Spektraläquivalente Vorkonditionierung lokaler Operatoren mittels $\mathcal{H}$-Matrizen und das Landau-Lifschitz-Modell als Anwendung

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Michael Bratsch

aus Dessau

Bonn, November 2012

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. Mario Bebendorf (betreuender Hochschullehrer)

2. Gutachter: Prof. Dr. Matthias Bollhöfer

Tag der Promotion: 11.03.2013
Erscheinungsjahr: 2013

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit Aspekten aus zwei verschiedenen Themengebieten. Zum einen betrachten wir hierarchische Matrizen zur Vorkonditionierung von Gleichungssystemen resultierend aus elliptischen Differentialgleichungen zweiter Ordnung. Zum anderen gehen wir auf die effiziente Berechnung minimaler Engergiepfade der mikromagnetischen Energie ein.

Hierarchische Matrizen zur Vorkonditionierung

Im ersten Teil dieser Arbeit beschäftigen wir uns mit der numerischen Lösung von elliptischen Differentialgleichungen zweiter Ordnung. Hierzu verwenden wir die Finite-Elemente-Methode und lösen die resultierenden Geichungssysteme mittels iterativer Methoden wie dem CG-Verfahren. Damit die Anzahl der Iterationsschritte nicht von der Kondition der Steifigkeitsmatrix abhängt, wird ein Vorkonditionierer verwendet. In dieser Arbeit beschränken wir uns dabei auf hierarchische Matrizen und betrachten die hierarchische LU- bzw. Cholesky-Zerlegung. Das Ziel dieser Arbeit war es, deren Vorkonditionierungseffekt genauer zu untersuchen und zu verbessern. Die originären Aspekte in diesem Teil der Arbeit umfassen:

(i) In Satz 4.7 wird ein spektraläquivalenter Vorkonditionierer präsentiert, bei dem nur der blockweise Fehler der hierarchischen Matrix-Approximation vorgegeben wird.

(ii) Es wird in Satz 3.3 bewiesen, dass in der hierarchischen Matrix-Arithmetik der blockweise Fehler der Gesamtapproximation durch die Aproximation der Schurkomplemente auf den einzelnen Blöcken bestimmt wird.

(iii) Das in hierarchischen Matrizen vorkommende Matrix-Schurkomplement wurde mit Hilfe des Poincaré-Steklov-Operators [KW04] dargestellt. Aus dieser Darstellung ergibt sich eine asymptotische Schranke des Matrix-Schurkomplements in der Spektralnorm, siehe Satz 3.10.

(iv) Die Approximationsgüte von hierarchischer LU- und Cholesky-Zerlegung wird in Satz 3.11 bzw. Satz 3.15 abgeschätzt. Numerische Resultate in Abschnitt 3.4 zeigen, dass die gewonnenen Abschätzungen scharf sind. Des Weiteren stellt die Abschätzung eine Verbesserung zu [Beb08, Theorem 4.35] dar.

(v) Erstmalig wird der blockweise Erhalt von Nebenbedingungen bei hierarchischen Matrizen eingesetzt. In Abschnitt 4.3 werden Methoden präsentiert, die diesen Ansatz effizient und numerisch stabil umsetzen.

(vi) Es wird in den Sätzen 4.8, 4.12 und 4.15 gezeigt, dass der Erhalt von Nebenbedingungen bei hierarchischen Matrizen mit konstanter Approximationsgenauigkeit zu einer Verbesserung der Ordnung des Vorkonditionierungseffektes von $\mathcal{O}(h^{-2})$ auf $\mathcal{O}(h^{-1})$ führt. Des Weiteren werden hinreichende blockweise Approximationsgenauigkeiten angegeben, die einen spektraläquivalenten Vorkonditionierer garantieren. Numerische Ergebnisse in Abschnitt 4.4 belegen die Abschätzungen.

(vii) In Satz 4.16 und Satz 4.17 wird gezeigt, dass die vorgestellten Vorkonditionierer bis auf logarithmische Terme einen linearen Speicherverbrauch in Bezug auf die Anzahl der Unbekannten besitzen.

Teile der genannten Resultate wurden oder werden veröffentlicht in:

> Bebendorf, M. ; Bollhöfer, M. ; Bratsch, M.: *Hierarchical matrix approximation with blockwise constraints.* Preprint No. 494, SFB 611, Universität Bonn (April 2011)
>
> Bebendorf, M. ; Bollhöfer, M. ; Bratsch, M.: *On the Spectral Equivalence of Hierarchical Matrix Preconditioners for Elliptic Problems.* geplante Veröffentlichung in 2012

Das Landau-Lifschitz-Modell und die Berechnung minimaler Energiepfade

Das Landau-Lifschitz-Modell bestimmt für eine gegebene Magnetisierung eines Objekts die mikromagnetische Energie. Dadurch erhält man eine Energielandschaft für verschiedene Konfigurationen der Magnetisierung. In diesem Teil der Arbeit konzentrieren wir uns auf die Berechnung minimaler Energiepfade. Diese verbinden zwei lokale Minima der mikromagnetischen Energie und minimieren die dazwischenliegende Energiebarriere.

Anspruchsvoll bei der numerischen Berechnung der mikromagnetischen Energie ist zum einen die Behandlung der punktweisen Normierung der Magnetisierung auf die Länge eins und zum anderen die Bestimmung der Streufeldenergie, dem nichtlokalen Anteil der mikromagnetischen Energie. Die wesentlichen Aspekte des zweiten Teils umfassen:

(i) Zur Berechnung der mikromagnetischen Energie wird die Finite-Elemente- und die Randelemente-Methode verwendet. Dabei wird ein in [GC07] präsentierten Ansatz zur Berechnung der Streufeldenergie umgesetzt. Desweiteren demonstriert diese Arbeit die effiziente Verwendung von hierarchischen Matrizen zur Lösung resultierender Gleichungssysteme.

(ii) Es werden Duffy-Transformationen zur Berechnung von singulären Integralen, resultierend aus der Randelente-Methode, angegeben.

(iii) Aktuelle Verfahren zur Berechnung des minimalen Energiepfades wurden implementiert, wie zum Beispiel ein Minimierungsverfahren von Alouges [Alo97, Alo01] oder die String-Methode [ERVE02, ERVE07]. Des Weiteren wird in Abschnitt 5.4 die Implementierung zur Minimierung der mikromagnetischen Energie an einem Referenzbeispiel getestet und ein minimaler Energiepfad berechnet.

Teilaspekte aus diesem Abschnitt werden veröffentlicht in:

> Bartels, S. ; Bebendorf, M. ; Bratsch, M.: *A fast and accurate numerical method for the computation of unstable micromagnetic configurations.* geplante Veröffentlichung in 2012

Danksagung

Zuallererst danke ich Prof. Dr. Mario Bebendorf für die Betreuung dieser Arbeit, insbesondere für die zahlreichen Anregungen und wissenschaftlichen Diskussionen. Des Weiteren bin ich Prof. Dr. Matthias Bollhöfer dankbar für die Übernahme des Zweitgutachtens und seine stets hilfreichen Hinweise. Mein Dank gilt außerdem Prof. Dr. Sören Bartels für die gute Zusammenarbeit. Danken möchte ich darüber hinaus allen Mitarbeitern des Instituts für Numerische Simulation, besonders meinen Kollegen aus der AG Bebendorf. Vor allem haben Christian Kuske und Raoul Venn diese Arbeit durch ihre vielen Korrekturvorschläge verbessert.

Ein großer Dank gilt auch meinen Eltern, die mich immer unterstützt haben. Weiterhin danke ich meiner Freundin, meinem Sohn und meinen Freunden für die gute Zeit.

Inhaltsverzeichnis

Einleitung

Motivation und Zielstellung

Diese Arbeit behandelt zwei Themengebiete. Erstens betrachten wir hierarchische Matrizen zur Vorkonditionierung von Gleichungssystemen resultierend aus der Diskretisierung von elliptischen Differentialgleichungen zweiter Ordnung. Zweitens präsentieren wir Methoden zur Berechnung eines minimalen Energiepfades bzgl. der mikromagnetischen Energie.

Das erste Teilgebiet betrifft die Lösung von elliptischen Differentialgleichungen, welche typischerweise im Zusammenhang mit stationären Problemen auftreten und oftmals Zustände minimaler Energie beschreiben. Da für diese Klasse von Problemstellungen nur in bestimmten Fällen (spezielle Gebiete oder Operatoren) analytische Lösungen bekannt sind, ist man gezwungen, numerische Methoden zur näherungsweisen Berechnung von Lösungen zu verwenden.

Ein Standard-Verfahren in diesem Bereich ist die Finite-Elemente-Methode (FE-Methode), die zum Beispiel im Ingenieurwesen verwendet wird. Dabei zerlegt man das Berechnungsgebiet in endlich viele Teilgebiete und definiert auf diesen Funktionen zur Approximation der Lösung. Somit wird das Problem durch eine endliche Anzahl von Parametern beschrieben wobei mit einer steigenden Anzahl der Unbekannten die Lösung angenähert wird.

Obwohl in den letzten Jahrzehnten die Rechenleistung und der verfügbare Speicherbedarf von Computern stetig zugenommen hat, ist es bis heute nur bedingt möglich, die bei der FE-Methode auftretenden Gleichungssysteme direkt zu lösen. Somit ist man auf schnelle Verfahren angewiesen, die diese bis auf eine vorgegebene Genauigkeit lösen. Dabei ist eine lineare Komplexität wünschenswert, welches es ermöglicht, die vorhandene Rechenleistung optimal auszunutzen und Problemgrößen zu rechnen, die ansonsten nicht denkbar wären.

Aus der FE-Methode resultieren lineare Gleichungssysteme, die wir mit iterativen Methoden wie dem CG-Verfahren lösen wollen. Diese Methoden haben den Nachteil, dass die Zahl der Iterationsschritte bei den von uns betrachteten Gleichungssystemen von der Anzahl der Unbekannten abhängt. Zwar bietet zum Beispiel das CG-Verfahren für symmetrisch, positiv-definite Matrizen sehr gute Eigenschaften hinsichtlich Stabilität und Konvergenzgeschwindigkeit, jedoch ist es möglich, diese für elliptische Problemstellungen weiter zu verbessern.

Eine Möglichkeit iterative Verfahren zu beschleunigen, ist die Verwendung eines Vorkonditionierers. Als Vorkonditionierer werden für elliptische Probleme zum Beispiel Mehrgitterverfahren, Gebietszerlegungs-Methoden oder ILU-Zerlegungen (Incomplete-LU) verwendet. Sie entsprechen einer Approxiation an die Inverse der Systemmatrix und können so die Konvergenzrate des Lösers beschleunigen.

Als Vorkonditionierer werden in dieser Arbeit die hierarchischen Matrizen verwendet. Mit Hilfe dieser Matrizen ist es möglich, mit quasi-linearem Aufwand (linear bis auf logarithmische Terme in Bezug auf die Anzahl der Unbekannten) die Iterationschritte des CG-Verfahrens durch eine Konstante nach oben zu beschränken. Dadurch ist man in der Lage, elliptische Differentialgleichungen effizient zu lösen. Vorteilhaft für praktische Anwendungen ist, dass man hierarchische Matrizen zum einen auf Grund ihrer Robustheit für eine Vielzahl von Problemstellungen anwenden kann und sie sich zum anderen als Black-box-Verfahren eignen, da sie im Wesentlichen nur die Approximationsgenauigkeit als Steuerparameter besitzen.

Die hierarchischen Matrizen stellen ein relativ neues Verfahren zur Vorkonditionierung dar. Somit ist es nicht verwunderlich, dass noch Lücken in der vorhandenen Theorie existieren. Aus diesem Grunde beschäftigen wir uns zum einen damit, bekannte Abschätzungen zu verbessern und zum anderen, die Wirkungsweise hierarchischer Matrizen genauer zu verstehen. Weiterhin wird mit dem Erhalt von Nebenbedingungen eine Methode zur Verbesserung der Vorkonditionierungseigenschaften präsentiert, welche anschließend analysiert wird.

Im zweiten Teil dieser Arbeit wenden wir uns dem Landau-Lifschitz-Modell zur Berechnung der mikromagnetische Energie zu. Sie ordnet einer Magnetisierung eines bestimmten Objektes eine Energie zu und kann zur Simulation von stationären mikromagnetischen Phänomenen genutzt werden. Da ferromagnetische Materialien in der Regel zwei stabile Zustände besitzen und zwischen diesen durch Einfluß eines magnetischen Feldes gewechselt werden kann, eignen sie sich zum Beispiel zur Speicherung von Informationen.

Ziel unserer Betrachtungen ist es, einen Weg aufzuzeigen, wie man mit Hilfe hierarchischer Matrizen einen minimalen Energiepfad bzgl. der mikromagnetischen Energie bestimmen kann, der zwei lokale Minima verbindet und die dazwischenliegende Energiebarriere minimiert. Minimale Energiepfade sind unter anderem hilfreich, um ungewollte Schaltvorgänge in Speichermedien vorherzusagen, die durch eine zunehmende Miniaturisierung ein nicht unerhebliches Problem darstellen.

Problematisch bei der numerischen Berechnung der minimalen Energiepfade ist zum einen die Nichtlokalität der Streufeldenergie, einem Teil der mikromagnetischen Energie und zum anderen die punktweise Normierung der Magnetisierung auf die Länge eins. In der vorliegenden Arbeit verwenden wir aktuelle Verfahren zur Behandlung dieser Probleme und präsentieren im Anschluss numerische Ergebnisse.

Aufbau

Die vorliegende Arbeit unterteilt sich in fünf Kapitel. Die ersten beiden behandeln Grundlagen, die als Voraussetzung für weitere Betrachtungen benötigt werden. So geht es in Kapitel 1 darum, elliptische Differentialgleichungen einzuführen und wichtige Eigenschaften zu erläutern. Außerdem gehen wir auf die FE-Methode und die Verwendung von vorkonditionierten iterativen Lösern ein. In Kapitel 2 führen wir die hierarchischen Matrizen ein und untersuchen wichtige Eigenschaften.

Das dritte Kapitel ist der hierarchischen LU- und Cholesky-Zerlegung von FE-Steifigkeitsmatrizen gewidmet. Zuerst beschreiben wir die Algorithmen anhand von Matrix-Schurkomplementen. Anschließend wird die Norm der auftretenden Schurkomplemente asymptotisch nach oben abgeschätzt und Aussagen über die Approximationsgüte der hierarchischen LU-Zerlegung bewiesen.

Im vierten Kapitel wird eine spezielle hierarchische Cholesky-Zerlegung vorgestellt, die bestimmte Nebenbedingungen erhält. Es wird gezeigt, dass der Vorkonditionierungseffekt gegenüber der gewöhnlichen hierarchischen Cholesky-Zerlegung davon wesentlich profitieren kann, obwohl sich die Anzahl der zur Berechnung nötigen Operationen von der Ordnung her nicht vergrößert. Des Weiteren wird gezeigt, wie diese Nebenbedingungen stabil und effizient implementiert werden können. Numerische Ergebnisse vergleichen die neue Methode mit bisher verwendeten.

Im fünften Kapitel geht es darum, die mikromagnetische Energie und entsprechende minimale Energiepfade mit Hilfe hierarchischer Matrizen zu berechnen. Dabei wird auf bestimmte Zwischenschritte

wie die Berechnung von singulären Integralen durch die Duffy-Transformation näher eingegangen. Im Anschluss werden wir unsere Implementierungen an einem Referenzbeispiel verifizieren. Ein minimaler Energiepfad wird für eine ausgewählte Geometrie berechnet.

Das letzte Kapitel beinhaltet denkbare Verbesserungen oder Fortführungen dieser Arbeit.

1 Grundlagen

Dieses Kapitel beinhaltet grundlegende Definitionen und Eigenschaften, die im Allgemeinen bekannt sind, jedoch für die späteren Betrachtungen zur Lösung elliptischer Differentialgleichungen benötigt werden. Im Wesentlichen führen wir zuerst geeignete Funktionenräume ein. Anschließend konkretisieren wir die zu betrachtenden elliptischen Differentialgleichungen, diskretisieren diese unter Verwendung der Finite-Elemente-Methode und geben hinreichende Bedingungen zur Beschränkung der Schrittzahlen ausgewählter iterativer Löser an.

Im Rahmen dieser Arbeit seien alle Gebiete $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, polygonale Lipschitz-Gebiete in zwei oder drei Raumdimensionen. Dabei ist ein Lipschitz-Gebiet ein beschränktes Gebiet, bei dem für alle $x \in \partial\Omega$ ein $r > 0$ und $\varphi \in C^{0,1}(\partial\Omega)$ existiert, so dass bis auf Koordinatentransformation,

$$\Omega \cap B_r(x) = \{y \in B_r(x) : y_d > \varphi(y_1, \ldots, y_{d-1})\}$$

gilt. In den Abbildungen 1 und 2 werden Beispiele für Lipschitz- und nicht Lipschitz-Gebiete gezeigt.

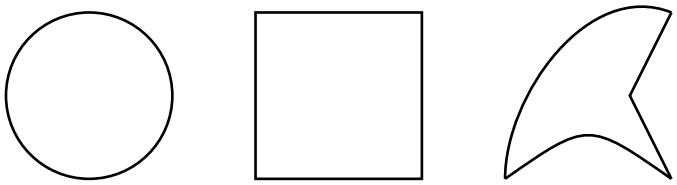


Abbildung 1: Beispiele für Lipschitz-Gebiete.

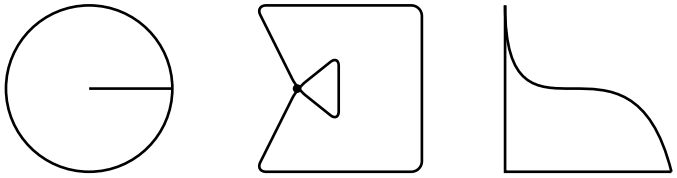


Abbildung 2: Beispiele für nicht Lipschitz-Gebiete.

1.1 Funktionalanalytische Hilfsmittel

Zur späteren Lösung von Differentialgleichungen führen wir die Sobolev-Räume ein, welche auf den Lebesgue-Räumen basieren. Anschließend definieren wir den Spur-Raum, um benötigte Betrachtungen von Funktionen auf dem Rand von Ω zu ermöglichen. Ein Standardwerk zu diesem Thema ist [Ada75].

1.1.1 Lebesgue- und Sobolev-Räume

Zur Definition des Lebesgue-Raums für Funktionen $u : \Omega \to \mathbb{R}$ benötigen wir

$$\|u\|_{L^p(\Omega)} := \left(\int_\Omega |u(x)|^p \, \mathrm{d}x \right)^{1/p}, \quad p \in [1, \infty),$$

und

$$\|u\|_{L^\infty(\Omega)} := \operatorname{ess\,sup}\{|u(x)| : x \in \Omega\},$$

wobei das wesentliche Supremum gegeben ist als $\operatorname{ess\,sup}\{|u(x)| : x \in \Omega\} := \inf\{\|u|_{\Omega \setminus S}\|_\infty : S \text{ Nullmenge}\}$. Im Falle von stetigen Funktionen stimmt das wesentliche Supremum mit der Supremumsnorm überein.

Somit können wir die Menge aller reellwertigen Funktionen $u : \Omega \to \mathbb{R}$ bestimmen, welche über Ω p-fach Lebesgue-integrierbar sind.

Definition 1.1 (Lebesgue-Raum)**.** *Sei $p \in [1, \infty]$. Dann ist der Lebesgue-Raum gegeben durch*

$$L^p(\Omega) = \tilde{L}^p(\Omega) \,/\, \sim,$$

wobei $u \sim v \Leftrightarrow \|u - v\|_{L^p(\Omega)} = 0$ und $\tilde{L}^p(\Omega) := \{u : \|u\|_{L^p(\Omega)} < \infty\}$.

Um den Begriff der Ableitungen zu verallgemeinern, benötigen wir den Begriff der lokal integrierbaren Funktionen.

Definition 1.2 (lokal integrierbare Funktionen)**.** *Die Menge der lokal integrierbaren Funktionen ist gegeben durch*

$$L^1_{\text{loc}}(\Omega) := \{u : u \in L^1(K) \text{ für alle kompakten Teilmengen } K \subset \Omega\}$$

und

$$L^p_{\text{loc}}(\Omega) := \{u : |u|^p \in L^1_{\text{loc}}(\Omega)\}.$$

Es folgt sofort, dass $L^1(\Omega) \subset L^1_{\text{loc}}(\Omega)$.

Sei $C_0^\infty(\Omega)$ die Menge aller Funktionen aus C^∞, die kompaktem Träger haben, so definieren wir den Begriff der schwachen Ableitung. Dabei verwenden wir für einen gegebenen Multiindex $\alpha \in \mathbb{N}^d$ die folgende Notation

$$\partial^\alpha u := \frac{\partial^{|\alpha|}}{\partial_{x_1}^{\alpha_1} \dots \partial_{x_d}^{\alpha_d}} u, \quad \text{für alle } u \in C^{|\alpha|}(\Omega).$$

Definition 1.3 (schwache Ableitung)**.** *Seien $u, v \in L^1_{\text{loc}}(\Omega)$ und $\alpha \in \mathbb{N}^d$. Falls*

$$(v, \varphi)_{L^2(\Omega)} = (-1)^{|\alpha|} (u, \partial^\alpha \varphi)_{L^2(\Omega)} \quad \text{für alle } \varphi \in C_0^\infty$$

gilt, dann ist v die α-te schwache Ableitung von u. Für v schreibt man auch $\partial^\alpha u$.

Ist $u \in C^{|\alpha|}(\Omega)$, so stimmen schwache und klassische Ableitung überein.

Mit diesen Hilfsmitteln können wir den Raum der Funktionen bestimmen, dessen k-te schwache Ableitung in $L^p(\Omega)$ ist.

Definition 1.4 (Sobolev-Raum)**.** *Sei $k \in \mathbb{N}_0$ und $p \in [1, \infty]$. Dann ist der Sobolev-Raum auf Ω definiert durch*

$$W_p^k(\Omega) := \big\{u : \partial^\alpha u \in L^p(\Omega) \text{ für alle } |\alpha| \leq k\big\}.$$

Als Spezialfall betrachten wir $H^k(\Omega) := W_2^k(\Omega)$.

Die zu W_p^k passende Norm bzw. Semi-Norm ist gegeben als

$$\|u\|_{W_p^k(\Omega)} := \Bigg(\sum_{|\alpha|\leq k} \|\partial^\alpha u\|_{L^p(\Omega)}^p\Bigg)^{1/p} \quad \text{und} \quad |u|_{W_p^k(\Omega)} := \Bigg(\sum_{|\alpha|= k} \|\partial^\alpha u\|_{L^p(\Omega)}^p\Bigg)^{1/p}.$$

Offensichtlich gilt, dass $L^p(\Omega) = W_p^0(\Omega)$ und $W_p^1(\Omega) \subset L^p(\Omega)$. Man erhält jedoch noch weitere Inklusionen mit Hilfe des Sobolevschen Einbettungssatz. Dies ist möglich, da die Existenz schwacher Ableitungen eine stärkere Integrabilitätsbedingung impliziert.

Definition 1.5. *Seien $(X, \|\cdot\|_X)$ und $(Y, \|\cdot\|_Y)$ normierte Vektorräume. Dann definieren wir die folgenden Eigenschaften:*

- *Eine Abbildung $f : X \to Y$ heißt kompakt, falls das Bild der abgeschlossenen X-Einheitskugel $f(\overline{B_1(0)})$ in Y kompakt ist;*
- *X ist stetig in Y eingebettet ($X \hookrightarrow Y$), falls $X \subset Y$ und die kanonische Injektion $X \to Y$ stetig ist;*
- *X ist kompakt in Y eingebettet ($X \overset{c}{\hookrightarrow} Y$), falls $X \subset Y$ und die kanonische Injektion $X \to Y$ kompakt ist.*

Aus Definition 1.5 wird sofort ersichtlich, dass eine kompakte Einbettung immer stetig ist. Weiterhin folgt aus $X \hookrightarrow Y$, dass eine Konstante $c > 0$ existiert, so dass $\|u\|_Y \leq c\|u\|_X$ für alle $u \in X$.

Satz 1.6 (Sobolevscher Einbettungssatz)**.** *Sei ein d-dimensionales Gebiet Ω mit Lipschitz-Rand gegeben. Dann gilt*

(i) *Sei $p < d$. Dann gilt $W_p^k(\Omega) \hookrightarrow W_q^{k-1}(\Omega)$ für alle $q \in \Big[1, \frac{pd}{d-p}\Big]$. Weiterhin gilt $W_p^k(\Omega) \overset{c}{\hookrightarrow} W_p^{k-1}(\Omega)$ für alle $q \in \Big[1, \frac{pd}{d-p}\Big)$.*

(ii) *Sei $p = d$. Dann gilt $W_p^k(\Omega) \hookrightarrow W_q^{k-1}(\Omega)$ für alle $q \in [1, \infty)$.*

(iii) *Sei $k > d/p$. Dann gilt $W_p^k(\Omega) \overset{c}{\hookrightarrow} C^\ell(\Omega)$ für alle $\ell \in \mathbb{N}$ mit $0 \leq \ell < k - d/p$.*

Beweis. Für einen Beweis siehe [Ada75]. □

In einer späteren Abschätzung benötigen wir den Normierungssatz von Sobolev. Dieser stellt eine Verallgemeinerung der Poincaré-Friedrich-Ungleichung dar.

Satz 1.7 (Normierungssatz von Sobolev)**.** *Sei $c \in \mathbb{R}$ und $f : H^1(\Omega) \to \mathbb{R}$ ein beschränktes lineares Funktional mit*

$$0 \leq |f(v)| \leq c_f \|v\|_{H^1(\Omega)} \quad \text{für alle } v \in H^1(\Omega).$$

Falls aus $f(c) = 0$ stets $c = 0$ folgt, so ist

$$\|v\|_{H^1(\Omega),f} := \{|f(v)|^2 + \|\nabla v\|^2_{L^2(\Omega)}\}^{1/2}$$

eine zu $\|v\|_{H^1(\Omega)}$ äquivalente Norm.

Beweis. Für einen Beweis siehe [Ste03]. □

Aufgrund von Skalierungsargumenten hängt die Konstante in Satz 1.7 für die Äquivalenz der Normen $\|v\|_{H^1(\Omega),f}$ und $\|v\|_{H^1(\Omega)}$ von dem Durchmesser des Gebietes Ω ab.

1.1.2 Spur-Raum

Um Funktionen auf dem Rand des Gebiets Ω zu betrachten, benötigt man geeignete Räume. So folgt aus $u \in H^1(\Omega)$ zum Beispiel nicht $u|_{\partial\Omega} \in H^1(\partial\Omega)$, da $\partial\Omega$ eine Nullmenge ist.

Eine geeignete Wahl ist der Raum $H^{1/2}(\partial\Omega)$ definiert mit Hilfe der Sobolev-Slobodeckij-Norm

$$H^{1/2}(\partial\Omega) := \big\{u \in L^2(\partial\Omega) : \|u\|_{H^{1/2}(\partial\Omega)} < \infty\big\},$$

wobei

$$\|u\|_{H^{1/2}(\partial\Omega)} := \left(\frac{1}{\operatorname{diam}\partial\Omega}\|u\|^2_{L^2(\Omega)} + \int_{\partial\Omega}\int_{\partial\Omega} \frac{|u(x)-u(y)|^2}{|x-y|^n}\,\mathrm{d}x\,\mathrm{d}y\right)^{1/2}.$$

Es folgt sofort, dass $H^{1/2}(\partial\Omega) \subset L^2(\partial\Omega)$.

Für Betrachtungen von $H^1(\Omega)$ Funktionen auf dem Rand definieren wir für ein $x \in \partial\Omega$ die Dirichlet-Spur $\gamma_0 : H^1(\Omega) \to H^{1/2}(\partial\Omega)$ gegeben durch

$$\gamma_0 u(x) := \lim_{\Omega \ni \tilde{x} \to x \in \partial\Omega} u(\tilde{x}).$$

Weiterhin lässt sich die $H^{1/2}$-Norm in der folgende Weise charakterisieren, vgl. [Ada75],

$$\|u\|_{H^{1/2}(\partial\Omega)} = \inf_{\substack{v \in H^1(\Omega) \\ \gamma_0(v) = u}} \|v\|_{H^1(\Omega)}.$$

Aus diesem Grunde wird $H^{1/2}$ als Spur-Raum von $H^1(\Omega)$ bezeichnet.

Die Dirichlet-Spur kann genutzt werden, um den Raum $H^1_0(\Omega)$ konstruktiv zu bestimmen

$$H^1_0(\Omega) := \{u \in H^1(\Omega) : \gamma_0 u = 0 \text{ auf } \partial\Omega\},$$

siehe [Ada75]. Mit Hilfe der Friedrichschen Ungleichung folgt, dass für Elemente aus $H^1_0(\Omega)$ die H^1-Norm äquivalent zur H^1-Semi-Norm ist.

1.2 Elliptische Differentialgleichungen

Probleme der Elektrostatik und der Kontinuumsmechanik können oftmals durch elliptische Differentialgleichungen beschrieben werden. Die Lösung u der folgenden Problemstellung

$$\Delta u = 0 \quad \text{in } \Omega \subset \mathbb{R}^d \tag{1.1}$$

mit gegebenen Randdaten für u auf $\partial\Omega$ besitzt die physikalische Interpretation eines Potentials. So beschreibt u das magnetische Potential bei verschwindender Stromdichte, das elektrische Potential unter Abwesenheit von elektrischer Ladung in Ω, oder auch das Geschwindigkeitspotential.

Obige Gleichung (1.1) wurde zuerst 1752 von L. Euler erwähnt, wird jedoch meist als Laplace-Gleichung bezeichnet, zurückgehend auf das fünfbändige Werk „Mécanique céleste“ (1799-1825) von P.S. Laplace.

Die wesentlichen Betrachtungen aus diesem Abschnitt sind zum Beispiel in [GT83, Wlo82, Neč67, Hac86] nachzulesen.

1.2.1 Allgemeine Problemstellung

Motiviert durch die Laplace-Gleichung (1.1) und deren Anwendungen betrachten wir elliptische Differentialgleichungen zweiter Ordnung mit Nullrandbedingungen,

$$\begin{aligned} \mathcal{L}u &= f \qquad \text{in } \Omega, \\ u &= 0 \quad \text{auf } \partial\Omega, \end{aligned} \tag{1.2}$$

wobei

$$\mathcal{L}u := -\mathrm{div}\big(C\nabla u\big), \tag{1.3}$$

mit $c_{ij} \in L^\infty(\Omega)$, $i,j = 1,\dots,d$. Als Annahme für die Elliptizität von $\mathcal{L}$ stellen wir, dass für die Eigenwerte $\lambda(x)$ der symmetrischen Matrix $C(x) \in \mathbb{R}^{d\times d}$ Konstanten $\lambda_{\mathcal{L}}, \Lambda_{\mathcal{L}}$ existieren, so dass in fast allen $x \in \Omega$

$$0 < \lambda_{\mathcal{L}} \le \lambda(x) \le \Lambda_{\mathcal{L}} \tag{1.4}$$

gilt. Es sei angemerkt, dass eine elliptische Differentialgleichung wie oben mit nicht Nullrandbedingungen durch Transformation auf den Typ (1.2) zurückgeführt werden kann.

Im Allgemeinen besitzt (1.2) für Lipschitz-Gebiete und Koeffizienten in L^∞ keine eindeutige Lösung in C^2. Durch praktische Anwendungen motiviert, ist es sinnvoll den Lösungsraum zu erweitern, indem wir die elliptische Differentialgleichung als variationelles Problem interpretieren. Es ergibt sich durch Umformung aus (1.2) eine Variationsgleichung mit der Bilinearform $a : V \times V \to \mathbb{R}$

$$a(u,v) = \int_\Omega \nabla v^T C \nabla u \,\mathrm{d}x \tag{1.5}$$

und dem Funktional l

$$l(v) = \int_\Omega fv \,\mathrm{d}x.$$

Aufgrund der Nullrandbedingungen setzen wir $V \equiv H_0^1(\Omega)$.

Somit erhält man die folgende Aufgabenstellung:

$$\text{finde ein } u \in H_0^1(\Omega), \text{ so dass } a(u,v) = l(v) \text{ für alle } v \in H_0^1(\Omega). \tag{1.6}$$

Ein wesentlicher Vorteil der Formulierung (1.6) gegenüber (1.2) ist, dass wir uns nicht auf Lösungen $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ beschränken müssen. Gleichwohl ist jede Lösung von (1.2) auch eine von (1.6) mit $u = 0$ auf $\partial\Omega$, siehe [Hac86, Bra07]. Es folgt weiterhin mit [GT83, Theorem 8.3], dass (1.6) für jedes $f \in L^2(\Omega)$ eine eindeutige Lösung $u \in H_0^1(\Omega)$ besitzt.

