.63	.61	.52	.54
	$D_p^{[2]}$	.56	.36	.39	.37	.70	.47	.38	.37
	$D_p^{[3]}$	.73	.50	.31	.34	.73	.50	.43	.41

Tabelle 9.3.: Vergleich verschiedener Approximationen für die Interpolationsmatrix bei Verwendung von $M_s = M_s^{[2]}$. Hier wird das AMGp-Verfahren auf die Poisson-Gleichung auf einem 16×16 Gitter angewendet. Die Werte a, b werden approximiert.

Tabelle 9.3 zeigt, dass diese drei verschiedenen Möglichkeiten der Wahl von D_p nur geringen Einfluss auf den Spektralradius haben. Die nächste Berechnung vergleicht eine hpd Approxi-

mation M_s mit einer nicht hermiteschen Approximation. Dazu werden die erzielten Ergebnisse bei der Verwendung von $M_s^{[2]}$ mit denen durch $M_s^{[4]} := \text{triu}(A_{11})$, was der oberen Dreiecksmatrix von A_{11} entspricht, verglichen.

	φ	n_c	Glätter	ρ			
				$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 4$
16×16	0.65	126	$M_s^{[2]}$	.37	.16	.07	.04
			$M_s^{[4]}$	.17	.07	.06	.06
	0.6	126	$M_s^{[2]}$	.37	.28	.13	.04
			$M_s^{[4]}$	.17	.07	.06	.06
	0.55	98	$M_s^{[2]}$	.70	.50	.37	.27
			$M_s^{[4]}$	.56	.44	.39	.37
	0.51	98	$M_s^{[2]}$	.70	.86	.73	.58
			$M_s^{[4]}$	.56	.44	.39	.37
32×32	0.65	126	$M_s^{[2]}$	.37	.17	.07	.04
			$M_s^{[4]}$	.17	.07	.06	.06
	0.6	126	$M_s^{[2]}$	.38	.29	.13	.04
			$M_s^{[4]}$	.17	.07	.06	.06
	0.55	98	$M_s^{[2]}$	.70	.50	.37	.28
			$M_s^{[4]}$	.59	.45	.40	.38
	0.51	98	$M_s^{[2]}$	.70	.86	.73	.58
			$M_s^{[4]}$	.59	.45	.40	.38

Tabelle 9.4.: Vergleich nicht hermitescher Approximation $M_s^{[4]}$ mit hermitescher Approximation $M_s^{[2]}$. Das AMGp-Verfahren wird auf die Poisson-Gleichung auf einem 16×16 bzw. 32×32 Gitter angewendet. Dabei ist $D_p = D_p^{[2]}$ und die Werte a, b werden approximiert.

Tabelle 9.4 zeigt die Verbesserung durch die Verwendung von $M_s^{[4]} = \text{triu}(A_{11})$ im Vergleich zu den Diagonalmatrizen im Glätter auf. Hervorzuheben ist die Performance für kleiner werdendes φ. Das polynombasierte AMG mit hermitescher Approximation M_s, welches auf dem Intervall $[a, b]$ arbeitet, hängt stark von der Approximation dieses Intervalls ab. Dies ist durch die Wahl von $(1-t)^\nu$ im nicht hermiteschen Fall anders.

Um die bisher erhaltenen Ergebnisse zu verifizieren, wird eine modifiziere Poisson-Gleichung

$$(\Delta - m)u(x, y) = f(x, y)$$

mit verschiedenen Massen m betrachtet. Hierbei wird m so gewählt, dass das jeweils resultierende lineare Gleichungssystem die Gestalt

$$(A - cI)x = b$$

hat und $A - cI$ den kleinsten Eigenwert 10^{-s}, $s = 2, 3, 4, 5$ besitzt. A entspricht der zuvor betrachteten 2D Laplace-Matrix.

In Tabelle 9.5 wird die Performance des AMGp-Verfahrens für verschiedene Systemmatrizen mit kleiner werdenden Eigenwerten untersucht. Als Approximationen an A_{11} in der Interpolation wird $D_p = \text{diag}(A_{11})$ gewählt und das Grobgittersystem mit $A_c = P^H AP$ wird auf dem zweiten Level exakt gelöst. Für die Approximation M_s an A_{11} wird erneut zwischen der hermiteschen und der nicht hermiteschen Approximation unterschieden.

$\lambda_{\min}$	φ	Glätter	ρ			
			$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 4$
$1e-2$	0.65	$M_s^{[2]}$	.39	.17	.08	.06
		$M_s^{[4]}$	.19	.09	.07	.07
	0.60	$M_s^{[2]}$	.39	.29	.14	.04
		$M_s^{[4]}$	.19	.09	.07	.07
	0.55	$M_s^{[2]}$	.86	.70	.69	.67
		$M_s^{[4]}$	.80	.74	.72	.71
$1e-3$	0.65	$M_s^{[2]}$	.43	.17	.15	.14
		$M_s^{[4]}$	.26	.16	.15	.15
	0.60	$M_s^{[2]}$	.43	.29	.18	.12
		$M_s^{[4]}$	.26	.16	.15	.15
	0.55	$M_s^{[2]}$	.98	.94	.96	.93
		$M_s^{[4]}$	.97	.96	.96	.96
$1e-4$	0.65	$M_s^{[2]}$	.66	.53	.53	.53
		$M_s^{[4]}$	.59	.54	.54	.53
	0.60	$M_s^{[2]}$	.66	.48	.53	.52
		$M_s^{[4]}$	.59	.54	.54	.53
	0.55	$M_s^{[2]}$	.99	.99	.99	.99
		$M_s^{[4]}$	.99	.99	.99	.99
$1e-5$	0.65	$M_s^{[2]}$	.94	.91	.91	.91
		$M_s^{[4]}$	.93	.92	.92	.92
	0.60	$M_s^{[2]}$	.94	.91	.91	.91
		$M_s^{[4]}$	.93	.92	.92	.92
	0.55	$M_s^{[2]}$	.99	.99	.99	.99
		$M_s^{[4]}$	.99	.99	.99	.99

Tabelle 9.5.: Das AMGp-Verfahren, angewendet auf die verschobene Poisson-Gleichung auf einem 16×16 Gitter. Hierbei werden die Ergebnisse bei der Verwendung einer hermiteschen ($M_s^{[2]}$) und einer nicht hermiteschen ($M_s^{[4]}$) Approximation verglichen. Außerdem ist $D_p = D_p^{[2]}$ und die Werte a, b werden approximiert.

In Tabelle 9.5 ist zu erkennen, dass eine gute Performance für ein fast singuläres System nur schwer zu sichern ist. Insbesondere wird die „schlechte“ Wahl der Interpolationsmatrix und damit auch der Grobgittermatrix im Vergleich zu der Performance des Glätters dominant. Weitere Glättungsschritte haben keinen weiteren Effekt.

Wie in Kapitel 6.4 festgestellt wurde, kann ein adaptives Verfahren Abhilfe schaffen. In diesem Abschnitt wird eine Abwandlung dieses Verfahrens verwendet. Dazu sei bemerkt, dass die hier betrachtete Matrix eine M-Matrix ist. D_p wird so gewählt, dass D_p diagonal und $D_p w_1 = A_{11} w_1$ erfüllt ist, wobei $w = \left[\, w_1^T \; w_2^T \,\right]^T > 0$ der Perron-Vektor ist, also $Aw > 0$ gilt.

$\lambda_{\min}$	φ	n_c	Glätter	ρ			
				$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 4$
$1e-4$	0.6	98	$M_s^{[2]}$	.61	.52	.59	.56
			$M_s^{[4]}$	.57	.57	.57	.57
$1e-5$	0.6	98	$M_s^{[2]}$	.93	.92	.93	.92
			$M_s^{[4]}$	.93	.93	.93	.93

Tabelle 9.6.: Das AMGp-Verfahren, angewendet auf die verschobene Poisson-Gleichung auf einem 16×16 Gitter, wobei D_p die Gleichung $D_p w_1 = A_{11} w_1$ erfüllt. Außerdem wird eine hermitesche ($M_s^{[2]}$) mit einer nicht hermiteschen Approximation ($M_s^{[4]}$) verglichen. Die Werte a, b werden approximiert.

In Tabelle 9.6 ist die numerische Berechnung aus Tabelle 9.5 auf dem 16×16 Gitter mithilfe des Eigenvektors D_p wiederholt worden. Die Matrix des betrachteten Systems besitzt den kleinsten Eigenwert 10^{-4} bzw. 10^{-5} und für den Greedy-Coarser wird $\varphi = 0.6$ gesetzt. Der adaptive Prozess bringt in diesem Beispiel keine Verbesserung, was an der schlechten Wahl von D_p liegt, denn es gilt

$$\alpha = (\lambda_{\max}(D_p^{-1}A_{11}))^{-1} = 3 \cdot 10^{-6},$$

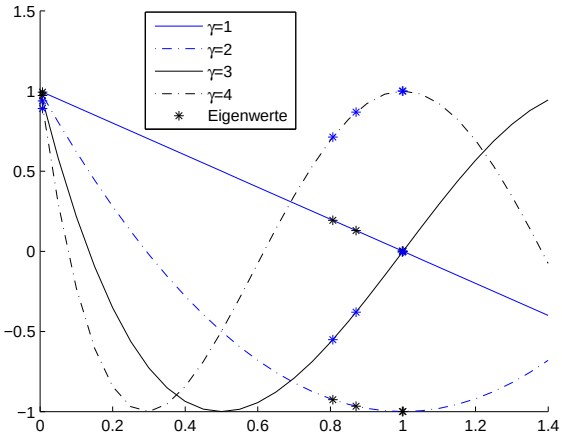
was zu einer schlechten Abschätzung der A-Norm von $I - B_{\mathrm{amg}_p}^{-1}A$ führt, siehe Satz 7.11.

In Abschnitt 9.2 wird die Variante des adaptiven Verfahrens aus Abschnitt 6.4, Algorithmus 7, verwendet, wodurch eine starke Verbesserung eintritt. Doch auch in dem hier besprochenen Fall kann die Performance durch eine kleine Manipulation verbessert werden. Da durch die Wahl von D_p die Voraussetzungen von Satz 7.11 durch Anwendung von Satz 6.10 erfüllt sind, können der Spektralradius von $I - B_{\mathrm{amg}_p}^{-1}A$ und die Eigenwerte von $B_{\mathrm{amg}_{p,s}}^{-1}A$, siehe Korollar 7.15, abgeschätzt werden. Im Folgenden wird das symmetrisierte polynombasierte AMG betrachtet. Durch die Verwendung eines auf 1 normierten Tschebysheff Polynoms P_γ auf $[\lambda_{\min}, \lambda_{\max}]$, wobei $[\lambda_{\min}, \lambda_{\max}]$ die Eigenwerte von $B_{\mathrm{amg}_{p,s}}^{-1}A$ einschließt, kann das symmetrische AMGp-Verfahren beschleunigt werden. Diese Beschleunigung entspricht derjenigen, die in (3.9) beschrieben, und deren Wirkung in Korollar 5.15 untersucht wurde. Die Ergebnisse werden mit denen bei der Verwendung des γ-Zyklus', also $\widetilde{P}_\gamma(t) = (1-t)^\gamma$, was einer γ-fachen Anwendung des Verfahrens entspricht, verglichen. Außerdem wird ein drittes Polynom, was eine Kombination der beiden darstellt, verwendet. Dieses Polynom $\widehat{P}_\gamma$ entspricht der Multiplikation des Tschebysheff Polynoms vom Grad $\gamma - 1$ mit $1 - t$.

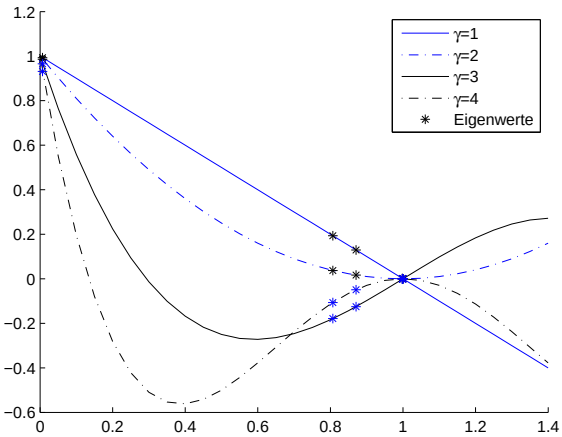
$\lambda_{\min}$	Polynom	γ	ρ			
			$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 4$
$1e{-}5$	Tschebysheff	1	.93	.93	.93	.93
	γ-Zyklus		.93	.93	.93	.93
	Kombination		.93	.93	.93	.93
	Tschebysheff	2	.99	.99	.99	.99
	γ-Zyklus		.88	.86	.86	.86
	Kombination		.86	.86	.86	.86
	Tschebysheff	3	.55	.40	.40	.40
	γ-Zyklus		.89	.80	.80	.80
	Kombination		.70	.67	.67	.66
	Tschebysheff	4	.99	.99	.99	.99
	γ-Zyklus		.77	.74	.74	.73
	Kombination		.44	.38	.37	.37
$1e{-}6$	Tschebysheff	1	.99	.99	.99	.99
	γ-Zyklus		.99	.99	.99	.99
	Kombination		.99	.99	.99	.99
	Tschebysheff	2	.99	.99	.99	.99
	γ-Zyklus		.99	.98	.98	.98
	Kombination		.99	.99	.99	.99
	Tschebysheff	3	.94	.93	.93	.93
	γ-Zyklus		.98	.98	.98	.98
	Kombination		.97	.96	.96	.96
	Tschebysheff	4	.99	.99	.99	.99
	γ-Zyklus		.97	.97	.97	.97
	Kombination		.93	.92	.92	.92

Tabelle 9.7.: Das symmetrische AMGp-Verfahren, angewendet auf die verschobene Poisson-Gleichung auf einem 16×16 Gitter. Dargestellt wird die Beschleunigung durch Verwendung von Tschebysheff Polynomen vom Grad γ. Außerdem wird $M_s = \mathrm{diag}(A_{11})$ gewählt. Die Werte a, b werden durch den Greedy-Coarser und $[\lambda_{\min}, \lambda_{\max}]$ durch Satz 7.15 approximiert. D_p erfüllt die Gleichung $D_p w_1 = A_{11} w_1$ und es ist $\varphi = 0.6$.

Aus den Ergebnissen aus Tabelle 9.7 wird deutlich, dass eine Verbesserung durch Tschebysheff Polynome möglich ist, was vom Grad des Polynoms und von den Eigenwertschranken abhängt. Außerdem fällt auf, dass der γ-Zyklus, was einer mehrmaligen Ausführung des Verfahrens entspricht, besser ist, als die Nutzung der optimalen Tschebysheff Polynome. Abbildung 9.2 zeigt die Graphen der Tschebysheff Polynome aus der obigen Berechnung für $\lambda_{\min} = 10^{-5}$ und verdeutlicht die Problematik bei der Verwendung dieser. Die Eigenwerte von $P_\gamma(B^{-1}_{amg_{p,s}}A)$ sollen möglichst nahe bei Null liegen. Durch die starke Oszillation der Tschebysheff Polynome kann es passieren, dass die Bilder der Eigenwerte nahe der Eins liegen. Anders verhält es sich bei der Kombination von Tschebysheff Polynomen und dem γ-Zyklus.



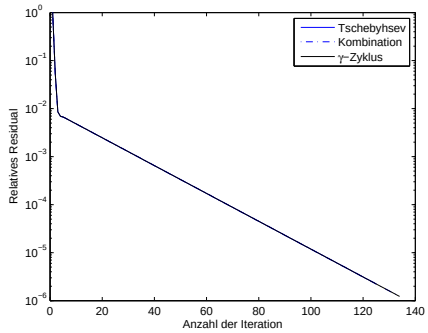
(a) Tschebysheff Polynome



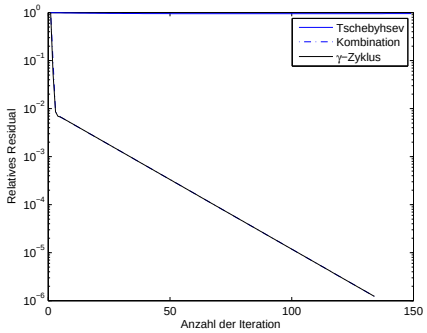
(b) Kombinierte Polynome

Abbildung 9.2.: Vergleich der Tschebysheff Polynome bzw. kombinierten Polynome vom Grad γ auf dem durch Satz 7.15 gegebenen Intervall. $*$ markiert jeweils das Bild der Eigenwerte von $B^{-1}_{\mathrm{amg}_{p,s}}A$.