Wichtig für spätere Abschnitte ist die Stetigkeit der Bilinearform (1.5). Diese ergibt sich mit

Lemma 1.8 (Stetigkeit der Bilinearform)**.** *Seien $u, v \in H^1(\Omega)$. Dann gilt für die Bilinearform* (1.5)*, dass*

$$|a(u,v)| \leq \Lambda_{\mathcal{L}} |u|_{H^1(\Omega)} |v|_{H^1(\Omega)}. \tag{1.7}$$

Beweis. Die Bilinearform kann abgeschätzt werden mit Hilfe der Hölderschen Ungleichung, so dass

$$\left| \int_\Omega \nabla v^T C \nabla u \,\mathrm{d}x \right| \leq \int_\Omega |\nabla v||C\nabla u| \,\mathrm{d}x \leq \Lambda_{\mathcal{L}} |u|_{H^1(\Omega)} |v|_{H^1(\Omega)}.$$

□

Eine wesentliche Eigenschaft der von uns betrachteten Probleme ist, dass die Bilinearform (1.5) im folgenden Sinne nach unten abgeschätzt werden kann.

Lemma 1.9 (Koerzivität)**.** *Für alle $v \in H_0^1(\Omega)$ gilt, dass*

$$a(v,v) \geq \lambda_{\mathcal{L}} |v|^2_{H^1(\Omega)}. \tag{1.8}$$

Beweis. Mit Hilfe von Eigenschaft (1.4) erhält man

$$a(v,v) = \int_\Omega \nabla v^T C \nabla v \,\mathrm{d}x \geq \lambda_{\mathcal{L}} |v|^2_{H^1(\Omega)}.$$

□

1.2.2 Darstellungsformel

In diesem Abschnitt stellen wir harmonische Funktionen durch deren Randwerte dar. Wie sich später zeigt, kann dies für die Abschätzung des in $\mathcal{H}$-Matrizen vorkommenden Schurkomplements verwendet werden. Zuerst brauchen wir hierfür einige Hilfsmittel.

Durch partielle Integration bzgl. der Bilinearform (1.5) ergibt sich die erste Greensche Formel

$$a(u,v) = \int_\Omega \nabla u(y)^T C \nabla v(y)\,\mathrm{d}y = \int_\Omega (\mathcal{L}u)(y)v(y)\,\mathrm{d}y + \int_{\partial\Omega} \gamma_1 u(y)v(y)\,\mathrm{d}s_y$$

für hinreichend glatte u, v. Hierbei bezeichnet γ_1 die Konormalenableitung auf $\partial\Omega$ mit

$$\gamma_1 u(y) := \nu^T C \nabla u(y) \quad \text{für } y \in \partial\Omega,\ u \in H^2(\Omega),$$

wobei ν die äußere Normale bezeichnet. Allgemeiner betrachtet ist die Konormalenableitung eine stetige lineare Abbildung $\gamma_1 : H^1(\Omega) \to H^{-1/2}(\partial\Omega)$. Dabei ist $H^{-1/2}(\partial\Omega)$ der Dual-Raum des Spur-Raumes $H^{1/2}(\partial\Omega)$. Für Details siehe zum Beispiel [Cos88].

Aufgrund der Symmetrie der Bilinearform, $a(u,v) = a(v,u)$, erhält man die zweite Greensche Formel

$$\int_\Omega (\mathcal{L}v)(y)u(y)\,\mathrm{d}y = -\int_{\partial\Omega} \gamma_1 v(y)u(y)\,\mathrm{d}s_y + \int_{\partial\Omega} \gamma_1 u(y)v(y)\,\mathrm{d}s_y + \int_\Omega (\mathcal{L}u)(y)v(y)\,\mathrm{d}y. \tag{1.9}$$

Sei $\mathfrak{S}(x,y)$ die zu dem Operator (1.3) gehörige Fundamentallösung. Diese ist definiert als distributionelle Lösung von

$$\mathcal{L}_y \mathfrak{S}(x,y) = \delta(y-x) \quad \text{für alle } x, y \in \mathbb{R}^d,$$

wobei δ die Delta-Distribution bezeichnet. Somit ergibt sich, dass

$$\int_\Omega \mathcal{L}_y \mathfrak{S}(x,y)u(y)\,\mathrm{d}y = u(x) \quad \text{für alle } x \in \Omega.$$

Wählt man $v(y) := \mathfrak{S}(x,y)$ in (1.9), dann erhält man für $x \in \Omega$ die folgende Darstellungsformel,

$$u(x) = \int_{\partial\Omega} \gamma_1 u(y)\mathfrak{S}(x,y) - \gamma_{1,y}\mathfrak{S}(x,y)u(y)\,\mathrm{d}s_y + \int_\Omega f(y)\mathfrak{S}(x,y)\,\mathrm{d}y. \tag{1.10}$$

Wie wir in (1.10) sehen, so wird $u(x)$ in Ω nur durch dessen Randwerte dargestellt falls $f = 0$ in Ω gilt. In den nächsten Schritten geht es darum, die Abhängigkeit von der Konormalenableitung bzgl. u zu eliminieren.

Wir definieren die Greensche Funktion $\mathfrak{G}(x,y) : \Omega \times \Omega \setminus \{x \times x : x \in \Omega\} \to \mathbb{R}$ als Summe der Fundamentallösung und einer harmonischen Funktion $\mathfrak{H}$

$$\mathfrak{G}(x,y) = \mathfrak{S}(x,y) - \mathfrak{H}(x,y), \quad x, y \in \Omega,\ x \neq y.$$

Die harmonische Funktion $\mathfrak{H}(x,.) \in H^2(\Omega)$ sei eine Lösung von

$$\begin{aligned} \mathcal{L}_y \mathfrak{H}(x,y) &= 0 && \text{in } \Omega, \\ \mathfrak{H}(x,\cdot) &= \mathfrak{S}(x,\cdot) && \text{auf } \partial\Omega. \end{aligned} \tag{1.11}$$

Somit ergibt sich für $v(y) := \mathfrak{H}(x,y)$ in (1.9), dass

$$0 = \int_{\partial\Omega} \gamma_1 u(y)\mathfrak{H}(x,y) - \gamma_{1,y}\mathfrak{H}(x,y)u(y)\,\mathrm{d}s_y + \int_\Omega f(y)\mathfrak{H}(x,y)\,\mathrm{d}y. \tag{1.12}$$

Die Addition von (1.10) und (1.12) liefert aufgrund der Linearität der Konormalenableitung eine Darstellung von u in Ω mit Hilfe der Greenschen Funktion

$$u(x) = \int_\Omega f(y)\mathfrak{G}(x,y)\,\mathrm{d}y - \int_{\partial\Omega} \gamma_{1,y}\mathfrak{G}(x,y)u(y)\,\mathrm{d}s_y, \quad x \in \Omega. \tag{1.13}$$

Somit sind wir in der Lage, eine in Ω harmonische Funktion nur durch dessen Dirichlet-Randdaten und der entsprechenden Greenschen Funktion darzustellen.

Wir betrachten im Folgenden harmonische Funktionen, die durch den Fortsetzungsoperator

$$E_D : H^{1/2}(\partial\Omega) \to H^1(\Omega)$$

bzgl. der Bilinearform (1.5) mit den Dirichletranddaten $\varphi \in H^{1/2}(\partial\Omega)$ und $u = \varphi$ auf $\partial\Omega$ gegeben sind. Dieser setzt Randdaten in das Innere des Gebiets fort und ist bestimmt durch die folgende Aufgabenstellung:

$$\text{finde ein } u = E_D\varphi \in H^1(\Omega), \text{ so dass } a(u,v) = 0 \text{ für alle } v \in H_0^1(\Omega). \tag{1.14}$$

Im folgenden Lemma, wird $E_D\varphi$ explizit dargestellt durch dessen Randdaten gefaltet mit der Konormalenableitung der Greenschen Funktion welches für kommende Abschätzungen des $\mathcal{H}$-Matrix Schurkomplements benötigt wird.

Lemma 1.10. *Sei $\varphi \in H^{1/2}(\partial\Omega)$. Dann gilt, dass*

$$\gamma_{1,x}E_D\varphi(x) = -\int_{\partial\Omega} \gamma_{1,x}\gamma_{1,y}\mathfrak{G}(x,y)\varphi(y)\,\mathrm{d}s_y \quad \text{für alle } x \in \partial\Omega,$$

wobei $\gamma_1 E_D\varphi \in H^{-1/2}(\partial\Omega)$.

Beweis. Sei $\tilde{x} \in \Omega$. Da

$$0 = a(E_D\varphi, \mathfrak{G}(\tilde{x},\cdot)) = \int_\Omega f(y)\mathfrak{G}(\tilde{x},y)\,\mathrm{d}y,$$

folgt mit Hilfe der Darstellungsformel (1.13)

$$E_D\varphi(\tilde{x}) = -\int_{\partial\Omega} \gamma_{1,y}\mathfrak{G}(\tilde{x},y)\varphi(y)\,\mathrm{d}s_y.$$

Mit Hilfe von Lemma 6.9 aus [Ste03] ergibt sich die Behauptung. □

1.3 Finite-Elemente-Methode

Eines der gebräuchlichsten Verfahren zur numerischen Behandlung von elliptischen Differentialgleichungen ist die Finite-Elemente-Methode. Sie erlaubt die näherungsweise Lösung der Variationsformulie-

rung (1.6). Standardwerke zu diesem Thema sind zum Beispiel [BS02, Hac86, Bra07, SF73].

1.3.1 Ritz-Galerkin-Verfahren

Im Gegensatz zum Differenzenverfahren, bei dem der Differentialoperator diskretisiert wird, beruht das Ritz-Galerkin-Verfahren darauf, den unendlichdimensionalen Raum $H_0^1(\Omega)$ des Variationsproblems (1.6) durch eine Folge von passenden endlichdimensionalen Räumen V_h zu ersetzen. Dieser Ansatz geht zurück auf [Rit08].

Mit der Bedingung $V_h \subset H_0^1(\Omega)$ ergibt sich das folgende konforme Variationsproblem:

$$\text{finde ein } u_h \in V_h, \text{ so dass } a(u_h, v_h) = l(v_h) \text{ für alle } v_h \in V_h. \tag{1.15}$$

Wir bezeichnen h als Diskretisierungsparameter und Aufgabenstellung (1.15) als Ritz-Galerkin-Gleichung. In diesem Zusammenhang definieren wir die Ritz-Projektion $P_h : V \to V_h$, welche kein lokaler Operator ist, durch

$$P_h u \in V_h \text{ und } a(P_h u, v_h) = a(u, v_h) \quad \text{für alle } v_h \in V_h, u \in V. \tag{1.16}$$

Es folgt aus (1.16) für $u \in V_h$, dass $P_h u = u$ und somit $P_h^2 = P_h$.

Lemma 1.11. *Seien $u, v \in V$. Dann gilt*

$$a(P_h u, v) = a(u, P_h v).$$

Beweis. Da $P_h v \in V_h$, so folgt durch (1.16), dass $a(P_h u, P_h v) = a(u, P_h v)$. Weiterhin ergibt sich durch die Symmetrie der Bilinearform und Vertauschung von u und v, dass

$$a(P_h u, v) = a(v, P_h u) = a(P_h v, P_h u) = a(P_h u, P_h v).$$

Somit folgt die Behauptung. □

Eine weitere wichtige Eigenschaft, um die exakte und die approximative Lösung in Beziehung zu setzen, ist die Galerkin-Orthogonalität.

Lemma 1.12 (Galerkin-Orthogonalität)**.** *Sei $u \in H_0^1(\Omega)$ eine Lösung von* (1.6) *und $u_h \in V_h$ eine Lösung von* (1.15)*, so gilt*

$$a(u - u_h, v_h) = 0, \qquad v_h \in V_h.$$

Beweis. Da $V_h \subset H_0^1(\Omega)$, gilt $a(u, v_h) = l(v_h)$ und $a(u_h, v_h) = l(v_h)$. Aufgrund der Linearität von a folgt die Behauptung. □

Sei $\Phi_h := (\varphi_i)_{i \in I}$, mit der Indexmenge $I = \{1, \ldots, n\}$, eine Basis von V_h. So erhält man die

natürliche Injektion $\mathcal{J} : \mathbb{R}^n \to V_h$ durch

$$\mathcal{J}x := \sum_{i=1}^{n} x_i \varphi_i, \qquad x \in \mathbb{R}^n. \tag{1.17}$$

Bezüglich der Basis Φ_h ist der Diskretisierungsparameter h definiert als $h := \max_{i \in I} \operatorname{diam} X_i$ wobei $X_i := \text{supp}\varphi_i$, $i \in I$ die Träger der linearen Ansatzfunktionen sind. Für $t \subset I$ gilt analog $X_t := \cup_{i \in t} \text{supp}\varphi_i$.

Entlehnt aus der Kontinuumsmechanik bezeichnet im Folgenden $A \in \mathbb{R}^{n \times n}$ die Steifigkeitsmatrix

$$a_{ij} := a(\varphi_j, \varphi_i), \quad i, j \in I, \tag{1.18}$$

$M \in \mathbb{R}^{n \times n}$ die Massenmatrix

$$m_{ij} := (\varphi_i, \varphi_j)_{L^2(\Omega)}, \quad i, j \in I \tag{1.19}$$

und $b \in \mathbb{R}^n$ den Lastvektor

$$b_i := l(\varphi_i).$$

Aufgrund der eindeutigen Basisdarstellung existiert zu $u_h \in V_h$ ein eindeutig bestimmtes $x \in \mathbb{R}^n$, so dass $\mathcal{J}x = u_h$. Somit ist das Variationsproblem (1.15) äquivalent zu:

$$\text{finde ein } x \in \mathbb{R}^n, \text{ so dass } Ax = b. \tag{1.20}$$

Mit Hilfe der Symmetrie und Koerzivität der Bilinearform (1.8) folgt, dass die Matrix A symmetrisch, positiv-definit (s.p.d.) ist. Denn es gilt für ein $\mathcal{J}x = u_h \in V_h$, dass

$$\lambda_{\mathcal{L}} |u_h|^2_{H^1(\Omega)} \leq a(u_h, u_h) = \sum_{i,j \in I} x_i a(\varphi_i, \varphi_j) x_j = x^T A x, \tag{1.21}$$

wobei $x \neq 0 \Leftrightarrow u_h \neq 0$. Somit ist die eindeutige Lösbarkeit des Problems (1.20) gegeben.

Wir wenden uns der Konstruktion des endlichdimensionalen Raumes V_h zu. Hierfür benötigt man eine zulässige Zerlegung $\mathcal{T}_h$ von Ω. Dies bedeutet, dass für eine Zerlegung $\mathcal{T}_h = \{\tau_1, \ldots, \tau_m\}$ von Ω in Dreiecke ($d = 2$) oder Tetraeder ($d = 3$), wobei $\text{int}(\tau)$ das Innere von $\tau \in \mathcal{T}_h$ bezeichne, gilt:

(i) $\bar{\Omega} = \cup_{i=1}^{m} \tau_i$ und $\text{int}(\tau_i) \cap \text{int}(\tau_j) = \emptyset$ für $i \neq j$;

(ii) für alle $\tau_i, \tau_j \in \mathcal{T}_h$ mit $i \neq j$, ist $\tau_i \cap \tau_j$ entweder eine Seite, eine Kante oder ein Punkt von τ_i und τ_j oder $\tau_i \cap \tau_j$ ist leer;

(iii) für alle $\tau \in \mathcal{T}_h$ ist $\tau \cap \partial\Omega$ eine Seite, eine Kante, ein Punkt oder ist leer.

Die Elemente von $\mathcal{T}_h$ werden als Finite Elemente (FE) bezeichnet. Des Weiteren ist $\mathcal{N}_h$ die Menge aller Knoten in $\mathcal{T}_h$.

Um später die natürliche Injektion abschätzen zu können, werden wir im Folgenden immer von einer quasi-uniformen Zerlegung $\mathcal{T}_h$ ausgehen. Dies bedeutet, dass eine Konstante $c > 0$ existiert, so dass

jedes $\tau \in \mathcal{T}_h$ eine Kugel mit Radius $\rho_\tau \geq c\,\mathrm{diam}(\tau)$ enthält.

Für eine gegebene zulässige Zerlegung definieren wir die folgenden Spline-Räume $\mathcal{M}^1$ und $\mathcal{M}^1_{0,0}$.

$$\begin{aligned}\mathcal{M}^1 &:= \{u \in C(\bar{\Omega}) : u|_\tau \in \Pi_1^d \text{ für alle } \tau \in \mathcal{T}_h\}\\ \mathcal{M}^1_{0,0} &:= \mathcal{M}^1 \cap H_0^1(\Omega)\end{aligned}$$

Hierbei bezeichne Π_1^d alle d-dimensionalen Polynome ersten Grades.

Per Definition gilt, dass $\mathcal{M}^1_{0,0} \subset H_0^1(\Omega)$. Somit können wir den Raum der stückweise linearen Funktionen mit Nullrandbedingungen $\mathcal{M}^1_{0,0}$ als Ansatz- und Testraum für (1.15) wählen, d.h. $V_h \equiv \mathcal{M}^1_{0,0}$.

1.3.2 Eigenschaften der Massen- und Steifigkeitsmatrix

In späteren Abschnitten werden wiederholt verschiedene Eigenschaften der Massen- und Steifigkeitsmatrix verwendet. Grundlegend dabei ist, dass die Massenmatrix gutartig konditioniert ist, wohingegen die Kondition der Steifigkeitsmatrix von der Feinheit der Diskretisierung abhängt.

Massenmatrix: Aus der Definition (1.19) folgt direkt, dass die Massenmatrix symmetrisch, positv definit ist. Denn für ein $\mathcal{J}x = u_h \in V_h$ gilt, dass

$$\|u_h\|^2_{L^2(\Omega)} = (\mathcal{J}x, \mathcal{J}x)_{L^2(\Omega)} = x^T M x$$

wobei $u_h = 0 \Leftrightarrow x = 0$.

Sei $\mathcal{J}^* : V_h \to \mathbb{R}^I$ der adjungierte Operator bzgl. des Skalarproduktes $(\cdot,\cdot)_{L^2(\Omega)}$ definiert durch $(\mathcal{J}^* u, v)_{L^2(\Omega)} = (u, \mathcal{J}v)_{L^2(\Omega)}$ für alle $u, v \in L^2(\Omega)$. Dann folgt, dass

$$M = \mathcal{J}^*\mathcal{J}.$$

Wesentlich zur Abschätzung der Norm der Massenmatrix und deren Inverse ist die folgende Schranke. Für eine quasi-uniforme Triangulierung gilt nach [Hac86, Satz 8.8.1]

$$c_J \|\mathcal{J}x\|_{L^2(\Omega)} \leq h^{d/2}\|x\|_2 \leq \frac{1}{c_J}\|\mathcal{J}x\|_{L^2(\Omega)} \tag{1.22}$$

mit einer Konstante $c_J > 0$ welche unabhängig von h ist. Der Faktor $h^{d/2}$ in (1.22) ist ein Resultat der Skalierung der linearen Ansatzfunktionen auf $\max_{x\in\Omega} \varphi_i(x) = 1$ für alle $i \in I$.

Lemma 1.13. *Die Norm von M ist nach unten und oben beschränkt durch*

$$c_J^2 h^d \leq \|M\|_2 \leq c_J^{-2} h^d.$$

Beweis. Aus der Definition (1.19) und (1.22) ergibt sich, dass

$$\|M\|_2 = \sup_{x\in\mathbb{R}^I} \frac{x^T M x}{\|x\|_2^2} = \sup_{x\in\mathbb{R}^I} \frac{\|\mathcal{J}x\|_{L^2(\Omega)}^2}{\|x\|_2^2} \leq c_J^{-2} h^d.$$

Die untere Abschätzung folgt analog. □

Lemma 1.14. *Die Norm von M^{-1} ist nach unten und oben beschränkt durch*

$$c_J^2 h^{-d} \leq \|M^{-1}\|_2 \leq c_J^{-2} h^{-d}.$$

Beweis. Aus der Definition (1.19) und (1.22) ergibt sich, dass

$$\|M^{-1}\|_2 = \sup_{x\in\mathbb{R}^I} \frac{\|x\|_2^2}{x^T M x} = \sup_{x\in\mathbb{R}^I} \frac{\|x\|_2^2}{\|\mathcal{J}x\|_{L^2(\Omega)}^2} \leq c_J^{-2} h^{-d}.$$

Die untere Abschätzung folgt analog. □

Mit Hilfe der vorherigen beiden Lemmata können wir die Kondition der Massenmatrix beschränken.

Korollar 1.15 (Kondition der Massenmatrix)**.** *Die Kondition der Massenmatrix ist beschränkt durch eine Konstante.*

$$\mathrm{cond}_2(M) := \|M^{-1}\|_2 \|M\|_2 \leq c_J^{-4}$$

Beweis. Die Behauptung folgt direkt aus dem Lemma 1.13 und dem Lemma 1.14. □

Steifigkeitsmatrix: Wie bereits in (1.21) gezeigt wurde, ist die Steifigkeitsmatrix eine s.p.d. Matrix. Aus der Defintion (1.18) ergibt sich die folgende Darstellung

$$A = \mathcal{J}^* \mathcal{L} \mathcal{J}.$$

Um die Norm der Steifigkeitsmatrix zu beschränken, benötigen wir die inverse Ungleichung. Sei eine quasi-uniforme Triangulierung gegeben, so werden Elemente aus V_h in der stärkeren gegen die schwächere Norm abgeschätzt, d.h.

$$\|\nabla v\|_{L^2(\Omega)} \leq c_{\text{inv}} h^{-1} \|v\|_{L^2(\Omega)}, \quad v \in V_h, \tag{1.23}$$

siehe [Hac86, Satz 8.8.5]. Somit erhält man die folgenden Lemmata.

Lemma 1.16. *Die Norm von A ist beschränkt durch*

$$\|A\|_2 \leq c_A h^{-2} \|M\|_2,$$

wobei $c_A := c_{\text{inv}} \Lambda_{\mathcal{L}}$.

Beweis. Die Norm von A kann mit Hilfe von (1.7), (1.22) und (1.23) nach oben beschränkt werden durch

$$\|A\|_2 = \sup_{x\in\mathbb{R}^I} \frac{x^T A x}{\|x\|_2^2} \leq \|M\|_2 \sup_{x\in\mathbb{R}^I} \frac{a(\mathcal{J}x, \mathcal{J}x)}{\|\mathcal{J}x\|_{L^2(\Omega)}^2} \leq c_{\text{inv}} \Lambda_{\mathcal{L}} h^{-2} \|M\|_2.$$

□

Lemma 1.17. *Die Norm von A^{-1} ist beschränkt durch*

$$\|A^{-1}\|_2 \leq c_B \|M^{-1}\|_2,$$

wobei $c_B := c_{\text{PF}}^2 \lambda_{\mathcal{L}}^{-1} (\operatorname{diam}\Omega)^2$.

Beweis. Da $\mathcal{J}x \in H_0^1(\Omega)$, ergibt sich mit Hilfe von (1.22), (1.8) und der Poincaré-Friedrichsschen Ungleichung, dass

$$\begin{aligned}\|A^{-1}\|_2 &= \sup_{x\in\mathbb{R}^I} \frac{\|x\|_2^2}{x^T A x} \leq \|M^{-1}\|_2 \sup_{x\in\mathbb{R}^I} \frac{\|\mathcal{J}x\|_{L^2(\Omega)}^2}{a(\mathcal{J}x, \mathcal{J}x)} \\ &\leq \lambda_{\mathcal{L}}^{-1} \|M^{-1}\|_2 \sup_{x\in\mathbb{R}^I} \frac{\|\mathcal{J}x\|_{L^2(\Omega)}^2}{|\mathcal{J}x|_{H^1(\Omega)}^2} \leq c_{\text{PF}}^2 \lambda_{\mathcal{L}}^{-1} \|M^{-1}\|_2 (\operatorname{diam}\Omega)^2.\end{aligned}$$

□

Es folgt direkt die Kondition der Steifigkeitsmatrix.

Korollar 1.18 (Kondition der Steifigkeitsmatrix)**.** *Die Kondition der Steifigkeitsmatrix A ist beschränkt durch*

$$\operatorname{cond}_2(A) = \|A^{-1}\|_2 \|A\|_2 \leq c_A c_B c_J^{-4} h^{-2}.$$

Beweis. Der Beweis folgt direkt aus Lemma 1.16, Lemma 1.17 und Korollar 1.15. □

Vergleicht man Korollar 1.15 mit Korollar 1.18, so sieht man direkt, dass sich die Kondition der Steifigkeitsmatrix mit kleiner werdendem h verschlechtert, nicht jedoch die Kondition der Massenmatrix.

1.4 Iterative Löser und Vorkonditionierung

Zur Lösung großdimensionierter linearer Gleichungssysteme, wie etwa in (1.20), sind direkte Löser oftmals aufgrund ihrer hohen Komplexität nicht geeignet. So benötigt etwa ein Eliminationsverfahren welches die Bandbreite von $\mathcal{O}(n^{(d-1)/d})$ ausnutzt mindestens $\mathcal{O}(n^{(3d-2)/d})$ Operationen.

Im Gegensatz dazu profitieren iterative Löser wesentlich von der Schwachbesetztheit der Systemmatrix da diese auf einer schnellen Matrix-Vektor-Multiplikation beruhen. Wie sich jedoch zeigt, erhöhen sich die Iterationsschritte für eine vorgegebene Lösungsgenauigkeit durch eine größer werdende Kondition. Dies geschieht zum Beispiel bei einer steigenden Anzahl der Unbekannten, vgl. Korollar 1.18. Um die Iterationsschritte der iterativen Löser dennoch konstant zu halten, bietet sich ein Vorkonditionierer an.

1.4.1 Iterative Löser

In diesem Abschnitt wenden wir uns zuerst einfachen iterativen Lösern zu. Anschließend betrachten wir das PCG-Verfahren (preconditioned conjugate gradient), welches für die Klasse der s.p.d. Matrizen ein Standardverfahren darstellt. Die wesentlichen Resultate aus diesem Abschnitt können in [Axe94, Meu99, Gre97, QSS01] nachgelesen werden.

Einfache iterative Löser: Ein lineares Gleichungssystem der Form (1.20) kann durch folgende Rekursion näherungsweise gelöst werden:

$$x_{k+1} = x_k - \omega \tilde{A}^{-1}(Ax_k - b). \tag{1.24}$$

Dabei bezeichnet ω den Relaxationsparameter und $\tilde{A}^{-1}$ die exakte Inverse einer Approximation von A.

Für die Wahl von $\tilde{A}$ gibt es verschiedene Möglichkeiten. Die gebräuchlichsten sind:

- $\tilde{A} = \mathrm{Id}$, Richardson-Verfahren;
- $\tilde{A} = \operatorname{diag} A$, ω-Jacobi Verfahren;
- $\tilde{A} = \operatorname{diag} A + \omega L$, SOR-Verfahren (wobei L die untere Dreiecksmatrix von A ist).

Sei x die exakte Lösung des Gleichungssystems $Ax = b$, dann ergibt sich aus (1.24) folgende Beziehung

$$x_{k+1} - x = (\mathrm{Id} - \omega \tilde{A}^{-1} A)(x_k - x).$$

Die Konvergenzrate des vorkonditionierten Systems lässt sich somit bestimmen durch

$$\sup_{x_k \neq 0} \frac{\|x_{k+1} - x\|_2}{\|x_k - x\|_2} = \rho(\mathrm{Id} - \omega \tilde{A}^{-1} A). \tag{1.25}$$

Hierbei bezeichnet $\rho(\mathrm{Id} - \omega \tilde{A}^{-1} A)$ den Spektralradius, d.h. den betragsmäßig größten Eigenwert von $\mathrm{Id} - \omega \tilde{A}^{-1} A$.

Es ist ersichtlich, dass die betrachteten iterativen Löser nicht gegen die exakte Lösung konvergieren, falls $\rho(\mathrm{Id} - \omega \tilde{A}^{-1} A) \geq 1$ gilt.

Entscheidend für die Konvergenzrate ist die Wahl des Relaxationsparameters. Im Falle von s.p.d. Matrizen A und $\tilde{A}^{-1}$ ist der optimale Relaxationsparamter gegeben als

$$\omega_{\text{opt}} = \frac{2}{\lambda_{\min}(\tilde{A}^{-1} A) + \lambda_{\max}(\tilde{A}^{-1} A)},$$

siehe [Axe94, Meu99, Gre97]. Wichtig dabei ist, dass $\tilde{A}^{-1} A$ ausschließlich positive, reelle Eigenwerte besitzt. Somit ergibt sich für die Konvergenzrate

$$\rho(\mathrm{Id} - \omega_{\text{opt}} \tilde{A}^{-1} A) = 1 - \omega_{\text{opt}} \lambda_{\min}(\tilde{A}^{-1} A) = \frac{\lambda_{\max}(\tilde{A}^{-1} A) - \lambda_{\min}(\tilde{A}^{-1} A)}{\lambda_{\max}(\tilde{A}^{-1} A) + \lambda_{\min}(\tilde{A}^{-1} A)} = \frac{\kappa(\tilde{A}^{-1} A) - 1}{\kappa(\tilde{A}^{-1} A) + 1},$$

mit der Konditionszahl $\kappa(\tilde{A}^{-1} A) := \lambda_{\max}(\tilde{A}^{-1} A) / \lambda_{\min}(\tilde{A}^{-1} A)$. Wählt man also für (1.25) den opti-

malen Relaxationsparameter, erhält man nach k Schritten

$$\|x_k - x\|_2 \leq \left(\frac{\kappa(\tilde{A}^{-1}A) - 1}{\kappa(\tilde{A}^{-1}A) + 1}\right)^k \|x_0 - x\|_2, \qquad \text{für } k = 1, 2, \ldots. \tag{1.26}$$

Mit Hilfe der Abschätzung (1.26) sieht man, dass wir an einem Vorkonditionierer $\tilde{A}$ interessiert sind, bei dem der Ausdruck $\kappa(\tilde{A}^{-1}A)$ möglichst nahe bei eins ist.

Weiterhin wird aus der Darstellung (1.24) ersichtlich, dass es genügt, die Anwendung von $\tilde{A}^{-1}$ auf einen Vektor schnell zu berechnen. Im Allgemeinen ist es nicht nötig die Matrix $\tilde{A}^{-1}$ aufzustellen.

Von praktischem Interesse ist die Anzahl der benötigten Schritte k_ε, um eine Genauigkeit von $\|x_k - x\|_2 \leq \varepsilon\|x_0 - x\|_2$ zu erreichen. Da $1 - 1/y < \log y$ für $y > 1$ gilt, erhält man durch Umformung von (1.26),

$$k_\varepsilon \leq \kappa(\tilde{A}^{-1}A)|\log \varepsilon|.$$

Wichtig bei dieser Schranke ist, dass die Konditionszahl linear in die Anzahl der Schritte k_ε eingeht.

Das PCG-Verfahren: Ein effizientes, iteratives Verfahren für groß-dimensionierte, s.p.d. Matrizen ist das PCG-Verfahren. Es handelt sich dabei um ein Krylov-Unterraum-Verfahren und wurde ursprünglich von Hestenes und Stiefel entwickelt, siehe [HS52]. In unserem Fall verwenden wir es zur Lösung des linearen Gleichungssystems (1.20) mit der s.p.d. Steifigkeitsmatrix A.

Der Algorithmus des PCG-Verfahrens kann mit Hilfe eines s.p.d.Vorkonditionierers $\tilde{A}^{-1}$ in folgender Weise implementiert werden:

Sei $x_0 \in \mathbb{R}^n$ ein beliebiger Startvektor.
Berechne $r_0 = b - Ax_0,\ d_0 = h_0 = \tilde{A}^{-1}r_0$
for k = 0,1,...
  $x_{k+1} = x_k + \alpha_k d_k,\ \alpha_k = \frac{r_k^T h_k}{d_k^T A d_k}$
  $r_{k+1} = r_k - \alpha_k A d_k$
  if $\|r_{k+1}\|_2 < \varepsilon$ **then stop**
  $h_{k+1} = \tilde{A}^{-1} r_{k+1}$
  $d_{k+1} = h_{k+1} + \beta_k d_k,\ \beta_k = \frac{r_{k+1}^T h_{k+1}}{r_k^T h_k}$

Algorithmus 4.1: Das PCG-Verfahren

Das Verfahren bricht ab, wenn die Norm des Residuums $r_{k+1} = b - Ax_{k+1}$ unter eine Schranke von ε fällt. Somit ist x_{k+1} unsere Näherungslösung an $x = A^{-1}b$.

Eine übliche Abschätzung in der Energienorm für die Konvergenz des PCG-Verfahren ist gegeben durch

$$\|x_k - x\|_A \leq 2\left(\frac{\sqrt{\kappa(\tilde{A}^{-1}A)} - 1}{\sqrt{\kappa(\tilde{A}^{-1}A)} + 1}\right)^k \|x_0 - x\|_A,$$

siehe zum Beispiel [Axe94, Meu99, Gre97].

Für die Anzahl k_ε an Iterationsschritten zum Erreichen einer relativen Genauigkeit ε in der Energienorm ergibt sich somit:

$$k_\varepsilon \leq \sqrt{\kappa(\tilde{A}^{-1}A)}|\log(\varepsilon/2)|.$$

Im Gegensatz zu einfachen iterativen Lösern geht bei dieser Schranke lediglich die Wurzel der Konditionszahl des vorkonditionierten Systems ein.

Insgesamt können wir für die Verwendung von iterativen Verfahren zwei wesentliche Bedingungen an unseren Vorkonditionierer stellen:

- schnelle Anwendbarkeit von $\tilde{A}^{-1}$ auf einen Vektor;
- Beschränkung von $\kappa(\tilde{A}^{-1}A)$.

1.4.2 Vorkonditionierung

In diesem Abschnitt betrachten wir hinreichende Bedingungen, um die Konditionszahl $\kappa(\tilde{A}^{-1}A)$ des vorkonditionierten Systems zu beschränken. Dies wird benötigt, um die Konvergenzrate der iterativen Löser aus Abschnitt 1.4.1 abzuschätzen.

Sei $A \in \mathbb{R}^{I\times I}$ eine s.p.d. Matrix und $\tilde{A} \in \mathbb{R}^{I\times I}$ eine symmetrische Approximation im folgenden Sinne

$$\|A - \tilde{A}\|_2 \leq \varepsilon\|A\|_2, \qquad \varepsilon > 0. \tag{1.27}$$

Wie sich im nächsten Lemma zeigt, so ist $\tilde{A}$ bei hinreichender Approximationsgenauigkeit ebenfalls eine s.p.d. Matrix. Dies ist zum Beispiel für das Gelingen der Cholesky-Zerlegung wichtig.

Satz 1.19. *Sei $A \in \mathbb{R}^{I\times I}$ eine s.p.d. Matrix mit einer symmetrischen Approximation $\tilde{A} \in \mathbb{R}^{I\times I}$. Falls die Bedingung* (1.27) *und $\varepsilon\,\mathrm{cond}_2(A) < 1$ erfüllt sind, dann gilt die Abschätzung*

$$\|\operatorname{Id} - A^{-1/2}\tilde{A}A^{-1/2}\|_2 \leq \delta < 1. \tag{1.28}$$

Weiterhin folgt, dass $\tilde{A}$ s.p.d. ist.

Beweis. Unter Verwendung von $\|A^{-1/2}\|_2^2 = \|A^{-1}\|_2$ und $\delta := \varepsilon\,\mathrm{cond}_2(A)$ ergibt sich

$$\|\operatorname{Id} - A^{-1/2}\tilde{A}A^{-1/2}\|_2 \leq \|A^{-1}\|_2\|A - \tilde{A}\|_2 \leq \delta < 1$$

und somit folgt die erste Behauptung.

Sei z ein normierter Eigenvektor zu einem beliebigen Eigenwert λ von $A^{-1/2}\tilde{A}A^{-1/2}$ dann zeigt sich, dass

$$|1 - \lambda| = \|(\operatorname{Id} - A^{-1/2}\tilde{A}A^{-1/2})z\|_2 \leq \|\operatorname{Id} - A^{-1/2}\tilde{A}A^{-1/2}\|_2 \leq \delta.$$

Somit erhält man unter anderem für den kleinsten Eigenwert

$$\lambda_{\min}(A^{-1/2}\tilde{A}^{1/2}A^{-1/2}) \geq 1 - \delta.$$

Sei weiterhin $x \neq 0$ und $y := A^{1/2}x$. Da $A^{1/2}$ regulär ist, folgt $y \neq 0$, und es gilt

$$x^T\tilde{A}x = y^TA^{-1/2}\tilde{A}A^{-1/2}y^T \geq (1-\delta)\|y\|_2^2 > 0.$$

Hieraus folgt die zweite Behauptung. □

Es folgt ein Resultat zur Beschränkung der Konditionszahl des vorkonditionierten Systems. Hierbei benutzen wir die Neumannsche Reihe, welche eine Verallgemeinerung der geometrischen Reihe ist. Sei $T \in \mathbb{R}^{I\times I}$. Falls die Reihe $\sum_{k=0}^{\infty} T^k$ konvergiert, dann ist

$$(\mathrm{Id} - T)^{-1} = \sum_{k=0}^{\infty} T^k. \tag{1.29}$$

Somit ergibt sich der folgende Satz zur Beschränkung der Konditionszahl.

Satz 1.20. *Seien $A, \tilde{A} \in \mathbb{R}^{I\times I}$ s.p.d. Matrizen, die die Bedingung* (1.28) *erfüllen, dann gilt für die Konditionszahl des vorkonditionierten Systems*

$$\kappa(\tilde{A}^{-1}A) \leq \frac{1+\delta}{1-\delta}.$$

Beweis. Da $\tilde{A}^{-1}A$ nur positive Eigenwerte besitzt, folgt für die Konditionszahl

$$\begin{aligned}\kappa(\tilde{A}^{-1}A) &= \lambda_{\max}(A^{-1}\tilde{A})\lambda_{\max}(\tilde{A}^{-1}A) = \lambda_{\max}(A^{-1/2}\tilde{A}A^{-1/2})\lambda_{\max}(A^{1/2}\tilde{A}^{-1}A^{1/2}) \\ &= \|A^{-1/2}\tilde{A}A^{-1/2}\|_2\|A^{1/2}\tilde{A}^{-1}A^{1/2}\|_2.\end{aligned} \tag{1.30}$$

Die Behauptung ergibt sich durch Verwendung der Neumannschen Reihe, so dass

$$\kappa(\tilde{A}^{-1}A) \leq \left(1 + \|\,\mathrm{Id} - A^{-1/2}\tilde{A}A^{-1/2}\|_2\right)\left(\sum_{k=0}^{\infty}\|\,\mathrm{Id} - A^{-1/2}\tilde{A}A^{-1/2}\|_2^k\right) \leq \frac{1+\delta}{1-\delta}.$$

□

Im Folgenden heißt eine Matrix $\tilde{A}$ spektraläquivalent zu A, falls eine Konstante $c \geq 1$ existiert, so dass $\kappa(\tilde{A}^{-1}A) \leq c$ gilt.

Bemerkung 1.21. Die Inverse einer Approximation kann auch dazu genutzt werden, um das Gleichungssystem (1.20) direkt zu lösen und zwar mit relativer Genauigkeit

$$\|x - \tilde{A}^{-1}b\|_2 \leq \varepsilon'\|x\|_2.$$

Hierbei ist $\| \operatorname{Id} - \tilde{A}^{-1} A\|_2 \leq \varepsilon'$ eine hinreichende Bedingung, da

$$\|x - \tilde{A}^{-1} b\|_2 = \|(\operatorname{Id} - \tilde{A}^{-1} A)x)\|_2 \leq \|(\operatorname{Id} - \tilde{A}^{-1} A)\|_2 \|x\|_2 \leq \varepsilon' \|x\|_2.$$

Nachteilig bei dieser Methode ist, dass die Genauigkeit ε' in der Größenordnung des FE-Fehlers gewählt werden muss. Weiterhin erweist es sich in der Praxis oftmals als schwierig, einen Vorkonditionierer mit einer bestimmten Approximationsgenauigkeit zu berechnen. In der Regel wird die Güte eines Vorkonditionierers über einen Steuerparameter geregelt. In diesem sind verschiedene Konstanten enthalten, die meist a priori nicht bekannt sind.

Im nächsten Schritt werden wir unsere Wahl des Vorkonditionierers konkretisieren. Dabei verwenden wir das Format der hierarchischen Matrizen, um eine hierarchische LU- oder Cholesky-Zerlegung zu erstellen. Diese erfüllen Voraussetzung (1.27) für einen elliptischen Operator der Form (1.3) und erlauben eine schnelle Anwendung auf einen Vektor.

2 Hierarchische Matrizen

In vielen Anwendungen ist es nötig, vollbesetzte Matrizen effizient zu behandeln. Diese entstehen zum Beispiel durch Inversion oder LU-Faktorisierung von Finite-Elemente-Diskretisierung bzgl. elliptischer partieller Differentialoperatoren, siehe (1.20). Weiterhin treten vollbesetzte Matrizen bei der Diskretisierung nicht-lokaler Operatoren auf, zum Beispiel durch Finite-Elemente- oder Randelemente-Diskretisierungen von Integraloperatoren.

Zum Einsatz für die oben genannten Zwecke entwickelten Tyrtyshnikov [Tyr98] und Hackbusch et al. [Hac99, HK00] die Mosaic-Skeleton-Methode bzw. die hierarchischen Matrizen ($\mathcal{H}$-Matrizen). ähnliche Ansätze sind Tree-Code [BH86], schnelle Multipol-Methoden [GR87, GR97] und das Panel-Clustering [HN89], welche auf die schnelle Multiplikation mit einem Vektor ausgerichtet sind.

Sämtliche wesentlichen Eigenschaften aus diesem Abschnitt bzgl. $\mathcal{H}$-Matrizen sind nachzulesen in den Übersichtswerken [Hac09, Beb08].

2.1 Niedrigrangmatrizen und Matrix-Paritionierung

Die hierarchischen Matrizen basieren auf zwei Prinzipien, einer Unterteilung der Gesamtmatrix in geeignete Blöcke und der blockweisen Niedrigrangapproximation.

2.1.1 Niedrigrangapproximation

Die Menge der Matrizen mit höchstens Rang k sei gegeben durch

$$\mathbb{R}_k^{m\times n} := \{B \in \mathbb{R}^{m\times n} : \operatorname{rank} B \leq k\}.$$

Folgender elementarer Satz ohne Beweis erläutert den Zusammenhang zur äußeren Produktform.

Satz 2.1. *Eine Matrix $B \in \mathbb{R}^{m\times n}$, gehört zu $\mathbb{R}_k^{m\times n}$ genau dann, wenn Matrizen $U \in \mathbb{R}^{m\times k}$ und $V \in \mathbb{R}^{n\times k}$ existieren, so dass*

$$B = UV^T. \tag{2.1}$$

Nun die Definition der Niedrigrangmatrizen.

Definition 2.2 (Niedrigrangmatrix)**.** *Eine Matrix $B \in \mathbb{R}_k^{m\times n}$ wird als Niedrigrangmatrix bezeichnet, falls*

$$k(m+n) < mn.$$

Somit hat eine Niedrigrangmatrix in der äußeren Produktform geringere Speicheranforderungen als bei einer eintragsweisen Speicherung.

Im Wesentlichen können sämtliche $\mathcal{H}$-Matrix-Operationen auf die Multiplikation und Addition von Niedrigrangmatrizen zurückgeführt werden. Bei der Multiplikation von zwei gegebenen Niedrigrang-

matrizen $B_1 = U_1 V_1^T \in \mathbb{R}_{k_1}^{m\times p}$ und $B_2 = U_2 V_2^T \in \mathbb{R}_{k_2}^{p\times n}$ gibt es zur Berechnung von $B_1 B_2 = \hat{U}\hat{V}^T$ folgende Optionen:

(i) $\hat{U} := U_1(V_1^T U_2)$ und $\hat{V} := V_2$, falls $k_1 \geq k_2$;

(ii) $\hat{U} := U_1$ und $\hat{V} := V_2(U_2^T V_1)$, falls $k_1 < k_2$.

Es ist sofort ersichtlich, dass $B_1 B_2 \in \mathbb{R}_{\min(k_1,k_2)}^{m\times n}$.

Weiterhin sieht man leicht, dass die Matrizen mit höchstens Rang k keinen linearen Raum bilden, da sich bei der Addition im Allgemeinen der Rang erhöht.

$$B_1 + B_2 = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^T =: UV^T \in \mathbb{R}_{k_1+k_2}^{m\times n}. \tag{2.2}$$

Um dieser Rangerhöhung entgegenzuwirken, kann eine approximative Addition verwendet werden. Dabei geht man in der Regel in der folgenden Weise vor. Man berechnet eine QR-Zerlegung von $U = Q_U R_U$ und $V = Q_V R_V$ und zerlegt das Produkt

$$R_U R_V^T = W \Sigma Z^T \in \mathbb{R}^{k'\times k'}, \quad \Sigma := \operatorname{diag}(\sigma_1, \ldots, \sigma_{k'}), \tag{2.3}$$

mittels Singulärwertzerlegung (SVD, singular value decomposition), wobei $k' := k_1 + k_2$. Somit erhält man

$$UV^T = Q_U R_U R_V^T Q_V^T = (Q_U W)\Sigma(Q_V Z)^T. \tag{2.4}$$

Als Approximation an UV^T erhalten wir die Rang-k-Matrix $\tilde{U}\tilde{V}^T$ mit

$$\tilde{U} := (Q_U W)\Sigma_k, \quad \tilde{V} := Q_V Z \tag{2.5}$$

und $\Sigma_k := \operatorname{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) \in \mathbb{R}^{k'\times k'}$. Die Approximation $\tilde{U}\tilde{V}^T \approx B_1 + B_2$ kann mit linearer Komplexität $\mathcal{O}\big((k')^2(m+n)\big)$ berechnet werden.

Die Rang-k-Matrix $\tilde{U}\tilde{V}^T$ ist eine Bestapproximation in $\mathbb{R}_k^{m\times n}$ an UV^T, denn mit Hilfe von Mirsky [Mir60] folgt

$$\min_{M\in\mathbb{R}_k^{m\times n}} \|UV^T - M\|_2 = \|UV^T - \tilde{U}\tilde{V}^T\|_2 = \|\Sigma - \Sigma_k\|_2 = \sigma_{k+1}.$$

Somit muss, um eine relative Genauigkeit $\varepsilon > 0$ zu garantieren, der Rang folgendermaßen gewählt werden

$$k(\varepsilon) := \min\{k \in \mathbb{N} : \sigma_{k+1} < \varepsilon\sigma_1\}. \tag{2.6}$$

Zur Approximation von Niedrigrangmatrizen werden im Wesentlichen die QR-Zerlegung und die SVD verwendet, welche sich durch ihre numerische Stabilität auszeichnen. Somit bilden sie ein solides Fundament für eine robuste Approximation.

2.1.2 Partitionierung

Im Allgemeinen kann man bei den von uns betrachteten Beispielen nicht davon ausgehen, dass die gesamte Matrix approximiert werden kann. Aus diesem Grunde partitionieren wir die Matrix $A \in \mathbb{R}^{I \times J}$.

Definition 2.3 (Partition). *Sei $\mathcal{P}$ Teilmenge der Potenzmenge von $I \times J$. Dann wird $\mathcal{P}$ als Partition von $I \times J$ bezeichnet, wenn*

$$I \times J = \cup_{b \in \mathcal{P}} b$$

und wenn aus $b_1 \cap b_2 \neq \emptyset$, $b_1 = b_2$ für alle $b_1, b_2 \in \mathcal{P}$ folgt.

Im Folgenden bezeichnet A_b die Einschränkung der Matrix A auf einen Block $b \in \mathcal{P}$. Um die Approximierbarkeit von Blöcken mit Niedrigrangmatrizen zu gewährleisten, folgt nun eine abstrakte Zulässigkeitsbedingung.

Definition 2.4 (abstrakte Zulässigkeitsbedingung). *Ein Block $b = t \times s \in \mathcal{P}$ heißt zulässig, falls folgende Bedingungen gelten:*

- *die Singulärwerte von A_b fallen exponentiell ab;*
- *jede Teilmenge $b' \subset b$ ist zulässig;*
- *die Zulässigkeit kann in $\mathcal{O}(|t| + |s|)$ Operationen geprüft werden.*

Für Problemstellungen resultierend aus elliptischen Operatoren, siehe zum Beispiel (1.20), kann eine konkrete Zulässigkeitsbedingung angegeben werden. Ein Block $A_{t \times s}$, $t \times s \in \mathcal{P}$, ist zulässig, falls

$$\max(\operatorname{diam} X_t, \operatorname{diam} X_s) \leq \eta \operatorname{dist}(X_t, X_s), \quad \eta > 0, \tag{2.7}$$

gilt. Für Betrachtungen zur Approximierbarkeit von zulässigen Blöcken im Sinne von Eigenschaft (2.7), siehe [BH03, Beb08].

Wesentlich ist, dass die Zulässigkeitsbedingung (2.7) für die Steifigkeitsmatrix, LU-Zerlegung und Inverse gleichermaßen gilt. Die resultierende einheitliche Partitionierung ermöglicht eine effiziente $\mathcal{H}$-Matrix-Arithmetik mit Blöcken im Niedrigrangformat.

Auf Basis der Zulässigkeitsbedingung ergibt sich die Definition der zulässigen Partition.

Definition 2.5 (zulässige Partition). *Eine Partition $\mathcal{P}$ heißt zulässig, wenn alle Blöcke $b = t \times s \in \mathcal{P}$ zulässig oder klein sind, i.e. $|t| < n_{\min}$ oder $|s| < n_{\min}$. Hierbei bezeichnet $n_{\min}$ die minimale Blockgröße.*

Im Folgenden werden wir nur von zulässigen Partitionen sprechen.

Für die Konstruktion einer zulässigen Partitionierung benötigt man in der Regel einen Clusterbaum.

Definition 2.6 (Clusterbaum). *Sei $I \subset \mathbb{N}$. Ein Baum $T_I = (V, E)$ mit Knoten V und Kanten E, wird als Clusterbaum bezeichnet, falls*

- *I die Wurzel von T_I ist;*

- $\emptyset \neq t = \cup_{t' \in \mathcal{S}(t)} t', \quad \forall t \in V \setminus \mathcal{L}(T_I)$;
- $|\mathcal{S}_I(t)| \geq 2, \quad \forall t \in V \setminus \mathcal{L}(T_I)$.

Es sei dabei die disjunkte Menge der Söhne definiert als $\mathcal{S}_I(t) := \{t' \in V : (t, t') \in E\}$ für alle $t \in V$ und die Menge der Blätter von T_I gegeben durch $\mathcal{L}(T_I) := \{t \in V : \mathcal{S}_I(t) = \emptyset\}$.

Die einzelnen Cluster werden für eine effiziente Implementierung lediglich soweit unterteilt, bis sie eine minimale Blockgröße $n_{\min}$ erreicht haben. Somit lassen sich Cache-Effekte in einer natürlichen Weise ausnutzen.

Das Level eines Blockes entspricht dessen Abstand zur Wurzel des Clusterbaums. Somit ist die Tiefe eines Baumes $L(T_I)$ gegeben durch das maximale Level der Blätter um eins erhöht. Weiterhin sei $T_I^{(\ell)}$ die Einschränkung von T_I auf Blöcke aus dem ℓ-ten Level. Wir bezeichnen einen Clusterbaum als balanciert, wenn für jedes Level $\ell = 0, \ldots, L(T_I) - 1$ Konstanten $c_{D_1}, c_{D_2} > 0$ existieren, so dass

$$c_{D_1} 2^{-\ell/d} \leq \operatorname{diam} X_t \leq c_{D_2} 2^{-\ell/d}, \quad t \in T_I^{(\ell)}, \tag{2.8}$$

gilt. Im Folgenden werden wir immer von balancierten Clusterbäumen ausgehen, zu dessen genauer Konstruktion siehe [Beb08].

Einen Spezialfall des Clusterbaums stellt der Blockclusterbaum dar.

Definition 2.7 (Blockclusterbaum)**.** *Ein Clusterbaum $T_{I\times J}$ wird als Blockclusterbaum bezeichnet, falls folgende Bedingung für die Sohnabbildung $\mathcal{S}_{I\times J}(t \times s)$ gilt:*

$$\mathcal{S}_{I\times J}(t \times s) = \begin{cases} \emptyset, & \text{falls } t \times s \text{ zulässig ist, } \mathcal{S}_I(t) = \emptyset \text{ oder } \mathcal{S}_J(s) = \emptyset, \\ \mathcal{S}_I(t) \times \mathcal{S}_J(s), & \text{sonst.} \end{cases}$$

Die Blätter von $T_{I\times J}$ entsprechen einer zulässigen Partition, $\mathcal{L}(T_{I\times J}) = \mathcal{P}$. Weiterhin bezeichnet $L(T_{I\times J})$ die Tiefe des Blockclusterbaumes.

Nun zur abschließenden Definition der $\mathcal{H}$-Matrizen (hierarchische Matrizen).

Definition 2.8 (Hierarchische Matrizen)**.** *Die Menge der hierarchischen Matrizen bzgl. der Partition $\mathcal{P}$ mit höchstens Rang k ist gegeben durch*

$$\mathcal{H}(\mathcal{P}, k) := \{A \in \mathbb{R}^{I\times J} : \operatorname{rank} A_b \leq k \text{ für alle } b \in \mathcal{P}\}.$$

In Abbildung 3 ist eine hierarchische Matrix mit entsprechender Rangverteilung abgebildet. Die roten Blöcke entsprechen vollbesetzten Matrizen.

Bezüglich der Menge der hierarchischen Matrizen können wichtige Matrixoperationen mit effizienter, approximativer Arithmetik definiert werden. Als Beispiele seien die Inversion, LU- und Cholesky-Zerlegung genannt. Auf die LU- und Cholesky-Zerlegung gehen wir in einem späteren Abschnitt genauer ein. Für weitere Details siehe [Hac09, Beb08].

Analog zu den Definitionen und Betrachtungen in diesem Abschnitt können die hierarchischen Matrizen auch über dem Körper der komplexen Zahlen definiert werden.

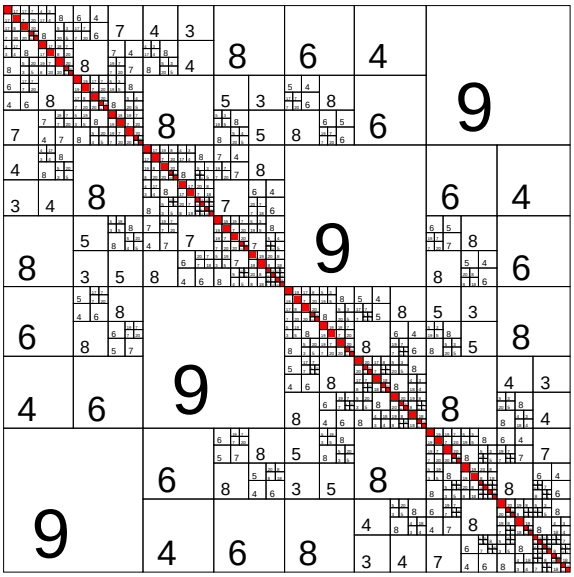


Abbildung 3: Eine $\mathcal{H}$-Matrix mit entsprechender Rangverteilung.

2.2 LU-Zerlegung und verschieden Eigenschaften

Die hierarchische LU- oder Cholesky-Zerlegung ist eine mögliche Wahl der Approximation an die Steifigkeitsmatrix in Abschnitt 1.4.2 zur Verwendung als Vorkonditionierer. Aus diesem Grunde betrachten wir zuerst die allgemeine Vorgehensweisen zur Berechnung solcher Zerlegungen und geben anschließend Abschätzungen zum Speicherbedarf und zur Laufzeit an.

In diesem Abschnitt gehen wir nur auf die $\mathcal{H}$-LU-Zerlegung ein, da sich sämtliche Aussagen direkt auf die $\mathcal{H}$-Cholesky-Zerlegung übertragen.

2.2.1 Allgemeine Beschreibung der hierarchischen LU-Zerlegung

Die Vorgehensweise der hierarchischen LU-Zerlegung lässt sich rekursiv beschreiben, vgl. [Beb08, Hac09]. Sei hierfür ein hierarchischer Block A_{tt}, $t \in T_I \backslash \mathcal{L}(T_I)$, gegeben. Dieser wird in der folgenden Weise zerlegt

$$A_{tt} = \begin{bmatrix} A_{t_1t_1} & A_{t_1t_2} \\ A_{t_2t_1} & A_{t_2t_2} \end{bmatrix} = \begin{bmatrix} L_{t_1t_1} & \\ L_{t_2t_1} & L_{t_2t_2} \end{bmatrix} \begin{bmatrix} U_{t_1t_1} & U_{t_1t_2} \\ & U_{t_2t_2} \end{bmatrix},$$

wobei $t_1, t_2 \in T_I$ die Söhne von t in T_I sind. Somit kann die LU-Zerlegung des Blockes A_{tt} auf die folgenden vier Probleme reduziert werden:

(i) Berechne $L_{t_1t_1}$ und $U_{t_1t_1}$ aus der LU-Zerlegung von $L_{t_1t_1}U_{t_1t_1} = A_{t_1t_1}$;

(ii) Berechne $U_{t_1t_2}$ aus $L_{t_1t_1}U_{t_1t_2} = A_{t_1t_2}$;

(iii) Berechne $L_{t_2t_1}$ aus $L_{t_2t_1}U_{t_1t_1} = A_{t_2t_1}$;

(iv) Berechne $L_{t_2t_2}$ und $U_{t_2t_2}$ aus der LU-Zerlegung von $L_{t_2t_2}U_{t_2t_2} = A_{t_2t_2} - L_{t_2t_1}U_{t_1t_2}$.

Für die Fälle (i) und (iv) sind zwei hierarchische LU-Zerlegungen der halben Größe vom selben Typ zu berechnen. Falls ein Diagonal-Block $t \times t \in \mathcal{L}(T_{I\times I})$ ein Blatt ist, so wenden wir die übliche LU-Zerlegung mit Pivotisierung an.

Um den Fall (ii) zu bestimmen, müssen wir eine Block-Vorwärtssubstitution der Form $L_{tt}B_{ts} = A_{ts}$ nach B_{ts} lösen, mit $t \times s \in T_{I\times I}$. Falls der Block $t \times s \in \mathcal{L}(T_{I\times I})$ ein Blatt ist, so lösen wir mit der gewöhnlichen Vorwärtssubstitution. Ansonsten führen wir eine Zerlegung in folgende Unterblöcke durch

$$\begin{bmatrix} L_{t_1t_1} & \\ L_{t_2t_1} & L_{t_2t_2} \end{bmatrix} \begin{bmatrix} B_{t_1s_1} & B_{t_1s_2} \\ B_{t_2s_1} & B_{t_2s_2} \end{bmatrix} = \begin{bmatrix} A_{t_1s_1} & A_{t_1s_2} \\ A_{t_2s_1} & A_{t_2s_2} \end{bmatrix},$$

wobei $t_1, t_2 \in T_I$ und $s_1, s_2 \in T_I$ die Söhne von t und s sind. Somit erhält man die entsprechenden vier Unterprobleme:

(i)* Berechne $B_{t_1s_1}$ aus $L_{t_1t_1}B_{t_1s_1} = A_{t_1s_1}$;

(ii)* Berechne $B_{t_1s_2}$ aus $L_{t_1t_1}B_{t_1s_2} = A_{t_1s_2}$;

(iii)* Berechne $B_{t_2s_1}$ aus $L_{t_2t_2}B_{t_2s_1} = A_{t_2s_1} - L_{t_2t_1}B_{t_1s_1}$;

(iv)* Berechne $B_{t_2s_2}$ aus $L_{t_2t_2}B_{t_2s_2} = A_{t_2s_2} - L_{t_2t_1}B_{t_1s_2}$.

Diese sind wiederum vom Typ (ii). Analog dazu können wir (iii) durch eine rekursive Block-Rückwärtssubstitution lösen.

2.2.2 Schwachbesetztheit, Speicherbedarf und Laufzeit

Um blockweise Abschätzungen einer Matrix auf die gesamte Matrix zu übertragen, bezeichnen wir die Anzahl der auftretenden Blöcke $t \times s \in \mathcal{P}$ je Blockzeile bzw. -spalte mit

$$c_{\text{sp}}^Z(t) := |\{s \subset J : t \times s \in \mathcal{P}\}| \qquad t \in T_I,$$
$$c_{\text{sp}}^S(s) := |\{t \subset I : t \times s \in \mathcal{P}\}| \qquad s \in T_J.$$

Wie in [GH03] gezeigt wurde, garantiert die Zulässigkeitsbedingung (2.7), dass $c_{\text{sp}}^Z(t)$ und $c_{\text{sp}}^S(s)$ jeweils für alle $t \in T_I$ und $s \in T_J$ durch eine Konstante unabhängig von I und J nach oben beschränkt sind. Somit ergibt sich die Schwachbesetztheitskonstante

$$c_{\text{sp}} := \max_{t\in T_I, s\in T_J} \{c_{\text{sp}}^Z(t), c_{\text{sp}}^S(s)\} \tag{2.9}$$

welche nicht von I und J abhängt.

Von großer Bedeutung in späteren Beweisen ist die Möglichkeit, die Spektralnorm durch die maximale Norm der einzelnen Blöcken abzuschätzen. Durch geschicktes Ausnutzen der levelweisen Tensorstruktur und der Schwachbesetztheitskonstante ist dies möglich, vgl. [Beb08, Hac09].

Für den Beweis benötigen wir folgendes Lemma.

Lemma 2.9. *Sei eine $r \times r$ Blockmatrix gegeben*

$$A = \begin{bmatrix} A_{11} & \dots & A_{1r} \\ \vdots & & \vdots \\ A_{r1} & \dots & A_{rr} \end{bmatrix},$$

mit $A_{ij} \in \mathbb{R}^{m_i \times n_j}$, $i, j = 1, \dots, r$. Dann gilt, dass

$$\max_{i,j=1,\dots,r} \|A_{ij}\|_2 \leq \|A\|_2 \leq \left(\max_{i=1,\dots,r} \sum_{j=1}^{r} \|A_{ij}\|_2 \right)^{1/2} \left(\max_{j=1,\dots,r} \sum_{i=1}^{r} \|A_{ij}\|_2 \right)^{1/2}.$$

Beweis. Die untere Abschätzung folgt, da die Spektralnorm eines Teilblocks einer Matrix durch die Spektralnorm der Gesamtmatrix abgeschätzt werden kann.

Für die obere Schranke sei $x = [x_1, \dots, x_r]^T \in \mathbb{R}^n$, wobei $x_j \in \mathbb{R}^{n_j}$, $j = 1, \dots, r$. Weiterhin definieren wir $\hat{a}_{ij} := \|A_{ij}\|_2$, $i, j = 1, \dots, r$ und $\hat{x}_j := \|x_j\|_2$, $j = 1, \dots, r$. Hieraus ergibt sich folgende Abschätzung

$$\|Ax\|_2^2 = \sum_{i=1}^{r} \| \sum_{j=1}^{r} A_{ij} x_j \|_2^2 \leq \sum_{i=1}^{r} \left(\sum_{j=1}^{r} \|A_{ij}\|_2 \|x_j\|_2 \right)^2 = \|\hat{A}\hat{x}\|_2^2.$$

Da $\|\hat{A}\|_2^2 \leq \|\hat{A}\|_1 \|\hat{A}\|_\infty$ gilt, folgt

$$\|\hat{A}\hat{x}\|_2^2 \leq \|\hat{A}\|_1 \|\hat{A}\|_\infty \|\hat{x}\|_2^2 = \|\hat{A}\|_1 \|\hat{A}\|_\infty \|x\|_2^2$$

und somit die Behauptung. □

Daraus ergibt sich mit dem Lemma 2.9 die Abschätzung der Spektralnorm einer $\mathcal{H}$-Matrix.

Satz 2.10. *Sei $A \in \mathcal{H}(\mathcal{P}, k)$, dann gilt*

$$\max_{b \in \mathcal{P}} \|A_b\|_2 \leq \|A\|_2 \leq c_{\text{sp}} L(T_I) \max_{b \in \mathcal{P}} \|A_b\|_2.$$

Beweis. Sei A_ℓ die Einschränkung von A auf Blöcke aus dem ℓ-ten Level.

$$A_\ell := \sum_{b \in \mathcal{P} \cap T_{I \times J}^{(\ell)}} A_b \quad \in \mathbb{R}^{I \times J}.$$

Da A_ℓ maximal c_{sp} viele Blöcke pro Zeile oder Spalte hat, siehe Definition (2.9), folgt mit Lemma 2.9, dass

$$\|A_\ell\|_2 \leq c_{\text{sp}} \max_{b \in \mathcal{P} \cap T_{I \times J}^{(\ell)}} \|A_b\|_2$$

und somit

$$\|A\|_2 \leq \sum_{\ell=0}^{L(T_I)-1} \|A_\ell\|_2 \leq c_{\text{sp}} \sum_{\ell=0}^{L(T_I)-1} \max_{b \in \mathcal{P} \cap T_{I\times I}^{(\ell)}} \|A_b\|_2 \leq c_{\text{sp}} L(T_I) \max_{b\in\mathcal{P}} \|A_b\|_2.$$

Die untere Abschätzung folgt aus Lemma 2.9. □

Wir untersuchen den Speicherverbrauch N_{sp} einer $\mathcal{H}$-Matrix $A \in \mathcal{H}(T_{I\times J}, k)$. Dieser lässt sich ebenfalls mit Hilfe der Schwachbesetztheitskonstante abschätzen.