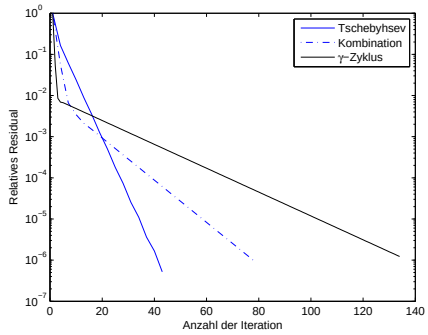
In der Grafik 9.3 wird die Konvergenzgeschwindigkeit unter Verwendung der drei Polynome dargestellt. Dazu wird die modifizierte Laplace-Matrix mit dem kleinsten Eigenwert $\lambda_{\min} = 10^{-5}$, die rechte Seite $b = Ae$ mit $e = [1 \ldots 1]^T$ und der Startvektor $x_0 = [0 \ldots 0]^T$ betrachtet. Die anderen Größen sind identisch zu dem Setting aus der Berechnung aus Tabelle 9.7.



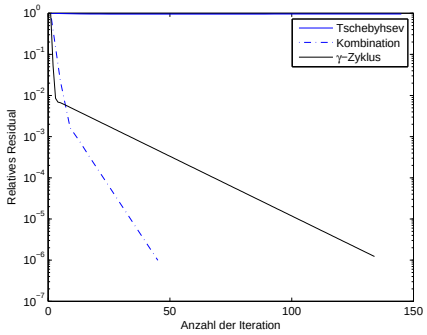
(a) Bei $\gamma = 1$ sind die drei Polynome identisch.



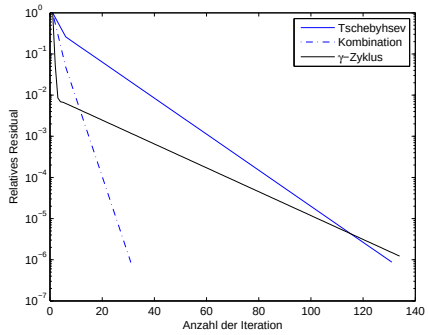
(b) Bei $\gamma = 2$ stimmen γ-Zyklus und das kombinierte Polynom noch überein.



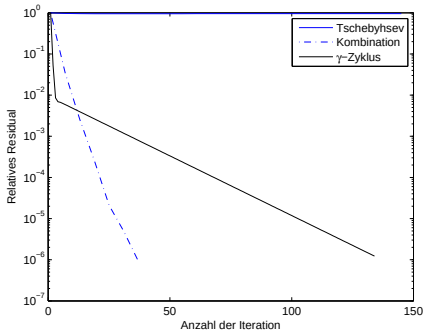
(c) Bei $\gamma = 3$ erzielt das Tschebysheff Polynom das beste Resultat.



(d) Bei $\gamma = 4$ erkennt man bereits den Vorteil der Kombination.



(e) $\gamma = 5$.



(f) $\gamma = 6$.

Abbildung 9.3.: Konvergenz für die drei Polynome. Man erkennt die gute Performance bei Verwendung der Kombination.

Die polynomielle Beschleunigung kann auf jedem Level angewendet werden und sichert somit eine gute Konvergenzrate für den Multilevelfall.

Anhand der Berechnungen für die Poisson-Gleichung wurde festgestellt, dass das polynombasierte AMG im Allgemeinen schnellere Konvergenz als das AMGr-Verfahren aufweist. Des Weiteren erlaubte die für das AMGp-Verfahren allgemeine Theorie die Verwendung und den Vergleich verschiedener und sogar nicht hermitescher Approximationen M_s.

Trotzdem hat das AMGp-Verfahren, angewandt auf ein fast singuläres Problem, seine Schwierigkeiten. Hier kann aber mithilfe des adaptiven Verfahrens und einer Beschleunigung durch Polynome Abhilfe geschaffen werden.

Dabei ist zu beachten, dass in diesem Abschnitt eine spezielle Struktur der Laplace-Matrix nicht ausgenutzt wurde. Die Matrix kann durch einen geeigneten Coarser (odd-even-Reduktion) auf die Hälfte ihrer Größe reduziert werden. Die gleiche Eigenschaft besitzt auch die Gauge-Laplace-Matrix, die im nächsten Abschnitt untersucht wird.

9.2

Gauge-Laplace. Eine größere Herausforderung als die Poisson-Gleichung stellt die Gauge-Laplace-Gleichung aus Abschnitt 2.2 dar. Bevor numerische Berechnungen durchgeführt werden, werden einige wichtige Vereinbarungen bezüglich der Gauge-Laplace-Matrix A wiedergegeben.

1. Die Matrix A wird in der Form $A = I - \kappa D$ mit $\kappa = \frac{1}{4+m}$ geschrieben.
2. Die Einträge der Matrix A hängen von der Temperatur $\beta \in (0, \infty)$ ab.
3. Eine Partitionierung der Matrix durch eine Aufteilung in gerade und ungerade Punkte im Gitter führt zu einer Reduktion der Größe des Systems. Teilt man dazu die Gitterpunkte in

   $$O := \Big\{(k,l) \,:\, k+l \text{ ungerade (odd)}\Big\}, \quad E := \Big\{(k,l) \,:\, k+l \text{ gerade (even)}\Big\}$$

   auf und ordnet man die ungeraden Variablen vor die geraden, dann ist A eine Permutation von

   $$A = \begin{bmatrix} I & A_{oe} \\ A_{eo} & I \end{bmatrix}$$

   mit dem Schurkomplement $S_{ee} := I - A_{eo}A_{oe}$. Im Folgenden wird dieses Schurkomplement als Systemmatrix betrachtet, also $A = S_{ee}$ und daher

   $$S_{ee} = A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

4. In den numerischen Berechnungen wird der Hopping-Parameter κ so variiert, dass die Matrix A hpd und fast singulär ist.

In der ersten Berechnungen zur Gauge-Laplace-Matrix wird $D_p = \mathrm{diag}(A_{11})$ und $A_c = P^H AP$ mit $P = \begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix}$ gesetzt. Zunächst wird der Zweilevelfall betrachtet und als Systemmatrix das Schurkomplement nach der oben angesprochenen Reduktion verwendet. Des Weiteren wird das polynombasierte AMG für hermitesches $M_s = \mathrm{diag}(A_{11})$ (mit approximierten

Schranken) und für nicht hermitesches $M_s = \mathrm{triu}(A_{11})$ betrachtet. Der Spektralradius wird abhängig von dem Greedy-Parameter φ, der Temperatur β, der Anzahl der Glätter ν und dem kleinsten Eigenwert $\lambda_{\min}(A)$ der Matrix vor der odd-even-Reduktion bestimmt. Zum Ende dieses Abschnittes wird das AMGp-Verfahren als Vorkonditionierer des CG-Verfahrens verwendet. Daher werden die Berechnungen alle anhand des symmetrisierten polynombasierten AMGs durchgeführt.

$\lambda_{\min}$	φ	n_c	Glätter	ρ			
				$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 4$
1e-1	.70	250	$\mathrm{diag}(A_{11})$	.16	.06	.06	.06
			$\mathrm{triu}(A_{11})$	.11	.06	.06	.06
	.65	199	$\mathrm{diag}(A_{11})$	.23	.09	.09	.09
			$\mathrm{triu}(A_{11})$	.17	.09	.09	.09
	.60	135	$\mathrm{diag}(A_{11})$	35	.13	.13	.13
			$\mathrm{triu}(A_{11})$	.27	.14	.13	.13
	.55	83	$\mathrm{diag}(A_{11})$	.44	.25	.16	.14
			$\mathrm{triu}(A_{11})$	.32	.15	.13	.12
1e-2	.70	280	$\mathrm{diag}(A_{11})$	.52	.47	.47	.47
			$\mathrm{triu}(A_{11})$	.49	.47	.47	.47
	.65	258	$\mathrm{diag}(A_{11})$	.53	.48	.48	.48
			$\mathrm{triu}(A_{11})$	.50	.48	.48	.44
	.60	204	$\mathrm{diag}(A_{11})$	.77	.70	.70	.70
			$\mathrm{triu}(A_{11})$	.74	.70	.70	.70
	.55	146	$\mathrm{diag}(A_{11})$	.85	.77	.77	.77
			$\mathrm{triu}(A_{11})$	.82	.78	.77	.77
1e-3	.70	281	$\mathrm{diag}(A_{11})$	.91	.90	.90	.90
			$\mathrm{triu}(A_{11})$	.91	.90	.90	.90
	.65	262	$\mathrm{diag}(A_{11})$	.92	.91	.91	.91
			$\mathrm{triu}(A_{11})$	.91	.91	.91	.91
	.60	208	$\mathrm{diag}(A_{11})$	.97	.96	.96	.96
			$\mathrm{triu}(A_{11})$	.97	.96	.96	.96
	.55	157	$\mathrm{diag}(A_{11})$	.98	.97	.97	.97
			$\mathrm{triu}(A_{11})$	.98	.97	.97	.97
1e-4	.70	281	$\mathrm{diag}(A_{11})$	.99	.99	.99	.99
			$\mathrm{triu}(A_{11})$	.99	.99	.99	.99
	.65	263	$\mathrm{diag}(A_{11})$	.99	.99	.99	.99
			$\mathrm{triu}(A_{11})$	.99	.99	.99	.99
	.60	208	$\mathrm{diag}(A_{11})$	.99	.99	.99	.99
			$\mathrm{triu}(A_{11})$	.99	.99	.99	.99
	.55	158	$\mathrm{diag}(A_{11})$	.99	.99	.99	.99
			$\mathrm{triu}(A_{11})$	.99	.99	.99	.99

Tabelle 9.8.: Die Konvergenzrate des AMGp-Verfahrens, angewendet auf das Schurkomplement $A = S_{ee} \in \mathbb{C}^{512,512}$ der Gauge-Laplace-Matrix auf einem 32×32 Gitter für $\beta = 1$.

In Tabelle 9.8 erkennt man, dass das AMGp-Verfahren eine gute Performance vorweist, wenn der kleinste Eigenwert nicht zu nahe an der Null liegt.

In der folgenden Berechnung wird ein adaptiver Prozess verwendet, um die Matrix D_p zu konstruieren. D_p soll dabei diagonal sein und die Gleichung

$$\begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix} w_2 = \begin{bmatrix} -A_{11}^{-1}A_{12} \\ I \end{bmatrix} w_2$$

erfüllen, wobei $w = \begin{bmatrix} w_1^H & w_2^H \end{bmatrix}^H$ ein Eigenvektor zum kleinsten Eigenwert von A ist. Bei dieser Konstruktion ist zu beachten, dass D_p Einträge enthält, die nicht reell sind, also gilt $D_p \neq D_p^H$. Diese Konstruktion und auch einige der weiteren Ideen wurden in [33] dargestellt. Ein wesentlicher Unterschied besteht darin, dass in [33] der Gauß-Seidel-Glätter auf allen Unbekannten verwendet wird. Hier wird der Glätter nur auf F-Punkten definiert, außerdem wird gezeigt, dass eine Diagonalmatrix als Glätter ausreicht, und kein Gauß-Seidel-Verfahren für die Glättung verwendet werden muss. In dem in [68] und [33] eingeführten adaptiven Prozess muss ein

System mit A_{11} gelöst werden. In diesem Abschnitt wird gezeigt, dass dies nicht notwendig ist. Stattdessen kann, basierend auf Satz 7.9, eine Approximation von A_{11} verwendet werden. Das adaptive AMGp-Verfahren wird in Anlehnung an [68] als αAMGp bezeichnet.

$\lambda_{\min}$	φ	n_c	Glätter	ρ			
				$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 4$
1e-3	.70	281	diag(A_{11})	.49	.47	.47	.47
			triu(A_{11})	.48	.47	.47	.47
	.65	262	diag(A_{11})	.74	.73	.73	.73
			triu(A_{11})	.73	.73	.73	.73
	.60	208	diag(A_{11})	.77	.76	.75	.75
			triu(A_{11})	.75	.75	.75	.75
	.55	157	diag(A_{11})	.92	.93	.92	.92
			triu(A_{11})	.92	.92	.92	.92
1e-4	.70	281	diag(A_{11})	.50	.48	.48	.48
			triu(A_{11})	.49	.48	.48	.48
	.65	263	diag(A_{11})	.76	.74	.74	.74
			triu(A_{11})	.75	.74	.74	.74
	.60	208	diag(A_{11})	.78	.77	.76	.76
			triu(A_{11})	.77	.76	.76	.76
	.55	158	diag(A_{11})	.93	.94	.93	.92
			triu(A_{11})	.93	.92	.92	.92
1e-5	.70	281	diag(A_{11})	.50	.48	.48	.48
			triu(A_{11})	.49	.48	.48	.48
	.65	263	diag(A_{11})	.76	.75	.74	.74
			triu(A_{11})	.75	.74	.74	.74
	.60	208	diag(A_{11})	.78	.77	.76	.76
			triu(A_{11})	.77	.76	.76	.76
	.55	158	diag(A_{11})	.93	.94	.93	.92
			triu(A_{11})	.93	.92	.92	.92
1e-6	.70	281	diag(A_{11})	.50	.48	.48	.48
			triu(A_{11})	.49	.48	.48	.48
	.65	263	diag(A_{11})	.76	.75	.74	.74
			triu(A_{11})	.75	.74	.74	.74
	.60	208	diag(A_{11})	.78	.77	.76	.76
			triu(A_{11})	.77	.76	.76	.76
	.55	158	diag(A_{11})	.93	.94	.93	.92
			triu(A_{11})	.93	.92	.92	.92

Tabelle 9.9.: Die Konvergenzrate des αAMGp-Verfahrens, angewendet auf das Schurkomplement $A = S_{ee} \in \mathbb{C}^{512,512}$ der Gauge-Laplace-Matrix auf einem 32×32 Gitter für $\beta = 1$, wobei D_p mit Hilfe eines Eigenvektors bestimmt wird.

Mit Hilfe der Tabelle 9.9 wird ersichtlich, dass eine Matrix D_p konstruiert werden kann, so dass das AMGp Verfahren auch für fast singuläre Systeme schnell konvergiert. Um dieses D_p zu bestimmen, wurde ein Eigenvektor der Matrix A berechnet und ein System $A_{11}x_1 = b_1$ gelöst. Im Folgenden werden Möglichkeiten angegeben, diese beiden Prozesse zu umgehen.

Anstatt das System mit A_{11} zu lösen, kann dieses approximiert werden. Nun stellt sich die Frage, wie solch eine Approximation aussehen soll. Die Aufgabe besteht darin, eine Matrix H_p zu finden, so dass

$$\begin{bmatrix} -H_p^{-1}A_{12} \\ I \end{bmatrix} \approx \begin{bmatrix} -A_{11}^{-1}A_{12} \\ I \end{bmatrix}. \tag{9.1}$$

Die Matrix H_p muss aber im Gegensatz zu der Matrix D_p nicht notwendigerweise dünn besetzt sein, daher kann Satz 7.9 verwendet und H_p durch (7.19) definiert werden. Da diese Matrix von der Anzahl der Glätter abhängt, wird in der nächsten Berechnung die Anzahl der Glätter erhöht. Des Weiteren wird der numerisch günstigere Fall $M_s = \text{diag}(A_{11})$ betrachtet, da die Performance für $M_s = \text{triu}(A_{11})$ und $M_s = \text{diag}(A_{11})$ in etwa gleich ist.