Satz 2.11. *Sei die Schwachbesetztheitskonstante c_{sp} des Blockclusterbaumes $T_{I\times J}$ gegeben, dann kann der Speicherbedarf N_{sp} von $A \in \mathcal{H}(T_{I\times J}, k)$ abgeschätzt werden durch*

$$N_{sp} \leq c_{\text{sp}} \max(k, n_{\min})(L(T_I)|I| + L(T_J)|J|).$$

Beweis. Der Speicherbedarf lässt sich durch Summation über die einzelnen Blöcke in folgender Weise beschränken

$$\begin{aligned}
N_{sp} &\leq \sum_{t\times s\in\mathcal{P}} \max(k, n_{\min})(|t| + |s|) \\
&\leq \max(k, n_{\min}) \left(\sum_{t\in T_I} \sum_{\{s\subset J : t\times s\in\mathcal{P}\}} |t| + \sum_{s\in T_J} \sum_{\{t\subset I : t\times s\in\mathcal{P}\}} |s| \right) \\
&\leq c_{\text{sp}} \max(k, n_{\min}) \left(\sum_{t\in T_I} |t| + \sum_{s\in T_J} |s| \right) \\
&= c_{\text{sp}} \max(k, n_{\min})(L(T_I)|I| + L(T_J)|J|).
\end{aligned}$$

□

Sämtliche $\mathcal{H}$-Matrixoperationen, die wir benötigen, können in logarithmisch-linearer Zeit ausgeführt werden. Im folgenden Lemma sind einige davon aufgeführt.

Satz 2.12. *Sei $A \in \mathcal{H}(T_{I\times J}, k)$ dann haben folgende Operationen in der $\mathcal{H}$-Matrix Arithmetik eine Komplexität von*

- $\mathcal{O}\big(k(L(T_I)|I| + L(T_J)|J|)\big)$ *für eine Matrix-Vektor-Multiplikation oder für eine Vorwärts-, Rückwärtssubstitution bzgl. einer gegebenen $\mathcal{H}$-LU-Zerlegung;*
- $\mathcal{O}\big(k^2(L(T_I)^2|I| + L(T_J)^2|J|)\big)$*, zur Berechnung der $\mathcal{H}$-Inversen oder $\mathcal{H}$-LU-Zerlegung von A.*

Beweis. Für einen Beweis siehe [Beb08, Hac09]. □

Aufgrund der logarithmisch-linearen Komplexität eignen sich $\mathcal{H}$-Matrizen zum Vorkonditionieren von großdimensionierten Gleichungssystemen. Jedoch muss im weiteren Verlauf geklärt werden, ob deren Approximationseigenschaften für einen spektraläquivalenten Vorkonditionierer ausreichend sind.

3 Approximationsgüte der hierarchischen LU- und Cholesky-Zerlegung[1]

Wie bereits in Abschnitt 1.4.2 gezeigt wurde, ist die Güte der Approximation (1.27) entscheidend für die Konvergenz des CG-Verfahrens. In unserem Fall betrachten wir die $\mathcal{H}$-LU- und $\mathcal{H}$-Cholesky-Zerlegung als Approximation an die Steifigkeitsmatrix. Diese Zerlegungen gelten aufgrund ihrer Stabilitäts- und Laufzeiteigenschaften als Standardmethode zur Vorkonditionierung mittels hierarchischer Matrizen.

3.1 Hierarchische LU-Zerlegung

In diesem Abschnitt beschreiben wir mittels Matrix-Schurkomplemente eine rekursive Vorgehensweise der hierarchischen LU-Zerlegung. Eine allgemeine Beschreibung ist in Abschnitt 2.2.1 oder [Beb08, Hac09] vorhanden. Es sei angemerkt, dass sich sämtliche Aussagen aus diesem Abschnitt natürlich auch auf die $\mathcal{H}$-Cholesky-Zerlegung übertragen.

3.1.1 Das Auftreten von Schurkomplementen

Sei eine reguläre Matrix $A \in \mathbb{R}^{I\times I}$ und dessen exakte LU-Zerlegung $A = LU$ gegeben. Diese wird durch das Produkt der hierarchischen unteren und oberen Dreiecksmatrizen $\tilde{L}$ und $\tilde{U}$ approximiert.

$$A \approx \tilde{L}\tilde{U} =: \tilde{A}$$

Bezüglich der Mengen $t, s \subset I$ definieren wir das approximative Matrix-Schurkomplement eines Blocks $t \times s$ einer Matrix A. Es sei

$$S_a(t,s) := A_{ts} - \tilde{L}_{tI_1}\tilde{U}_{I_1 s} = A_{ts} - \tilde{A}_{tI_1}\tilde{A}^{-1}_{I_1 I_1}\tilde{A}_{I_1 s}, \tag{3.1}$$

wobei

$$I_1 := \{i \in I : i < \min(\min t, \min s)\}.$$

Das approximative Schurkomplement beinhaltet dabei alle Approximationen, die bis zu diesem Zeitpunkt der Zerlegung gemacht wurden. Zur besseren Veranschaulichung der Indexmengen siehe Abbildung 4.

Die Schurkomplement-Darstellung vereinfacht sich zu $S_a(t,s) = A_{ts}$, falls $I_1 = \emptyset$. Weiterhin ist zu beachten, dass $S_a(t,s)$ mit $t, s \in T_I$ im Allgemeinen keine Restriktion des Schurkomplements ist sondern eine Matrixfunktion.

Analog zur Beschreibung der $\mathcal{H}$-LU-Zerlegung in Abschnitt 2.2.1 wird die hierarchische Blockstruktur in natürlicher Weise rekursiv ausgenutzt. Dabei starten wir im ersten Schritt per Definition mit $A = S_a(I,I)$.

Sei ein approximatives Schurkomplementes $S_a(t,t)$, $t \times t \in T_{I\times I} \setminus \mathcal{P}$, wobei $t_1, t_2 \in T_I$ die Söhne

[1] Teile aus diesem Abschnitt werden veröffentlicht in [BBBb].

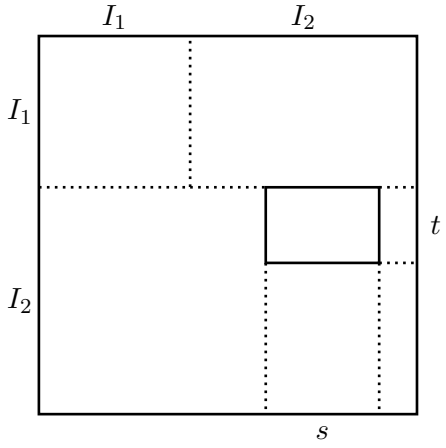


Abbildung 4: Indexmengen verwendet zur Definition des Schurkomplements bzgl. $t \times s \in \mathcal{P}$.

von $t \in T_I$ sind, gegeben und zerlegt in die entsprechenden Unterblöcke, so dass

$$S_a(t,t) = \begin{bmatrix} S_{t_1t_1} & S_{t_1t_2} \\ S_{t_2t_1} & S_{t_2t_2} \end{bmatrix} = \begin{bmatrix} \tilde{L}_{t_1t_1} & \\ \tilde{L}_{t_2t_1} & \tilde{L}_{t_2t_2} \end{bmatrix} \begin{bmatrix} \tilde{U}_{t_1t_1} & \tilde{U}_{t_1t_2} \\ & \tilde{U}_{t_2t_2} \end{bmatrix}^T + E(t,t). \tag{3.2}$$

Es bezeichnet dabei die Matrixfunktion $E(t,t) \in \mathbb{R}^{t \times t}$ den Fehler des jeweiligen Zerlegungsschrittes bzgl. des Schurkomplements $S_a(t,t)$. In unserem Fall genügt es, diesen später auf den Blöcken der Partition näher zu bestimmen.

Die Approximationen $\tilde{L}_{tt}$ und $\tilde{U}_{tt}$ werden aus dem Schurkomplement-Block $S_a(t,t)$ in der folgenden Weise berechnet:

(i) Berechne $\tilde{L}_{t_1t_1}$ und $\tilde{U}_{t_1t_1}$ aus $S_a(t_1,t_1) = S_{t_1t_1} = \tilde{L}_{t_1t_1}\tilde{U}_{t_1t_1} + E(t_1,t_1)$;

(ii) Berechne $\tilde{U}_{t_1t_2}$ aus $S_a(t_1,t_2) = S_{t_1t_2} = \tilde{L}_{t_1t_1}\tilde{U}_{t_1t_2} + E(t_1,t_2)$;

(iii) Berechne $\tilde{L}_{t_2t_1}$ aus $S_a(t_2,t_1) = S_{t_2t_1} = \tilde{L}_{t_2t_1}\tilde{U}_{t_1t_1} + E(t_2,t_1)$;

(iv) Berechne $\tilde{L}_{t_2t_2}$ und $\tilde{U}_{t_1t_1}$ aus $S_a(t_2,t_2) = S_{t_2t_2} - \tilde{L}_{t_2t_1}\tilde{U}_{t_1t_2} = \tilde{L}_{t_2t_2}\tilde{U}_{t_1t_1} + E(t_2,t_2)$.

Die Schritte (i) und (iv) können wieder rekursiv aufgerufen werden. Dies geschieht bis man einen Block $t \times t \in \mathcal{P}$ erreicht und die gewöhnliche pivotisierte LU-Zerlegung durchführt

$$S_a(t,t) = \tilde{L}_{tt}\tilde{U}_{tt}. \tag{3.3}$$

Hierbei ist offensichtlich, dass $E(t,t) = 0$ mit $t \times t \in \mathcal{P}$ gilt.

Im nächsten Lemma zeigen wir, dass bei der Berechnung der Diagonalblöcke von (3.2) wieder approximative Schurkomplemente entstehen und somit eine Rekursion wie oben möglich ist.

Lemma 3.1. *Sei eine Schurkomplement-Zerlegung von $S_a(t,t)$, $t \times t \in T_{I\times I} \setminus \mathcal{P}$, wie in* (3.2) *gegeben.*

Es folgt, dass

$$\begin{aligned} S_a(t_1, t_1) &= S_{t_1 t_1}, \\ S_a(t_1, t_2) &= S_{t_1 t_2}, \\ S_a(t_2, t_1) &= S_{t_2 t_1}, \\ S_a(t_2, t_2) &= S_{t_2 t_2} - \tilde{L}_{t_2 t_1} \tilde{U}_{t_1 t_2}. \end{aligned}$$

Beweis. In [Beb08, Lemma 4.33], wurden Beziehungen für das exakte Schurkomplement bewiesen, welche auch in gleicher Weise für das Approximative gelten. Somit erhält man

$$\begin{aligned} S_{t_1 t_1} &= S_a(t_1, t_1), \\ S_{t_1 t_2} &= S_a(t_1, t_2), \\ S_{t_2 t_1} &= S_a(t_2, t_1) \end{aligned}$$

und es folgt, dass

$$S_{t_2 t_2} - \tilde{L}_{t_2 t_1} \tilde{U}_{t_1 t_2} = A_{t_2 t_2} - \tilde{L}_{t_2 I_1} \tilde{U}_{I_1 t_2} - \tilde{L}_{t_2 t_1} \tilde{U}_{t_1 t_2} = S_a(t_2, t_2).$$

□

Zur Berechnung von $\tilde{U}_{t_1 t_2}$ in (ii), betrachten wir die rekursive Zerlegung von Außerdiagonalblöcken $t \times s \in T_{I \times I} \setminus \mathcal{P}$. Es sei angemerkt, dass die Zerlegung vom Typ (iii) in analoger Weise durchgeführt werden kann.

Sei $S_a(t, s)$, $t \times s \in T_{I \times I} \setminus \mathcal{P}$ zerlegt in der folgenden Weise:

$$S_a(t, s) = \begin{bmatrix} S_{t_1 s_1} & S_{t_1 s_2} \\ S_{t_2 s_1} & S_{t_2 s_2} \end{bmatrix} = \begin{bmatrix} \tilde{L}_{t_1 t_1} & \\ \tilde{L}_{t_2 t_1} & \tilde{L}_{t_2 t_2} \end{bmatrix} \begin{bmatrix} \tilde{U}_{t_1 s_1} & \tilde{U}_{t_1 s_2} \\ \tilde{U}_{t_2 s_1} & \tilde{U}_{t_2 s_2} \end{bmatrix} + E(t, s), \tag{3.4}$$

wobei $t_1, t_2 \in T_I$ und $s_1, s_2 \in T_I$ jeweils Söhne von $t, s \in T_I$ sind.

Die Approximation $\tilde{U}_{ts}$ wird aus dem Schurkomplement-Block $S_a(t, s)$, $t \times s \in T_{I \times I} \setminus \mathcal{P}$ in der folgenden Weise berechnet:

(i)* Berechne $\tilde{U}_{t_1 s_1}$ aus $S_a(t_1, s_1) = S_{t_1 s_1} = \tilde{L}_{t_1 t_1} \tilde{U}_{t_1 s_1} + E(t_1, s_1)$;

(ii)* Berechne $\tilde{U}_{t_1 s_2}$ aus $S_a(t_1, s_2) = S_{t_1 s_2} = \tilde{L}_{t_1 t_1} \tilde{U}_{t_1 s_2} + E(t_1, s_2)$;

(iii)* Berechne $\tilde{U}_{t_2 s_1}$ aus $S_a(t_2, s_1) = S_{t_2 s_1} - \tilde{L}_{t_2 t_1} \tilde{U}_{t_1 s_1} = \tilde{L}_{t_2 t_1} \tilde{U}_{t_1 s_1} + E(t_2, s_1)$;

(iv)* Berechne $\tilde{U}_{t_2 s_2}$ aus $S_a(t_2, s_2) = S_{t_2 s_2} - \tilde{L}_{t_2 t_1} \tilde{U}_{t_1 s_2} = \tilde{L}_{t_2 t_1} \tilde{U}_{t_1 s_2} + E(t_2, s_2)$;

Sämtliche Schritte können rekursiv aufgerufen werden. Falls ein Block $t \times s \in \mathcal{P}$ erreicht wird, approximiert man das Schurkomplement $S_a(t, s)$ durch eine Niedrigrangapproximation $\tilde{S}_a(t, s)$ wobei $S_a(t, s) - \tilde{S}_a(t, s) = E(t, s)$ gilt und berechnet anschließend mittels Vorwärtssubstitution $\tilde{U}_{ts}$.

$$S_a(t, s) - E(t, s) = \tilde{S}_a(t, s) = \tilde{L}_{tt} \tilde{U}_{ts} \tag{3.5}$$

Das folgende Lemma zeigt, dass in der Rekursion zur Berechnung der Außerdiagonalblöcke wieder

Schurkomplemente entstehen.

Lemma 3.2. *Sei eine Schurkomplement-Zerlegung von $S_a(t,s)$, $t \times s \in T_{I\times I} \setminus \mathcal{P}$ und $t \neq s$, wie in* (3.4) *gegeben. Es folgt, dass*

$$\begin{aligned} S_a(t_1,s_1) &= S_{t_1t_1},\\ S_a(t_1,s_2) &= S_{t_1t_2},\\ S_a(t_2,s_1) &= S_{t_2t_1} - \tilde{L}_{t_2t_1}\tilde{U}_{t_1s_1},\\ S_a(t_2,s_2) &= S_{t_2t_2} - \tilde{L}_{t_2t_1}\tilde{U}_{t_1s_2}. \end{aligned}$$

Beweis. Mit Hilfe von [Beb08, Lemma 4.33] ergibt sich, dass

$$\begin{aligned} S_{t_1t_1} &= S_a(t_1,s_1),\\ S_{t_1t_2} &= S_a(t_1,s_2). \end{aligned}$$

Somit erhält man

$$S_{t_2t_1} - \tilde{L}_{t_2t_1}\tilde{U}_{t_1s_1} = A_{t_2s_1} - \tilde{L}_{t_2I_1}\tilde{U}_{I_1s_1} - \tilde{L}_{t_2t_1}\tilde{U}_{t_1s_1} = S_a(t_2,s_1).$$

Der Beweis für $S_a(t_2,s_2)$ erfolgt analog. Somit ergibt sich die Behauptung. □

Somit sind wir in der Lage, eine hierarchische LU-Zerlegung rekursiv durchzuführen und mittels Matrix-Schurkomplementen zu beschreiben.

3.1.2 Blockweiser Fehler

Im folgenden Satz zeigen wir, dass der blockweise Fehler $A_{ts} - \tilde{A}_{ts}$, $t \times s \in \mathcal{P}$ in der betrachteten $\mathcal{H}$-LU-Zerlegung nur durch Kürzung des approximativen Schurkomplements entsteht. Dies ist eine für die folgenden Betrachtungen wesentliche Aussage, da sie zu Abschätzungen der Approximationsgenauigkeit von $\tilde{A}$ führt.

Satz 3.3. *Sei $\tilde{S}_a(t,s)$, $t \times s \in \mathcal{P}$ eine Approximation des Schurkomplements $S_a(t,s)$ dann gilt für die $\mathcal{H}$-LU-Zerlegung, dass*

$$(\tilde{L}\tilde{U})_{ts} - A_{ts} = \tilde{S}_a(t,s) - S_a(t,s).$$

Beweis. Ohne Beschränkung der Allgemeinheit können wir davon ausgehen, dass sich $t \times s \in \mathcal{P}$ in der oberen Dreiecks-Hälfte befindet. Mit Definition (3.1) und den Zerlegungen (3.3) und (3.5), folgt

$$(\tilde{L}\tilde{U})_{ts} - A_{ts} = \tilde{L}_{tt}\tilde{U}_{ts} + \tilde{L}_{tI_1}\tilde{U}_{I_1s} - A_{ts} = \tilde{S}_a(t,s) + \tilde{L}_{tI_1}\tilde{U}_{I_1s} - A_{ts} = \tilde{S}_a(t,s) - S_a(t,s).$$

Somit ergibt sich die Behauptung. □

Satz 3.3 ergibt sich, da bei der $\mathcal{H}$-Matrix-Arithmetik nur approximiert wird, wenn zwei Niedrigrangmatrizen auf den Blöcken der Partition addiert werden. Somit entstehen Störungen lediglich bei der Approximation des Schurkomplements.

Im Allgemeinen verwendet man für hierarchische Matrizen auf einem zulässigen Block der Partition $t \times s \in \mathcal{P}$ folgendes relatives Approximationskriterium

$$\|S_a(t,s) - \tilde{S}_a(t,s)\|_2 \leq \varepsilon \|S_a(t,s)\|_2, \quad \varepsilon \in [0,1). \tag{3.6}$$

Dieses findet zum Beispiel in der $\mathcal{H}$-Matrix-Bibliothek A$\mathcal{H}$MED[2] Verwendung. Somit bestimmt das asymptotische Verhalten der Norm des approximativen Schurkomplement-Blocks den blockweisen Fehler.

3.2 FE-$\mathcal{H}$-Matrix-Schurkomplement

Um den blockweisen Fehler der $\mathcal{H}$-LU-Zerlegung mit Hilfe von Satz 3.3 und Approximationskriterium (3.6) zu beschränkten, muss die Norm des approximativen Schurkomplements abgeschätzt werden. Dies erweist sich jedoch als schwierig, da per se nicht klar ist, inwiefern die bisherigen Approximationen die Norm beeinträchtigen.

Zur Ermöglichung von asymptotischen Aussagen betrachten wir das exakte Schurkomplement einer s.p.d. Matrix A definiert durch

$$S(t,s) := A_{ts} - L_{tI_1} U_{I_1 s} = A_{ts} - A_{tI_1} A_{I_1 I_1}^{-1} A_{I_1 s}, \tag{3.7}$$

wobei

$$I_1 = \{i \in I : i < \min(\min t, \min s)\} \quad \text{und} \quad I_2 := I \setminus I_1. \tag{3.8}$$

In den weiteren Betrachtungen gehen wir von folgender Stabilitätsannahme aus: für alle zulässigen Blöcke $t \times s \in \mathcal{P}$ existiert eine Konstante $c_S > 0$, so dass

$$\|S_a(t,s)\|_2 \leq c_S \|S(t,s)\|_2. \tag{3.9}$$

Diese Annahme ist für die $\mathcal{H}$-Matrix-Arithmetik schwer zu prüfen, da eine Störungsanalyse durch den rekursiven Charakter der Zerlegung erschwert wird. Dem Autor sind vergleichbare Aussagen in der Fachliteratur unbekannt. Beobachtungen legen jedoch nahe, dass die Norm des approximativen und des exakten Schurkomplements dicht beieinander liegen auch für große Toleranzen ε in (3.6).

Das folgende Lemma gibt eine Schranke für die Norm des exakten Schurkomplements an, welches für den Beweis nur Argumente aus der linearen Algebra verwendet.

Lemma 3.4. *Sei A eine s.p.d. Matrix und die Mengen I_1, I_2 gegeben wie in* (3.8)*. Dann gilt, dass*

$$\|S(t,s)\|_2 \leq \|A_{I_2 I_2}\|_2 \leq \|A\|_2.$$

Beweis. Die Matrix $A_{I_1 I_1}$ ist s.p.d., da Hauptabschnittsmatrizen von s.p.d. Matrizen wieder s.p.d. sind. Somit ist auch $A_{I_2 I_2} - S(I_2, I_2) = A_{I_2 I_1} A_{I_1 I_1}^{-1} A_{I_1 I_2}$ symmetrisch positiv semidefinit und es folgt, dass

$$\|S(I_2, I_2)\|_2 \leq \|A_{I_2 I_2}\|_2.$$

Da $S(t,s)$ ein Teilblock von $S(I_2, I_2)$ ist, siehe Lemma 4.3.3 in [Beb08], folgt die Behauptung. □

[2] `http://bebendorf.ins.uni-bonn.de/AHMED.html`

Für $t = s$ liefert Lemma 3.4 eine scharfe Abschätzung (man betrachte als Beispiel die Einheitsmatrix). Im Falle disjunkter Gebiete X_t und X_s, wie sie für $\mathcal{H}$-Matrix-Abschätzungen gebraucht werden, lässt sie sich jedoch weiter verbessern. Damit werden wir uns im Rest dieses Abschnittes beschäftigen.

3.2.1 Alternative Formulierung des Schurkomplements

Bisherige Abschätzungen für das $\mathcal{H}$-Matrix-Schurkomplement basieren auf Schranken für die Inverse der Steifigkeitsmatrix, vgl. [Beb08]. Um die Beweisidee zu skizzieren, sei $\mathcal{N}(t)$ die Menge der Nachbarn von $t \subset I$ definiert durch

$$\mathcal{N}(t) := \{i \in I : \operatorname{dist}(i,t) \leq 1\}. \tag{3.10}$$

In diesem Zusammenhang bezeichnet dist den Abstand im Matrixgraphen. Somit kann das Schurkomplement aus (3.7) für $t \times s \in \mathcal{P}$ mit $\mathcal{N}'(t) := \mathcal{N}(t) \cap I_1$ umgeschrieben werden zu

$$S(t,s) = A_{tI_1} A_{I_1 I_1}^{-1} A_{I_1 s} = \sum_{i,j \in I_1} A_{ti} \big(A_{I_1 I_1}^{-1}\big)_{ij} A_{js} = \sum_{i \in \mathcal{N}'(t), j \in \mathcal{N}'(s)} A_{ti} \big(A_{I_1 I_1}^{-1}\big)_{ij} A_{js}.$$

Somit muss man den folgenden Ausdruck beschränken

$$\|S(t,s)\|_2 \leq \|A\|_2^2 \|(A_{I_1 I_1}^{-1})_{\mathcal{N}'(s),\mathcal{N}'(t)}\|_2.$$

Problematisch bei diesem Ansatz ist, dass die Inverse der Massenmatrix zu einem Verschmieren bei blockweisen Einschränkungen führt, da man das Abklingverhalten der Greenschen Funktion abschätzt. Somit werden scharfe Schranken erschwert.

In dieser Arbeit wählen wir einen anderen Ansatz und werden das Schurkomplement durch Betrachtungen von harmonischen Funktionen auf Randabschnitte zurückführen. ähnlich Vorgehensweisen werden bei Abschätzungen von Gebietszerlegungs-Methoden verwendet, siehe [Bre99]. Diese können jedoch nicht ohne weiteres übertragen werden, da wir Gebiete mit nicht überlappenden Trägern betrachten.

Im Folgenden werden Schurkomplement-Blöcke mit speziell gewählten Basisfunktionen durch die Bilinearform dargestellt. Für diese Betrachtungen benötigen wir die Einschränkung der natürlichen Injektion (1.17). Sei φ_i, $i \in I$, eine Basis des Finite-Elemente-Raumes V_h. Wir definieren $\mathcal{J}_\sigma : \mathbb{R}^\sigma \to V_h$ als

$$\mathcal{J}_\sigma x := \sum_{i \in \sigma} x_i \varphi_i$$

und dessen Bild $V_{h,\sigma} := \operatorname{Im} \mathcal{J}_\sigma$ bzgl. einer Menge $\sigma \subset I$. Der Träger von Funktionen in $V_{h,\sigma}$ sei gegeben durch $X_\sigma := \bigcup_{i \in \sigma} \operatorname{supp} \varphi_i$.

Im Rest dieses Abschnittes zeigen wir wie V_{h,I_1} und sein a-orthogonales Komplement bzgl. $V_{h,I_1 \cup \sigma}$, mit $\sigma \subset I_2$, den Matrix-Schurkomplement-Block (3.7) beschreiben. Hierfür benötigen wir die folgende Menge:

$$V_{h,\sigma}^{\perp} := \{v \in V_{h,I_1 \cup \sigma} : a(v,w) = 0 \text{ für alle } w \in V_{h,I_1}\}, \tag{3.11}$$

wobei $\sigma \subset I_2$.

Das nächste Lemma stellt eine Beziehung zwischen der Menge aus (3.11) und der Steifigkeitsmatrix

her.

Lemma 3.5. *Eine Funktion $v = \mathcal{J}_{I_1} x_{I_1} + \mathcal{J}_\sigma x_\sigma \in V_{h,I_1\cup\sigma}$ mit $\sigma \subset I_2$, gehört zu $V_{h,\sigma}^\perp$ genau dann, wenn $A_{I_1 I_1} x_{I_1} + A_{I_1\sigma} x_\sigma = 0$.*

Beweis. Sei $w := \sum_{i\in I_1} y_i \varphi_i \in V_{h,I_1}$. Somit ergibt sich aus (3.11) für $v \in V_{h,\sigma}^\perp$, dass

$$0 = a(v,w) = \begin{bmatrix} x_{I_1} \\ x_\sigma \end{bmatrix}^T \begin{bmatrix} A_{I_1 I_1} & A_{I_1\sigma} \\ A_{\sigma I_1} & A_{\sigma\sigma} \end{bmatrix} \begin{bmatrix} y_{I_1} \\ 0 \end{bmatrix} = x_{I_1}^T A_{I_1 I_1} y_{I_1} + x_\sigma^T A_{\sigma I_1} y_{I_1}.$$

Da $y_{I_1} \in \mathbb{R}^{I_1}$ beliebig gewählt werden kann, folgt die Behauptung. □

Seien $e^{(\ell)} \in \mathbb{R}^\sigma$, $\ell \in \sigma$, die kanonischen Einheitsvektoren und folgende Funktionen definiert vermöge

$$\varphi_\ell^\perp := \varphi_\ell - \mathcal{J}_{I_1}\big(A_{I_1 I_1}^{-1} A_{I_1\sigma} e^{(\ell)}\big) \quad \text{für alle } l \in \sigma \subset I_2. \tag{3.12}$$

Wie sich im nächsten Lemma zeigt bilden diese Funktionen eine Basis von $V_{h,\sigma}^\perp$ mit $\sigma \subset I_2$.

Lemma 3.6. *Die Menge der Funktionen $(\varphi_\ell^\perp)_{\ell\in\sigma}$ definiert wie in* (3.12)*, bilden eine Basis von $V_{h,\sigma}^\perp$.*

Beweis. Aus (3.12) ergibt sich für ein $\ell \in \sigma$, dass

$$\varphi_\ell^\perp = -\mathcal{J}_{I_1}\big(A_{I_1 I_1}^{-1} A_{I_1\sigma} e^{(\ell)}\big) + \mathcal{J}_\sigma e^{(\ell)} = \mathcal{J}_{I_1} x^{(\ell)} + \mathcal{J}_\sigma e^{(\ell)},$$

wobei $x^{(\ell)} := -A_{I_1 I_1}^{-1} A_{I_1\sigma} e^{(\ell)}$. Da

$$A_{I_1 I_1} x^{(\ell)} + A_{I_1\sigma} e^{(\ell)} = -A_{I_1 I_1} A_{I_1 I_1}^{-1} A_{I_1\sigma} e^{(\ell)} + A_{I_1\sigma} e^{(\ell)} = 0,$$

folgt mit Lemma 3.5, dass $\varphi_\ell^\perp \in V_{h,\sigma}^\perp$.

Nun bleibt zu zeigen, dass die Funktionen $\varphi_\ell^\perp$, $\ell \in \sigma$, eine Basis von $V_{h,\sigma}^\perp$ bilden. Offensichtlich sind die kanonischen Einheitsvektoren $(e^{(\ell)})_{\ell\in\sigma}$ eine Basis von $\mathbb{R}^\sigma$. Da $x^{(\ell)}$ wie oben durch $e^{(\ell)}$ dargestellt werden kann ist $(\varphi_\ell^\perp)_{\ell\in\sigma}$ eine Basis von $V_{h,\sigma}^\perp$. □

Als direkte Folge aus Lemma 3.6 ergibt sich, dass $\dim(V_{h,\sigma}^\perp) = |\sigma|$. Weiterhin ist ersichtlich, dass sämtliche Funktionen aus $V_{h,\sigma}^\perp$ bereits durch die Ansatzfunktionen φ_i, $i \in \sigma \subset I_2$, eindeutig bestimmt sind.

Nun stellen wir einen Zusammenhang her zwischen den in (3.12) definierten Funktionen und dem blockweisen Schurkomplement.

Lemma 3.7. *Seien $(\varphi_i^\perp)_{i\in t}$ und $(\varphi_j^\perp)_{j\in s}$ definiert wie in* (3.12)*. Dann ist das blockweise Schurkomplement aus Definition* (3.7) *gegeben durch*

$$S(t,s) = [a(\varphi_i^\perp, \varphi_j^\perp)]_{i\in t, j\in s}, \quad t \times s \in \mathcal{P}.$$

Beweis. Für die Basisfunktionen ergibt sich, dass

$$\begin{aligned} a(\varphi_i^{\perp}, \varphi_j^{\perp}) &= a(\varphi_i, \varphi_j) - a(\varphi_i, \mathcal{J}_{I_1}\big(A_{I_1 I_1}^{-1} A_{I_1 s} e^{(j)}\big)) - a(\mathcal{J}_{I_1}\big(A_{I_1 I_1}^{-1} A_{I_1 t} e^{(i)}\big), \varphi_j) \\ &\quad + a(\mathcal{J}_{I_1}\big(A_{I_1 I_1}^{-1} A_{I_1 t} e^{(i)}\big), \mathcal{J}_{I_1}\big(A_{I_1 I_1}^{-1} A_{I_1 s} e^{(j)}\big)) \\ &= (e^{(i)})^T A_{ts} e^{(j)} - (e^{(i)})^T A_{t I_1} A_{I_1 I_1}^{-1} A_{I_1 s} e^{(j)} = [S(t,s)]_{i,j}, \end{aligned}$$

für $i \in t$ und $j \in s$. □

In dem nächsten Abschnitt werden wir das in hierarchischen Matrizen auftretende Schurkomplement (3.7) nach oben bzgl. der Spektralnorm abschätzen.

3.2.2 Asymptotische Schranke

Abschätzungen für das Schurkomplement bei Gebietszerlegungs-Methoden basieren in der Regel auf dem Spursatz, siehe [Bre99]. Dieses Prinzip lässt sich jedoch nicht direkt auf hierarchische Matrizen übertragen. Aufgrund von nicht überlappenden Trägern nutzen wir das Abklingverhalten der Greenschen Funktion.