$\lambda_{\min}$	φ	n_c	Glätter	ρ			
				$\nu = 4$	$\nu = 5$	$\nu = 6$	$\nu = 7$
1e-4	.70	281	diag(A_{11})	.48	.48	.48	.48
	.65	262	diag(A_{11})	.75	.74	.74	.74
	.60	208	diag(A_{11})	.78	.77	.76	.76
	.55	157	diag(A_{11})	.94	.94	.93	.93
1e-5	.70	281	diag(A_{11})	.49	.48	.48	.48
	.65	263	diag(A_{11})	.75	.74	.74	.74
	.60	208	diag(A_{11})	.78	.77	.76	.76
	.55	158	diag(A_{11})	.99	.94	.93	.93
1e-6	.70	281	diag(A_{11})	.48	.48	.48	.48
	.65	263	diag(A_{11})	.75	.74	.74	.74
	.60	208	diag(A_{11})	.97	.84	.76	.76
	.55	158	diag(A_{11})	.99	.99	.93	.93

Tabelle 9.10.: Die Konvergenzrate des αAMGp-Verfahrens, angewendet auf das Schurkomplement $A = S_{ee} \in \mathbb{C}^{512,512}$ der Gauge-Laplace-Matrix auf einem 32×32 Gitter für $\beta = 1$, wobei D_p mit Hilfe eines Eigenvektors und der Matrix H_p bestimmt wird.

In Tabelle 9.10 erkennt man, dass die Konstruktion von D_p mittels H_p qualitativ ähnliche Ergebnisse liefert, wie bei der Verwendung von A_{11}. Dabei wurde H_p durch

$$H_p = A_{11}\Big[I - P_\nu(M_s^{-1}A_{11})(I - \widetilde{D}_p^{-1}A_{11})\Big]^{-1} \tag{9.2}$$

bestimmt, wobei $\widetilde{D}_p = \mathrm{diag}(A_{11})$ gesetzt wurde und nicht der Matrix D_p, welche in einem AMG Schritt für die Interplationsmatrix verwendet wird, entspricht. Der Grad ν wurde mit der Anzahl der Glätter gleichgesetzt. Dies wird im Weiteren entkoppelt und es wird der Zweilevel V(2,2)-Zyklus betrachtet, also $\nu = 2$. Der Grad des Polynoms in (9.2) wird zur besseren Differenzierung mit ν_a bezeichnet.

Um eine möglichst leicht zu berechnende Approximation H_p zu bestimmen, wird $M_s = \widetilde{D}_p = \mathrm{diag}(A_{11})$ gesetzt. $-H_p^{-1}A_{12}w_2$ kann durch den folgenden Algorithmus günstig bestimmt werden. Dazu sei $Q_{\nu_a} \in \mathbb{R}_{\leq \nu_a}[t]$ durch

$$Q_{\nu_a}(t) := \tfrac{1-P_{\nu_a}(t)(1-t)}{t} := \sum_{k=0}^{\nu_a} q_k t^k$$

definiert, wobei P_{ν_a} durch (3.13) mit $a = 2 - \frac{1}{\varphi}$ und $b = \frac{1}{\varphi}$ (siehe Satz 6.6) gegeben ist.

Algorithmus 8: $w_1 = -H_p^{-1}A_{12}w_2 = -Q_{\nu_a}(M_s^{-1}A_{11})M_s^{-1}A_{12}w_2$

Input : $w_2, Q_{\nu_a}(t)$
Output : $w_1 = -H_p^{-1}A_{12}w_2$

1 $r = -A_{12}w_2$,
2 $d = q_{\nu_a}r$,
3 **for** $k = 0, \ldots, \nu_a - 1$ **do**
4 löse $M_s w_1 = d$,
5 setze $d = A_{11}w_1 + q_{\nu_a-1-k}r$,
6 löse $M_s d = w_1$.

In der folgenden Tabelle wird aufgezeigt, wie sich der Spektralradius in Abhängigkeit von ν_a ändert. Als Referenz wird das adaptive Verfahren mit der Verwendung von A_{11} gewählt.

Zudem wird das Gitter auf 64×64 vergrößert und die Systemmatrix so skaliert, dass $\lambda_{\min} = 10^{-6}$ ist.

		ρ					
		mit A_{11}^{-1}	mit H_p^{-1}				
φ	β		$\nu_a = 2$	$\nu_a = 4$	$\nu_a = 6$	$\nu_a = 8$	$\nu_a = 10$
.70	0.1	.76	.93	.76	.76	.76	.76
	0.2	.83	.84	.83	.83	.83	.83
	1	.97	.97	.97	.97	.97	.97
.65	0.1	.53	.98	.72	.53	.53	.53
	0.2	.84	.99	.84	.84	.84	.83
	1	..97	.97	.97	.97	.97	.97
0.60	0.1	.86	.99	.97	.86	.86	.86
	0.2	.87	.99	.97	.87	.87	.87
	1	.99	.99	.99	.99	.99	.99
0.55	0.1	.92	.99	.99	.98	92	.91
	0.2	.91	.99	.99	.98	.91	.91
	1	.98	.99	.99	.98	.98	.98

Tabelle 9.11.: Das Zweilevel V(2,2) αAMGp-Verfahren, angewandt auf das Schurkomplement $A = S_{ee} \in \mathbb{C}^{2048,2048}$ der Gauge-Laplace-Matrix auf einem 64×64 Gitter. Hier wird das original adaptive Verfahren mit dem hierarchisch motivierten verglichen. Der kleinste Eigenwert liegt bei 1e-6 .

In Tabelle 9.11 wird der Vorteil bei der Verwendung von H_p^{-1} deutlich. Hier muss kein System mit A_{11} gelöst werden und man benötigt lediglich ν_a Matrix-Vektor-Multiplikationen und ν_a Lösungen eines Gleichungssystems mit $M_s = \mathrm{diag}(A_{11})$, wie aus Algorithmus 8 deutlich wird. Trotz dieser numerisch günstigeren Variante sind die Ergebnisse ab $\nu_a = 6$ identisch.

Des Weiteren erkennt man in Tabelle 9.11, dass für größer werdendes β die asymptotische Konvergenzgeschwindigkeit abnimmt. Durch solch eine Wahl von β nimmt die Zufälligkeit in dem System zu, was zu einer größeren Unstrukturiertheit führt. Dadurch werden die glatten Vektoren nicht mehr ausreichend durch Kern-nahe Vektoren repräsentiert, siehe dazu [33].

Eine Möglichkeit, dieses Problem zu umgehen, wird ebenfalls in [33] vorgestellt. Anstatt den Eigenvektor zum kleinsten Eigenwert für die adaptive Konstruktion wie in (9.2) zu nehmen, wird stattdessen das homogene System $Ax = 0$ mit dem Gauß-Seidel-Verfahren, also

$$x^{[k+1]} = x^{[k]} - \big(\mathrm{triu}(A)\big)^{-1} A x^{[k]}, \quad k = 0, 1, 2, \ldots, k_{\max}$$

approximiert. Der Startvektor $x^{[0]}$ wird dabei als Prototyp bezeichnet.

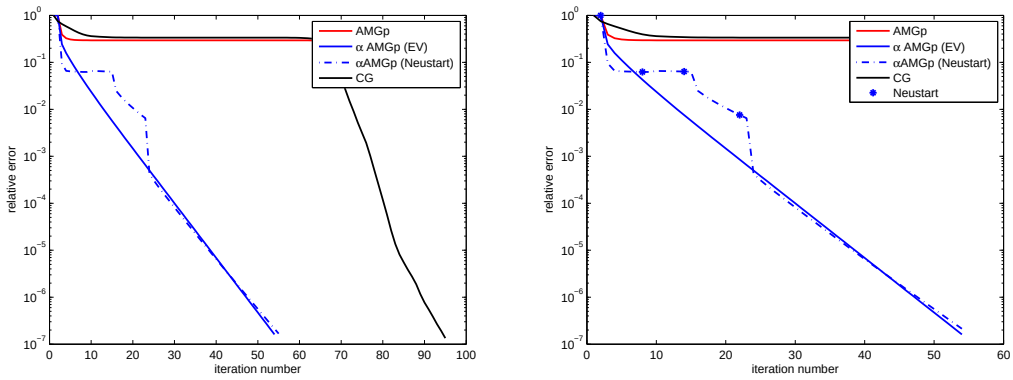
	$\rho(\beta)$					
	64×64			**128×128**		
$k_{\max}$	$\beta = 0.1$	$\beta = 0.2$	$\beta = 1$	$\beta = 0.1$	$\beta = 0.2$	$\beta = 1$
50	.99	.99	.99	.99	.99	.95
100	.99	.99	.98	.99	.99	.98
150	.97	.98	.90	.99	.99	.98
200	.98	.95	.82	.99	.99	.97
250	.98	.83	.89	.99	.95	.96
300	.98	.92	.80	.99	.99	.99
350	.95	.77	.83	.99	.90	.99
400	.89	.82	.93	.99	.98	.99
450	.67	.88	.92	.99	.93	.99
500	.79	.82	.95	.99	.90	.99

Tabelle 9.12.: Das Zweilevel V(2,2) αAMGp Verfahren, angewandt auf das Schurkomplement nach der odd-even-Reduktion der Gauge-Laplace-Matrix auf einem 64×64 und einem 128×128 Gitter. Der kleinste Eigenwert liegt bei 1e-6 und es ist $\varphi = .65$ sowie $\nu_a = 10$.

In der Tabelle 9.12 erkennt man, dass die Verwendung des Gauß-Seidel-Verfahrens zur Konstruktion von D_p eine Verkleinerung des Spektralradius' bewirkt, jedoch hängt dies stark von lokalen Größen des Problems ab. Die adaptive Konstruktion von D_p hat einen lokalen Charakter und weist daher global nicht die gleichen Eigenschaften wie A_{11} oder auch H_p auf. Mit den lokalen Eigenschaften lassen sich nichtlineare AMG's konstruieren, siehe dazu [33]. Dazu wird parallel zu dem System $Ax = b$ auch das System $Ax = 0$ gelöst und bei langsamer Konvergenzgeschwindigkeit des homogenen Systems wird $x^{[k]}_{\text{hom}}$ als Prototyp für den adaptiven Prozess genommen. Es wird AMGp mit $D_p = \text{diag}(A_{11})$ mit dem αAMGp verglichen. Bei dem adaptiven Verfahren wird zwischen der Konstruktion durch den Eigenvektor und dem nichtlinearen Verfahren durch einen Neustart differenziert. Das CG-Verfahren dient hier als Richtgröße, um zu verdeutlichen, dass mit größer werdender Matrix auch die Anzahl der Iterationen zunimmt. Die AMG-Verfahren werden in dieser numerischen Berechnung als V(2,2) Zyklus verwendet. Das Grobgittersystem wird exakt gelöst, sobald es kleiner als 10×10 ist. Als Startvektor wird $x^{[0]} = [1 \ldots 1]^H$, also Lösung $x = [x_j]$ mit $x_j = j$ und als rechte Seite $b = S_{ee}x$ gewählt. Die Iteration wird abgebrochen, wenn entweder $\|x^{[k]} - x\|_2 < 1e-4$ ist oder mehr als 1000 Iterationen benötigt werden. Der Neustart bei dem nichtlinearen Verfahren wird ausgelöst, sobald $\|x^{[k]} - x\|_2 - \|x^{[k+1]} - x\|_2 < \frac{1}{5} \cdot \|x^{[k]} - x\|_2$ ist, wobei mindestens fünf Iterationen vor jedem Neustart durchlaufen werden müssen.

β	Gitter	AMGp #It.	αAMGp mit EV #It.	αAMPp mit Neustart #It.	#Neustart	CG #It.
0.1	32×32	-	35	33	3	99
	64×64	-	170	44	6	194
	128×128	-	201	87	14	388
	256×256	-	-	68	7	446
1	32×32	-	53	54	4	94
	64×64	-	453	53	6	156
	128×128	-	-	63	8	178
	256×256	-	-	62	5	201

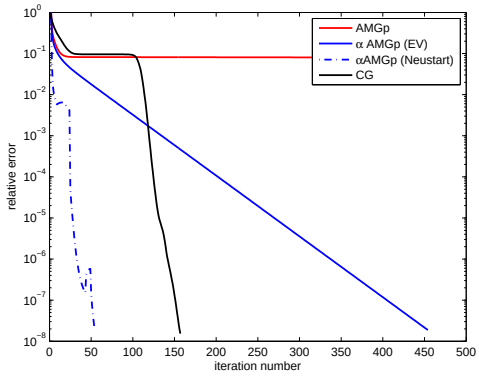
Tabelle 9.13.: Das V(2,2) αAMGp Verfahren, angewandt auf das Schurkomplement der Gauge-Laplace-Matrix auf verschiedenen Gittern für $\beta = 0.1$ und $\beta = 1$.



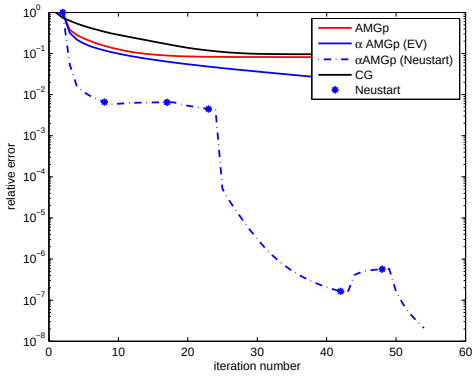
(a) 32×32 Gitter - kompletter Bereich

(b) 32×32 Gitter - Ausschnitt

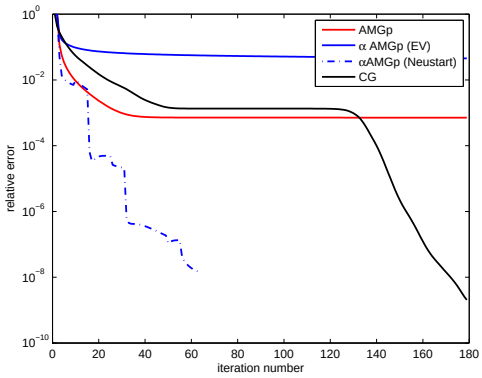
Abbildung 9.4.: Der Konvergenzverlauf von αAMGp. Auf der rechten Seite befindet sich jeweils ein Ausschnitt. $*$ markieren die Stellen, an denen neu gestartet wurde.



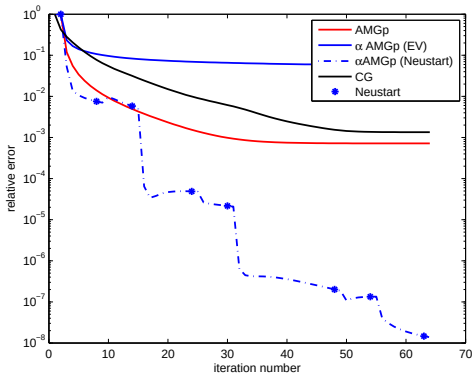
(c) 64×64 Gitter - kompletter Bereich



(d) 64×64 Gitter - Ausschnitt



(e) 128×128 Gitter - kompletter Bereich



(f) 128×128 Gitter - Ausschnitt

Abbildung 9.4.: Der Konvergenzverlauf von αAMGp. Auf der rechten Seite befindet sich jeweils ein Ausschnitt. $*$ markieren die Stellen, an denen neu gestartet wurde.

Zum Abschluss dieser Arbeit wird das AMGp-Verfahren als Vorkonditonierer für das CG-Verfahren verwendet. Dabei ist $D_p = M_s = \text{diag}(A_{11})$ und das AMGp-Verfahren wird als V(2,2) Mehrlevelverfahren verwendet. Der Greedy-Parameter wird als $\varphi = 0.55$ bzw. $\varphi = 0.65$ gewählt.

β	Verfahren	φ	#It. 32×32	64×64	128×128	256×256
0.1	AMGp	0.55	28	56	108	126
		0.65	22	39	75	86
	αAMGp	0.55	16	25	50	55
		0.65	8	11	19	35
	CG	-	99	194	388	446
1	AMGp	0.55	25	41	47	54
		0.65	17	27	30	34
	αAMGp	0.55	20	42	97	179
		0.65	11	21	55	91
	CG	-	94	156	178	201

Tabelle 9.14.: V(2,2) AMGp Verfahren angewandt als Vorkonditionierer für das CG-Verfahren auf das Schurkomplement S der Gauge-Laplace-Matrix auf verschiedenen Gittern für $\beta = 0.1$ und $\beta = 1$.