Im nächsten Lemma zeigt sich, dass sich nur bestimmte Teile des Randes ∂X_{I_1} vom Träger X_{I_1} auf das Schurkomplement auswirken. Dies resultiert aus der Lokalität der Ansatzfunktionen und den disjunkten Trägern der betrachteten Funktionen.

Sei G die Menge der Knoten auf ∂X_{I_1}, vgl. hierzu Abbildung 5. Weiterhin definieren wir die

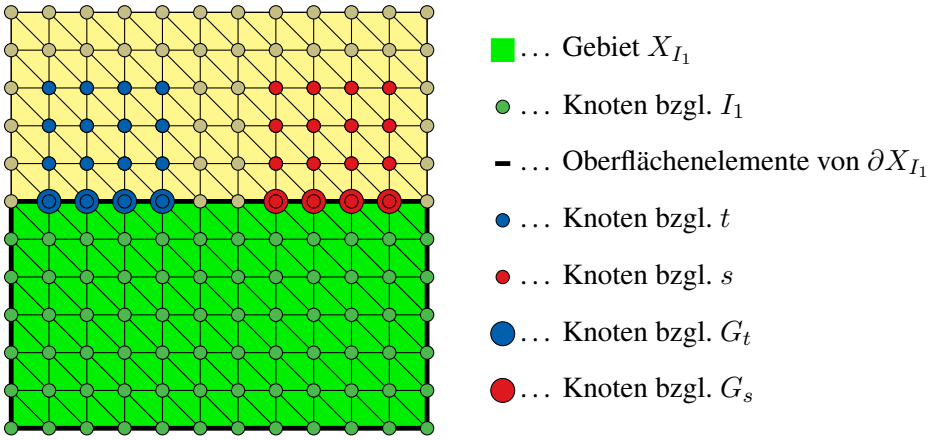


Abbildung 5: Verschiedene Mengen zur Beschreibung des Schurkomplementes.

Teilmengen $G_t := G \cap t$ und $G_s := G \cap s$. Im Folgenden bezeichnen wir v_σ als Einschränkung von $v \in V_h$ auf die Basisfunktionen φ_i, $i \in \sigma \subset I$.

Lemma 3.8. *Seien die Träger von $t, s \subset I_2$ disjunkt, so dass $X_t \cap X_s = \emptyset$. Für Funktionen $v \in V_{h,t}^{\perp}$, $w \in V_{h,s}^{\perp}$ gilt, dass*

$$a(v, w) = a(v_{G_t} + v_{I_1}, w_{G_s} + w_{I_1})$$

und $v_{G_t} + v_{I_1} \in V_{h,G_t}^{\perp}$, $w_{G_s} + w_{I_1} \in V_{h,G_s}^{\perp}$.

Beweis. Aufgrund der disjunkten Träger X_t und X_s erhält man, dass

$$a(v_t, w_s) = a(v_{G_t}, w_{G_s}) = 0.$$

Es ergibt sich die folgende Umformung

$$a(v, w) = a(v_t + v_{I_1}, w_s + w_{I_1}) = a(v_{G_t}, w_{G_s}) + a(v_t, w_{I_1}) + a(v_{I_1}, w_s) + a(v_{I_1}, w_{I_1}).$$

Da die Ansatzfunktionen einen lokalen Träger besitzen, folgt

$$a(v_t, w_{I_1}) = a(v_{G_t}, w_{I_1}) \quad \text{und} \quad a(v_{I_1}, w_s) = a(v_{I_1}, w_{G_s}).$$

Somit erhält man, dass

$$a(v, w) = a(v_{G_t} + v_{I_1}, w_{G_s} + w_{I_1}).$$

Nun bleibt zu zeigen, dass $v_{G_t} + v_{I_1} \in V_{h,G_t}^{\perp}$ und $w_{G_s} + w_{I_1} \in V_{h,G_s}^{\perp}$. Dies folgt direkt aus Definition (3.11) und dem lokalen Träger der Ansatzfuktionen

$$0 = a(v, u) = a(v_t + v_{I_1}, u) = a(v_{G_t} + v_{I_1}, u), \quad u \in V_{h,I_1}.$$

Die gleichen Argumente können auch für $w_{G_s} + w_{I_1}$ verwendet werden und es folgen die Behauptungen. □

Das nächste Lemma beschränkt den Ausdruck auf der rechten Seite von Lemma 3.8, jedoch für die Bilinearform

$$a_\Delta(v, w) := \int_\Omega \nabla v \cdot \nabla w \, \mathrm{d}x.$$

Dabei wird das Abklingverhalten der Greenschen Funktion genutzt. Weiterhin haben die disjunkten Träger von $v \in V_{h,G_t}^{\perp}$, $w \in V_{h,G_s}^{\perp}$ den Effekt, dass wir auf der rechte Seite die Funktionen in der L^2- anstelle der $H^{1/2}$-Norm abschätzen können. Mit einer Verallgemeinerung des folgenden Lemmas werden wir später das Schurkomplement nach oben beschränken.

Lemma 3.9. *Seien die Träger $t, s \subset I_2$ disjunkt, so dass $\operatorname{dist}(X_t, X_s) > 0$. Dann existiert für alle $v \in V_{h,G_t}^{\perp}$, $w \in V_{h,G_s}^{\perp}$ eine Konstante $c_G > 0$, so dass*

$$a_\Delta(v, w) \leq \frac{c_G}{\operatorname{dist}(X_t, X_s)} \|v\|_{L^2(\partial X_{I_1})} \|w\|_{L^2(\partial X_{I_1})}.$$

Beweis. Um Funktionen der Menge $V_{h,G}^{\perp}$ zu beschränken, definieren wir den diskreten Poincaré-Steklov-Operator

$$\mathfrak{P}_h : V_{h,G} \to V_{h,G}$$

durch eine diskrete a_Δ-harmonische Fortsetzung $E_{I_1,h} : V_{h,G} \to V_{h,I_1\cup G}$, so dass für alle $u \in V_{h,G}$

$$\begin{aligned}(\mathfrak{P}_h u, u')_{\partial X_{I_1}} &= a_\Delta(E_{I_1,h}u, u'), \quad && u' \in V_{h,I_1\cup G},\\ a_\Delta(E_{I_1,h}u, u'') &= 0, && u'' \in V_{h,I_1}\end{aligned}$$

gilt, wobei $E_{I_1,h}u(\xi) = u(\xi)$ für alle $\xi \in \partial X_{I_1}$. Aufgrund von (3.11) folgt, dass $E_{I_1,h}v = v$ für alle $v \in V_{h,G_t}^\perp$. Somit erhält man für $v \in V_{h,G_t}^\perp$, $w \in V_{h,G_s}^\perp$, dass

$$a_\Delta(v,w) = a_\Delta(E_{I_1,h}v, w) = (\mathfrak{P}_h v, w)_{L^2(\partial X_{I_1})}.$$

Durch den beschränkten Träger von $w \in V_{h,G_s}^\perp$, ergibt sich, dass

$$a_\Delta(v,w) = (\mathfrak{P}_h v, w)_{L^2(\partial X_{I_1}\cap X_s)} \le \|\mathfrak{P}_h v\|_{L^2(\partial X_{I_1}\cap X_s)}\|w\|_{L^2(\partial X_{I_1})}.$$

Um den diskreten Poincaré-Steklov-Operator zu beschränken, verwenden wir das Abklingverhalten der Greenschen Funktion $\mathfrak{G}(\xi,\zeta)$, wobei $\xi,\zeta \in X_{I_1}$. Hierfür, definieren wir den hypersingulären Integraloperator

$$(\mathcal{D}v)(\xi) := \int_{\partial X_{I_1}} \gamma_{1,\xi}\gamma_{1,\zeta}\mathfrak{S}(\xi,\zeta)v(\zeta)\,\mathrm{d}s_\zeta, \quad \xi \in \partial X_{I_1}\cap X_s,\ v \in V_{h,G_t}^\perp,$$

wobei

$$\gamma_{1,\xi}\gamma_{1,\zeta}\mathfrak{S}(\xi,\zeta) = \begin{cases} -\frac{1}{2\pi}\left(\frac{(\nu_\xi,\nu_\zeta)}{|\xi-\zeta|^2} - 2\frac{(\xi-\zeta,\nu_\xi)(\xi-\zeta,\nu_\zeta)}{|\xi-\zeta|^3}\right), & d=2,\\ \frac{1}{4\pi}\left(\frac{(\nu_\xi,\nu_\zeta)}{|\xi-\zeta|^3} - 3\frac{(\zeta-\xi,\nu_\xi)(\zeta-\xi,\nu_\zeta)}{|\xi-\zeta|^5}\right), & d=3.\end{cases}$$

Der exakte Poincaré-Steklov-Operator bildet die Dirichlet-Randdaten auf die Neumann-Randdaten ab, siehe [KW04]. Somit gilt $\mathfrak{P}v = \gamma_1 v$ und durch Verwendung von Lemma 1.10, folgt

$$(\mathfrak{P}v)(\xi) = -\int_{\partial X_{I_1}} \gamma_{1,\xi}\gamma_{1,\zeta}\mathfrak{G}(\xi,\zeta)v(\zeta)\,\mathrm{d}s_\zeta.$$

Für die diskrete Version des Poincaré-Steklov-Operators können wir annehmen, dass eine Konstante $c > 0$ existiert, so dass

$$\begin{aligned}\|\mathfrak{P}_h v\|_{L^2(\partial X_{I_1}\cap X_s)} &\le c\|\mathfrak{P}v\|_{L^2(\partial X_{I_1}\cap X_s)} = c\left\|\int_{\partial X_{I_1}} \gamma_{1,\cdot}\gamma_{1,\zeta}\mathfrak{G}(\cdot,\zeta)v(\zeta)\,\mathrm{d}s_\zeta\right\|_{L^2(\partial X_{I_1}\cap X_s)}\\ &\le c\|\mathcal{D}v\|_{L^2(\partial X_{I_1}\cap X_s)} + c\left\|\int_{\partial X_{I_1}} \gamma_{1,\cdot}\gamma_{1,\zeta}\mathfrak{H}(\cdot,\zeta)v(\zeta)\,\mathrm{d}s_\zeta\right\|_{L^2(\partial X_{I_1}\cap X_s)}. \qquad (3.13)\end{aligned}$$

Der erste Term auf der rechten Seite von (3.13) kann aufgrund der disjunkten Träger von v und w beschränkt werden und man erhält

$$\|\mathcal{D}v\|_{L^2(\partial X_{I_1}\cap X_s)} \le \frac{c'}{\operatorname{dist}(X_t,X_s)}\left\|\int_{\partial X_{I_1}} \frac{|v(\zeta)|}{|\cdot-\zeta|^{d-1}}\,\mathrm{d}s_\zeta\right\|_{L^2(\partial X_{I_1}\cap X_s)}.$$

Da der Calderón-Zygmund operator [Cos88] in L^2 stetig ist, erhält man

$$\left\| \int_{\partial X_{I_1}} \frac{|v(\zeta)|}{|\cdot - \zeta|^{d-1}} \, \mathrm{d}s_\zeta \right\|_{L^2(\partial X_{I_1} \cap X_s)} \leq \left\| \int_{\partial X_{I_1}} \frac{|v(\zeta)|}{|\cdot - \zeta|^{d-1}} \, \mathrm{d}s_\zeta \right\|_{L^2(\partial X_{I_1})} \leq c'' \|v\|_{L^2(\partial X_{I_1})}.$$

Somit gilt die Schranke

$$\|\mathcal{D}v\|_{L^2(\partial X_{I_1} \cap X_s)} \leq \frac{c' c''}{\operatorname{dist}(X_t, X_s)} \|v\|_{L^2(\partial X_{I_1})}.$$

Da $\mathfrak{H}(\cdot, \zeta) \in H^1_\Delta(X_{I_1}) := \{u \in H^1(X_{I_1}) : \Delta u \in L^2(X_{I_1})\}$ für alle $\zeta \in X_{I_1}$ (vgl. [Cos88]), kann der zweite Term auf der rechten Seite von (3.13) beschränkt werden durch

$$\begin{aligned} \left\| \int_{\partial X_{I_1}} \gamma_{1,\cdot} \gamma_{1,\zeta} \mathfrak{H}(\cdot, \zeta) v(\zeta) \, \mathrm{d}s_\zeta \right\|_{L^2(\partial X_{I_1} \cap X_s)} &\leq \left\| \int_{\partial X_{I_1}} \gamma_{1,\cdot} \gamma_{1,\zeta} \mathfrak{H}(\cdot, \zeta) v(\zeta) \, \mathrm{d}s_\zeta \right\|_{L^2(\partial X_{I_1})} \\ &\leq c''' \|v\|_{L^2(\partial X_{I_1})} \end{aligned}$$

mit einer Konstante $c''' > 0$ unabhängig von X_t und X_s. □

Wir nehmen im Folgenden an, dass Lemma 3.9 verallgemeinert werden kann auf elliptische Operatoren mit variablen Koeffizienten, so dass

$$a(v, w) \leq \frac{c_G}{\operatorname{dist}(X_t, X_s)} \|v\|_{L^2(\partial X_{I_1})} \|w\|_{L^2(\partial X_{I_1})} \tag{3.14}$$

für alle $v \in V^\perp_{h,G_t}$, $w \in V^\perp_{h,G_s}$.

Für den nächsten Satz ist es wichtig, dass eine quasi-uniforme Triangulierung von Ω eingeschränkt auf ∂X_{I_1} wieder eine quasi-uniforme Triangulierung ergibt. Somit existiert nach [Hac86, Satz 8.8.1] eine Konstante $c_{J,b} > 0$, so dass

$$\|\mathcal{J}x\|_{L^2(\partial X_{I_1})} \leq c_{J,b} h^{(d-1)/2} \|x\|_2 \leq \hat{c}_J h^{(d-1)/2} \|x\|_2, \quad x \in \mathbb{R}^G, b \in \mathcal{P} \tag{3.15}$$

gilt, wobei $\hat{c}_J := \max_{b \in \mathcal{P}} c_{J,b}$.

Somit erhalten wir mit den bisherigen Aussagen den folgenden Satz welcher eine blockweise Abschätzung für das Matrix-Schurkomplement liefert.

Satz 3.10. *Das Matrix-Schurkomplement für einen zulässigen Block $t \times s \in \mathcal{P}$ kann beschränkt werden durch*

$$\|S(t, s)\|_2 \leq \frac{c_G \hat{c}_J^2 c_J^{-2} \|M\|_2}{h \operatorname{dist}(X_t, X_s)}.$$

Beweis. Mit dem Lemma 3.7, Lemma 3.8 und (3.14) folgt, dass

$$\begin{aligned}\|S(t,s)\|_2 &= \sup_{\substack{x\in\mathbb{R}^{|t|},\\ y\in\mathbb{R}^{|s|}}} \frac{x^T S(t,s) y}{\|x\|_2\|y\|_2} = \sup_{\substack{x\in\mathbb{R}^{|t|},\\ y\in\mathbb{R}^{|s|}}} \frac{a\big(\sum_{i\in t} x_i\varphi_i^{\perp}, \sum_{j\in s} y_j\varphi_j^{\perp}\big)}{\|x\|_2\|y\|_2} \\ &\leq \sup_{\substack{x\in\mathbb{R}^{|t|},\\ y\in\mathbb{R}^{|s|}}} \frac{c_G\|\mathcal{J}x\|_{L^2(\partial X_{I_1})}\|\mathcal{J}y\|_{L^2(\partial X_{I_1})}}{\operatorname{dist}(X_t,X_s)\|x\|_2\|y\|_2}.\end{aligned}$$

Die Behauptung erhält man durch (3.15) und Lemma 1.13. □

Die in diesem Abschnitt gewonnene Schranke für das Matrix-Schurkomplement wollen wir im nächsten dazu nutzen, die Approximationsgüte der hierarchischen Cholesky-Zerlegung abzuschätzen.

3.3 Bestimmung der Approximationsgüte

Wie in Satz 1.19 und 1.20 gezeigt wurde, ist es hinreichend die Steifigkeitsmatrix im Sinne von (1.27) zu approximieren um die Konditionszahl des vorkonditionierten Systems zu beschränken. Wir betrachten nun speziell die Approximationseigenschaften der hierarchischen LU- und Cholesky-Zerlegung. Dabei gehen wir auf ein vereinfachtes und ein praxisnahes Approximationskriterium ein.

3.3.1 Vereinfachte Approximationsbedingung

In [Beb08, Theorem 4.35] wurde für folgende blockweise Approximationsbedingung

$$\|S(t,s) - \tilde{S}(t,s)\|_2 \leq \varepsilon\|A\|_2 \tag{3.16}$$

die Approximationsgüte der $\mathcal{H}$-LU-Zerlegung abgeschätzt durch

$$\|A - \tilde{L}\tilde{U}\|_2 \leq c\,\varepsilon L(T_I)\operatorname{cond}_2(A)\|L\|_2\|U\|_2, \quad c > 0. \tag{3.17}$$

Dies bedeutet, dass in jedem Schritt davon ausgegangen wird, dass das exakte Schurkomplement $S(t,s)$ vorliegt und anschliessend durch $\tilde{S}(t,s)$ im Sinne von (3.16) approximiert wird. Es ist wichtig zu erkennen, dass in diesem Fall nicht mehr Satz 3.3 gilt.

In unserem Fall wählen wir eine leicht andere Bedingung, bei der wir wie in Abschnitt 3.1 die bisherigen Approximationen berücksichtigen. Sei eine hierarchische LU-Zerlegung $\tilde{L}\tilde{U} \approx LU = A$ einer regulären Matrix A gegeben, wobei Approximationsbedingung (3.6) bzw. (3.16) durch die folgende ersetzt wird

$$\|S_a(t,s) - \tilde{S}_a(t,s)\|_2 \leq \varepsilon\|A\|_2. \tag{3.18}$$

Die Abschätzung (3.17) lässt sich für die $\mathcal{H}$-LU-Zerlegung mit den bisher gezeigten Resultaten wie folgt verbessern.

Satz 3.11. *Sei eine $\mathcal{H}$-LU-Zerlegung gegeben, die die Approximationsbedingung* (3.18) *verwendet und*

bei der die blockweise LU-Zerlegung ohne Approximation hinreichend stabil ist. Dann folgt, dass

$$\|A - \tilde{L}\tilde{U}\|_2 \leq c_{\text{sp}} \varepsilon L(T_I) \|A\|_2,$$

mit der Schwachbesetztheitskonstante c_{sp} *aus* (2.9).

Beweis. Mit Hilfe von (2.9) und Satz 3.3 ergibt sich, dass

$$\|A - \tilde{L}\tilde{U}\|_2 \leq c_{\text{sp}} \sum_{\ell=0}^{L(T_I)-1} \max_{b \in \mathcal{P} \cap T_{I\times I}^{(\ell)}} \|E_b\|_2 \leq c_{\text{sp}} \varepsilon L(T_I) \|A\|_2.$$

□

Satz 3.11 ist in zweierlei Hinsicht eine Verbesserung von Abschätzung (3.17). Zum einen hängt die obere Schranke nicht von der Kondition von A ab. Zum anderen befindet sich auf der rechten Seiten nur die Norm von A und nicht die Norm von den Faktoren L und U. Dies unterstreicht die Stabilität der $\mathcal{H}$-LU-Zerlegung.

Bemerkung 3.12. Sei A eine s.p.d. Matrix. Dann folgt mit Hilfe der Annahme (3.9) und Lemma 3.4 aus Approximationskriterium (3.6) die vereinfachte Kürzung (3.18). Somit gilt Satz 3.11 auch mit dem Approximationskriterium (3.6). Bemerkenswert hierbei ist, dass der Beweis nur auf Argumenten der linearen Algebra beruht.

Im nächsten Abschnitt wird an Stelle von Approximationsbedinung (3.18) das Kriterium (3.6), gewählt welches zum Beispiel in der $\mathcal{H}$-Matrix-Bibliothek A$\mathcal{H}$MED Verwendung findet.

3.3.2 Praxisnahe Approximationsbedingung

In diesem Abschnitt gehen wir von einer $\mathcal{H}$-Cholesky-Zerlegung einer s.p.d. Matrix A aus, welche auf den Blöcken der Partition die Approximationsbedingung (3.6) verwendet. Mit Hilfe von Lemma 3.3 und Annahme (3.9) kann die Schurkomplement-Abschätzung aus Satz 3.10 verwendet werden, um den blockweisen Fehler $E_b := A_b - \tilde{A}_b$, $b \in \mathcal{P}$, einer $\mathcal{H}$-Cholesky-Zerlegung abzuschätzen.

Lemma 3.13. *Sei* (3.9) *erfüllt und eine $\mathcal{H}$-Cholesky-Zerlegung mit blockweiser Approximationsgenauigkeit ε_b in* (3.6) *gegeben. Dann ergibt sich folgende Schranke für den blockweisen Fehler*

$$\|E_b\|_2 \leq \frac{\varepsilon_b c_E \|M\|_2}{h \max\big(\operatorname{diam} X_t, \operatorname{diam} X_s\big)}, \quad b = t \times s \in \mathcal{P},$$

wobei $c_E := c_G \hat{c}_J^2 c_J^{-2} c_S \eta$.

Beweis. Sei $b \in \mathcal{P}$ ein zulässiger Block. Mit (3.6), (3.9) und Satz 3.3 folgt sofort, dass

$$\|E_b\|_2 \leq c_S \varepsilon_b \|S(t,s)\|_2.$$

Satz 3.10 und die Zulässigkeitsbedingung (2.7) ergeben die folgende Abschätzung

$$\begin{aligned}\|E_b\|_2 &\leq \varepsilon_b c_G \hat{c}_J^2 c_J^{-2} c_S \|M\|_2 h^{-1} \big(\operatorname{dist}(X_t, X_s)\big)^{-1} \\ &\leq \varepsilon_b c_G \hat{c}_J^2 c_J^{-2} c_S \eta \|M\|_2 h^{-1} \big(\max\big(\operatorname{diam} X_t, \operatorname{diam} X_s\big)\big)^{-1}.\end{aligned}$$

Sei $b \in \mathcal{P}$ ein nicht zulässiger Block. Da hier keine Approximation stattfindet, folgt mit Satz 3.3, dass $\|E_b\|_2 = 0$ und somit die Behauptung. □

Die Abschätzung des blockweisen Fehlers im vorherigen Lemma können wir nutzen, um die Approximationsgenauigkeit der $\mathcal{H}$-Cholesky-Zerlegung zu bestimmen. Zuvor jedoch ein Lemma zur Beschränkung einer Reihe. Hierfür definieren wir zur einfacheren Lesbarkeit

$$\operatorname{mindiam}_\ell := \min_{t \in T_I^{(\ell)}} \operatorname{diam} X_t$$

für alle $\ell = 0, \ldots, L(T_I) - 1$.

Lemma 3.14. *Sei $d \in \{2, 3\}$ und ein balancierter Clusterbaum gegeben. Dann ist folgende Reihe beschränkt durch*

$$\sum_{\ell=0}^{L(T_I)-1} \operatorname{mindiam}_\ell^{-1} \leq c_b h^{-1},$$

wobei $c_b := 5c_{D_2}/c_{D_1}$.

Beweis. Umformulieren der Summe mit (2.8) und die Beschränkung der geometrischen Reihe ergeben

$$\begin{aligned}\sum_{\ell=0}^{L(T_I)-1} \operatorname{mindiam}_\ell^{-1} &\leq c_{D_1}^{-1} \sum_{\ell=0}^{L(T_I)-1} 2^{\ell/d} = \frac{2^{(L(T_I)-1)/d}}{c_{D_1}} \sum_{\ell=0}^{L(T_I)-1} 2^{(\ell - L(T_I)+1)/d} \\ &= \frac{2^{(L(T_I)-1)/d}}{c_{D_1}} \sum_{\ell=0}^{L(T_I)-1} 2^{-\ell/d} \leq \frac{5 \cdot 2^{(L(T_I)-1)/d}}{c_{D_1}}.\end{aligned}$$

Sei $\hat{t} \in T_I^{(L(T_I)-1)}$ der Cluster aus dem untersten Level, in dem sich der Index $\arg\max_{i \in I} \operatorname{diam} X_i$ befindet. Dadurch folgt mit (2.8) die Abschätzung

$$h = \max_{i \in I} \operatorname{diam} X_i \leq \operatorname{diam} X_{\hat{t}} \leq c_{D_2} 2^{-(L(T_I)-1)/d}$$

und man erhält die Behauptung. □

Die Approximationsgenauigkeit der $\mathcal{H}$-Cholesky-Zerlegung wird in dem folgenden Satz abgeschätzt.

Satz 3.15. *Seien die Voraussetzungen wie in Lemma 3.13 und ein balancierter Clusterbaum gegeben. Dann gilt für die $\mathcal{H}$-Cholesky-Zerlegung mit einer festen Approximationsgenauigkeit $\varepsilon > 0$, dass*

$$\|A - \tilde{L}\tilde{L}^T\|_2 \leq c_e \varepsilon h^{-2} \|M\|_2,$$

wobei $c_e := c_{\mathrm{sp}} c_E c_b$.

Beweis. Mit Hilfe von Lemma 3.13 und (2.9) folgt, dass

$$\|A - \tilde{L}\tilde{L}^T\|_2 \leq c_{\mathrm{sp}} \sum_{\ell=0}^{L(T_I)-1} \max_{b \in \mathcal{P} \cap T_{I\times I}^{(\ell)}} \|E_b\|_2 \leq c_{\mathrm{sp}} c_E \varepsilon h^{-1} \|M\|_2 \sum_{\ell=0}^{L(T_I)-1} \operatorname{mindiam}_\ell^{-1}.$$

Die Behauptung ergibt sich mit Lemma 3.14. □

Vergleicht man die vorherige Aussage mit der schwächeren Approximationsbedingung (3.18) in Satz 3.11, so hängt diese nicht mehr von der Tiefe des Clusterbaumes ab.

In den folgenden numerischen Ergebnissen wollen wir belegen, dass die Abschätzung in Satz 3.15 scharf ist.

3.4 Numerische Ergebnisse

Sämtliche Tests in dieser Arbeit wurden auf einem einzelnen Kern eines Intel Xeon X5482 Prozessors mit 3.2 GHz und 64 GB Arbeitsspeicher durchgeführt. Als Programmierbasis wurde die $\mathcal{H}$-Matrix-Bibliothek A$\mathcal{H}$MED verwendet.

Das Approximationsresultat aus Satz 3.15 soll anhand des folgenden akademischen Beispiels belegt werden. Sei $\Omega = (0,1)^3$ der Einheitswürfel und das Dirichletproblem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{auf } \partial\Omega \end{aligned}$$

gegeben. Für die Diskretisierung wurden lineare Ansatzfunktionen gewählt. Bei der $\mathcal{H}$-Matrix-Partitionierung kam die Bounding-Box-Methode zum Einsatz [Gie01] und die minimale Blockgröße wurde auf $n_{\min} = 20$ gesetzt. Anschließend wurde die Steifigkeitsmatrix durch eine hierarchische Cholesky-Zerlegung approximiert.

Wie sich in Abbildung 6 zeigt, schwankt die relative Approximationsgenauigkeit $\|A - \tilde{L}\tilde{L}^T\|_2 / \|A\|_2$, welche mittels Vektoriteration berechnet wurde, um einen Wert von ca. $0{,}035$. Die Abweichungen können zum Beispiel durch leichte Veränderungen in den Blockclusterbäumen entstehen. Die numerischen Ergebnisse jedoch zeigen, dass kein besseres Verhalten als in Satz 3.15 zu erwarten ist.

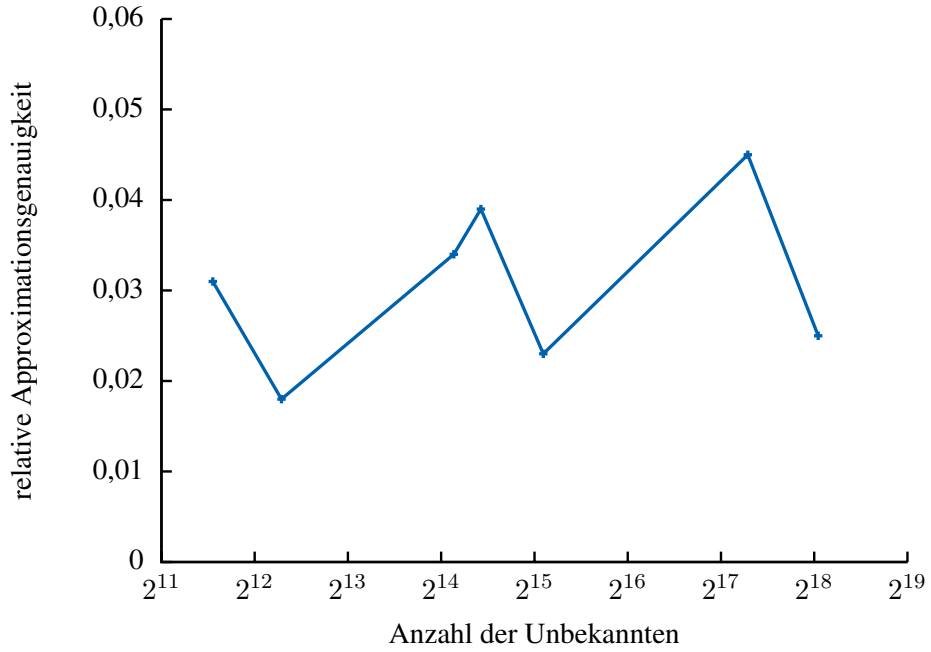


Abbildung 6: Die relative Approximationsgenauigkeit $\|A - \tilde{L}\tilde{L}^T\|_2/\|A\|_2$ der $\mathcal{H}$-Cholesky-Zerlegung mit blockweiser Approximationsgenauigkeit $\varepsilon = 0{,}1$ in Abhängigkeit von der Anzahl der Unbekannten.

4 Hierarchische Cholesky-Zerlegung mit Erhalt von Nebenbedingungen[3]

In diesem Kapitel betrachten wir Vorkonditionierer mit dem exakten Erhalt von gewissen Unterräumen auf der gesamten Matrix oder einzelnen Blöcken der Partition. Dabei ist das Ziel, die Konditionszahl des vorkonditionierten Systems zu verbessern. Diese Erkenntnisse werden anschließend auf die $\mathcal{H}$-Cholesky-Zerlegung angewendet.

Unter Nebenbedingungen verstehen wir:

Definition 4.1 (starke Nebenbedingungen)**.** *Seien $X \in \mathbb{R}^{J\times r}$, $Y \in \mathbb{R}^{I\times r}$ und $A, \tilde{A} \in \mathbb{R}^{I\times J}$. Die Matrix $\tilde{A}$ erfüllt bzgl. A die starken Nebenbedingungen, falls*

$$AX = \tilde{A}X \quad \text{und} \quad A^T Y = \tilde{A}^T Y \tag{4.1}$$

gilt.

Definition 4.2 (schwache Nebenbedingungen)**.** *Seien $X \in \mathbb{R}^{J\times r}$, $Y \in \mathbb{R}^{I\times r}$ und $A, \tilde{A} \in \mathbb{R}^{I\times J}$. Die Matrix $\tilde{A}$ erfüllt bzgl. A die schwachen Nebenbedingungen, falls*

$$Y^T AX = Y^T \tilde{A}X \tag{4.2}$$

gilt.

Es ist offensichtlich, dass die starken Nebenbedingungen die schwachen implizieren. Im Falle symmetrischer Matrizen können wir o.B.d.A. $X = Y$ annehmen.

Eine spezielle Wahl der Nebenbedingungen ist $X, Y = (1, \ldots, 1)^T$. Es ergibt sich aus (4.1) der Erhalt der Zeilen- bzw. Spaltensumme. Die Bedingung (4.2) liefert den Erhalt der Summe aller Matrixeinträge.

4.1 Bekannte Verfahren mit dem Erhalt von Nebenbedingungen

Die unvollständige Cholesky-Zerlegung (ICC, incomplete Cholesky decomposition) für den Standard-Fünf-Punkte-Stern der Laplace-Gleichung profitiert wesentlich von dem Erhalt der Zeilen- und Spaltensumme. Inspiriert dadurch untersuchen wir, ob es sinnvoll ist, dieses Prinzip auf $\mathcal{H}$-Matrizen zu übertragen. Anschließend betrachten wir die Technik der geglätteten Aggregation (smoothed aggregation), verwendet bei dem algebraische Mehrgitterverfahren (AMG-Verfahren).

4.1.1 ICC mit Nebenbedingungen

Die ICC ist eine fehlerbehaftete Zerlegung einer s.p.d. Matrix $A \in \mathbb{R}^{I\times I}$ in die Faktoren $A \approx \tilde{A} = \tilde{L}\tilde{L}^T$, wobei $\tilde{L}$ eine untere Dreiecksmatrix ist. Im Allgemeinen werden bei der Berechnung dieser approximativen Zerlegung gewisse Matrixeinträge unterdrückt, um eine effiziente Berechnung zu gewährleisten; siehe [Var60, Saa03].

[3] Teile aus diesem Abschnitt wurden veröffentlicht in [BBB11, BBBb].

Bei einer modifizierten Variante, siehe [MV77], werden die unterdrückten Elemente auf die Hauptdiagonale verschoben. Es gelang Gustafsson in [Gus78] für den Standard-Fünf-Punkte-Stern des Laplace-Operators zu zeigen, dass die Konditionszahl des vorkonditionierten Systems von $\mathcal{O}(h^{-2})$ auf $\mathcal{O}(h^{-1})$ reduziert werden kann, falls die Approximationsbedingung (4.1) mit $X, Y = (1, \dots, 1)^T$ für die Systemmatrix gilt.

Diese Idee der Erhaltung von Nebenbedingungen wollen wir im nächsten Abschnitt aufgreifen und auf hierarchische Matrizen anwenden.

4.1.2 Hierarchische Matrizen mit globalen Nebenbedingungen

Eine Möglichkeit, Bedingung (4.1) für $\mathcal{H}$-Matrizen zu erfüllen, wurde in [Hac09, Kapitel 6.8.1] präsentiert. Bei diesem Ansatz korrigiert man eine $\mathcal{H}$-Matrix-Approximation $\hat{A}$ von A mit einer Rang-r-Matrix

$$\tilde{A} := \hat{A} + \delta A, \quad \delta A := (AX - \hat{A}X)(X^T X)^{-1} X^T.$$

Es ist ersichtlich, dass

$$\tilde{A}X = AX.$$

Da der Rang von δA durch r beschränkt ist, kann die Approximation $\tilde{A}$ im Format $\mathcal{H}(\mathcal{P}, k + r)$ abgespeichert werden, falls $\hat{A} \in \mathcal{H}(\mathcal{P}, k)$.

Der Vorteil dieser Methode ist die einfache Implementierbarkeit. Es ist möglich, eine Rang-r-Matrix effizient auf eine $\mathcal{H}$-Matrix zu addieren, falls r klein ist. Der Aufwand erhöht sich jedoch, falls man dies auf die $\mathcal{H}$-LU-Zerlegung von A übertragen möchte. Da δA nicht einfach auf die Approximation addiert werden kann, benötigt man zum Beispiel ein Update der Faktoren L und U wie in [BC07].

Im folgenden Lemma untersuchen wir den Effekt von globalen Nebenbedingungen auf die Konditionszahl des vorkonditionierten Systems.

Lemma 4.3. *Seien $A, \tilde{A} \in \mathbb{R}^{n \times n}$ s.p.d. Matrizen, die die starken Nebenbedingungen* (4.1) *mit $X = Y$ erfüllen. Falls für $0 \leq \varepsilon < \tau$ zusätzlich gilt:*

$$\|\tilde{A} - A\|_2 \leq \varepsilon \|A\|_2$$

und

$$\|(\mathrm{Id} - XX^T)A^{-1}(\mathrm{Id} - XX^T)\|_2 \leq \frac{1}{\tau \|A\|_2},$$

dann erhält man

$$\kappa(\tilde{A}^{-1}A) \leq \frac{1 + \varepsilon/\tau}{1 - \varepsilon/\tau}.$$

Beweis. Da $\tilde{A} - A = (\tilde{A} - A)(\mathrm{Id} - XX^T)$, ergibt sich $\tilde{A} - A = (\mathrm{Id} - XX^T)^T(\tilde{A} - A)(\mathrm{Id} - XX^T)$.

Somit erhält man

$$\begin{aligned}
\| \operatorname{Id} - A^{-1/2}\tilde{A}A^{-1/2}\|_2 &= \|A^{-1/2}(\tilde{A}-A)A^{-1/2}\|_2 \\
&= \|A^{-1/2}(\operatorname{Id} - XX^T)(\tilde{A}-A)(\operatorname{Id} - XX^T)A^{-1/2}\|_2 \\
&\leq \|A^{-1/2}(\operatorname{Id} - XX^T)\|_2 \, \|\tilde{A}-A\|_2 \, \|(\operatorname{Id} - XX^T)A^{-1/2}\|_2 \\
&= \|A^{-1/2}(\operatorname{Id} - XX^T)\|_2^2 \, \|\tilde{A}-A\|_2 \\
&= \|(\operatorname{Id} - XX^T)A^{-1}(\operatorname{Id} - XX^T)\|_2 \, \|\tilde{A}-A\|_2 \\
&\leq \frac{1}{\tau\|A\|_2}\,\varepsilon\|A\|_2 = \frac{\varepsilon}{\tau} < 1.
\end{aligned}$$

Mit Satz 1.20 folgt die Behauptung. □

Bisher wurde die zu erhaltende Matrix X nicht konkret angegeben. Die Eigenvektoren zu den kleinsten Eigenwerten von A sind die optimale Wahl um den Ausdruck $\|(\operatorname{Id} - XX^T)A^{-1}(\operatorname{Id} - XX^T)\|_2$ für ein festen Rang zu minimieren. Dies kann konstruktiv mittels der Singulärwertzerlegung gezeigt werden.

Seien λ_i, $i = 1, \ldots, n$, die absteigend sortierten Eigenwerte von A und v_i, $i = 1, \ldots, n$, die dazugehörigen orthonormalen Eigenvektoren.

Lemma 4.4. *Sei $k \in \{1, \ldots, n\}$. Dann gilt, dass*

$$\min_{X \in \mathbb{R}^{n\times k}} \|(\operatorname{Id} - XX^T)A^{-1}(\operatorname{Id} - XX^T)\|_2 = \|(\operatorname{Id} - V_kV_k^T)A^{-1}(\operatorname{Id} - V_kV_k^T)\|_2 = \lambda_{n-k}^{-1},$$

wobei $V_k := (v_n, \ldots, v_{n-k+1})$.

Beweis. Da A eine s.p.d. Matrix ist, lässt sich die Inverse mithilfe der Eigenwerte und -vektoren von A darstellen als

$$A^{-1} = (v_1, \ldots, v_n) \begin{pmatrix} \lambda_1^{-1} & & \\ & \ddots & \\ & & \lambda_n^{-1} \end{pmatrix} (v_1, \ldots, v_n)^T.$$

Aus der Orthogonalität der Eigenvektoren ergibt sich, dass

$$\|(\operatorname{Id} - V_kV_k^T)A^{-1}(\operatorname{Id} - V_kV_k^T)\|_2 = \lambda_{n-k}^{-1}.$$

Die optimale Wahl der Basis ergibt sich aus der Bestapproximationseigenschaft der Singulärwertzerlegung. □

Wir untersuchen nun, wie sich die Konditionszahl bei optimaler Wahl der zu erhaltenden Basis $X = V_k$ beschränken lässt.

Korollar 4.5. *Seien die Voraussetzungen wie in Lemma 4.3 mit $X = Y = (v_n, \ldots, v_{n-k+1})$. Dann ergibt sich, dass*

$$\kappa(\tilde{A}^{-1}A) \leq \frac{1 + \frac{\varepsilon\|A\|_2}{\lambda_{n-k}}}{1 - \frac{\varepsilon\|A\|_2}{\lambda_{n-k}}}.$$

Beweis. Die Behauptung folgt direkt aus Lemma 4.3 und Lemma 4.4. □

Vergleicht man Korollar 4.5 mit Satz 1.19 und 1.20, dann zeigt sich, dass der Erhalt von konstant vielen Vektoren keine Verbesserung der Ordnung der Konditionszahl liefert. Es müssten $\mathcal{O}(n)$ viele Vektoren erhalten werden, um $\kappa(\tilde{A}^{-1}A)$ gegen eine Konstante abzuschätzen. Dies überschreitet jedoch die Speicher- und Laufzeitanforderungen der hierarchischen Matrizen.

Beispiel 4.6. Wir analysieren den Effekt einer globalen Rang-Eins-Korrektur für einen $\mathcal{H}$-Matrix-Vorkonditionierer. Hierfür betrachten wir ein zweidimensionales Poisson-Problem mit Dirichlet-Randdaten auf dem Einheitsquadrat $\Omega = (0,1)^2$:

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega, \\ u &= 0 \quad \text{auf } \partial\Omega. \end{aligned}$$

Das aus der FE-Methode resultierende Gleichungssystem wird mit dem PCG-Verfahren bis auf eine Genauigkeit von $1e-10$ gelöst, wobei die hierarchische Cholesky-Zerlegung mit einer blockweisen Approximationsgenauigkeit von $0{,}1$ als Vorkonditionierer dient.

In unserem Test wird die gewöhnliche $\mathcal{H}$-Cholesky-Zerlegung mit der korrigierten $\mathcal{H}$-Cholesky-Zerlegung verglichen. Dabei erhält die korrigierte Version die starken Nebenbedingungen (4.1) für den Vektor $X := [1, \ldots, 1]^T$. Dieser Vektor bietet sich an, da die Eigenvektoren zu den kleinen Eigenwerten niederfrequente Schwingungen sind und gut durch einen konstanten Vektor approximiert werden können.

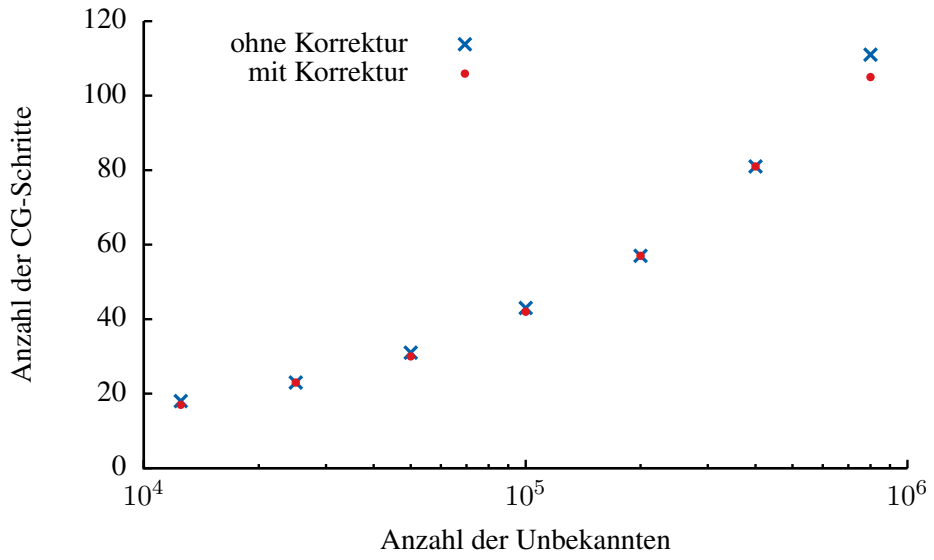


Abbildung 7: Hierarchische Cholesky-Zerlegung als Vorkonditionierer, mit und ohne globaler Korrektur.

Bereits an diesem einfachen Problem wird ersichtlich, dass sich das asymptotische Verhalten der $\mathcal{H}$-Matrix-Vorkonditionierung nicht ändert, siehe Abbildung 7. Die Iterationszahlen verbessern sich trotz des erhöhten Aufwands nur unwesentlich.

4.1.3 AMG-Verfahren mit geglätteter Aggregation

Mehrgitterverfahren sind iterative Verfahren zur Lösung von elliptischen Differentialgleichungen, siehe [TOS00]. Die wesentliche Idee dabei ist, einfache iterative Verfahren zu verwenden und diese durch globale Korrekturen zu beschleunigen, welche aus den Lösungen von vergröberten Problemen entstanden sind. Eine wichtige Eigenschaft von Mehrgitterverfahren ist, dass die Anzahl der Iterationsschritte nicht von der Kondition der Systemmatrix abhängt. Somit eignen sich diese Verfahren zum Beispiel als spektraläquivalenter Vorkonditionierer.

Eine bestimmte Klasse von Mehrgitterverfahren sind die algebraischen Mehrgitterverfahren, welche keine Informationen der Geometrie des zugrundeliegenden Problems verwenden, sondern nur die Systemmatrix benötigen, siehe [Sha03]. Diese Methoden zeichnen sich vor allem durch ihren black-box-artigen Charakter und ihre schnellen Laufzeiteigenschaften aus.

Mehrgitterverfahren verwenden eine Gitter-Hierarchie, für die ein Restriktions- und Prolongationsoperator definiert werden muss. Eine Möglichkeit, den Prolongationsoperator für AMG-Verfahren zu wählen, ist die geglättete Aggregation, vorgestellt in [Van92, Van95]. Diese entspricht einer diskreten konstanten Interpolation, wobei eine Menge von Einträgen der Systemmatrix als Aggregation zusammengefasst wird.

Für die geglättete Aggregation konnte in [VBM01, BVV12] gezeigt werden, dass die Konditionszahl des vorkonditionierten Systems nur polynomiell von der Tiefe der Gitter-Hierarchie abhängt und nicht exponentiell. Inspiriert durch die Aggregation wollen wir auf den Blöcken der Partition einer $\mathcal{H}$-Matrix verschiedene Nebenbedingungen erhalten.

4.2 Lokaler Erhalt von starken und schwachen Nebenbedingungen

Die Approximationsbedingung (3.6) soll durch den Erhalt von zusätzlichen Vektoren im Sinne von (4.1) und (4.2) ergänzt werden. Wir beschränken uns dabei auf die Vektoren

$$\mathbf{1}_s := \begin{cases} 1, & i \in s, \\ 0, & \text{sonst}, \end{cases} \in \mathbb{R}^I$$

für $s \subset I$. So soll die Approximation des Schurkomplementes bzgl. eines Blockes $t \times s \in \mathcal{P}$ zusätzlich die starken

$$(\tilde{S}_a(t,s) - S_a(t,s))\mathbf{1}_s = 0 \quad \text{und} \quad (\tilde{S}_a(t,s) - S_a(t,s))^T\mathbf{1}_t = 0 \tag{4.3}$$

bzw. schwachen Nebenbedingungen

$$\mathbf{1}_t^T(\tilde{S}_a(t,s) - S_a(t,s))\mathbf{1}_s = 0 \tag{4.4}$$

erhalten.

Die einzigen nicht exakten Operationen in der $\mathcal{H}$-Matrix-Arithmetik treten bei der Approximation der Schurkomplemente auf. Mit Hilfe von Satz 3.3 folgt aus (4.3) und (4.4), dass

$$(\tilde{L}\tilde{L}^T)_{ts}\mathbf{1}_s = A_{ts}\mathbf{1}_s, \qquad (\tilde{L}\tilde{L}^T)_{ts}^T\mathbf{1}_t = A_{ts}^T\mathbf{1}_t$$

bzw.

$$\mathbf{1}_t^T(\tilde{L}\tilde{L}^T)_{ts}\mathbf{1}_s = \mathbf{1}_t^T A_{ts}\mathbf{1}_s.$$

Dadurch erhält die so gewonnene Approximation blockweise die entsprechenden Nebenbedingungen.

In dem folgenden Abschnitt werden wir zeigen, dass solche Nebenbedingungen zu einer wesentlichen Verbesserung des Vorkonditionierungseffektes führen können.

4.2.1 Spektraläquivalenz

In diesem Abschnitt betrachten wir zuerst die gewöhnliche $\mathcal{H}$-Cholesky-Zerlegung und deren Vorkonditionierungseffekt. Anschließend untersuchen wir Zerlegungen mit dem Erhalt von Nebenbedingungen. Dabei unterscheiden wir jeweils zwei Fälle, zum einen eine konstante Approximationsgenauigkeit und zum anderen inwiefern die Approximationsgenauigkeit angepasst werden muss um Spektraläquivalenz zu garantieren.

Der blockweise Rang einer $\mathcal{H}$-Matrix-Approximation für elliptische partielle Differentialgleichungen hängt nur logarithmisch von der Approximationsgenauigkeit ab, siehe [Beb08, Hac09] oder den folgenden Abschnitt 4.2.2. Somit ist eine polynomielle Verkleinerung der Genauigkeit in Abhängigkeit des Diskretisierungsparameters h möglich, ohne die Gesamtkomplexität der $\mathcal{H}$-Matrix-Arithmetik zu zerstören. Um jedoch einen möglichst effizienten Vorkonditionierer zu erhalten, sollte die Approximationsgenauigkeit auf allen Blöcken optimal gewählt werden.

Zunächst folgt ein allgemeines Lemma zur Beschränkung der Konditionszahl des vorkonditionierten Systems. Sei hierfür der Fehler levelweise aufgespaltet

$$A - \tilde{A} = E = \sum_{\ell=0}^{L(T_I)-1} E_\ell,$$

mit $E_\ell := \sum_{b\in\mathcal{P}\cap T_{I\times I}^{(\ell)}} E_b$, $\ell = 0,\dots,L(T_I)-1$. Wir fordern, dass die Approximation $\tilde{A}$ von A im folgenden Sinne auf gewissen Unterräumen $P_\ell \in \mathbb{R}^{I\times I}$ exakt ist

$$P_\ell^T E_\ell P_\ell = E_\ell, \quad \ell = 0,\dots,L(T_I)-1. \tag{4.5}$$

Satz 4.7. *Seien $A, \tilde{A}$ s.p.d. Matrizen mit Eigenschaft* (4.5)*. Falls eine Konstante $0 \le \delta < 1$ existiert, so dass*

$$\sum_{\ell=0}^{L(T_I)-1} \|P_\ell A^{-1} P_\ell^T\|_2 \|E_\ell\|_2 \le \delta,$$

dann ergibt sich

$$\kappa(\tilde{A}^{-1}A) \le \frac{1+\delta}{1-\delta}.$$

Beweis. Durch eine levelweise Aufspaltung des Fehlers erhält man folgende Abschätzung

$$\begin{aligned}\| \operatorname{Id} - A^{-1/2}\tilde{A}A^{-1/2}\|_2 = \|A^{-1/2}EA^{-1/2}\|_2 &\leq \sum_{\ell=0}^{L(T_I)-1} \|A^{-1/2}E_\ell A^{-1/2}\|_2 \\ &= \sum_{\ell=0}^{L(T_I)-1} \|A^{-1/2}P_\ell^T E_\ell P_\ell A^{-1/2}\|_2 \leq \sum_{\ell=0}^{L(T_I)-1} \|E_\ell\|_2 \|P_\ell A^{-1} P_\ell^T\|_2 \leq \delta.\end{aligned}$$

Mit Satz 1.20 folgt die Behauptung. □

Im Folgenden werden wir verschiedene $\mathcal{H}$-Cholesky-Zerlegungen als Vorkonditionierer untersuchen. Dies bedeutet eine unterschiedliche Wahl der Projektoren P_ℓ und verschiedene blockweise Approximationsgenauigkeiten. Dabei beschränken wir uns im Wesentlichen auf die Typen:

- `Prec0`: die gewöhnliche $\mathcal{H}$-Cholesky-Zerlegung mit fester Approximationsgenauigkeit;
- `Prec0_S`: `Prec0` mit adaptiver blockweiser Genauigkeit;
- `PrecS`: $\mathcal{H}$-Cholesky-Zerlegung mit blockweisem Erhalt der starken Nebenbedingungen (4.3);
- `PrecS_S`: `PrecS` mit adaptiver blockweiser Genauigkeit;
- `PrecW`: $\mathcal{H}$-Cholesky-Zerlegung mit blockweisem Erhalt der schwachen Nebenbedingungen (4.4);
- `PrecW_S`: `PrecW` mit adaptiver blockweiser Genauigkeit.

Die levelweisen Genauigkeiten der Vorkonditionierer `Prec0_S`, `PrecS_S` und `PrecW_S` werden jeweils in den Sätzen 4.8, 4.12 und 4.15 festgelegt, so dass ein spektraläquivalenter Vorkonditionierer, in Bezug auf die Freiheitsgrade, garantiert ist.

Zuerst betrachten wir die $\mathcal{H}$-Cholesky-Zerlegung ohne Erhalt von Nebenbedingungen (`Prec0` und `Prec0_S`).

Keine Nebenbedingungen (`Prec0`, `Prec0_S`): Der folgende Satz gibt eine Beschränkung der Konditionszahl für einen gewöhnlichen $\mathcal{H}$-Cholesky-Vorkonditionierer an. Offensichtlich gilt in diesem Fall $P_\ell \equiv \operatorname{Id}$ für Bedingung (4.5).

Satz 4.8. *Sei eine gewöhnliche $\mathcal{H}$-Cholesky-Zerlegung gegeben. Dann gelten die folgenden Aussagen:*

- *(`Prec0`) Sei $\varepsilon_\ell \equiv \varepsilon$ mit $\ell = 0, \ldots, L(T_I) - 1$ und $0 \leq \varepsilon < h^2/c_\alpha$. Dann folgt, dass*

$$\kappa\big((\tilde{L}\tilde{L}^T)^{-1}A\big) \leq \frac{1 + c_\alpha \varepsilon h^{-2}}{1 - c_\alpha \varepsilon h^{-2}},$$

  wobei $c_\alpha := c_B c_b c_E c_{\mathrm{sp}} c_J^{-4}$.

- *(`Prec0_S`) Sei $\varepsilon_\ell \equiv \varepsilon L(T_I)^{-1} h \operatorname{mindiam}_\ell$ mit $\ell = 0, \ldots, L(T_I) - 1$ und $0 \leq \varepsilon < 1/c_\beta$. Dann*

folgt, dass

$$\kappa\big((\tilde{L}\tilde{L}^T)^{-1}A\big) \leq \frac{1+c_\beta\varepsilon}{1-c_\beta\varepsilon},$$

wobei $c_\beta := c_B c_E c_{\text{sp}} c_J^{-4}$.

Beweis. Mit Hilfe von Lemma 1.17 und (2.9) folgt, dass

$$\sum_{\ell=0}^{L(T_I)-1} \|A^{-1}\|_2 \|E_\ell\|_2 \leq c_B c_{\text{sp}} \|M^{-1}\|_2 \sum_{\ell=0}^{L(T_I)-1} \max_{b\in\mathcal{P}\cap T_{I\times I}^{(\ell)}} \|E_b\|_2.$$

Somit ergibt sich durch Lemma 3.13 und Korollar 1.15 die folgende Abschätzung

$$\sum_{\ell=0}^{L(T_I)-1} \|A^{-1}\|_2 \|E_\ell\|_2 \leq c_B c_E c_{\text{sp}} c_J^{-4} h^{-1} \sum_{\ell=0}^{L(T_I)-1} \varepsilon_\ell \text{mindiam}_\ell^{-1}.$$

Die Behauptungen folgen mit Lemma 3.14 und Satz 4.7. □

Satz 4.8 zeigt, dass mit einer konstanten Approximationsgenauigkeit auf den Blöcken die Konditionszahl des vorkonditionierten Systems auf $\mathcal{O}(h^{-2})$ beschränkt ist. Bei einer spektraläquivalenten Version muss die Genauigkeit bei den kleinen Blöcke mit $\mathcal{O}(h^2)$ und bei den großen mit $\mathcal{O}(h)$ umskaliert werden.

Starke Nebenbedingungen (`PrecS`, `PrecS_S`): Nun betrachten wir die $\mathcal{H}$-Cholesky-Zerlegung mit dem blockweisen Erhalt der starken Nebenbedingungen.

Für jedes Level definieren wir die folgende Matrix

$$Q_\ell := \sum_{r\in T_I^{(\ell)}} \frac{\mathbf{1}_r \mathbf{1}_r^T}{|r|} \in \mathbb{R}^{I\times I}, \quad \ell = 0,\dots,L(T_I)-1.$$

Aus den starken Nebenbedingungen (4.3) und Lemma 3.3 folgt, dass

$$E_\ell Q_\ell = \sum_{b\in\mathcal{P}\cap T_{I\times I}^{(\ell)}} E_b Q_\ell = 0, \quad \ell = 0,\dots,L(T_I)-1.$$

Somit gilt Bedingung (4.5) für die Wahl $P_\ell^S := \text{Id} - Q_\ell$.

Um δ in Satz 4.7 explizit anzugeben, benötigen wir eine Schranke für den Ausdruck $\|P_\ell^S A^{-1} P_\ell^S\|_2$.

Lemma 4.9. *Falls für* P_ℓ^S *ein* $\varepsilon_\ell > 0$ *existiert, so dass* $\|\mathcal{J} P_\ell^S x\|_{L^2(\Omega)} \leq \varepsilon_\ell \|\nabla \mathcal{J} x\|_{L^2(\Omega)}$ *für alle* $x \in \mathbb{R}^I$*, dann gilt*

$$\|P_\ell^S A^{-1} P_\ell^S\|_2 \leq \varepsilon_\ell^2 \lambda_{\mathcal{L}}^{-1} \|M^{-1}\|_2.$$

Beweis. Mit $\|P_\ell^S A^{-1} P_\ell^S\|_2 = \|P_\ell^S A^{-1/2}\|_2^2$ und Lemma (1.8) erhält man

$$\begin{aligned}\|P_\ell^S A^{-1} P_\ell^S\|_2 &= \sup_{x\neq 0} \frac{\|P_\ell^S A^{-1/2} x\|_2^2}{\|x\|_2^2} = \sup_{y\neq 0} \frac{\|P_\ell^S y\|_2^2}{y^T A y} = \sup_{y\neq 0} \frac{\|P_\ell^S y\|_2^2}{a(\mathcal{J}y, \mathcal{J}y)} \\ &\leq \lambda_{\mathcal{L}}^{-1} \sup_{y\neq 0} \frac{\|P_\ell^S y\|_2^2}{\|\nabla \mathcal{J} y\|_{L^2}^2} \leq \lambda_{\mathcal{L}}^{-1} \|M^{-1}\|_2 \sup_{y\neq 0} \frac{\|\mathcal{J} P_\ell^S y\|_{L^2(\Omega)}^2}{\|\nabla \mathcal{J} y\|_{L^2(\Omega)}^2}.\end{aligned}$$

Somit ergibt sich die Behauptung. □

Mit dem nächsten Lemma können wir eine obere Schranke für ε_ℓ in Lemma 4.9 angeben. Der Beweis beruht dabei im Wesentlichen auf dem Sobolevschen Normierungssatz, wird jedoch dadurch verkompliziert, dass konstante Funktionen im Allgemeinen kein Element aus unserem endlichdimensionalen Ansatzraum V_h sind. Inspiriert wurde die Äquivalenz der verschiedenen Massenmatrizen durch [VBM01].

Um weitere Betrachtungen zu vereinfachen, definieren wir analog zu mindiam_ℓ die Variable

$$\text{maxdiam}_\ell := \max_{t\in T_I^{(\ell)}} \operatorname{diam} X_t$$

für alle $\ell = 0, \ldots, L(T_I) - 1$.

Lemma 4.10. *Sei P_ℓ^S gegeben wie oben. Dann existiert eine Konstante $c_V > 0$, so dass*

$$\|\mathcal{J} P_\ell^S x\|_{L^2(\Omega)} \leq c_V \text{maxdiam}_\ell \|\nabla \mathcal{J} x\|_{L^2(\Omega)}$$

für alle $x \in \mathbb{R}^I$ und $\ell = 0, \ldots, L(T_I) - 1$.

Beweis. Für alle $t \in T_I^{(\ell)}$ definieren wir

$$\mathring{X}_t := \{\xi \in X_t : (\mathcal{J}\mathbf{1}_t)(\xi) = 1\}.$$

Diese Menge entspricht X_t ohne eine Randschicht der Größenordnung h. Somit gilt $\mathring{X}_t \cap \operatorname{int} X_i = \emptyset$ für alle $i \notin t$.

Um in den folgenden Abschätzungen Überschneidungen von Trägern der Ansatzfunktionen für verschiedene Cluster zu vermeiden, benötigen wir eine natürliche Injektion welche nur auf $\mathring{X}_t$ agiert. Sei $\mathring{\mathcal{J}}_t^* : L^2(\Omega) \to \mathbb{R}^t$ definiert als

$$x^T \mathring{\mathcal{J}}_t^* v = (\mathcal{J}x, v)_{L^2(\mathring{X}_t)} \quad \text{für alle } x \in \mathbb{R}^t,\ v \in L^2(\Omega).$$

Somit ist $\mathring{M}_t := \mathring{\mathcal{J}}_t^* \mathcal{J} \in \mathbb{R}^{t\times t}$ die Massenmatrix eingeschränkt auf $\mathring{X}_t$ und hat die Einträge

$$\mathring{m}_{ij} := (\varphi_i, \varphi_j)_{L^2(\mathring{X}_t)}, \quad i, j \in t.$$

Da eine quasi-uniforme Triangulierung von Ω eingeschränkt auf $\mathring{X}_t$ wieder eine quasi-uniforme Triangu-

lierung ist, gibt es analog zu (1.22) eine Konstante $\bar{c}_J$, so dass

$$h^{d/2}\|x\|_2 \leq \frac{1}{\bar{c}_J}\|\mathcal{J}x\|_{L^2(\mathring{X}_t)}.$$

Mit Hilfe dieser Schranke können wir die Norm von $\mathring{M}_t^{-1}$ abschätzen durch

$$\|\mathring{M}_t^{-1}\|_2 = \sup_{x\in\mathbb{R}^t}\frac{x^Tx}{x^T\mathring{M}_tx} = \sup_{x\in\mathbb{R}^t}\frac{x^Tx}{(\mathcal{J}x,\mathcal{J}x)_{L^2(\mathring{X}_t)}} \leq \bar{c_J}^{-2}h^{-d}.$$

Weiterhin bezeichnen wir M_t als Massenmatrix bzgl. X_t. Somit erhalten wir mit Lemma 1.13 für $x \in \mathbb{R}^t$, dass

$$\begin{aligned}\|JP_\ell^Sx\|^2_{L^2(X_t)} &= \|M_t^{1/2}P_\ell^Sx\|_2^2 \leq \|M_t\|_2\|P_\ell^Sx\|_2^2 \leq \|M_t\|_2\|\mathring{M}_t^{-1}\|_2\|\mathring{M}_t^{1/2}P_\ell^Sx\|_2^2\\ &= \|M_t\|_2\|\mathring{M}_t^{-1}\|_2\|JP_\ell^Sx\|^2_{L^2(\mathring{X}_t)} \leq c_J^{-2}\bar{c}_J^{-2}\|JP_\ell^Sx\|^2_{L^2(\mathring{X}_t)}.\end{aligned}$$

Durch Umformung ergibt sich

$$JP_\ell^Sx = \sum_{t\in T_I^{(\ell)}} J\left(x_t - \sum_{r\in T_I^{(\ell)}}\frac{\mathbf{1}_r^Tx_t}{|r|}\mathbf{1}_r\right) = \sum_{t\in T_I^{(\ell)}} Jx_t - \frac{\mathbf{1}_t^Tx_t}{|t|}J\mathbf{1}_t, \quad x\in\mathbb{R}^t.$$

Sei ein lineares Funktional $F : H^1(\Omega) \to \mathbb{R}$ gegeben als $F(v) := \frac{1}{|t|}\mathbf{1}_t^T\mathring{M}_t^{-1}\mathring{J}_t^*v$. Für die konstante Funktion c gilt, dass

$$x^T\mathring{M}_t^{-1}\mathring{J}_t^*c = (J\mathring{M}_t^{-1}x, c)_{L^2(\mathring{X}_t)} = c(J\mathring{M}_t^{-1}x, J\mathbf{1}_t)_{L^2(\mathring{X}_t)} = c\,x^T\mathbf{1}_t, \quad x\in\mathbb{R}^t$$

und somit $\mathring{M}_t^{-1}\mathring{J}_t^*c = c\,\mathbf{1}_t$. Somit erhält man aus $F(c) = 0$ sofort $c = 0$. Aus diesem Grunde kann der Normierungssatz von Sobolev, siehe Satz 1.7, auf $\mathring{X}_t$, $t \in T_I^{(\ell)}$ angewendet werden. Somit existiert eine Konstante $c > 0$, so dass

$$\begin{aligned}\left\|Jx_t - \frac{\mathbf{1}_t^Tx_t}{|t|}J\mathbf{1}_t\right\|_{L^2(\mathring{X}_t)} &= \|Jx_t - F(Jx_t)\|_{L^2(\mathring{X}_t)} \leq c\,\mathrm{diam}\,\mathring{X}_t\,\|\nabla Jx_t\|_{L^2(\mathring{X}_t)}\\ &= c\,\mathrm{diam}\,\mathring{X}_t\,\|\nabla Jx\|_{L^2(\mathring{X}_t)}.\end{aligned}$$

Mit dieser Schranke erhalten wir unsere gewünschte Abschätzung

$$\begin{aligned}\|\mathcal{J}P_\ell^S x\|_{L^2(\Omega)}^2 &\leq \sum_{t\in T_I^{(\ell)}} \|\mathcal{J}P_\ell^S x\|_{L^2(X_t)}^2 \leq c_J^{-2}\bar{c}_J^{-2} \sum_{t\in T_I^{(\ell)}} \|\mathcal{J}P_\ell^S x\|_{L^2(\mathring{X}_t)}^2 \\ &= c_J^{-2}\bar{c}_J^{-2} \sum_{t\in T_I^{(\ell)}} \left\| \mathcal{J}x_t - \frac{\mathbf{1}_t^T x_t}{|t|}\mathcal{J}\mathbf{1}_t \right\|_{L^2(\mathring{X}_t)}^2 \\ &\leq c_J^{-2}\bar{c}_J^{-2} c \sum_{t\in T_I^{(\ell)}} \big(\operatorname{diam}\mathring{X}_t\big)^2 \|\nabla\mathcal{J}x\|_{L^2(\mathring{X}_t)}^2 \\ &\leq c_J^{-2}\bar{c}_J^{-2} c\,\operatorname{maxdiam}_\ell^2 \|\nabla Jx\|_{L^2(\Omega)}^2.\end{aligned}$$

□

Das nächste Korollar zeigt die Auswirkung des Projektors P_ℓ^S auf die Norm von A^{-1}.