Man erkennt, dass die symmetrischen AMGp und αAMGp-Verfahren als Vorkonditionierer geeignet sind und die Anzahl der Iterationen deutlich senken. Auf die Frage des numerischen Aufwandes wird in dieser Arbeit nicht näher eingegangen, es sei dazu auf [67] verwiesen.

In diesem Abschnitt wurden die Ideen aus [33] aufgegriffen und so ein adaptives polynombasiertes AMG (αAMGp) konstruiert. Zusammenfassend kann festgehalten werden, dass eine F-Glättung qualitativ ähnliche Ergebnisse liefert, wie die Verwendung eines vollen Gauß-Seidel-Glätters, vergleiche dazu [33]. Eine Neuerung stellt die Approximation von A_{11} durch H_p, siehe (9.1) und (9.2), dar. Dies ermöglicht es, den adaptiven Prozess durchzuführen, ohne ein lineares Gleichungssystem mit A_{11} zu lösen.

Kapitel 10

Zusammenfassung

Diese Arbeit befasst sich mit zwei Varianten algebraischer Mehrgitterverfahren, dem Multilevel-Block-Faktorisierungs-Verfahren und dem polynombasierten AMG, einer Verallgemeinerung des von MacLachlan, Manteuffel und McCormick eingeführten AMGr-Verfahrens. Beide Verfahren können mithilfe generalisierter hierarchischer Basen umfassend analysiert werden.

Eines der wichtigsten Instrumente ist dabei durch die C.B.S. Konstante gegeben, die in einem direkten Zusammenhang zur spektralen Äquivalenz zwischen der Grobgittermatrix A_c und dem Schurkomplement $S(A, A_{11})$ steht, siehe Lemma 6.4. Daher wird eine Abschätzung dieser Konstante für die Analyse von MBF- aber auch anderen algebraischen Mehrgitterverfahren benötigt. Bis zum Zeitpunkt dieser Arbeit scheint das einzig anwendbare Resultat in diesem Zusammenhang Satz 6.3 von Notay zu sein. In dieser Arbeit wird mit Satz 6.11 nicht nur eine Abschätzung der C.B.S. Konstante unter anderen Voraussetzungen vorgestellt, sondern auch gezeigt, dass Satz 6.11 die Aussage von Notay verallgemeinert. Des Weiteren wird gezeigt, dass die Voraussetzungen von Satz 6.11 immer erfüllbar sind, sobald die C.B.S. Konstante von A kleiner als $\sqrt{\frac{4}{5}}$ ist, was wiederum bei vielen Anwendungen unter der Verwendung von einer sogenannten hierarchischen finiten Elemente Basis der Fall ist [18, 63, 83]. Des Weiteren kann für generelle elliptische 2D Probleme mit stückweisen linearen Basisfunktionen gezeigt werden, dass die C.B.S. Konstante von A kleiner als $\sqrt{\frac{3}{4}}$ ist [5, 72]. In [73] wurde dieselbe Schranke für 2D Elastizitäts-Probleme gefunden, siehe dazu [6].

In Abschnitt 6.6 wird gezeigt, dass eine k-fache rekursive Anwendung von generalisierten hierarchischen Basen auf ein spezielles polynombasiertes MBF-Verfahren führt, mit dessen Hilfe polynombasierte AMG's eingeführt (Definition 7.6) und analysiert (Satz 7.9) werden können. Es stellt sich heraus, dass das polynombasierte AMG bei der Verwendung von Tschebysheff Polynomen eine Verallgemeinerung des AMGr-Verfahrens von MacLachlan, Manteuffel und McCormick darstellt. In Satz 7.11 wird gezeigt, dass das AMGp-Verfahren unter den gleichen Voraussetzungen wie das AMGr-Verfahren konvergiert und eine bessere Konvergenzschranke mit sich bringt. Diese Analyse führt des Weiteren zu einer neuen Interpretation des AMGr-Verfahrens mittels Tschebysheff Polynomen und der C.B.S. Konstante.

Ein weiterer Bestandteil ist die Analyse, unter welchen Voraussetzungen die Interpolationsmatrix $P = \begin{bmatrix} -D_p^{-1}A_{12} \\ I \end{bmatrix}$ eine gute Approximation an die optimale Matrix P_{opt} aus (4.5) darstellt. Obwohl diese Voraussetzungen für eine Matrix $A \in \mathbb{C}^{n,n}$ im Allgemeinen nur schwer mit einer dünn besetzten Matrix D_p erfüllt werden können, kann gezeigt werden, dass für diagonaldominante Matrizen (dies wurde erstmals in [70] nachgewiesen) und für M-Matrizen (Satz 6.10) solch eine Matrix D_p angegeben werden kann. Für die anderen Fälle wird mit Satz 7.9 eine Möglichkeit vorgestellt, eine Interpolationsmatrix P mittels der Matrix H_p aus (7.19) zu

konstruieren, so dass die Konditionszahl des AMGp-Verfahrens abgeschätzt werden kann. Die Grobittermatrix ist dann durch $A_c = P^H AP$ bzw. (7.13) gegeben.

In Kapitel 9 wird diese Interpolation verwendet, um die Wirkung von P_{opt} auf einen Vektor w_2 mittels einer Approximation $H_p \approx A_{11}$ und Algorithmus 8 darzustellen, ohne ein System mit A_{11} lösen zu müssen. Dies führt auf die Erstellung eines adaptiven AMGp-Verfahrens (αAMGp). Dafür wird D_p nicht länger als globale Approximation von A_{11} betrachtet, sondern es genügt, dass $-D_p^{-1}A_{12}w_2 = -A_{11}^{-1}A_{12}w_2 = -H_p^{-1}A_{12}w_2$ für einen geeigneten Vektor w_2 erfüllt ist. Resultierend aus dem adaptiven Prozess ist die Matrix D_p nicht länger hpd, jedoch wird mit Lemma 7.14 ein Hilfsmittel für diesen Fall angegeben.

Zusammenfassend stellen die erzielten Ergebnisse dieser Arbeit eine ausführliche Analyse algebraischer Mehrgitterverfahren mit F-Glättung mit dem Fokus auf dem Block-Faktorisierungs- und dem AMGr-Verfahren dar. Diese Verfahren wurden in den letzten zehn Jahren separat betrachtet und untersucht, wobei die wichtigsten Ergebnisse von Notay [81] (2005) für das MBF-Verfahren und MacLachlan, Manteuffel und McCormick [68] (2006) für das AMGr-Verfahren erzielt wurden. Die Ergebnisse dieser Arbeit verallgemeinern beide Resultate, zeigen neue Interpretationsmöglichkeiten und stellen die Zusammenhänge der beiden Verfahren dar.

Kapitel A

Eine allgemeine Abschätzung der C.B.S. Konstante

In diesem Abschnitt werden zwei Verallgemeinerungen der Sätze 6.3 sowie 6.11 angegeben. Zum einen kann durch eine geeignete Skalierung auf die Bedingung $H \preceq A_{11}$ bzw. $H \succeq A_{11}$ verzichtet werden.

Satz A.1. *Sei A hpd und partitioniert wie in* (4.3). *Weiter sei $\widehat{A}$ gegeben durch* (6.2) *mit $H \in \mathbb{R}^{n_1,n_1}$ hpd, so dass $\sigma(H^{-1}A_{11}) \subseteq [a,b]$ für $b > \frac{1}{a}$ und für*

$$\overline{\delta} := 1 - \tfrac{(1-a)(b-1)}{a+b-2}$$

die Matrix

$$\begin{bmatrix} \overline{\delta}H & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

hpsd ist. Dann erfüllt die C.B.S. Konstante von $\widehat{A}$ die Abschätzung

$$\gamma(\widehat{A}) \leq \sqrt{1 - \tfrac{1}{a \cdot b}} < 1.$$

Beweis. Mit (5.5) folgt

$$1 - \gamma(\widehat{A})^2 = \min_{x_2 \neq 0} \frac{x_2^H S(\widehat{A}, \widehat{A}_{11}) x_2}{x_2^H \widehat{A}_{22} x_2},$$

wobei

$$S(\widehat{A}, \widehat{A}_{11}) = S(A, A_{11}) = A_{22} - A_{21}A_{11}^{-1}A_{12} \text{ und } \widehat{A}_{22} = A_{22} - A_{21}(2H^{-1} - H^{-1}A_{11}H^{-1})A_{12}$$

sind.

Es ist zu zeigen, dass $x_2^H(S(\widehat{A}, \widehat{A}_{11}) - \frac{1}{a \cdot b}\widehat{A}_{22})x_2 \geq 0$ für alle $x_2 \in \mathbb{C}^{n_2}$ erfüllt ist. Dies führt schließlich auf die Abschätzung

$$1 - \gamma(\widehat{A})^2 = \min_{x_2 \neq 0} \frac{x_2^H S(\widehat{A}, \widehat{A}_{11}) x_2}{x_2^H \widehat{A}_{22} x_2} \geq \tfrac{1}{a \cdot b}$$

und damit auf

$$\gamma(\widehat{A})^2 \leq 1 - \tfrac{1}{a \cdot b}.$$

Mit der Annahme $b > \frac{1}{a}$ ist $1 - \frac{1}{a \cdot b} > 0$.

Durch die Abschätzung

$$a+b-2 > a+\tfrac{1}{a}-2 = \tfrac{1}{2}(a^2-2a+1) = \tfrac{(1-a)^2}{a} > 0$$

folgt

$$\overline{\delta} = 1 - \tfrac{(1-a)(b-1)}{a+b-2} = \tfrac{a+b-2+1-a-b+ab}{a+b-2} = \tfrac{ab-1}{a+b-2} > 0.$$

Folglich ist die Matrix $\overline{\delta}H$ hpd und mit der Voraussetzung, dass

$$\begin{bmatrix} \overline{\delta}H & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

hpsd ist, folgt für dessen Schurkomplement $A_{22} - \overline{\delta}^{-1} A_{21} H^{-1} A_{12} \succeq 0$ mit $\overline{\delta}^{-1} = \frac{a+b-2}{a\cdot b-1}$. Daher ist für $x_2 \in \mathbb{C}^{n_2}$:

$$(1-\tfrac{1}{a\cdot b})x_2^H A_{22} x_2 = \tfrac{a\cdot b-1}{a\cdot b} x_2^H A_{22} x_2 \geq \tfrac{a+b-2}{a\cdot b} x_2^H A_{21} H^{-1} A_{12} x_2. \tag{A.1}$$

Betrachte für alle $x_2 \in \mathbb{C}^{n_2}$

$$\begin{aligned}
& x_2^H (S(\widehat{A}, \widehat{A}_{11}) - \frac{1}{a\cdot b}\widehat{A}_{22}) x_2 \\
= \; & x_2^H \left(A_{22} - A_{21} A_{11}^{-1} A_{12} - \tfrac{1}{a\cdot b}\Big(A_{22} - A_{21}(2H^{-1} - H^{-1}A_{11}H^{-1})A_{12} \Big) \right) x_2 \\
= \; & x_2^H \left((1-\tfrac{1}{a\cdot b})A_{22} - A_{21}A_{11}^{-1}A_{12} + \tfrac{1}{a\cdot b} A_{21}(2H^{-1} - H^{-1}A_{11}H^{-1})A_{12} \right) x_2 \\
\overset{(A.1)}{\geq} \; & x_2^H \left(\tfrac{a+b-2}{a\cdot b} A_{21}H^{-1}A_{12} - A_{21}A_{11}^{-1}A_{12} + \tfrac{1}{a\cdot b} A_{21}(2H^{-1} - H^{-1}A_{11}H^{-1})A_{12} \right) x_2 \\
= \; & x_2^H \left(\tfrac{a+b}{a\cdot b} A_{21}H^{-1}A_{12} - A_{21}A_{11}^{-1}A_{12} - \tfrac{1}{a\cdot b} A_{21}H^{-1}A_{11}H^{-1}A_{12} \right) x_2 \\
= \; & x_2^H A_{21} A_{11}^{-\frac{1}{2}} \left(\tfrac{a+b}{a\cdot b} A_{11}^{\frac{1}{2}} H^{-1} A_{11}^{\frac{1}{2}} - I - \tfrac{1}{a\cdot b} A_{11}^{\frac{1}{2}} H^{-1} A_{11} H^{-1} A_{11}^{\frac{1}{2}} \right) A_{11}^{-\frac{1}{2}} A_{21} x_2.
\end{aligned}$$

Der letzte Term ist nicht negativ, wenn das Polynom $p(t) = \frac{a+b}{a\cdot b}t - 1 - \frac{1}{a\cdot b}t^2$ nicht negativ auf dem Spektrum von $\sigma(H^{-1}A_{11}) \subseteq [a,b]$ ist. Es gilt

$$p(t) = (\tfrac{1}{a}+\tfrac{1}{b})t - 1 - \tfrac{1}{a\cdot b}t^2 = (\tfrac{t}{a}-1)(1-\tfrac{t}{b})$$

und damit $p(t) \geq 0$ für $t \in [a,b]$. □

Die zweite Verallgemeinerung bezieht sich auf Satz 6.11. Man kann zeigen, dass die C.B.S. Konstante auch unter schwächeren Voraussetzungen abgeschätzt werden kann, was eine Verschlechterung der Schranke mit sich bringt. Die Voraussetzung von Satz 6.11 ist, dass $H \succeq A_{11}$ und $\begin{bmatrix} H(H+HA_{11}^{-1}H-A_{11})^{-1}H & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \succeq 0$. Die zweite Bedingung ist äquivalent zu

$$A_{22} - A_{21}(H^{-1} + A_{11}^{-1} - H^{-1}A_{11}H^{-1})A_{12} \succeq 0.$$

Unter Umständen ist dies für die gegebene Matrix nur schwer mit einer dünn besetzten Matrix H zu erfüllen. Es gilt aber in jedem Fall, dass

$$A_{22} - A_{21}A_{11}^{-1}A_{12} \succeq 0.$$

Diese beiden Bedingungen werden zu der folgenden kombiniert

$$\begin{aligned}
&s\Big(A_{22} - A_{21}(H^{-1} + A_{11}^{-1} - H^{-1}A_{11}H^{-1})A_{12}\Big) + (1-s)\Big(A_{22} - A_{21}A_{11}^{-1}A_{12}\Big)\\
&= A_{22} - A_{21}\Big(s(H^{-1} + A_{11}^{-1} - H^{-1}A_{11}H^{-1}) + (1-s)A_{11}^{-1}\Big)A_{12}\\
&= A_{22} - A_{21}\Big(s(H^{-1} - H^{-1}A_{11}H^{-1}) + A_{11}^{-1}\Big)A_{12}.
\end{aligned}$$

Hierbei ist $s \in [0,1]$ ein Parameter zur Steuerung der Voraussetzung. Damit kann folgendes Resultat gezeigt werden.

Satz A.2. *Sei A hpd und partitioniert wie in* (4.3)*. Weiter sei $\widehat{A}$ definiert wie* (6.2) *mit H hpd, so dass*

1. *$A_{11} \preceq H \preceq \beta A_{11}$ und*
2. *die Matrix $\begin{bmatrix} H\big(s(H-A_{11})+HA_{11}^{-1}H\big)^{-1}H & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ hpsd ist.*

Dann erfüllt die C.B.S. Konstante von $\widehat{A}$ die Abschätzung:

$$\gamma(\widehat{A}) \leq \sqrt{1 - \tfrac{s}{s+\beta-1}}. \tag{A.2}$$

Beweis. Der Beweis verläuft analog zum Beweis des Satzes 6.11. Die Voraussetzung 2 ist äquivalent zu

$$x_2^H A_{22} x_2 \geq x_2^H A_{21}\big(s(H^{-1} - H^{-1}A_{11}H^{-1}) + A_{11}\big)A_{12}x_2 \quad \text{für alle } x_2 \in \mathbb{C}^{n_2}. \tag{A.3}$$

Sei $c = \frac{\beta-1}{s}$, dann gilt für $x_2 \in \mathbb{C}^{n_2}$

$$\begin{aligned}
&(1+c)x_2^H S(\widehat{A}, \widehat{A}_{11})x_2 - x_2^H \widehat{A}_{22}x_2\\
&= (1+c)x_2^H(A_{22} - A_{21}A_{11}^{-1}A_{12})x_2 - x_2^H(A_{22} - A_{21}(2H^{-1} - H^{-1}A_{11}H^{-1})A_{12})x_2\\
&\overset{(A.3)}{\geq} cx_2^H A_{21}(s(H^{-1} - H^{-1}A_{11}H^{-1}) + A_{11})A_{12}x_2\\
&\quad + x_2^H A_{21}(2H^{-1} - (1+c)A_{11}^{-1} - H^{-1}A_{11}H^{-1})A_{12}x_2\\
&= x_2^H A_{21}A_{11}^{-\frac{1}{2}}\left((2+cs)A_{11}^{\frac{1}{2}}H^{-1}A_{11}^{\frac{1}{2}} - I - (1+cs)A_{11}^{\frac{1}{2}}H^{-1}A_{11}H^{-1}A_{11}^{\frac{1}{2}}\right)A_{11}^{-\frac{1}{2}}A_{12}x_2.
\end{aligned}$$

Dieser Ausdruck ist nicht negativ, wenn das Polynom

$$p(t) := (2+cs)t - 1 - (1+cs)t^2$$

auf $\sigma(H^{-1}A_{11}) \subseteq [\frac{1}{\beta}, 1]$ nicht negativ ist. Dies ist erfüllt, da

$$p(t) = -\Big(1+cs\Big)\Big(t^2 - \tfrac{2+cs}{1+cs}t + \tfrac{1}{1+cs}\Big) = -\Big(1+cs\Big)\Big(t-1\Big)\Big(t - \tfrac{1}{1+cs}\Big),$$

auf $\Big[\frac{1}{1+cs}, 1\Big] = \Big[\frac{1}{\beta}, 1\Big]$ nicht negativ ist.

Damit ergibt sich

$$\tfrac{1}{1-\gamma(\widehat{A})^2} \leq 1 + c = \tfrac{s+\beta-1}{s}$$

und

$$\gamma(\widehat{A})^2 \leq 1 - \tfrac{s}{s+\beta-1}.$$

□

Mithilfe von Satz A.2 kann Satz 6.13 ebenfalls verallgemeinert werden.

Satz A.3. *Sei A hpd und partitioniert wie in* (4.3)*. Weiter sei $\gamma(A)$ die C.B.S. Konstante von A und H eine Approximation A_{11} mit $H \succeq A_{11}$. Falls eine der folgenden Bedingungen*

1. $\gamma(A) \leq \sqrt{\frac{4}{4+s}}$,
2. $\lambda_{\min}(H^{-1}A_{11}) \geq \frac{\sqrt{(4+s)\gamma(A)^2-4}+\sqrt{s}\gamma(A)}{2\sqrt{s}\gamma(A)}$ *für* $\gamma(A) > \sqrt{\frac{4}{4+s}}$

erfüllt ist, erfüllt H die Voraussetzung 2 von SatzA.2.

Beweis. Analog zum Beweis von Satz 6.13 kann eine Bedingung angegeben werden, so dass das Schurkomplement hpsd ist. Betrachtet man dazu

$$\begin{aligned} & x_2^H(A_{22} - A_{21}(s(H^{-1} - H^{-1}A_{11}H^{-1}) + A_{11}^{-1})A_{12})x_2 \\ \geq & x_2^H A_{21} A_{11}^{-\frac{1}{2}} \left((\tfrac{1}{\gamma(A)^2} - 1)I + s(X^2 - X) \right) A_{11}^{-\frac{1}{2}} A_{12} x_2 \end{aligned} \tag{A.4}$$

mit $X = A_{11}^{\frac{1}{2}} H^{-1} A_{11}^{\frac{1}{2}}$, so ist (A.4) nicht negativ, wenn $p(t) := t^2 - t + \frac{1}{s}(\frac{1}{\gamma(A)^2} - 1) \geq 0$ auf $\sigma(H^{-1}A_{11})$ gilt.

Dies ist erfüllt, wenn

$$\begin{aligned} & p(t) := t^2 - t + (\frac{1}{\gamma(A)^2} - 1) = (t - \tfrac{1}{2})^2 - \tfrac{1}{4} + s^{-1} \cdot \tfrac{1-\gamma(A)^2}{\gamma(A)^2} \geq 0 \\ \Leftrightarrow \quad & (t - \tfrac{1}{2})^2 \geq \tfrac{(4+s)\gamma(A)^2-4}{4s\gamma(A)^2} \end{aligned} \tag{A.5}$$

für $t \in \sigma(H^{-1}A_{11})$.

Falls $\gamma(A) \leq \sqrt{\frac{4}{4+s}}$ ist, ist die rechte Seite von (A.5) kleiner gleich Null. Andernfalls erhält man die Bedingung für die Eigenwerte von $H^{-1}A_{11}$. □

Kapitel B

Eine kurze Einführung in die Quantenchromodynamik

Im Folgenden wird eine kurze Einführung in die Quantenchromodynamik (QCD) gegeben. Die Quantenchromodynamik ist eine Eichfeldtheorie, die vor allem auf der Quantenelektrodynamik (QED) basiert. Sie ist der Teilchenphysik zuzuordnen und beschreibt *starke Wechselwirkungen* zwischen Quarks und Gluonen. Im Laufe dieses Abschnittes wird der Leser durch die verschiedenen Aspekte der Quantenchromodynamik geführt, wozu einige Grundkenntnisse zum Aufbau der Materie nötig sind.

B.1

Das Quarkmodell. In diesem kurzen Abschnitt werden die wichtigsten Grundbegriffe der Teilchenphysik besprochen, um eine Vorstellung zu erlangen, woraus Materie besteht. Insbesondere wird die Frage beantwortet, warum sich Protonen in einem Atomkern trotz ihrer positiven Ladung nicht abstoßen. Diese Fragestellung führt auf den Begriff der starken Wechselwirkungen, welche zwischen sogenannten *Quarks* herrschen, die die Grundbausteine der Materie darstellen. Zuvor wird beschrieben, wie es überhaupt zur Suche nach diesen Quarks kam. Hierfür werden Quantenzahlen eingeführt.

Die Quantenzahlen charakterisieren den Zustand eines Elektrons. Zunächst wird zwischen vier verschiedenen Quantenzahlen unterschieden.

1. Die *Hauptquantenzahl* n bezeichnet das grundlegende Energieniveau bzw. im Schalenmodell, die Schale, auf der sich das Elektron mit einer Wahrscheinlichkeit von 90% befindet. Diese Zahl kann Werte $n \in \{1, 2, 3, \dots\}$ annehmen.

2. Die *Nebenquantenzahl* l, welche auch Drehimpulsquantenzahl genannt wird, kennzeichnet die Form des Orbitals in einem Atom. Die Zahl l kann die Werte $l \in \{0, 1, \dots, n-1\}$ annehmen.

3. Die *Magnetquantenzahl* m beschreibt die räumliche Orientierung des Elektron-Bahndrehimpuls'. Sie nimmt Werte $m \in \{-l, -l+1, \dots, -1, 0, 1 \dots, l-1, l\}$ an.

4. Die *Spinquantenzahl* s des Elektrons beschreibt die Orientierung des Spins des Elektrons, d.h. den Eigendrehimpuls und kann nur Werte $s \in \{-\frac{1}{2}, \frac{1}{2}\}$ annehmen.

Eines der wichtigsten Prinzipien der Teilchenphysik ist das *Pauli-Prinzip*[1] oder auch *Pauli'sches Ausschlussprinzip*, welches besagt:

[1] Formuliert im Jahre 1925 von Wolfgang Pauli (1900 - 1958), einem österreichischer Physiker, der im Jahr 1945 den Nobelpreis der Physik für das Ausschlussprinzip erhielt.

„Eine Wellenfunktion eines Teilchens muss beim Vertauschen von Elektronen total antisymmetrisch sein.“

In Formeln ausgedrückt, bedeutet das, dass für zwei Elektronen, die durch ihren Ort x bzw. $\widetilde{x}$ und ihre Ladung q bzw. $\widetilde{q}$ beschrieben werden, und ihre assoziierte Wellenfunktion ψ die Gleichung

$$\psi(x, q, \widetilde{x}, \widetilde{q}) = -\psi(\widetilde{x}, \widetilde{q}, x, q)$$

erfüllt ist.

Mithilfe der Spinquantenzahl lassen sich Teilchen in verschiedene Gruppen unterteilen.

Definition B.1. Hat ein Teilchen

1. einen halbzahligen Spin, so gehört es zu der Gruppe der *Fermionen*.
2. einen ganzzahligen Spin, so gehört es zu der Gruppe der *Bosonen*.

Die Menge der Fermionen und Bosonen wird unter dem Begriff der *Hadronen* zusammengefasst.

Bei dieser Definition ist zu beachten, dass das Pauli-Prinzip nur bei den Fermionen Anwendung findet.

Durch verbesserte Messapparaturen und Teilchenbeschleuniger wurden Teilchen gefunden, die sich nicht mehr mit dem bestehenden Modell erklären ließen. Das veranlasste, nach weiteren Fermionen und Bosonen zu suchen, was zu weiteren Teilchen, den sogenannten Quarks führte. Diese haben den Spin $\frac{1}{2}$ und sechs Freiheitsgrade: *Up, Down, Charm, Strange, Top* und *Bottom*, die als *flavours* bezeichnet werden. Zum Beispiel besteht ein Proton aus zwei Up- und einem Down-Quark, siehe Abbildung B.1.

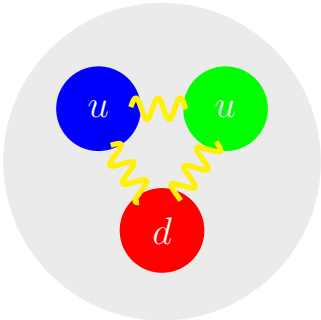


Abbildung B.1.: Der Aufbau eines Protons bestehend aus zwei Up-Quarks und einem Down-Quark.

In der Natur treten nur bestimmte Kombinationen von Hadronen, bestehend aus zwei bzw. drei Quarks, auf.

Definition B.2. Ein Hadron bestehend aus drei Quarks heißt *Meson*. Bilden ein Quark und ein zugehöriges Antiquark ein Hadron, so nennt man dieses *Baryon*.

Durch weitere Forschung wurden Teilchen gefunden, die sog. Δ^{++}, dargestellt in Abbildung B.2, welche aus drei Up-Quarks bestehen und deren Spin kollinear zueinander ausgerichtet ist. Die Wellenfunktion solch eines Teilchens ist in dem vorhandenen Modell symmetrisch unter Vertauschung zweier Quarks, was gegen das Pauli-Prinzip verstößt. Da man von diesem Prinzip nicht abweichen wollte, wurde eine weitere Quantenzahl postuliert und mit „Farbe (chrom)“ bezeichnet. Um das Pauli-Prinzip zu erhalten, wurde die Existenz von drei Farben, *rot*, *blau* und *grün*, vorausgesagt. Dies führte auf eine neue Theorie: *Die Quantenchromodynamik.*

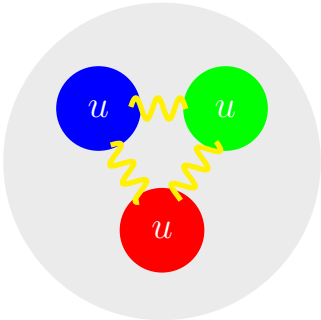


Abbildung B.2.: Das Teilchen Δ^{++} besteht aus drei Up-Quarks. Es veranlasste zur Suche nach einer weiteren Quantenzahl und führte auf die QCD.

B.2

Quantenchromodynamik. Wie oben angedeutet, behandelt die Quantenchromodynamik (QCD) Wechselwirkungen zwischen Quarks. Sie beschreibt den Übergang der Farbladung und wird als starke Wechselwirkung bezeichnet. Es wurde festgestellt, dass Teilchen in der Natur immer als *weiße* Teilchen auftreten, d.h. die Farbladung neutral ist. Daher ist es nicht möglich, einzelne Quarks mit nicht neutraler Farbe zu beobachten. Diese Wechselwirkungen sind dafür verantwortlich, dass sich Protonen trotz gleicher Ladung in einem Atomkern nicht abstoßen. Obwohl sie durch ihre jeweilige positive Ladung Abstoßungskräfte entwickeln, werden sie durch die deutlich stärkeren Verbindungen der Farbladungen zusammengehalten. Dieses Phänomen wird als *Confinement* bezeichnet. Hadronen bestehen entweder aus drei Quarks, die jeweils paarweise verschiedene Farbladungen besitzen („rot+blau+grün=weiß“), oder aus einem Quark - Antiquark - Paar. Erstere werden als *Baryonen* und die Pärchen als *Mesonen* bezeichnet. Dies wird in Abbildung B.3 veranschaulicht. Der Zusammenhang zwischen den drei verschiedenen Ladungen, die sich nur mit ihrer Antifarbe oder in einer Kombination aus drei verschiedenen Farben auslöschen, ist auch der Grund für die Namensgebung der QCD.

Eine weitere wichtige Größe in der QCD sind die sogenannten *Gluonen.* Sie übernehmen die Rolle der Protonen aus der Quantenelektrodynamik und dienen als eine Art Kleber zwischen den Quarks. Ein wichtiger Unterschied zu den Protonen aus der QED besteht darin, dass die Gluonen selbst auch Farbladungen besitzen, bestehend aus einer Farbe und einer Anti-Farbe. Daher wechselwirken diese mit den Quarks, was zu einer komplexeren Theorie führt und auch der Grund für die Stärke der Wechselwirkung ist.

Die QCD basiert zum großen Teil auf Beobachtungen und Anwendungen der theoretischen Konzepte der Quantenelektrodynamik. Daher lohnt sich ein Blick auf dieses verwandte Gebiet,

welches im nächsten Abschnitt besprochen wird.

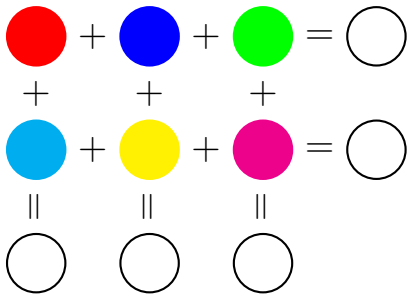


Abbildung B.3.: Die Farbwahl in der QCD. In der ersten Zeile sind die drei Grundfarben dargestellt, die die Quarks annehmen können, in der zweiten Zeile ihre Anti-Farben. In der Summe muss jeweils die Farbe weiß erreicht werden. Von oben nach unten erhält man die realisierbaren Mesonen und von links nach rechts die realisierbaren Baryonen.

B.3

Quantenelektrodynamik (QED). Dieser Abschnitt basiert auf dem Buch von Peter Schmüser, [87]. Zunächst wird erläutert, warum der Sprung von der klassischen Physik zur relativistischen Physik vollzogen wurde. Dazu wird auf die Herleitung der *Schrödinger-Gleichung*[2] zurückgegriffen. Die Grundideen kamen von Max Planck[3]. Ausgehend von der Annahme, dass Energie nur „gequantelt“ in der Form $E = \hbar\omega$ emittiert wird - hierbei ist $\hbar$ das *Plancksche Wirkungsquantum* und ω die Frequenz der Strahlung - wurde die Physik neu entwickelt.