Korollar 4.11. *Sei P_ℓ^S gegeben wie oben. Dann gilt*

$$\|P_\ell^S A^{-1} P_\ell^S\|_2 \leq c_P^2 \operatorname{maxdiam}_\ell^2 \|M^{-1}\|_2,$$

wobei $c_P := c_V/\sqrt{\lambda_{\mathcal{L}}}$.

Beweis. Die Behauptung folgt direkt aus Lemma 4.9 und Lemma 4.10. □

Der folgende Satz beinhaltet analoge Aussagen wie Satz 4.8 jedoch für die $\mathcal{H}$-Cholesky-Zerlegung mit dem Erhalt von starken Nebenbedingungen.

Satz 4.12. *Sei eine $\mathcal{H}$-Cholesky-Zerlegung mit dem Erhalt der starken Nebenbedingungen gegeben. Dann gelten die folgenden Aussagen:*

- *(`PrecS`) Sei $\varepsilon_\ell \equiv \varepsilon$ mit $\ell = 0, \ldots, L(T_I) - 1$ und $0 \leq \varepsilon < h/c_\delta$. Dann folgt, dass*

$$\kappa\big((\tilde{L}\tilde{L}^T)^{-1}A\big) \leq \frac{1 + c_\delta \varepsilon h^{-1}}{1 - c_\delta \varepsilon h^{-1}},$$

  mit $c_\delta := 5c_{\mathrm{sp}}c_P^2 c_E c_J^{-4} c_{D_2}$.

- *(`PrecS_S`) Sei $\varepsilon_\ell \equiv \varepsilon L(T_I)^{-1} h \operatorname{maxdiam}_\ell^{-1}$ mit $\ell = 0, \ldots, L(T_I) - 1$ und $0 \leq \varepsilon < 1/c_\gamma$. Dann folgt, dass*

$$\kappa\big((\tilde{L}\tilde{L}^T)^{-1}A\big) \leq \frac{1 + c_\gamma \varepsilon}{1 - c_\gamma \varepsilon},$$

  mit $c_\gamma := c_{\mathrm{sp}}c_P^2 c_E c_J^{-4}$.

Beweis. Mit Korollar 4.11 und (2.9) erhält man, dass

$$\sum_{\ell=0}^{L(T_I)-1} \|P_\ell^S A^{-1} P_\ell^S\|_2 \|E_\ell\|_2 \leq c_{\mathrm{sp}}c_P^2 \|M^{-1}\|_2 \sum_{\ell=0}^{L(T_I)-1} \operatorname{maxdiam}_\ell^2 \max_{b\in\mathcal{P}\cap T_{I\times I}^{(\ell)}} \|E_b\|_2.$$

Durch Lemma 3.13 und Korollar 1.15 ergibt sich die folgende Schranke

$$\sum_{\ell=0}^{L(T_I)-1} \|P_\ell^S A^{-1} P_\ell^S\|_2 \|E_\ell\|_2 \leq c_{\text{sp}} c_P^2 c_E c_J^{-4} h^{-1} \sum_{\ell=0}^{L(T_I)-1} \varepsilon_\ell \text{maxdiam}_\ell. \tag{4.6}$$

Die zweite Behauptung erhält man mit Satz 4.7.

Für den ersten Fall $\varepsilon_\ell \equiv \varepsilon$ erhält man mit (2.8) und für $d \in \{2, 3\}$, dass

$$\sum_{\ell=0}^{L(T_I)-1} \text{maxdiam}_\ell \leq c_{D_2} \sum_{\ell=0}^{L(T_I)-1} 2^{-\ell/d} \leq \frac{c_{D_2} \sqrt[d]{2}}{\sqrt[d]{2}-1} \leq 5 c_{D_2}.$$

und die erste Behauptung folgt mit Satz 4.7. □

Anhand von Satz 4.7 und Satz 4.12 sieht man, dass der Vorkonditionierungseffekt von `PrecS` um eine Ordnung besser ist als der von `Prec0`. Somit ist der Erhalt von Nebenbedingungen für hinreichend große Beispiele sinnvoll.

Weiterhin muss der spektraläquivalente Vorkonditionierer `PrecS_S` im Gegensatz zu `Prec0_S` nur mit dem Level skaliert werden und nicht mit dem Diskretisierungsparameter. So wird nur die Genauigkeit der großen Blöcke um $\mathcal{O}(h)$ skaliert. Die kleinen Blöcke sind bereits hinreichend gut approximiert.

Nun folgen die gleichen Betrachtungen für die schwachen Nebenbedingungen.

Schwache Nebenbedingungen (`PrecW`, `PrecW_S`): Die schwachen Nebenbedingungen sind, wie sich später zeigt, im Allgemeinen leichter zu implementieren als die starken und benötigen nur eine blockweise Rang-Eins statt eine Rang-Zwei-Korrektur. Aus diesem Grunde untersuchen wir nun deren Auswirkung auf die Konditionszahl des vorkonditionierten Systems.

Seien die schwachen Nebenbedingungen (4.4) für die s.p.d. Matrizen $A, \tilde{A}$ erfüllt. Dann folgt mit obiger Definition von Q_ℓ und Lemma 3.3, dass

$$Q_\ell E_\ell Q_\ell = \sum_{b \in \mathcal{P} \cap T_{I\times I}^{(\ell)}} Q_\ell E_b Q_\ell = 0.$$

Somit ergibt sich die folgende Aussage

$$E_\ell = (P_\ell^S + Q_\ell) E_\ell (P_\ell^S + Q_\ell) = P_\ell^S E_\ell P_\ell^S + Q_\ell E_\ell P_\ell^S + P_\ell^S E_\ell Q_\ell. \tag{4.7}$$

Um den Effekt der schwachen Nebenbedingungen abzuschätzen, müssen wir zuerst die Normen von Q_ℓ und P_ℓ^S beschränken.

Lemma 4.13. *Seien Q_ℓ und P_ℓ^S gegeben wie oben. Dann gilt, dass*

$$\|Q_\ell\|_2 \leq 1 \quad \text{und} \quad \|P_\ell^S\|_2 \leq 1.$$

Beweis. Die Matrix Q_ℓ ist symmetrisch und es gilt $Q_\ell = Q_\ell^2$. Somit erhält man für alle $x, y \in \mathbb{R}^{|I|}$, dass

$$(Q_\ell x, (\mathrm{Id} - Q_\ell)y) = (Q_\ell x, y) - (Q_\ell x, Q_\ell y) = 0.$$

Aufgrund der Orthogonalität ergibt sich

$$\|x\|_2^2 = \|Q_\ell x + (\mathrm{Id} - Q_\ell)x\|_2^2 = \|Q_\ell x\|_2^2 + \|(\mathrm{Id} - Q_\ell)x\|_2^2$$

und somit $\|Q_\ell x\|_2^2 \leq \|x\|_2^2$. Die zweite Behauptung für P_ℓ^S folgt analog. □

Mit Hilfe dieser Schranken ergibt sich der nächste Satz, welcher ähnlich ist zu Satz 4.7, jedoch auf der abgeschwächten Bedingung (4.7) basiert.

Satz 4.14. *Seien $A, \tilde{A}$ s.p.d. Matrizen mit Eigenschaft* (4.7)*. Falls eine Konstante $0 \leq \delta < 1$ existiert, so dass*

$$\sum_{\ell=0}^{L-1} 3\|A^{-1/2}\|_2 \|P_\ell^S A^{-1/2}\|_2 \|E_\ell\|_2 \leq \delta,$$

dann gilt

$$\kappa(\tilde{A}^{-1}A) \leq \frac{1+\delta}{1-\delta}.$$

Beweis. Mit Eigenschaft (4.7) ergibt sich, dass

$$\begin{aligned}\|\mathrm{Id} - A^{-1/2}\tilde{A}A^{-1/2}\|_2 = \|A^{-1/2}EA^{-1/2}\|_2 &\leq \sum_{\ell=0}^{L-1} \|A^{-1/2}E_\ell A^{-1/2}\|_2 \\ &= \sum_{\ell=0}^{L-1} \|A^{-1/2}\big(P_\ell^S E_\ell P_\ell^S + Q_\ell E_\ell P_\ell^S + P_\ell^S E_\ell Q_\ell\big)A^{-1/2}\|_2.\end{aligned}$$

Somit können die jeweiligen Terme aus der obigen Summe mit Lemma 4.13 beschränkt werden durch

$$\begin{aligned}\sum_{\ell=0}^{L-1} \|A^{-1/2}Q_\ell E_\ell P_\ell^S A^{-1/2}\|_2 &\leq \sum_{\ell=0}^{L-1} \|E_\ell\|_2 \|A^{-1/2}\|_2 \|P_\ell^S A^{-1/2}\|_2, \\ \sum_{\ell=0}^{L-1} \|A^{-1/2}P_\ell^S E_\ell P_\ell^S A^{-1/2}\|_2 &\leq \sum_{\ell=0}^{L-1} \|E_\ell\|_2 \|A^{-1/2}\|_2 \|P_\ell^S A^{-1/2}\|_2\end{aligned}$$

und es folgt

$$\|\mathrm{Id} - A^{-1/2}\tilde{A}A^{-1/2}\|_2 \leq \sum_{\ell=0}^{L-1} 3\|A^{-1/2}\|_2 \|P_\ell^S A^{-1/2}\|_2 \|E_\ell\|_2 \leq \delta.$$

Mit Satz 1.20 ergibt sich die Behauptung. □

Für eine $\mathcal{H}$-Cholesky-Zerlegung mit dem blockweisen Erhalt der schwachen Nebenbedingung wird im folgenden Satz die Konditionszahl des vorkonditionierten Systems abgeschätzt.

Satz 4.15. *Sei eine $\mathcal{H}$-Cholesky-Zerlegung mit dem Erhalt der schwachen Nebenbedingungen gegeben. Dann gelten folgende Aussagen:*

- *(`PrecW`) Sei $\varepsilon_\ell \equiv \varepsilon$ mit $\ell = 0, \dots, L(T_I) - 1$ und $0 \leq \varepsilon < h/(c_\mu L(T_I))$. Dann folgt, dass*

$$\kappa\big((\tilde{L}\tilde{L}^T)^{-1}A\big) \leq \frac{1 + c_\mu \varepsilon h^{-1} L(T_I)}{1 - c_\mu \varepsilon h^{-1} L(T_I)},$$

  mit $c_\mu := C_E c_P c_{\mathrm{sp}} c_B^{1/2} c_J^{-4}$.

- *(`PrecW_S`) Sei $\varepsilon_\ell \equiv \varepsilon h L(T_I)^{-1}$ mit $\ell = 0, \dots, L(T_I) - 1$ und $0 \leq \varepsilon < 1/c_\mu$. Dann folgt, dass*

$$\kappa\big((\tilde{L}\tilde{L}^T)^{-1}A\big) \leq \frac{1 + c_\mu \varepsilon}{1 - c_\mu \varepsilon}.$$

Beweis. Da $\|P_\ell^S A^{-1} P_\ell^S\|_2 = \|P_\ell^S A^{-1/2}\|_2^2$, ergibt sich mit Korollar 4.11, Lemma 1.17 und (2.9), dass

$$\sum_{\ell=0}^{L(T_I)-1} \|A^{-1/2}\|_2 \|P_\ell^S A^{-1/2}\|_2 \|E_\ell\|_2 \leq c_{\mathrm{sp}} c_P c_B^{1/2} \|M^{-1}\|_2 \sum_{\ell=0}^{L(T_I)-1} \mathrm{maxdiam}_\ell \max_{b \in \mathcal{P} \cap T_{I\times I}^{(\ell)}} \|E_b\|_2.$$

Durch Lemma 3.13 und 1.15 erhält man die folgende Schranke

$$\sum_{\ell=0}^{L(T_I)-1} \|A^{-1/2}\|_2 \|P_\ell^S A^{-1/2}\|_2 \|E_\ell\|_2 \leq c_E c_P c_{\mathrm{sp}} c_B^{1/2} c_J^{-4} h^{-1} \sum_{\ell=0}^{L(T_I)-1} \varepsilon_\ell.$$

Die Behauptungen folgen mit Satz 4.14. □

Die Konditionszahl des vorkonditionierten Systems einer $\mathcal{H}$-Cholesky-Zerlegung mit dem Erhalt von schwachen Nebenbedingungen `PrecW` ist bis auf logarithmische Terme von der gleichen Ordnung wie bei einer Zerlegung mit starken Nebenbedingungen `PrecS`.

Für einen spektraläquivalenten Vorkonditionierer muss die Approximationsgenauigkeit auf allen Blöcken mit $\mathcal{O}\big(hL(T_I)^{-1}\big)$ skaliert werden.

4.2.2 Komplexitätsanalyse

Der blockweise Rang k_b, $b \in \mathcal{P}$ einer $\mathcal{H}$-Cholesky-Zerlegung kann abgeschätzt werden durch

$$k_b \in \mathcal{O}\big(L(T_I)^\alpha |\log \varepsilon_b|^\beta\big), \tag{4.8}$$

mit Konstanten $\alpha, \beta > 0$. Für Details siehe [Beb07, Beb08].

Somit hängt der blockweise Rang nur logarithmisch von der Approximationsgenauigkeit ab. Dies ermöglicht eine levelweise Anpassung von ε_b, wie bei den Vorkonditionierern `Prec0_S`, `PrecS_S` und `PrecW_S`, ohne die quasi-lineare Komplexität der $\mathcal{H}$-Matrix-Arithmetik zu zerstören.

Der folgende Satz betrachtet $\mathcal{H}$-Cholesky-Vorkonditionierer mit konstanter Approximationsgenauigkeit.

Satz 4.16. *Der Gesamtspeicherbedarf N_{sp} der Vorkonditionierer `Prec0`, `PrecS` und `PrecW` beträgt*

$$N_{sp} \in \mathcal{O}\big(L(T_I)^{\alpha+1}|I|\big).$$

Beweis. Eine zusätzliche blockweise Erhaltung von konstant vielen Vektoren erhöht die Gesamtkomplexität der $\mathcal{H}$-Matrix-Arithmetik nicht, da sie nur eine konstante Rangerhöhung bedeutet. Somit folgt mit (4.8) und Satz 2.11 die Behauptung für die verschiedenen Vorkonditionierer. □

Für eine konstante Approximationsgenauigkeit sind somit die Vorkonditionierer `PrecS` und `PrecW` gegenüber `Prec0` für hinreichend große Beispiele zu bevorzugen, da der Vorkonditionierungseffekt um eine Ordnung besser ist, vgl. Abschnitt 4.2.1.

Nun betrachten wir die spektraläquivalenten Versionen der verschiedenen Vorkonditionierer.

Satz 4.17. *Der Gesamtspeicherbedarf der Vorkonditionierer `Prec0_S`, `PrecS_S` und `PrecW_S` beträgt*

$$N_{sp} \in \mathcal{O}(L(T_I)^{\alpha+\beta+1}|I|).$$

Beweis. Der Gesamtspeicherbedarf einer $\mathcal{H}$-Matrix ergibt sich durch die Summation über alle Blöcke der Partition

$$N_{sp} = \sum_{t\times s\in\mathcal{P}} k_{t\times s}(|t|+|s|) = \sum_{t\in T_I}\sum_{\{s\in T_I : t\times s\in\mathcal{P}\}} k_{t\times s}|t| + \sum_{s\in T_I}\sum_{\{t\in T_I : t\times s\in\mathcal{P}\}} k_{t\times s}|s|.$$

Für eine hinreichend große Anzahl an Unbekannten existiert mit (4.8) eine Konstante $c>0$, so dass

$$\sum_{t\in T_I}\sum_{\{s\in T_I : t\times s\in\mathcal{P}\}} k_{t\times s}|t| \le c\sum_{t\in T_I}\sum_{\{s\in T_I : t\times s\in\mathcal{P}\}} L(T_I)^{\alpha}|\log\varepsilon_{t\times s}|^{\beta}|t|.$$

Für die verschiedenen Vorkonditionierer `Prec0_S`, `PrecS_S` und `PrecW_S` gilt aufgrund der Balanciertheit des Clusterbaums, dass

$$|\log\varepsilon_{t\times s}|^{\beta} \sim |\log h|^{\beta} \sim |\log 2^{-L(T_I)/d}|^{\beta} \sim L(T_I)^{\beta}.$$

Somit folgt mit der Schwachbesetztheitskonstante (2.9) und einer hinreichend großen Konstante $\hat{c}>0$

$$\begin{aligned}\sum_{t\in T_I}\sum_{\{s\in T_I : t\times s\in\mathcal{P}\}} k_{t\times s}|t| &\le \hat{c}L(T_I)^{\alpha+\beta}\sum_{t\in T_I}\sum_{s\in T_I : t\times s\in\mathcal{P}}|t| \\ &\le \hat{c}c_{\mathrm{sp}}L(T_I)^{\alpha+\beta}\sum_{t\in T_I}|t| \le \hat{c}c_{\mathrm{sp}}L(T_I)^{\alpha+\beta+1}|I|\end{aligned}$$

die Behauptung. □

Bemerkung 4.18. Der Speicherbedarf der Vorkonditionierer `PrecS_S` und `PrecW_S` kann in Satz 4.17 nicht besser abgeschätzt werden, da die levelweise verschiedenen Genauigkeiten nur auf eine harmo-

nische und keine geometrische Reihe führen. Dies bedeutet, dass sich nur die Konstanten verbessern, jedoch nicht die Asymptotik.

Für eine spektraläquivalente Version der Vorkonditionierer `Prec0_S`, `PrecS_S` und `PrecW_S` mit Approximationsgenauigkeiten wie in Abschnitt 4.2.1 vorgeschlagen, ergeben sich keine asymptotischen Unterschiede im Speicherbedarf.

4.3 Methoden zum Erhalt von Nebenbedingungen

In diesem Abschnitt betrachten wir Methoden, um eine Niedrigrangmatrix $A \in \mathbb{R}_{k'}^{m\times n}$ durch eine weitere $\tilde{A} \in \mathbb{R}_{k}^{m\times n}$ mit $k' \geq k$ unter Erhalt der starken bzw. schwachen Nebenbedingungen mit $X \in \mathbb{R}^{n\times r}$, $Y \in \mathbb{R}^{m\times r}$, siehe (4.1) und (4.2), zu approximieren. Diese Methoden werden benötigt, um eine geeignete Approximation des Schurkomplements zu erstellen, siehe (3.6).

Im Wesentlichen erläutern wir zwei Ansätze. Der erste basiert auf der Gram-Schmidt-Methode und der zweite auf dem Householder-Verfahren.

4.3.1 Nebenbedingungen mittels Gram-Schmidt-Methode

Bei den Methoden aus diesem Abschnitt wird zuerst eine Approximation generiert und anschließend die Nebenbedingung durch eine Rang-r- bzw. Rang-$2r$-Korrektur erhalten. Für die folgenden Betrachtungen sei A durch ein äußeres Produkt $A = UV^T$ gegeben, wobei $U \in \mathbb{R}^{m\times k'}$ und $V \in \mathbb{R}^{n\times k'}$.

Starke Nebenbedingungen: Zum Erhalt der starken Nebenbedingungen (4.1) werden wir zuerst den Anteil von U und V senkrecht zu Y und X herausfiltern und ihn anschließend approximieren. Dazu definieren wir

$$U' := P(Y)U, \quad V' := P(X)V, \tag{4.9}$$

wobei

$$P(X) := \mathrm{Id} - X(X^TX)^{-1}X^T$$

ein orthogonaler Projektor ist mit $\|P(X)\|_2 \leq 1$, vgl. Lemma 4.13. Wichtig ist, dass das Produkt $U'V'^T$ bei gleichem Rang mindestens mit der selben Genauigkeit approximiert werden kann wie UV^T. Dies folgt aus

$$\begin{aligned}
\min_{B'\in\mathbb{R}_k^{m\times n}} \|U'V'^T - B'\|_2 &= \min_{\substack{B'\in\mathbb{R}_k^{m\times n}\\ B'=P(Y)B'P(X)}} \|U'V'^T - B'\|_2 \\
&= \min_{B\in\mathbb{R}_k^{m\times n}} \|P(Y)(UV^T - B)P(X)\|_2 \\
&\leq \min_{B\in\mathbb{R}_k^{m\times n}} \|P(Y)\|_2\|P(X)\|_2\|UV^T - B\|_2 \\
&\leq \min_{B\in\mathbb{R}_k^{m\times n}} \|UV^T - B\|_2.
\end{aligned}$$

Sei $\tilde{U}'\tilde{V}'^T$ eine Rang-k-Approximation von $U'V'^T$, definiert wie in (2.5). Da die Approximation mittels SVD den Kern erhält, folgt

$$Y^T U' = 0 = Y^T \tilde{U}', \quad X^T V' = 0 = X^T \tilde{V}'. \tag{4.10}$$

Nach der Approximation addieren wir zwei Rang-r-Matrizen und erhalten für $k + 2r < k'$ die Approximation $\tilde{A} := \tilde{U}\tilde{V}^T$, gegeben durch

$$\begin{aligned} \tilde{U} &:= [AX,\, Q(Y),\, \tilde{U}'] \in \mathbb{R}^{m\times(k+2r)}, \\ \tilde{V} &:= [Q(X),\, P(X)A^T Y,\, \tilde{V}'] \in \mathbb{R}^{n\times(k+2r)}, \end{aligned} \tag{4.11}$$

wobei $Q(X) := X(X^T X)^{-1}$. Für den Fall $k + 2r \geq k'$ ist eine Approximation in diesem Sinne nicht sinnvoll und wir setzen

$$\tilde{U} := U, \quad \tilde{V} := V.$$

Im folgenden Lemma zeigen wir, dass $\tilde{A} := \tilde{U}\tilde{V}^T \in \mathbb{R}^{m\times n}_{\min(k+2r,k')}$ die gewünschten Eigenschaften besitzt.

Lemma 4.19. *Sei $\tilde{A}$ konstruiert wie in* (4.11)*. Dann gilt $A - \tilde{A} = U'V'^T - \tilde{U}'\tilde{V}'^T$ und somit*

$$\|A - \tilde{A}\|_2 \leq \min_{B\in\mathbb{R}^{m\times n}_k} \|UV^T - B\|_2.$$

Weiterhin erfüllt $\tilde{A}$ die Bedingung (4.1)*.*

Beweis. Die erste Behauptung ergibt sich durch

$$\begin{aligned} A - \tilde{A} &= A - [AX,\, Q(Y),\, \tilde{U}'][Q(X),\, P(X)A^T Y,\, \tilde{V}']^T \\ &= A - AXQ(X)^T - Q(Y)Y^T A + Q(Y)Y^T AXQ(X)^T - \tilde{U}'\tilde{V}'^T \\ &= U'V'^T - \tilde{U}'\tilde{V}'^T. \end{aligned}$$

Mit Hilfe von Bedingung (4.10) werden die Vektoren erhalten, da

$$\tilde{A}X = \tilde{U}[Q(X),\, P(X)A^T Y,\, \tilde{V}']^T X = [AX,\, Q(Y),\, \tilde{U}'][\mathrm{Id},\, 0,\, 0]^T = AX$$

und

$$\begin{aligned} \tilde{A}^T Y &= \tilde{V}[AX,\, Q(Y),\, \tilde{U}']^T Y = [Q(X),\, P(X)A^T Y,\, \tilde{V}'][Y^T AX,\, \mathrm{Id},\, Y^T\tilde{U}']^T \\ &= Q(X)X^T A^T Y + A^T Y - Q(X)X^T A^T Y = A^T Y \end{aligned}$$

gilt. □

Schwache Nebenbedingungen: Es ist offensichtlich, dass aus

$$AX = \tilde{A}X \tag{4.12}$$

die schwachen Nebenbedingungen (4.2) folgen. Um solch ein $\tilde{A}$ zu generieren, werden wir die Approximation (4.11) abändern, so dass eine Rang-r-Korrektur genügt, um Bedingung (4.12) zu gewährleisten.

Sei V' definiert wie in (4.9). Weiterhin approximieren wir das Produkt UV'^T durch $U^*\tilde{V}^{*T}$ mittels der gewöhnlichen Kürzung aus (2.5). Die Approximationseigenschaften folgen analog zu denen der starken Nebenbedingung durch

$$\begin{aligned}
\min_{B' \in \mathbb{R}_k^{m\times n}} \|UV'^T - B'\|_2 &= \min_{\substack{B' \in \mathbb{R}_k^{m\times n} \\ B' = B'P(X)}} \|UV'^T - B'\|_2 \\
&= \min_{B \in \mathbb{R}_k^{m\times n}} \|(UV^T - B)P(X)\|_2 \\
&\leq \min_{B \in \mathbb{R}_k^{m\times n}} \|P(X)\|_2 \|UV^T - B\|_2 \\
&\leq \min_{B \in \mathbb{R}_k^{m\times n}} \|UV^T - B\|_2.
\end{aligned}$$

Somit erhält man für $k + r < k'$ die gewünschte Approximation $\tilde{A} := \tilde{U}\tilde{V}^T$ mittels einer Rang-r-Korrektur

$$\tilde{U} := [AX,\, U^*] \in \mathbb{R}^{m\times(k+r)}, \quad \tilde{V} := [Q(X),\, \tilde{V}^*] \in \mathbb{R}^{n\times(k+r)}, \tag{4.13}$$

wobei $Q(X) := X(X^TX)^{-1}$. Im Falle von $k + r \geq k'$ ist es nicht sinnvoll, zu approximieren und wir definieren

$$\tilde{U} := U, \quad \tilde{V} := V.$$

Das nachfolgende Lemma zeigt die gewünschten Eigenschaften.

Lemma 4.20. *Sei $\tilde{A}$ konstruiert wie in* (4.13)*. Dann gilt $A - \tilde{A} = UV'^T - U^*\tilde{V}^{*T}$ und somit*

$$\|A - \tilde{A}\|_2 \leq \min_{B \in \mathbb{R}_k^{m\times n}} \|UV^T - B\|_2.$$

Weiterhin erfüllt $\tilde{A}$ die Bedingung (4.2)*.*

Beweis. Die erste Behauptung ergibt sich durch

$$\begin{aligned}
A - \tilde{A} &= A - [AX,\, U^*][Q(X), \tilde{V}^*]^T \\
&= A - AXQ(X)^T - U^*\tilde{V}^{*T} \\
&= UV'^T - U^*\tilde{V}^{*T}.
\end{aligned}$$

Mit Hilfe von Bedingung (4.10) werden die Vektoren erhalten, da

$$\tilde{A}X = \tilde{U}[Q(X),\, \tilde{V}^*]^T X = [AX,\, U^*][\mathrm{Id},\, 0]^T = AX$$

gilt. Hieraus folgt die Bedingung (4.2). □

Sämtliche Betrachtungen aus diesem Abschnitt können anstelle von (4.12) für $A^TY = \tilde{A}^TY$ gemacht werden, da auch diese hinreichend für die schwachen Nebenbedingungen sind.

Ein wesentlicher Vorteil der Approximationen (4.11) und (4.13) ist die leichte Implementierbarkeit. Sie sind jedoch keine Bestapproximationen bzgl. der Nebenbedingungen (4.1) und (4.2), da immer noch Redundanzen zwischen dem orthogonal projizierten Anteil und den addierten Niedrigrangkorrekturen bestehen können. Ein weiterer Nachteil ist, dass dieses Verfahren die gleichen Stabilitätseigenschaften wie das Gram-Schmidt-Verfahren besitzt. Aus diesem Grund betrachten wir im nächsten Abschnitt Methoden, die auf dem Householder-Verfahren basieren.

4.3.2 Nebenbedingungen mittels Householder-Verfahren

In diesem Abschnitt präsentieren wir Vorgehensweisen zum Erhalt der starken und schwachen Nebenbedingungen (4.1) und(4.2), die bessere Eigenschaften als die Approximationen (4.11) und (4.13) in Bezug auf die Bestapproximation und Stabilität liefern, jedoch den Aufwand für die Implementierung erhöhen.

Für diesen Ansatz vergleichen wir die Zerlegung (2.4) einer Rang-k'-Matrix $A = UV^T \in \mathbb{R}^{m\times n}$ mit der Approximation (2.5) und erhalten

$$A = \tilde{U}\tilde{V}^T + EF^T. \tag{4.14}$$

Hierbei ist der Restterm EF^T eine Rang-$\hat{k}$-Matrix mit $\hat{k} := k' - k$. Dieser Term ist dafür verantwortlich, dass die Approximation $\tilde{U}\tilde{V}^T$ die Bedingungen (4.1) und (4.2) nicht erfüllt. Aus diesem Grund wollen wir nur den Anteil von EF^T verwerfen, welcher orthogonal zu den Nebenbedingungen ist.

Seien $Y = H_1B_1$ und $X = H_2B_2$ jeweils die QR-Zerlegungen von Y und X. Dabei sind H_1, H_2 Produkte von Householder-Matrizen und $B_1 \in \mathbb{R}^{m\times r}$, $B_2 \in \mathbb{R}^{n\times r}$ obere Dreiecksmatrizen. Diese Householder-Matrizen können verwendet werden, um den Restterm aus (4.14) zu transformieren:

$$H_1^T E =: \begin{bmatrix} C \\ G_1 \end{bmatrix}, \quad C \in \mathbb{R}^{r\times\hat{k}}, \qquad H_2^T F =: \begin{bmatrix} D \\ G_2 \end{bmatrix}, \quad D \in \mathbb{R}^{r\times\hat{k}}.$$

Somit ergibt sich durch die Orthogonalität von H_1 und H_2, dass

$$EF^T = H_1 \begin{bmatrix} C \\ G_1 \end{bmatrix} \begin{bmatrix} D \\ G_2 \end{bmatrix}^T H_2^T = H_1 \begin{bmatrix} CD^T & CG_2^T \\ G_1D^T & G_1G_2^T \end{bmatrix} H_2^T. \tag{4.15}$$

Das weitere Vorgehen wird im folgenden auf die starken bzw. schwachen Nebenbedingungen abgestimmt.

Starke Nebenbedingungen: Die Householder-Matrizen filtern den Anteil von EF^T senkrecht zu X bzw. Y heraus. Somit können wir den Term $G_1G_2^T$ in (4.15) Null setzen. Wie sich später zeigt, verletzt dies nicht die Bedingung (4.1).

Um spätere Rechenkosten zu senken, reduzieren wir die Größe der inneren Matrix mittels QR-Zerlegung $G_1D^T = Q_1R_1$ und $G_2C^T = Q_2R_2$ mit den Householder-Matrizen $Q_1 \in \mathbb{R}^{(m-r)\times(m-r)}$, $Q_2 \in \mathbb{R}^{(n-r)\times(n-r)}$ und den oberen Dreiecksmatrizen $R_1 := [\hat{R}_1^T, 0]^T \in \mathbb{R}^{(m-r)\times r}$ und $R_2 :=$

$[\hat{R}_2^T, 0]^T \in \mathbb{R}^{(n-r)\times r}$. Somit ergibt sich durch Umformung

$$H_1 \begin{bmatrix} CD^T & CG_2^T \\ G_1D^T & 0 \end{bmatrix} H_2^T = \underbrace{H_1 \begin{bmatrix} \mathrm{Id} & \\ & Q_1 \end{bmatrix}}_{=:\hat{H}_1} \begin{bmatrix} CD^T & \hat{R}_2^T & 0 \\ \hat{R}_1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \underbrace{\begin{bmatrix} \mathrm{Id} & \\ & Q_2 \end{bmatrix}^T H_2^T}_{=:\hat{H}_2^T}.$$

Nun ist es noch möglich, dass Redundanzen in der inneren Matrix existieren. Aus diesem Grund führen wir eine Singulärwertzerlegung auf der inneren $2r \times 2r$-Matrix durch. Mit Hilfe der SVD ergibt sich

$$\begin{bmatrix} CD^T & \hat{R}_2^T \\ \hat{R}_1 & 0 \end{bmatrix} =: \bar{W}S\bar{Z}^T$$

und wir definieren die Approximation

$$\tilde{A} := \tilde{U}\tilde{V}^T + \hat{H}_1 \begin{bmatrix} \bar{W}S\bar{Z}^T & 0 \\ 0 & 0 \end{bmatrix} \hat{H}_2^T = [\tilde{U}, \hat{W}] \begin{bmatrix} \mathrm{Id} & \\ & S \end{bmatrix} [\tilde{V}, \hat{Z}]^T, \tag{4.16}$$

wobei $\hat{W} := \hat{H}_1[\bar{W}^T, 0]^T$ und $\hat{Z} := \hat{H}_2[\bar{Z}^T, 0]^T$.