Louis-Victor de Broglie[4] schloss im Jahre 1924 sein Studium mit der Arbeit „Recherches sur la theorie des quanta“ ab. Mit den Ergebnissen von Albert Einstein[5] ($E = mc^2$) kam er zu dem Entschluss, dass, bei der von Planck postulierten Beziehung $E = \hbar\omega$, jeder Masse m eine Frequenz $\omega = \frac{mc^2}{\hbar}$ zugeordnet werden kann. Diese Frequenz ist nicht auf das Teilchenvolumen beschränkt, sondern begleitet das Teilchen in Form einer Welle auch in großen Raumbereichen. Daraus resultiert, dass der *Welle-Teilchen-Dualismus* nicht nur bei Photonen, sondern auch bei anderer Materie Anwendung findet, d.h. dass auch Elektronen Welleneigenschaften zugesprochen werden können. Für die Wellenlänge ergibt sich

$$\lambda = \frac{\hbar}{p}, \qquad \text{wobei } p = \frac{mv}{\sqrt{1-\left(\frac{v}{c}\right)^2}}$$

[2] Im Jahre 1926 formuliert von Erwin Rudolf Alexander Schrödinger (1887-1961), österreichischer Physiker und Wissenschaftstheoretiker. 1933 erhielt er zusammen mit Paul Dirac den Nobelpreis der Physik.

[3] Max Planck (1858-1947) war ein deutscher Physiker auf dem Gebiet der Theoretischen Physik. Für die Entdeckung des Planckschen Wirkungsquantums erhielt er im Jahre 1918 den Nobelpreis der Physik.

[4] Mit vollem Namen Louis-Victor Pierre Raymond de Broglie, 7. Herzog de Broglie (1892-1982), war ein französischer Physiker. Er erhielt im Jahre 1929 für die Entdeckung der Wellennatur des Elektrons den Nobelpreis der Physik.

[5] Albert Einstein (1879-1955), deutscher Physiker, erhielt im Jahre 1922 für seine Erklärung des photoelektrischen Effekts den Nobelpreis der Physik.

der relativistische Impuls des Teilchens ist. Hierbei ist m die Masse, v die Momentangeschwindigkeit des Teilchens und c die Lichtgeschwindigkeit.

Ein Elektron kann daher durch eine Wellenfunktion der Form

$$\Psi(x,t) = Ce^{ikx}e^{-i\omega t}$$

mit $C \in \mathbb{C}$ beschrieben werden. Die Variablen sind hierbei die Zeit t, der Ort x und k ist ein Parameter mit $k = \frac{2\pi}{\lambda}$.

Eine weitere Errungenschaft in der Quantentheorie ist die Darstellung physikalischer Größen durch Operatoren. Diese wirken auf die Wellenfunktionen, deren Eigenwerte mögliche Messwerte darstellen.

Definition B.3. Sei $\Psi(x,t) = Ce^{ikx}e^{-i\omega t}$ die Wellenfunktion eines Elektrons im Ort x zum Zeitpunkt t. Dann werden der *Energieoperator* $\mathbf{E}$ und der *Impulsoperator* $\mathbf{P}$ durch die Eigenwertgleichungen

$$\mathbf{E}\Psi = \hbar\omega\Psi \quad \text{sowie} \quad \mathbf{P}\Psi = \hbar k\Psi$$

definiert. Damit können diese Operatoren als Differentialoperatoren

$$\mathbf{E} = i\hbar\frac{\partial}{\partial t} \text{ und } \mathbf{P} = -i\hbar\nabla.$$

aufgefasst werden.

Im weiteren Verlauf werden Operatoren durch „fettgedruckte“ Buchstaben gekennzeichnet.

Die Quantentheorie basiert trotz aller Innovationen auf der klassischen Physik. Betrachtet man die Impuls-Energie-Beziehung aus der klassischen Physik $E = \frac{p^2}{2m}$, so erhält man durch Substitution mittels der Operatoren die *Schrödingergleichung*

$$i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar}{2m}\nabla^2\Psi.$$

Im relativistischen Fall lautet die Beziehung zwischen Energie, Impuls und Ruhemasse

$$E^2 = p^2c^2 + (mc^2)^2,$$

welche durch Einsetzen der Operatoren auf die Gleichung

$$-\hbar^2\frac{\partial^2\Psi}{\partial t^2} = \left(-\hbar^2c^2\nabla^2 + (mc^2)^2\right)\Psi$$

führt. Durch Umstellung gelangt man zur sogenannten *Klein-Gordon-Gleichung*[6]

$$\left[\frac{1}{c^2}\frac{\partial^2\Psi}{\partial t^2} - \nabla^2 + \left(\frac{mc}{\hbar}\right)^2\right]\Psi = 0.$$

Dies ist die relativistische Erweiterung der Schrödinger-Gleichung für ein Teilchen, dessen Spin den Wert 0 annimmt. Beim Lösen dieser Gleichung tritt das Ereignis auf, dass Zustände mit

[6] Benannt nach Oskar Klein (1894-1977), einem schwedischen Physiker und Walter Gordon (1983-1939), deutscher Physiker.

den Energien $E = \pm\sqrt{c^2p^2 + m^2c^4}$ existieren; insbesondere treten Zustände mit negativer Energie auf. Um dieses Problem zu umgehen, sollte eine Bewegungsgleichung konstruiert werden, bei der keine negativen Energien auftreten. Die Ursache für solche Energien ist bei der Klein-Gordon-Gleichung die zweite Ableitung in der Zeit.

Paul Dirac[7] stellte 1928 die nach ihm benannte Dirac-Gleichung

$$(i\partial\!\!\!/ - m)\Psi(x,t) = 0 \tag{B.1}$$

auf, wobei $\partial\!\!\!/$ die sogenannte *Feynman-Slash-Notation*[8] $\partial\!\!\!/ := \sum_{\mu=0}^{3} \gamma^\mu \partial_\mu$ verwendet. Die Terme γ^μ und ∂_μ sind sogenannte *Vierervektoren*

$$\gamma^\mu = \begin{bmatrix} \gamma^0 & \gamma^1 & \gamma^2 & \gamma^3 \end{bmatrix} \text{ und } \partial_\mu = \begin{bmatrix} \frac{\partial}{\partial t} & \nabla \end{bmatrix}$$

mit $\gamma_i \in \mathbb{C}^{4,4}$, $i = 0, \ldots, 3$, die die Anti-Vertauschungsrelation

$$\{\gamma^\mu, \gamma^\nu\} := \gamma^\mu\gamma^\nu + \gamma^\nu\gamma^\mu = 2g^{\mu\nu} I \tag{B.2}$$

erfüllen. Hierbei ist $g^{\mu\nu}$definiert durch $g^{\mu\nu} = 0$ für $\mu \neq \nu$ sowie $g^{11} = g^{22} = g^{33} = -1$ und $g^{00} = 1$. Z.B. erfüllen die Matrizen

$$\gamma^0 = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & -1 & \\ & & & -1 \end{bmatrix}, \gamma^1 = \begin{bmatrix} & & & 1 \\ & & 1 & \\ & -1 & & \\ -1 & & & \end{bmatrix}, \gamma^2 = \begin{bmatrix} & & & -i \\ & & i & \\ & -i & & \\ i & & & \end{bmatrix}, \gamma^3 = \begin{bmatrix} & & 1 & \\ & & & -1 \\ -1 & & & \\ & 1 & & \end{bmatrix} \tag{B.3}$$

die Gleichung (B.2) und werden als *Repräsentanten* bezeichnet.

Bemerkung B.4. Mathematisch formuliert, betrachtet man einen sogenannten *Minkowski-Raum*[9], d.h. einen $\mathbb{R}$-Vektorraum mit einer nicht ausgearteten Bilinearform $\beta(\cdot,\cdot)$, die die Grundlage der angesprochenen Analyse bildet. In diesem konkreten Fall handelt es sich um den 4D Minkowski-Raum mit dem inneren Produkt

$$v \cdot w := \beta(v,w) = v_0w_0 - v_1w_1 - v_2w_2 - v_3w_3, \text{ wobei } v = \begin{bmatrix} v_0 \\ v_1 \\ v_2 \\ v_3 \end{bmatrix}, w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix}.$$

Oft wird in der Literatur auch $v \cdot w := \beta(v,w) = -v_0w_0 + v_1w_1 + v_2w_2 + v_3w_3$ verwendet. Die erste Variable v_0 (bzw. w_0) ist eine Transformation der Zeit $v_0 = ct$.

Ausgeschrieben hat die Dirac-Gleichung die folgende Gestalt:

$$\begin{aligned} 0 = (i\partial\!\!\!/ - m)\Psi(x,t) &= \left[i\left(\gamma^0\frac{\partial}{\partial t} + \gamma^1\frac{\partial}{\partial x_1} + \gamma^2\frac{\partial}{\partial x_2} + \gamma^3\frac{\partial}{\partial x_3}\right) - mI \right]\Psi \\ &= \begin{bmatrix} i\frac{\partial}{\partial t} - m & 0 & i\frac{\partial}{\partial x_3} & i\frac{\partial}{\partial x_1} + \frac{\partial}{\partial x_2} \\ 0 & i\frac{\partial}{\partial t} - m & i\frac{\partial}{\partial x_1} - \frac{\partial}{\partial x_2} & -i\frac{\partial}{\partial x_3} \\ -i\frac{\partial}{\partial x_3} & -i\frac{\partial}{\partial x_1} - \frac{\partial}{\partial x_2} & -i\frac{\partial}{\partial t} - m & 0 \\ -i\frac{\partial}{\partial x_1} + \frac{\partial}{\partial x_2} & i\frac{\partial}{\partial x_3} & 0 & -i\frac{\partial}{\partial t} - m \end{bmatrix} \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \Psi_4 \end{bmatrix} = \begin{bmatrix} i\frac{\partial\Psi_1}{\partial t} - m\Psi_1 + i\frac{\partial\Psi_3}{\partial x_3} + i\frac{\partial\Psi_4}{\partial x_1} + \frac{\partial\Psi_4}{\partial x_2} \\ i\frac{\partial\Psi_2}{\partial t} - m\Psi_2 + i\frac{\partial\Psi_3}{\partial x_1} - \frac{\partial\Psi_3}{\partial x_2} - i\frac{\partial\Psi_4}{\partial x_3} \\ -i\frac{\partial\Psi_1}{\partial x_3} - i\frac{\partial\Psi_2}{\partial x_1} - \frac{\partial\Psi_2}{\partial x_2} - i\frac{\partial\Psi_3}{\partial t} - m\Psi_3 \\ -i\frac{\partial\Psi_1}{\partial x_1} + \frac{\partial\Psi_1}{\partial x_2} + i\frac{\partial\Psi_2}{\partial x_3} - i\frac{\partial\Psi_4}{\partial t} - m\Psi_4 \end{bmatrix}. \end{aligned}$$

[7] Paul Adrien Maurice Dirac (1902-1984) war ein britischer Physiker und erhielt im Jahre 1933 zusammen mit Schrödinger den Nobelpreis der Physik. Er gilt als Mitbegründer der Quantenphysik.

[8] Eingeführt von Richard Feynman (1918-1988), einem amerikanischen Physiker und Nobelpreisträger des Jahres 1965.

[9] Eingeführt im Jahre 1907 von Hermann Minkowski (1864-1909), deutscher Mathematiker und Physiker.

Für ein ruhendes Elektron hat diese Gleichung vier Lösungen

$$\Psi_1(x,t) = \begin{bmatrix} 1 \\ 0 \\ \frac{p_z}{E+m} \\ \frac{p_x+ip_y}{E+m} \end{bmatrix} \exp(-iEt)\exp(ip\cdot x), \quad \Psi_2(x,t) = \begin{bmatrix} 0 \\ 1 \\ \frac{p_x-ip_y}{E+m} \\ \frac{-p_z}{E+m} \end{bmatrix} \exp(-iEt)\exp(ip\cdot x),$$

$$\Psi_3(x,t) = \begin{bmatrix} \frac{p_x-ip_y}{E+m} \\ \frac{-p_z}{E+m} \\ 0 \\ 1 \end{bmatrix} \exp(iEt)\exp(-ip\cdot x), \quad \Psi_4(x,t) = \begin{bmatrix} \frac{p_z}{E+m} \\ \frac{p_x+ip_y}{E+m} \\ 1 \\ 0 \end{bmatrix} \exp(iEt)\exp(-ip\cdot x).$$

Hierbei ist $E = \sqrt{p^2+m^2}$. Die ersten beiden Zustände haben die positive Energie E, doch bei den Zuständen Ψ_3, Ψ_4 treten erneut negative Energien auf.

Im Jahr 1928 erklärte Dirac die Zustände mit negativer Energie durch die Existenz von Antiteilchen. Z.B. besitzt das Elektron ein Antielektron, das sogenannte Positron. Dieses Positron wurde 1932 von Carl David Anderson[10] experimentell nachgewiesen. Mit dieser Entdeckung ergibt sich, dass jedes Teilchen aus Symmetriegründen ein Antiteilchen mit identischer Masse besitzt.

Im Weiteren ist die Bewegungsgleichung von Teilchen, auf die eine äußere Kraft wirkt, von zentraler Bedeutung. Zum Beispiel hat die Dirac-Gleichung für ein Teilchen mit Ladung q im elektromagnetischen Feld die folgende Gestalt

$$(i\partial\!\!\!/ - m)\Psi(x,t) = q A\!\!\!/ \Psi \quad \text{mit} \quad A\!\!\!/ := \sum_{\mu=0}^{3} \gamma^\mu A_\mu. \tag{B.4}$$

Hierbei symbolisieren A_μ Wirkungen eines elektromagnetischen Feldes und werden *Eichfelder* genannt.

B.4

Eichtheorie. Eine der wichtigsten Bedingungen, die an eine physikalische Theorie gestellt werden, ist die von Albert Einstein eingeführte Invarianz bzgl. zueinander bewegter Inertialsysteme (Koordinatensysteme). Dies bedeutet, dass eine Bewegungsgleichung nicht von dem Beobachter abhängt bzw. die Bewegungsgleichungen in allen Inertialsystemen dieselbe Form haben. Die Zustände (Lösungen der Gleichungen) müssen folglich in Abhängigkeit vom Betrachter transformiert werden. Im Folgenden werden die Variablen x und t als gleichwertig betrachtet und zu der Variablen x zusammengefasst.

Seien also Σ und Σ' zwei zueinander gleichförmig bewegte Koordinatensysteme. Dann soll gelten

$$(i\partial\!\!\!/ - m)\Psi(x) = 0 \text{ in } \Sigma,$$
$$(i\partial\!\!\!/ - m)\Psi'(x) = 0 \text{ in } \Sigma'.$$

Hierbei soll es eine Vorschrift geben, wie sich $\Psi(x)$ in $\Psi'(x)$ transformiert. Anders ausgedrückt bedeutet das, dass die Zustände entsprechend zum Betrachter geeicht werden können. Der Begriff der Eichtheorie wurde 1929 von Herrmann Weyl[11] geprägt.

[10] Carl David Anderson (1905-1991) war amerikanischer Physiker. Er erhielt für die Entdeckung des Positrons im Jahr 1936 den Nobelpreis für Physik.

[11] Herrmann Weyl (1885-1955) war ein deutscher Mathematiker, Physiker und Philosoph. Er führte bereits 1918 das Konzept der Eichtheorie ein, hatte jedoch mit viel Widerstand, unter anderem von Einstein, zu kämpfen. 1929 gelang es ihm, diese Theorie soweit voranzutreiben, dass sie allgemein anerkannt wurde.

Zunächst wird untersucht, wie sich die Zustände der Dirac-Gleichung bei globalen Transformationen der Form $\Psi'(x) = e^{qi\varphi}\Psi(x)$ ändern, wobei q eine Ladung ist. Diese betrachtete Transformation ist eine *Phasentransformation*. Des Weiteren soll das Teilchen „frei “ sein, es sollen also keine äußeren Kräfte wirken. Dazu betrachtet man (B.1). Wird nun Ψ' in (B.1) eingesetzt, ergibt sich

$$(i\partial\!\!\!/ - m)\Psi'(x) = 0,$$

also erfüllt ein global phasentransformierter Zustand (Spinor) die Dirac-Gleichung. Es stellt sich die Frage, ob dies bei lokalen Transformationen, auch *Eichtransformationen* genannt, ebenfalls der Fall ist. Sei nun

$$\Psi'(x) = e^{qi\varphi(x)}\Psi(x) \tag{B.5}$$

eine transformierte Wellenfunktion mit Ladung q. Es ist zu untersuchen, wie die zu der Wellenfunktion Ψ' assoziierte Dirac-Gleichung aussieht. Betrachtet man dazu

$$\begin{aligned}(i\partial\!\!\!/ - m)\Psi'(x) =& (i\partial\!\!\!/ - m)e^{iq\varphi(x)}\Psi(x)\\ =& i\left[(\partial\!\!\!/ e^{iq\varphi(x)})\Psi(x) + e^{iq\varphi(x)}\partial\!\!\!/\Psi(x)\right] - me^{iq\varphi(x)}\Psi(x)\\ =& e^{iq\varphi(x)}\underbrace{(i\partial\!\!\!/ - m)\Psi(x)}_{=0} + i(\partial\!\!\!/ e^{iq\varphi(x)})\Psi = -q(\partial\!\!\!/\varphi(x))\Psi'(x)\end{aligned}$$

und definiert $A'_\mu := -\partial_\mu\varphi(x)$, dann gilt

$$(i\partial\!\!\!/ - m)\Psi'(x) = qA\!\!\!/'\Psi'(x).$$

Hierbei ist $A\!\!\!/'$ analog zu $A\!\!\!/$ in (B.4) definiert. Man sieht, dass die transformierte Wellenfunktion nicht mehr die homogene Dirac-Gleichung erfüllt. Dies bedeutet, es wurde durch die lokale Phasentransformation eine äußere Kraft erzeugt.

Wird stattdessen bei der Wellengleichung eine äußere Kraft $A_\mu(x)$ vorausgesetzt, die z.B. durch ein elektromagnetisches Feld gegeben ist, erfüllt die Wellenfunktion die Gleichung

$$(i\partial\!\!\!/ - m)\Psi(x) = qA\!\!\!/(x)\Psi(x). \tag{B.6}$$

Mit der Transformation (B.5) erhält man

$$(i\partial\!\!\!/ - m)\Psi'(x) = qA\!\!\!/'\Psi'(x), \tag{B.7}$$

wobei $\Psi'(x) = e^{iq\varphi(x)}\Psi(x)$ und $A'_\mu(x) = A_\mu(x) - \partial_\mu\varphi(x)$.

Das heißt, dass bei lokalen Transformationen äußere Kräfte ebenfalls transformiert werden und somit die Eichinvarianz erfüllt ist.

Die Gleichung (B.6) lässt sich auf eine homogene Gestalt überführen, indem die Ableitungen durch kovariante Ableitungen, Definition 2.2, ersetzt werden, also $D_\mu = \partial_\mu + iA_\mu$ und $D'_\mu = \partial_\mu + iA'_\mu$.

Die kovarianten Ableitungen berücksichtigen äußere Kräfte bzw. Krümmungen im Raum. Damit erfüllt die Dirac-Gleichung (B.1) die Eichinvarianz, d.h für Ψ und Ψ' gilt

$$(iD\!\!\!/ - m)\Psi(x) = (iD\!\!\!/' - m)\Psi'(x) = 0 \quad \text{mit} \quad D\!\!\!/ := \sum_{\mu=0}^{3}\gamma^\mu D_\mu \text{ und } D\!\!\!/' := \sum_{\mu=0}^{3}\gamma^\mu D'_\mu.$$

Es wurde damit gezeigt, dass Bewegungsgleichung eines Elektrons unter der Eichtransformation $\Psi'(x) = e^{qi\varphi(x)}\Psi(x)$ invariant ist.

Bemerkung B.5. Die Menge $\mathrm{U}(1) := \{e^{qi\varphi} \,|\, \varphi \in [0, 2\pi[\}$ beinhaltet die betrachteten Eichtransformationen. In der modernen Physik sagt man, es liegt eine *Symmetrie* zu Grunde, wenn gewisse Eigenschaften bei einer Transformation erhalten bleiben. Aus diesem Grund wird U(1) als die *Symmetriegruppe der QED* bezeichnet. In der QCD ist die zugrunde liegende Symmetriegruppe die $\mathrm{SU}(3) := \{U \in \mathbb{C}^{3,3} \,|\, UU^H = I, \det(U) = 1\}$. Dies führt auf Eichfelder $A_\mu \in \mathbb{C}^{3,3}$ und damit auf einen etwas anderen Dirac-Operator.

B.5

Dirac-Operator in der QCD. Analog zu den Ergebnissen der QED wird versucht, eine Theorie für die QCD zu entwickeln. Um die Dirac-Gleichung für die QCD anzugeben, muss erneut die lokale Eichinvarianz gefordert werden. Dies führt auf die kovariante Ableitung (2.7) mit dem Eichfeld A_μ. Dabei muss beachtet werden, dass $A_\mu \in \mathbb{C}^{3,3}$. Mit ähnlichen Überlegungen wie bei der QED erhält man den Dirac-Operator für die QCD

$$\mathbf{D} = \sum_{\mu=1}^{4} (\gamma_\mu \otimes (I_3\partial_\mu + iA_\mu)) - mI_{12},$$

wobei $I_{12} \in \mathbb{R}^{12,12}$ ist. Hier wurde bereits in Anlehnung an [57, 58] die Transformation vom Minkowski-Raum auf einen 4D euklidischen Raum durchgeführt. Dabei wurde anders als im letzten Abschnitt $x_4 = ict$ gesetzt. Die hier betrachteten Repräsentanten im Vergleich zu (B.3) sind gegeben durch

$$\gamma_1 = \begin{bmatrix} & & & i \\ & & i & \\ & -i & & \\ -i & & & \end{bmatrix}, \gamma_2 = \begin{bmatrix} & & & -1 \\ & & 1 & \\ & 1 & & \\ -1 & & & \end{bmatrix}, \gamma_3 = \begin{bmatrix} & & i & \\ & & & -i \\ -i & & & \\ & i & & \end{bmatrix}, \gamma_4 = \begin{bmatrix} & & & 1 \\ & & 1 & \\ & 1 & & \\ 1 & & & \end{bmatrix}.$$

Setzt man $D_\mu = I_3\partial_\mu + iA_\mu$, sieht der Dirac-Operator ausgeschrieben wie folgt aus

$$\mathbf{D} = \begin{bmatrix} -mI_3 & 0 & iD_3-D_4 & iD_1-D_2 \\ 0 & -mI_3 & iD_1+D_2 & -D_3-D_4 \\ iD_3-D_4 & -iD_1+D_2 & -mI_3 & 0 \\ -iD_1-D_2 & iD_3-D_4 & 0 & -mI_3 \end{bmatrix}.$$

Es stellt sich die Frage, wie dieser Operator analysiert werden kann, um die entsprechenden Wechselwirkungen zu verstehen. Bei der QED und den schwachen Wechselwirkungen kann man mithilfe der Störungstheorie und asymptotischer Analysis eine Lösung konstruieren. Dies funktioniert bei starken Wechselwirkungen nicht mehr. Daher wird versucht, die Lösung numerisch durch Diskretisierung auf einem Gitter zu bestimmen.

B.6

QCD auf einem Gitter. Um den Dirac-Operator in der QCD zu analysieren, schlug Wilson in [98] vor, die Wechselwirkungen zu simulieren, indem ein diskreter Dirac-Operator entwickelt wird. Dazu bediente er sich der Gittereichtheorie. Man betrachte ein Gitter im vierdimensionalen Raum mit $N_1 \times N_2 \times N_3 \times N_4$ Gitterpunkten, wobei $N_1, N_2, N_3, N_4 \in \mathbb{N}$ die Anzahl der Gitterpunkte in jeder Dimension sind, also $\Omega_h := \{x = (x_1, x_2, x_3, x_4) \,|\, x_j = j \cdot a,\ j = 0, \ldots, N_j - 1\}$, und periodischen Randbedingungen. Hierbei wird o.B.d.A. angenommen, dass das Gitter in den Ursprung $(0, 0, 0, 0)$ verschoben ist. Des Weiteren sei der Gitterabstand a, wobei oft $a = 1$ gefordert wird. Um den Dirac-Operator zu diskretisieren, muss zunächst eine diskrete Variante der Eichfelder (*gauge fields*) A_μ angegeben werden. Die Darstellung basiert auf [58].

Definition B.6. Sei $\mu \in \{1,2,3,4\}$ und $x \in \Omega_h$. Dann wird das *diskretisierte Eichfeld* durch

$$U_\mu^x := e^{-iA_\mu(x+\frac{1}{2}e_\mu)} \approx e^{-i\int_x^{x+e_\mu} A_\mu(s)\,ds}$$

definiert. Dabei ist e_μ der μ-te Einheitsvektor in $\mathbb{C}^4$, $\mu = 1,2,3,4$. Eine Ansammlung $\{U_\mu^x\}$ von diskretisierten Eichfeldern wird als eine *Gauge-Konfiguration* bezeichnet.

Die Gauge-Konfiguration kann als eine Approximation der Wirkung des Eichfeldes A_μ entlang der Kanten des Gitters interpretiert werden. Zu solch einer Konfiguration wird eine Temperatur β zugeordnet, die die Zufälligkeit des zugehörigen Eichfeldes wiedergibt. Analog zu Abschnitt 2.1 müssen die Ableitungen durch finite Differenzen ersetzt werden. Hier sehen die zentralen finiten Differenzen von ψ auf einem Gitterpunkt $x \in \Omega_h$ unter Berücksichtigung der speziellen Struktur der Ableitungen $D_\mu = I_3\partial_\mu + iA_\mu$ wie folgt aus:

$$D_\mu\psi_x \approx \tfrac{1}{2}\big(U_\mu^x\psi_{x+e_\mu} - (U_\mu^{x-e_\mu})^H\psi_{x-e_\mu}\big),$$

wobei $x + e_\mu$ der Nachbargitterpunkt in Richtung μ und $\psi_x = \psi(x)$ ist. Für weitere Details sei auf [58] verwiesen. Damit lässt sich der *diskretisierte Dirac-Operator* durch

$$\mathbf{D}_{\text{diskret}} := \tfrac{1}{2}\sum_{\mu=1}^{4}\Big(\gamma_\mu \otimes \big(U_\mu^x\psi_{x-e_\mu} - (U_\mu^{x-e_\mu})^H\psi_{x+e_\mu}\big)\Big) + mI \tag{B.8}$$

angeben. Wie in der Literatur üblich, wurde der Term mit -1 multipliziert, so dass die Masse als positiver additiver Term auftritt. Diese Art der Diskretisierung führt zu numerischen Instabilitäten, welche durch die Addition des Terms

$$\sum_{\mu=1}^{4} D_\mu^2 = \Delta_D \tag{B.9}$$

beseitigt werden, siehe [44, 98].

Die Aufgabenstellung ist daher: Finde zu einer Indexmenge I und einem Satz $\{\varphi_j \,|\, j \in I\}$ die Lösungen ψ_j von den linearen Gleichungssystemen

$$\mathbf{D}_w\psi_j := \Big(\tfrac{1}{2}\sum_{\mu=1}^{4}\Big(\gamma_u \otimes \big(U_\mu^x\psi_{x-e_\mu} - (U_\mu^{x-e_\mu})^H\psi_{x+e_\mu}\big) - D_u^2\Big) + mI\Big)\psi_j = \varphi_j.$$

Der Operator $\mathbf{D}_w = \mathbf{D}_{\text{diskret}} - \frac{1}{2}\Delta_D$ wird als *Wilson-Dirac-Operator* bezeichnet.

Anhand dieses Systems erkennt man schnell die Schwierigkeiten numerischer Berechnungen in der QCD. Die Größe der Systemmatrix nimmt durch den vierdimensionalen Raum schnell zu. Es ist $n = n_x n_y n_z n_t$, wobei n_x, n_y, n_z, n_t die Anzahl der inneren Gitterpunkte in den jeweiligen Koordinaten bezeichnet. Des Weiteren ist das System sehr unstrukturiert und komplexwertig. In der Physik ist man außerdem nicht nur an einer Lösung für eine rechte Seite interessiert, sondern möchte für einen ganzen Satz von rechten Seiten die Lösungen berechnen. Dies kann soweit gehen, dass der Dirac-Operator invertiert werden soll. Solch komplexe Berechnungen sind nur auf Supercomputern möglich. Um neue Verfahren für diese Problemstellung zu entwickeln und analysieren, wird in der Literatur [33, 45, 50, 57, 69] häufig als Modellproblem die Gauge-Laplace-Matrix aus Abschnitt **??** betrachtet.

Literaturverzeichnis

[1] Arnold, G. ; Cundy, N. ; Eshof, J. van d. ; Frommer, A. ; Krieg, S. ; Lippert, T. ; Schäfer, K.: Numerical methods for the QCD overlap operator. II. Optimal Krylov subspace methods. In: *QCD and numerical analysis III* Bd. 47. Berlin : Springer, 2005, S. 153–167

[2] Aubert, G. ; Kornprobst, P.: *Mathematical problems in image processing. Partial differential equations and the calculus of variations. Foreword by Olivier Faugeras. 2nd ed.* New York : Springer, 2006

[3] Axelsson, O.: On multigrid methods of the two-level type. In: *Multigrid methods (Cologne, 1981)* Bd. 960. Berlin : Springer, 1982, S. 352–367

[4] Axelsson, O.: *Iterative solution methods.* Cambridge : Cambridge University Press, 1994

[5] Axelsson, O.: Stabilization of algebraic multilevel iteration methods; additive methods. In: *Numer. Algorithms* 21 (1999), Nr. 1-4, S. 23–47

[6] Axelsson, O.: A survey of algebraic multilevel iteration (AMLI) methods. In: *BIT* 43 (2003), S. 863–879

[7] Axelsson, O. ; Gustafsson, I.: Preconditioning and two-level multigrid methods of arbitrary degree of approximation. In: *Math. Comp.* 40 (1983), Nr. 161, S. 219–242

[8] Axelsson, O. ; Neytcheva, M.: Algebraic multilevel iteration method for Stieltjes matrices. In: *Numer. Linear Algebra Appl.* 1 (1994), Nr. 3, S. 213–236

[9] Axelsson, O. ; Padiy, A.: On the additive version of the algebraic multilevel iteration method for anisotropic elliptic problems. In: *SIAM J. Sci. Comput.* 20 (1999), Nr. 5, S. 1807–1830

[10] Axelsson, O. ; Vassilevski, P. S.: Algebraic multilevel preconditioning methods. I. In: *Numer. Math.* 56 (1989), Nr. 2-3, S. 157–177

[11] Axelsson, O. ; Vassilevski, P. S.: Algebraic multilevel preconditioning methods. II. In: *SIAM J. Numer. Anal.* 27 (1990), Nr. 6, S. 1569–1590

[12] Bakhvalov, N. S.: On the convergence of a relaxation method under natural contraints on an elliptic operator. In: *Ž. Vyčisl. Mat. i Mat. Fiz.* 6 (1966), S. 861–883

[13] Bank, R. E. ; Dupont, T. F.: Analysis of a two-level scheme for solving finite element equations. 1980. – Forschungsbericht

[14] Bank, R. E. ; Dupont, T. F. ; Yserentant, H.: The hierarchical basis multigrid method. In: *Numer. Math.* 52 (1988), Nr. 4, S. 427–458

[15] Bärwolff, G.: *Numerik für Ingenieure, Physiker und Informatiker. Für Bachelor und Diplom.* Heidelberg : Elsevier/Spektrum Akademischer Verlag, 2007

[16] Benzi, M.: Preconditioning techniques for large linear systems: a survey. In: *J. Comput. Phys.* 182 (2002), Nr. 2, S. 418–477

[17] Berman, A. ; Plemmons, R. J.: *Classics in Applied Mathematics.* Bd. 9: *Nonnegative Matrices in the Mathematical Sciences.* Philadelphia, PA : Society for Industrial and Applied Mathematics (SIAM), 1994

[18] Blaheta, R. ; Margenov, S. ; Neytcheva, M.: Uniform estimate of the constant in the strengthened C.B.S. inequality for anisotropic non-conforming FEM systems. In: *Numer. Linear Algebra Appl.* 11 (2004), Nr. 4, S. 309–326

[19] Bollhöfer, M. ; Mehrmann, V.: Algebraic multilevel methods and sparse approximate inverses. In: *SIAM J. Matrix Anal. Appl.* 24 (2002), Nr. 1, S. 191–218 (electronic)

[20] Bollhöfer, M. ; Mehrmann, V.: *Numerische Mathematik.* Wiesbaden : Vieweg Studium: Grundkurs Mathematik, 2004

[21] Bolten, M. ; Brandt, A. ; Brannick, J. ; Frommer, A. ; Kahl, K. ; Livshits, I.: A bootstrap algebraic multilevel method for Markov chains. In: *SIAM J. Sci. Comput.* 33 (2011), Nr. 6, S. 3425–3446

[22] Bolten, M. ; Friedhoff, S. ; Frommer, A. ; Heming, M. ; Kahl, K.: Algebraic multigrid methods for Laplacians of graphs. In: *Linear Algebra Appl.* 434 (2011), Nr. 11, S. 2225–2243

[23] Bramble, J. H. ; Pasciak, J. E.: New estimates for multilevel algorithms including the V-cycle. In: *Math. Comp.* 60 (1993), Nr. 202, S. 447–471

[24] Bramble, J. H. ; Pasciak, J. E. ; Wang, J. P. ; Xu, J.: Convergence estimates for multigrid algorithms without regularity assumptions. In: *Math. Comp.* 57 (1991), Nr. 195, S. 23–45

[25] Bramble, J. H. ; Pasciak, J. E. ; Wang, J. P. ; Xu, J.: Convergence estimates for product iterative methods with applications to domain decomposition. In: *Math. Comp.* 57 (1991), Nr. 195, S. 1–21

[26] Brandt, A.: Multi–level adaptive technique (MLAT) for fast numerical solution to boundary value problems. In: *Proceedings of the Third International Conference on Numerical Methods in Fluid Mechanics* Bd. 18. Berlin : Springer, 1973 (Lecture Notes in Physics), S. 82–89

[27] Brandt, A.: General highly accurate algebraic coarsening. In: *Electron. Trans. Numer. Anal.* 10 (2000), S. 1–20

[28] Brandt, A.: Multiscale scientific computation: review 2001. In: *Multiscale and multiresolution methods* Bd. 20. Berlin : Springer, 2002, S. 3–95

[29] Brandt, A. ; Brannick, J. ; Kahl, K. ; Livshits, I.: Bootstrap AMG. In: *SIAM J. Sci. Comput.* 33 (2011), Nr. 2, S. 612–632

[30] Brandt, A. ; McCormick, S. ; Ruge, J.: Algebraic multigrid (AMG) for automatic multigrid solution with application to geodetic computations / Inst. Comp. Studies, Colo. State Univ. 1982. – Forschungsbericht

[31] Brannick, J. ; Brezina, M. ; Keyes, D. ; Livne, O. ; Livshits, I. ; MacLachlan, S. ; Manteuffel, T. ; McCormick, S. ; Ruge, J. ; Zikatanov, L.: Adaptive smoothed aggregation in lattice QCD. In: *Domain decomposition methods in science and engineering XVI* Bd. 55. Berlin : Springer, 2007, S. 505–512

[32] Brannick, J. ; Falgout, R. D.: Compatible relaxation and coarsening in algebraic multigrid. In: *SIAM J. Sci. Comput.* 32 (2010), Nr. 3, S. 1393–1416

[33] Brannick, J. ; Frommer, A. ; Kahl, K. ; MacLachlan, S. ; Zikatanov, L.: Adaptive reduction-based multigrid for nearly singular and highly disordered physical systems. In: *Electron. Trans. Numer. Anal.* 37 (2010), S. 276–295

[34] Brannick, J. ; Zikatanov, L.: Algebraic multigrid methods based on compatible relaxation and energy minimization. In: *Domain decomposition methods in science and engineering XVI* Bd. 55. Berlin : Springer, 2007, S. 15–26

[35] Brezina, M. ; Falgout, R. D. ; MacLachlan, S. ; Manteuffel, T. ; McCormick, S. ; Ruge, J.: Adaptive smoothed aggregation (αSA) multigrid. In: *SIAM Rev.* 47 (2005), Nr. 2, S. 317–346

[36] Brezina, M. ; Falgout, R. D. ; MacLachlan, S. ; Manteuffel, T. ; McCormick, S. ; Ruge, J.: Adaptive algebraic multigrid. In: *SIAM J. Sci. Comput.* 27 (2006), Nr. 4, S. 1261–1286

[37] Brezina, M. ; Falgout, R. D. ; MacLachlan, S. ; Manteuffel, T. ; McCormick, S. F. ; Ruge, J.: Adaptive smoothed aggregation (αSA). In: *SIAM J. Sci. Comput.* 25 (2004), Nr. 6, S. 1896–1920

[38] Briggs, W. L. ; Henson, V. E. ; McCormick, S. F.: *A multigrid tutorial.* Second. Philadelphia, PA : Society for Industrial and Applied Mathematics (SIAM), 2000

[39] Briggs, W. L. ; McCormick, S.: Introduction. In: McCormick, S. (Hrsg.): *Multigrid Methods* Bd. 3. Philadelphia, PA : SIAM, 1987, S. 1–30

[40] Bunse, W. ; Bunse-Gerstner, A.: *Numerische lineare Algebra.* Stuttgart : Teubner Studienbücher: Mathematik, 1985

[41] Chartier, T. ; Falgout, R. D. ; Henson, V. E. ; Jones, J. ; Manteuffel, T. ; McCormick, S. ; Ruge, J. ; Vassilevski, P. S.: Spectral AMGe (ρAMGe). In: *SIAM J. Sci. Comput.* 25 (2003), Nr. 1, S. 1–26

[42] Cheney, C.C.: *Introduction to Approximation Theory.* 2. New York : McGraw Hill, 1966

[43] Cundy, N. ; Krieg, S. ; Arnold, G. ; Frommer, A. ; Lippert, Th. ; Schilling, K.: Numerical methods for the QCD overlap operator. IV. Hybrid Monte Carlo. In: *Comput. Phys. Comm.* 180 (2009), Nr. 1, S. 26–54

[44] DeGrand, T. ; DeTar, C.: *Lattice methods for Quantum Chromodynamics.* Pte. Ltd., Hackensack, NJ : World Scientific Publishing, 2006

[45] Edwards, R.: Numerical simulations in lattice gauge theories and statistical mechanics. In: *PhD thesis, New York University* (1999)

[46] Eijkhout, V. ; Vassilevski, P. S.: The role of the strengthened Cauchy-Bunyakovski-Schwarz inequality in multilevel methods. In: *SIAM Rev.* 33 (1991), Nr. 3, S. 405–419

[47] Engeli, M. ; Ginsburg, T. ; Rutishauser, H. ; Stiefel, E.: Refined iterative methods for computation of the solution and the eigenvalues of self-adjoint boundary value problems. In: *Mitt. Inst. Angew. Math. Zürich. No.* 8 (1959), S. 107

[48] Falgout, R. D. ; Vassilevski, P. S. ; Zikatanov, L. T.: On two-grid convergence estimates. In: *Numer. Linear Algebra Appl.* 12 (2005), Nr. 5-6, S. 471–494

[49] Fedorenko, R. P.: On the speed of convergence of an iteration process. In: *Ž. Vyčisl. Mat. i Mat. Fiz.* 4 (1964), S. 559–564

[50] Frommer, A. (Hrsg.) ; Lippert, T. (Hrsg.) ; Medeke, B. (Hrsg.) ; Schilling, K. (Hrsg.): *Lecture Notes in Computational Science and Engineering.* Bd. 15: *Numerical challenges in lattice quantum chromodynamics.* Berlin : Springer, 2000

[51] Gossler, F.: *Zwei-Level-Verfahren zur Lösung linearer Gleichungssysteme mit nichtsingulärer M-Matrix*, TU Berlin, Diplomarbeit, 2007

[52] Gossler, F. ; Nabben, R.: Multilevel methods, C.B.S. constants and spectral equivalence - in Bearbeitung. (2013)

[53] Gossler, F. ; Nabben, R.: On AMG based on Chebyshev polynomials and their connection to the AMGr method - in Bearbeitung. (2013)

[54] Grossmann, C. ; Roos, H. G.: *Numerische Behandlung partieller Differentialgleichungen, 3rd revised and expanded ed.* Wiesbaden : Teubner Studienbücher Mathematik, 2005

[55] Hackbusch, W.: *Springer Series in Computational Mathematics.* Bd. 4: *Multigrid methods and applications.* Berlin : Springer, 1985

[56] Hestenes, M. R. ; Stiefel, E.: Methods of conjugate gradients for solving linear systems. In: *J. Research Nat. Bur. Standards* 49 (1952), S. 409–436 (1953)

[57] Kahl, K.: *An algebraic multilevel approach for a model problem for disordered physical systems*, Bergische Universität Wuppertal, Diplomarbeit, 2006

[58] Kahl, K.: *Adaptive Algebraic Multigrid for Lattice QCD Computations*, Bergische Universität Wuppertal, Diss., 2009

[59] Köckler, Norbert: *Mehrgittermethoden. Ein Lehr- und Übungsbuch.* Wiesbaden : Springer Spektrum, Vieweg+Teubner, 2012

[60] Kraus, J. K.: An algebraic preconditioning method for M-matrices: linear versus nonlinear multilevel iteration. In: *Numer. Linear Algebra Appl.* 9 (2002), Nr. 8, S. 599–618

[61] Kraus, J. K. ; Pillwein, V. ; Zikatanov, L.: Algebraic multilevel iteration methods and the best approximation to $1/x$ in the uniform norm. In: *RICAM-Report* (2009), Nr. 17

[62] Larsson, S. ; Thomée, V.: *Partielle Differentialgleichungen und numerische Methoden.* Berlin : Springer, 2005

[63] Lazarov, R. ; Margenov, S.: C.B.S. constants for Graph-Laplacians and application to multilevel methods for discontinuous Galerkin systems. In: *Journal of Complexity* 23(4-6) (2007), S. 498–515

[64] Liesen, J. ; Mehrmann, V.: *Lineare Algebra. Ein Lehrbuch über die Theorie mit Blick auf die Praxis.* 1. Aufl. Wiesbaden : Vieweg+Teubner, 2011

[65] Liesen, J. ; Strakos, Z.: *Krylov subspace methods. Principle and analysis.* Oxford : Numerical Mathematics and Scientific Computation. Oxford University Press, 2012

[66] Livne, O. E.: Coarsening by compatible relaxation. In: *Numer. Linear Algebra Appl.* 11 (2004), Nr. 2-3, S. 205–227

[67] MacLachlan, S.: *Improving Robustness in Multiscale Methods*, University of Colorado at Boulder. Boulder, Colorado, Diss., 2004

[68] MacLachlan, S. ; Manteuffel, T. ; McCormick, S.: Adaptive reduction-based AMG. In: *Numer. Linear Algebra Appl.* 13 (2006), Nr. 8, S. 599–620

[69] MacLachlan, S. ; Oosterlee, C.: Algebraic multigrid solvers for complex-valued matrices. In: *SIAM J. Sci. Comput.* 30 (2008), Nr. 3, S. 1548–1571

[70] MacLachlan, S. ; Saad, Y.: Greedy coarsening strategies for nonsymmetric problems. In: *SIAM J. Sci. Comput.* 29 (2007), Nr. 5, S. 2115–2143

[71] MacLachlan, S. ; Saad, Y.: A greedy strategy for coarse-grid selection. In: *SIAM J. Sci. Comput.* 29 (2007), Nr. 5, S. 1825–1853

[72] Maitre, J. F. ; Musy, F.: The contraction number of a class of two-level methods; an exact evaluation for some finite element subspaces and model problems. In: *Multigrid methods (Cologne, 1981)* Bd. 960. Berlin : Springer, 1982, S. 535–544

[73] Margenov, S. D.: Upper bound of the constant in the strengthened C.B.S. inequality for FEM 2D elasticity equations. In: *Numer. Linear Algebra Appl.* 1 (1994), Nr. 1, S. 65–74

[74] Medeke, Bjorn: On Algebraic Multilevel Preconditioners in Lattice Gauge Theory. In: *Numerical challenges in lattice quantum chromodynamics* Bd. 15. Berlin : Springer, 2000, S. 99–114

[75] Meister, A.: *Numerik linearer Gleichungssysteme.* 2. Aufl. Wiesbaden : Vieweg, 2005

[76] Mense, C.: *Konvergenzanalyse von algebraischen Mehr-Gitter-Verfahren für M-Matrizen*, TU Berlin, Diss., 2008

[77] Mense, C. ; Nabben, R.: On algebraic multi-level methods for non-symmetric systems—comparison results. In: *Linear Algebra Appl.* 429 (2008), Nr. 10, S. 2567–2588

[78] Mense, C. ; Nabben, R.: On algebraic multilevel methods for non-symmetric systems—convergence results. In: *Electron. Trans. Numer. Anal.* 30 (2008), S. 323–345

[79] Notay, Y.: Using approximate inverses in algebraic multilevel methods. In: *Numer. Math.* 80 (1998), Nr. 3, S. 397–417

[80] Notay, Y.: Robust parameter-free algebraic multilevel preconditioning. In: *Numer. Linear Algebra Appl.* 9 (2002), Nr. 6-7, S. 409–428

[81] Notay, Y.: Algebraic multigrid and algebraic multilevel methods: a theoretical comparison. In: *Numer. Linear Algebra Appl.* 12 (2005), Nr. 5-6, S. 419–451

[82] Plato, R.: *Concise Numerical Mathematics.* Bd. 57. Providence, Rhode Island : American Mathematical Society,, 2003

[83] Pultarová, I.: The strengthened C.B.S. inequality constant for second order elliptic partial differential operator and for hierarchical bilinear finite element functions. In: *Appl. Math.* 50 (2005), Nr. 3, S. 323–329

[84] Ruge, J. W. ; Stüben, K.: Efficient solution of finite difference and finite element equations. In: *Multigrid methods for integral and differential equations (Bristol, 1983)* Bd. 3. New York : Oxford Univ. Press, 1985, S. 169–212

[85] Ruge, J. W. ; Stüben, K.: Algebraic multigrid. In: *Multigrid methods* Bd. 3. Philadelphia, PA : SIAM, 1987, S. 73–130

[86] Saad, Y.: *Iterative Methods for Sparse Linear Systems.* 2. Aufl. Philadelphia, PA : Society for Industrial and Applied Mathematics, 2003

[87] Schmüser, P.: *Feynman-Graphen und Eichtheorien für Experimentalphysiker.* 2. Aufl. Heidelberg : Springer, 1995

[88] Smirnov, V. I. ; Lebedev, N. A.: *Functions of a complex variable: Constructive theory.* Cambridge, Mass. : The M.I.T. Press, 1968

[89] Stüben, K.: A review of algebraic multigrid. In: *J. Comput. Appl. Math.* 128 (2001), Nr. 1-2, S. 281–309

[90] Trottenberg, U. ; Oosterlee, C. W. ; Schüller, A.: *Multigrid.* San Diego, CA : Academic Press Inc., 2001

[91] Tveito, A. ; Winther, R.: *Einführung in partielle Differentialgleichungen. Ein numerischer Zugang.* Berlin : Springer-Lehrbuch. Springer, 2002

[92] Vassilevski, P. S.: On two ways of stabilizing the hierarchical basis multilevel methods. In: *SIAM Rev.* 39 (1997), Nr. 1, S. 18–53

[93] Vassilevski, P. S.: *Multilevel block factorization preconditioners.* New York : Springer, 2008

[94] Waldmann, K. ; Stocker, U. M.: *Stochastische Modelle. Eine anwendungsorientierte Einführung.* Berlin : EMILA-stat. Medienreihe zur angewandten Statistik. Springer, 2004

[95] Walsh, J. L.: *Interpolation and approximation by rational functions in the complex domain.* Providence, R.I. : American Mathematical Society, 1965

[96] Weickert, J.: *Anisotropic diffusion in image processing.* Stuttgart : European Consortium for Mathematics in Industry. Teubner, 1998

[97] Wesseling, Pieter: *An introduction to multigrid methods.* Chichester : John Wiley & Sons Ltd., 1992 (Pure and Applied Mathematics (New York))

[98] Wilson, K. G.: Confinement of Quarks. In: *Phys. Rev. D* 10 (1974), Nr. 8, S. 2445–2459