Das nächste Lemma beweist die gewünschten Eigenschaften für die Approximation (4.16).

Lemma 4.21. *Die Approximation* (4.16) *erfüllt die starken Nebenbedingungen* (4.1) *und es gilt die Schranke* $\operatorname{rank} \tilde{A} \leq k + 2r$. *Weiterhin ist* $A - \tilde{A} = \hat{E}\hat{F}^T$, *mit*

$$\hat{E} = H_1 \begin{bmatrix} 0 \\ G_1 \end{bmatrix}, \quad \hat{F} = H_2 \begin{bmatrix} 0 \\ G_2 \end{bmatrix},$$

wobei $\|\hat{E}\hat{F}^T\|_2 \leq \|EF^T\|_2$ *gilt.*

Beweis. Per Konstruktion erhält man

$$A = \tilde{A} + \hat{E}\hat{F}^T.$$

Somit genügt es zu zeigen, dass $\hat{E}^T Y = 0$ und $\hat{F}^T X = 0$. Dies folgt aus

$$\hat{E}^T Y = \begin{bmatrix} 0 \\ G_1 \end{bmatrix}^T H_1^T Y = \begin{bmatrix} 0 \\ G_1 \end{bmatrix}^T H_1^T H_1 B_1 = \begin{bmatrix} 0 \\ G_1 \end{bmatrix}^T B_1 = 0,$$

da in den Matrizen B_1 und B_2 höchstens die ersten r Zeilen ungleich Null sind. Analoge Betrachtungen gelten für $\hat{F}^T X = 0$. □

Schwache Nebenbedingungen: Im Gegensatz zu der vorherigen Approximation können wir alle inneren Terme von (4.15) bis auf CD^T gleich Null setzen. In dem Lemma 4.22 wird gezeigt, dass dies nicht die Bedingung (4.2) verletzt. Somit betrachten wir den Term

$$H_1 \begin{bmatrix} CD^T & 0 \\ 0 & 0 \end{bmatrix} H_2^T.$$

Um Redundanzen zu eliminieren, führen wir eine Singulärwertzerlegung auf der inneren $r \times r$-Matrix durch, $CD^T =: \bar{W}S\bar{Z}^T$. Es ergibt sich dadurch die folgende Approximation

$$\tilde{A} := \tilde{U}\tilde{V}^T + H_1 \begin{bmatrix} \bar{W}S\bar{Z}^T & 0 \\ 0 & 0 \end{bmatrix} H_2^T = [\tilde{U},\, \hat{W}] \begin{bmatrix} \mathrm{Id} & \\ & S \end{bmatrix} [\tilde{V},\, \hat{Z}]^T, \tag{4.17}$$

wobei $\hat{W} := H_1[\bar{W}^T,\, 0]^T$ und $\hat{Z} := H_2[\bar{Z}^T,\, 0]^T$.

Das nächste Lemma beweist die gewünschten Eigenschaften für die Approximation (4.17).

Lemma 4.22. *Die Approximation* (4.17) *erfüllt die schwachen Nebenbedingungen* (4.2) *und es gilt* $\operatorname{rank} \tilde{A} \leq k + r$. *Weiterhin erhält man*

$$A - \tilde{A} = H_1 \begin{bmatrix} 0 & CG_2^T \\ G_1D^T & G_1G_2^T \end{bmatrix} H_2^T,$$

wobei $\|A - \tilde{A}\|_2 \leq \|EF^T\|_2$ *gilt.*

Beweis. Per Konstruktion gilt

$$A = \tilde{A} + H_1 \begin{bmatrix} 0 & CG_2^T \\ G_1D^T & G_1G_2^T \end{bmatrix} H_2^T.$$

Somit genügt es, die schwachen Nebenbedingungen zu zeigen. Diese folgen aus

$$\begin{aligned} Y^T H_1 \begin{bmatrix} 0 & CG_2^T \\ G_1D^T & G_1G_2^T \end{bmatrix} H_2^T X &= B_1^T H_1^T H_1 \begin{bmatrix} 0 & CG_2^T \\ G_1D^T & G_1G_2^T \end{bmatrix} H_2^T H_2 B_2 \\ &= B_1^T \begin{bmatrix} 0 & CG_2^T \\ G_1D^T & G_1G_2^T \end{bmatrix} B_2 = 0, \end{aligned}$$

da bei den Matrizen B_1 und B_2 höchstens die ersten r Zeilen ungleich Null sind. □

Eine wesentliche Eigenschaft der Approximationen (4.16) und (4.17) ist, dass sie Bestapproximationen bzgl. der Spektralnorm unter Berücksichtigung der jeweiligen Nebenbedingungen sind. Dies wird ersichtlich, da die erste Approximation (4.14) mittels SVD berechnet wurde und der restliche Teil mittels einer weiteren SVD zu $\bar{W}S\bar{Z}^T$ gekürzt wird.

Weiterhin anzumerken ist, dass die Verfahren basierend auf der Householder-Zerlegung bessere Stabilitätseigenschaften als die Approximationen mittels der Gram-Schmidt-Methode besitzen.

4.4 Numerische Ergebnisse

Die Resultate aus Satz 4.8, Satz 4.12 und Satz 4.15 sollen zuerst an einem akademischen Beispiel belegt werden. Anschließend wählen wir anspruchsvollere Geometrien und springende Koeffizienten, um unsere vorgestellten Vorkonditionierer mit einer Implementierung des algebraischen Mehrgitterverfahrens zu vergleichen.

In unseren Tests wurden die linearen Gleichungssysteme bis auf eine relative Genauigkeit von 10^{-10} gelöst. Dabei kam das PCG-Verfahren mit der hierarchischen Cholesky-Zerlegung als Vorkonditionierer zum Einsatz.

Beispiel A: Sei unser Gebiet $\Omega = (0,1)^3$ der Einheitswürfel und das Dirichletproblem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{auf } \partial\Omega \end{aligned}$$

gegeben. Zur Diskretisierung wurden lineare Ansatzfunktionen verwendet und bei der $\mathcal{H}$-Matrix-Partitionierung kam die Bounding-Box-Methode zum Einsatz, vgl. [Gie01], wobei die minimale Blockgröße auf $n_{\min} = 20$ gesetzt wurde.

	`Prec0`		`PrecW`		`PrecS`	
N	$\kappa(\tilde{A}^{-1}A)$	$\star$	$\kappa(\tilde{A}^{-1}A)$	$\star$	$\kappa(\tilde{A}^{-1}A)$	$\star$
3 k	1,67	10	1,33	9	1,12	8
7 k	1,72	10	1,30	8	1,03	7
11 k	4,26	18	3,04	15	1,41	10
30 k	4,01	17	2,36	13	1,18	9
63 k	4,04	17	2,31	13	1,07	7
94 k	13,60	30	6,72	23	1,24	10
250 k	13,31	31	6,48	22	1,09	8

Tabelle 1: Numerische Resultate für Beispiel A und die jeweiligen Vorkonditionerer mit konstanter Approximationsgenauigkeit $\varepsilon = 0{,}1$, wobei $\star :=$ Anzahl der CG-Schritte.

Die blockweisen Approximationsgenauigkeiten der Vorkonditionierer `Prec0`, `PrecW` und `PrecS` beträgt für die Tests $\varepsilon = 0{,}1$. Zur Berechnung der Konditionszahl des Vorkonditionierten Systems wurde die Power-Iteration verwendet. Wie sich in der Tabelle 1 zeigt, steigt die Anzahl der CG-Schritte und die Konditionszahl des vorkonditionierten Systems für `PrecS` kaum, wohingegen `PrecW` einen leichten und `Prec0` den stärksten Anstieg mit zunehmender Anzahl der Unbekannten verzeichnet.

Um zu den Werten aus Tabelle 1 vergleichbare Resultate für die Vorkonditionierer `Prec0_S`, `PrecW_S` und `PrecS_S` zu erzielen, wurde die Approximationsgenauigkeit so skaliert, dass die Anzahl der CG-Schritte für das kleinste Beispiel mit 3000 Unbekannten bei acht liegt. Dies entspricht ungefähr der Anzahl der benötigten Schritte für `Prec0`, `PrecW` und `PrecS` bei gleicher Problemgröße. In Tabelle 2 sind die Ergebnisse der spektraläquivalenten Vorkonditionierer zu sehen, dabei zeigt `PrecW_S` mit zunehmender Anzahl der Unbekannten einen leichten Anstieg der CG-Schritte und der Konditionszahl. Die Vorkonditionierer `Prec0_S` und `PrecS_S` hingegen weisen ein konstantes Verhalten auf.

In Abbildung 8 sind die verwendeten Vorkonditionierer im Vergleich zueinander dargestellt. Insgesamt weist nur `PrecW_S` ein leicht abweichendes Verhalten zu den Resultaten aus Satz 4.8, Satz 4.12 und Satz 4.15 auf. Der moderate Anstieg der Konditionszahl und der CG-Schritte ab 63000 Unbekannten für den theoretisch spektraläquivalenten Vorkonditionierer `PrecW_S` kann nicht eindeutig interpretiert werden, da die Konditionszahlen sehr gering sind.

	Prec0_S		PrecW_S		PrecS_S	
N	$\kappa(\tilde{A}^{-1}A)$	$\star$	$\kappa(\tilde{A}^{-1}A)$	$\star$	$\kappa(\tilde{A}^{-1}A)$	$\star$
3 k	1,19	8	1,23	8	1,16	8
7 k	1,24	8	1,14	7	1,03	7
11 k	1,48	10	1,48	10	1,42	10
30 k	1,18	8	1,41	9	1,29	10
63 k	1,16	7	1,58	10	1,08	8
94 k	1,14	8	1,67	11	1,29	10
250 k	1,04	6	2,04	12	1,10	8

Tabelle 2: Numerische Resultate für Beispiel A und die jeweiligen Vorkonditionerer mit angepasster Approximationsgenauigkeit, wobei $\star :=$ Anzahl der CG-Schritte.

Beispiel B: Das Berechnungsgebiet Ω sind zwei parallele Zylinder, siehe Abbildung 9. Die Oberfläche Γ_1 bezeichnet die oberen kreisförmigen Flächen und Γ_2 die unteren. Die restliche Mantelfläche wird mit $\Gamma_3 := \partial\Omega\backslash(\Gamma_1 \cup \Gamma_2)$ bezeichnet. Der Koeffizient σ ist 10^{-7} und 10^{-4} in dem linken bzw. rechten Leiter. In dem Zwischenraum beträgt dieser Null.

Wir betrachten das folgende Randwertproblem

$$\begin{aligned} -\text{div}(\sigma\nabla u) &= 0 \quad \text{in } \Omega, \\ u &= 0 \quad \text{auf } \Gamma_1, \\ \frac{\partial u}{\partial \nu} &= I \quad \text{auf } \Gamma_2, \\ \frac{\partial u}{\partial \nu} &= 0 \quad \text{auf } \Gamma_3. \end{aligned} \tag{4.18}$$

Zur Diskretisierung wurden quadratische Ansatzfunktionen verwendet. Für die $\mathcal{H}$-Matrix-Partitionierung wurden Techniken basierend auf Nested-Dissection eingesetzt, siehe [BF11]. Die minimale Blockgröße beträgt $n_{\min} = 150$.

Auffällig bei dem Randwertproblem (4.18) ist, dass der Dirichlet-Rand nur durch die oberen kreisförmigen Flächen beschrieben wird und somit relativ zur Gesamtoberfläche klein ist. Dies führt zu einer hohen Konditionszahl der Systemmatrix, so dass ein robuster Vorkonditionierer von Nöten ist.

Um die verschiedenen Vorkonditionierer miteinander vergleichen zu können, werden diese in zwei Gruppen eingeteilt. Zum einen vergleichen wir in Tabelle 3 `Prec0`, `PrecW` und `PrecS` und in Tabelle 4 `Prec0_S`, `PrecW_S` und `PrecS_S`. Jeweils sind die Vorkonditionierer so gewählt, dass sie für die gleiche Anzahl an Unbekannten die gleiche Menge an Speicher verbrauchen (relative Differenz beträgt weniger als ein Prozent). Dies ist sinnvoll, da numerische Verfahren oftmals beschleunigt werden können, indem man mehr Speicher verwendet.

In Tabelle 3 wird die blockweise Genauigkeit für `PrecS` auf $\varepsilon = 5{,}6e-3$ gesetzt. Die Approximationsgenauigkeit von `Prec0` und `PrecW` wird entsprechend angepasst, so dass sie bei einer gleichen Anzahl der Unbekannten den selben Speicher wie `PrecS` verbrauchen. Wie sich in den numerischen Ergebnissen zeigt, führt der Erhalt von schwachen Nebenbedingungen auf eine marginale Verringerung der Anzahl

		Prec0			PrecW			PrecS		
n	Speicher	Chol.	CG	Summe	Chol.	CG	Summe	Chol.	CG	Summe
$2{,}0 \cdot 10^5$	0,7 GB	25 s	21 s (39)	46 s	27 s	28 s (51)	55 s	31 s	14 s (26)	45 s
$7{,}6 \cdot 10^5$	2,9 GB	133 s	124 s (53)	257 s	149 s	111 s (48)	260 s	173 s	68 s (29)	241 s
$1{,}3 \cdot 10^6$	5,2 GB	271 s	265 s (62)	536 s	302 s	245 s (58)	547 s	350 s	136 s (32)	486 s
$4{,}1 \cdot 10^6$	17,2 GB	942 s	1 228 s (87)	2 170 s	1 066 s	1 171 s (84)	2 237 s	1 252 s	480 s (34)	1 732 s

Tabelle 3: Vergleich der Vorkonditionierer `Prec0`, `PrecW` und `PrecS` bei gleichem Speicherbedarf.

		Prec0_S			PrecW_S			PrecS_S		
n	Speicher	Chol.	CG	Summe	Chol.	CG	Summe	Chol.	CG	Summe
$2{,}0 \cdot 10^5$	0,7 GB	23 s	25 s (48)	48 s	23 s	25 s (48)	48 s	27 s	14 s (27)	41 s
$7{,}6 \cdot 10^5$	2,9 GB	143 s	111 s (48)	254 s	144 s	109 s (47)	253 s	163 s	69 s (30)	232 s
$1{,}3 \cdot 10^6$	5,2 GB	310 s	218 s (51)	528 s	302 s	222 s (52)	524 s	360 s	137 s (32)	497 s
$4{,}1 \cdot 10^6$	18,0 GB	1 233 s	1 057 s (72)	2 290 s	1 216 s	1 122 s (77)	2 338 s	1 484 s	429 s (29)	1 913 s

Tabelle 4: Vergleich der Vorkonditionierer `Prec0_S`, `PrecW_S` und `PrecS_S` bei gleichem Speicherbedarf.

	PrecS				BoomerAMG			
n	Init.	Chol.	CG	Summe	Init.	CG	Summe	Einsparung
$2{,}0 \cdot 10^5$	12 s	31 s	14 s (26)	57 s	8 s	62 s (4)	70 s	13 s (23 %)
$7{,}6 \cdot 10^5$	51 s	173 s	68 s (29)	292 s	40 s	208 s (3)	248 s	−44 s (- 7 %)
$1{,}3 \cdot 10^6$	101 s	350 s	135 s (32)	587 s	81 s	396 s (3)	477 s	−110 s (-19 %)
$4{,}1 \cdot 10^6$	341 s	1 252 s	480 s (34)	2 073 s	321 s	1 537 s (3)	1 858 s	−215 s (-10 %)

Tabelle 5: Vergleich des $\mathcal{H}$-Matrix-Vorkonditionierers `PrecS` mit einem Löser basierend auf der algebraischen Mehrgittermethode.

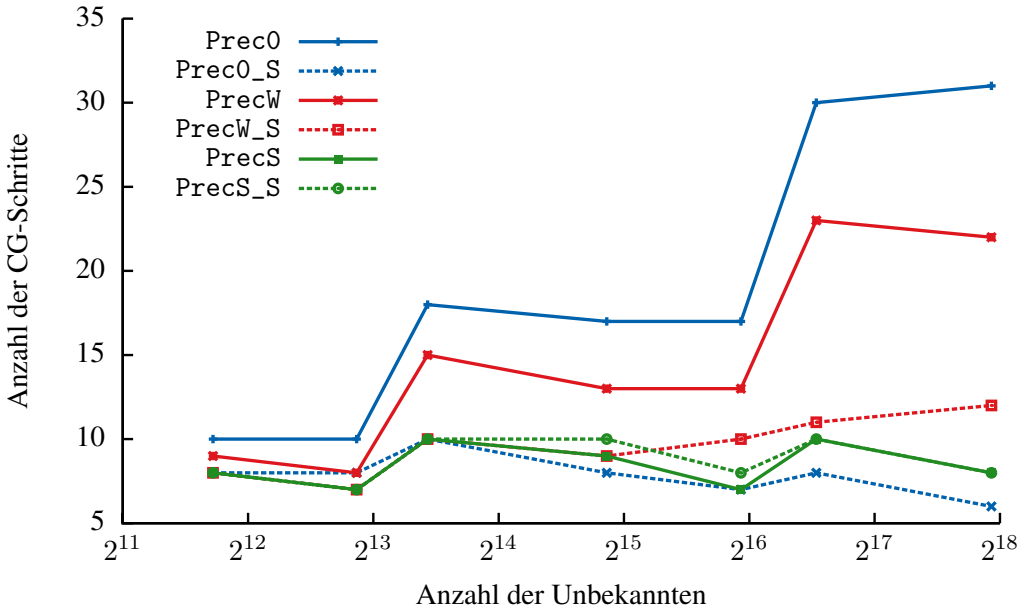


Abbildung 8: Vergleich der CG-Schritte für die verschiedenen $\mathcal{H}$-Cholesky-Vorkonditionierer in Abhängigkeit der Anzahl der Unbekannten für Beispiel A.

der CG-Schritte, wohingegen der Erhalt der starken Nebenbedingungen zu einer fast konstanten Anzahl der CG-Schritte führt. Da einzelne CG-Schritte schnell durchgeführt werden können, rechnet sich der höhere Rechenaufwand für einen besseren Vorkonditionierer nur bei `PrecS`.

Die spektraläquivalenten Vorkonditionierer werden in Tabelle 4 verglichen. Hierbei besitzt `Prec0_S` eine konstante Skalierung ε und die Genauigkeiten der anderen beiden Vorkonditionierer `PrecW_S` und `PrecS_S` wurden so angepasst, dass sie den gleichen Speicherbedarf besitzen. Es zeigt sich bei `PrecS_S` ein konstantes Verhalten der CG-Schritte. Hingegen ist bei `Prec0_S` und `PrecW_S` ein Anstieg der Schritte für die Problemgröße mit $4{,}1 \cdot 10^6$ Unbekannten zu verzeichnen. Möglicherweise kommt es bei solch einer großen Anzahl der Freiheitsgrade in der $\mathcal{H}$-Matrix-Arithmetik für `Prec0_S` und `PrecW_S` zu Instabilitäten, so dass Bedingung (3.9) nicht mehr erfüllt ist. Andererseits ist es auch möglich, dass

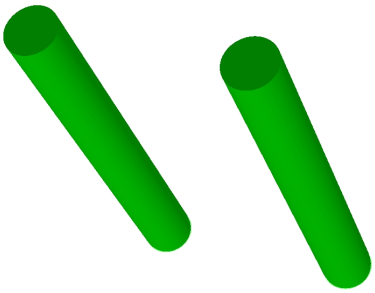


Abbildung 9: Die Geometrie (parallele Zylinder) verwendet in Beispiel B.

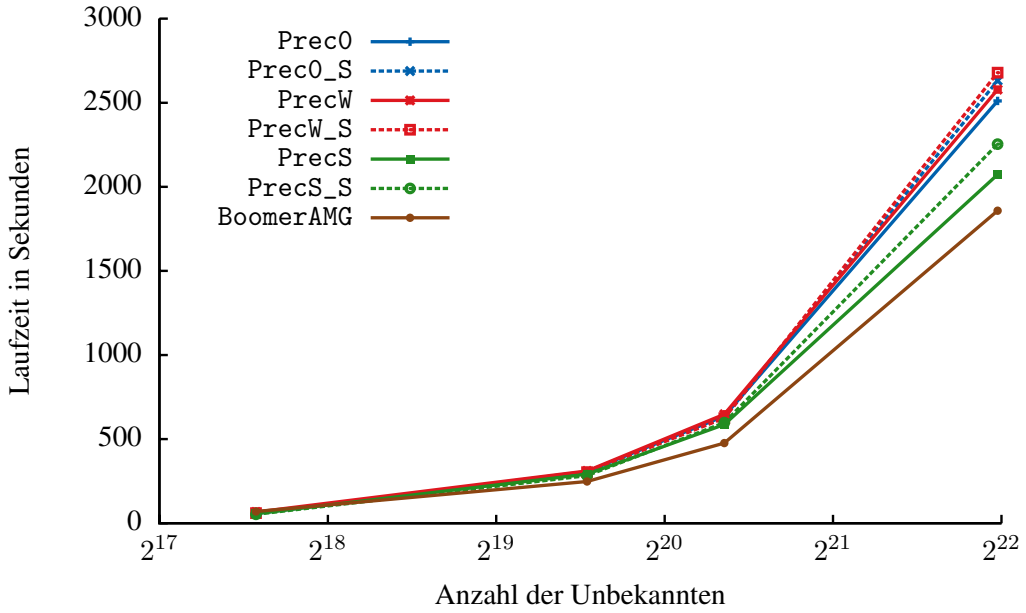


Abbildung 10: Übersicht der verschiedenen $\mathcal{H}$-Matrix-Vorkonditionierer und einem Löser basierend auf der algebraischen Mehrgittermethode (`BoomerAMG`).

schwer zu analysierende Effekte aus der Diskretisierung entstehen, da einzelne Dreiecke entartet sind.

In den betrachteten Tests in Tabelle 3 und Tabelle 4 zeigt `PrecS` die besten Laufzeiteigenschaften. Deshalb vergleichen wir diesen Vorkonditionierer mit einer Implementierung des algebraischen Mehrgitterverfahrens (`BoomerAMG` aus der Bibliothek HYPRE 2.0.0[4]). In Tabelle 5 zeigt sich, dass die Laufzeitunterschiede für dieses Beispiel bei $\pm 20\,\%$ liegen also im Bereich von Code-Optimierungen. Ein abschließender Vergleich der Laufzeiten aller betrachteten Vorkonditionierer ist in Abbildung 10 zu finden. So zeigt sich, dass nur der Erhalt von starken Nebenbedingungen zu besseren Laufzeiten im Vergleich zu dem gewöhnlichen Vorkonditionierer `Prec0` führt.

Beispiel C: In diesen numerischen Tests untersuchen wir die Effizienz der verschiedenen $\mathcal{H}$-Matrix-Vorkonditionierer für die Problemstellung (4.18), jedoch mit einer anspruchsvolleren Geometrie als in Beispiel B und stark springenden Zufallskoeffizienten. Wir betrachten als Berechnungsgebiet Ω eine Spule mit 12 Windungen, siehe Abbildung 11. Hierbei sind nur an einem der beiden Enden, bezeichnet jeweils mit Γ_1 und Γ_2, Dirichlet-Bedingungen gegeben. Die restliche Mantelfläche Γ_3 ist definiert durch $\Gamma_3 := \partial\Omega \setminus (\Gamma_1 \cup \Gamma_2)$. Weiterhin ist der Koeffizient σ zufällig in Ω aus dem Intervall $(0, 10^6)$ gewählt.

Wie in Beispiel B, wurden zur Diskretisierung quadratische Ansatzfunktionen verwendet und die $\mathcal{H}$-Matrix-Partitionierung mittels Nested-Dissection erstellt. Die minimale Blockgröße beträgt in allen Tests $n_{\min} = 150$.

Zuerst vergleichen wir die Vorkonditionierer `Prec0`, `Prec0_S`, `PrecW`, `PrecW_S`, `PrecS`, `PrecS_S`

[4] `http://acts.nersc.gov/hypre`

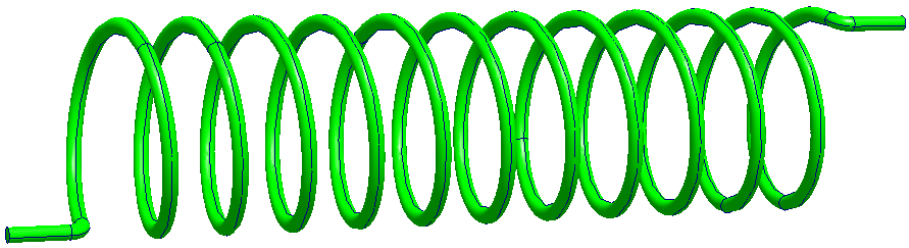


Abbildung 11: Die Geometrie (Spule) verwendet in Beispiel C.

und `BoomerAMG` in Bezug auf die benötigte Laufzeit, um das aus der FE-Methode resultierende Gleichungssystem mittels dem PCG-Verfahren zu lösen. Hierbei wurde für jede Problemgröße der optimale Approximationsparameter ε bestimmt. Wie sich in Abbildung 12 zeigt, schneiden `PrecS` und `PrecS_S` bei großen Beispielen im Vergleich zu `Prec0`, `Prec0_S`, `PrecW` und `PrecW_S` am besten ab. `BoomerAMG` ist verglichen mit den $\mathcal{H}$-Matrix-Vorkonditionierern mit Ausnahme einer Testinstanz langsamer. Im Vergleich zu `PrecS` und `PrecS_S` ist die Laufzeit ca. doppelt so hoch. Leider konnte für die größte Testinstanz kein Wert für `BoomerAMG` ermittelt werden, da dieser Vorkonditionierer aus unbekanntem Grund mit einem Speicherfehler abbricht.

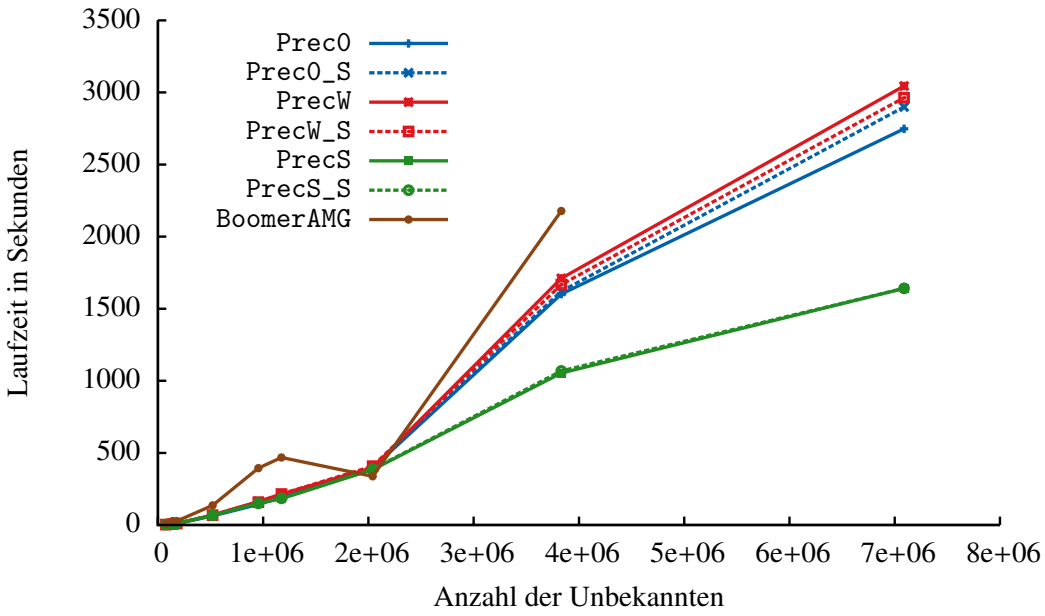


Abbildung 12: Die Laufzeit der verschiedenen Vorkonditionierer ist dargestellt in Abhängigkeit der Anzahl der Unbekannten für einen optimal gewählten Approximationsparameter ε.

Ein weiterer Testindikator für die Leistungsfähigkeit eines Vorkonditionierers ist das Produkt aus Speicherbedarf mit der benötigter Rechenzeit. Dies erscheint sinnvoll, da im Allgemeinen bekannt ist, dass mit mehr Speicher Rechenzeit einsparen werden kann. In Abbildung 13 werden die verschiedenen $\mathcal{H}$-Matrix-Vorkonditionierer miteinander verglichen. Leider war es nicht möglich den Speicherbedarf

von BoomerAMG zuverlässig zu bestimmen, so dass dieser Vorkonditionierer nicht in den Vergleich mit aufgenommen werden konnte. Wie oben zeigt sich in Abbildung 13, dass für große Instanzen PrecS und PrecS_S um einen Faktor zwei schneller sind.

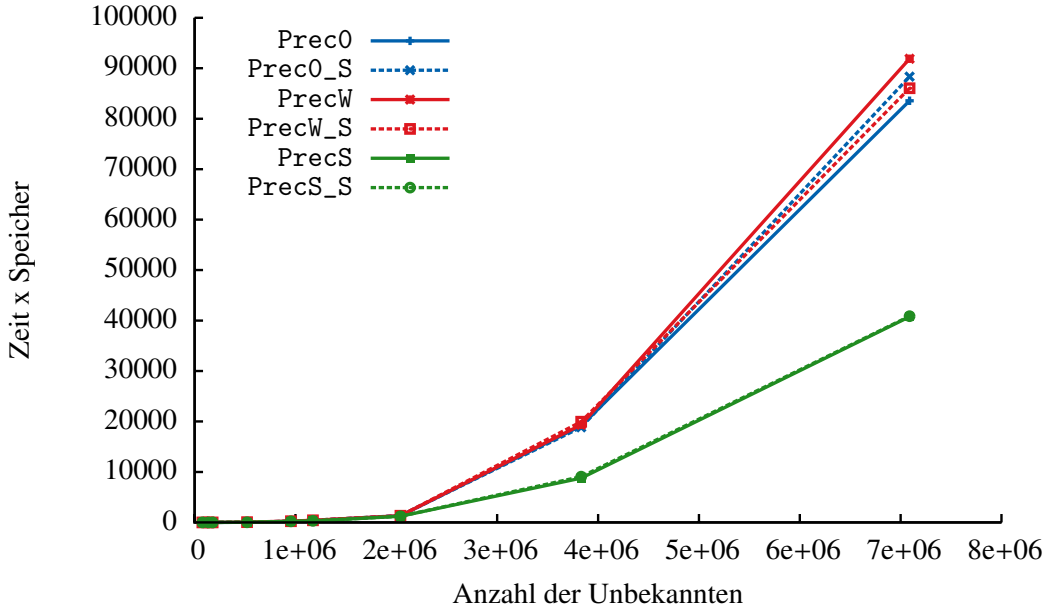


Abbildung 13: Das Produkt aus Laufzeit und Speicherbedarf ist für die verschiedenen Vorkonditionierer dargestellt, wobei der Approximationsparameter ε optimal gewählt wurde.

Um die sich stark unterscheidende Problemgrößen aus Abbildung 13 gut miteinander vergleichen zu können, wurden in Abbildung 14 die Werte für das Produkt von Speicherbedarf und Rechenzeit so skaliert, dass der beste Vorkonditionierer den Wert eins besitzt und die anderen umso schlechter sind, je näher sie an null sind. Wie in Abbildung 14 zu erkennen ist, schneiden die gewöhnlichen $\mathcal{H}$-Matrizen Prec0 bis zu einer Problemgröße von 10^6 sehr gut im Vergleich zu den anderen ab. Bei größer werdenden Beispielen dominieren jedoch die Vorkonditionierer, welche die starken Nebenbedingungen erhalten.

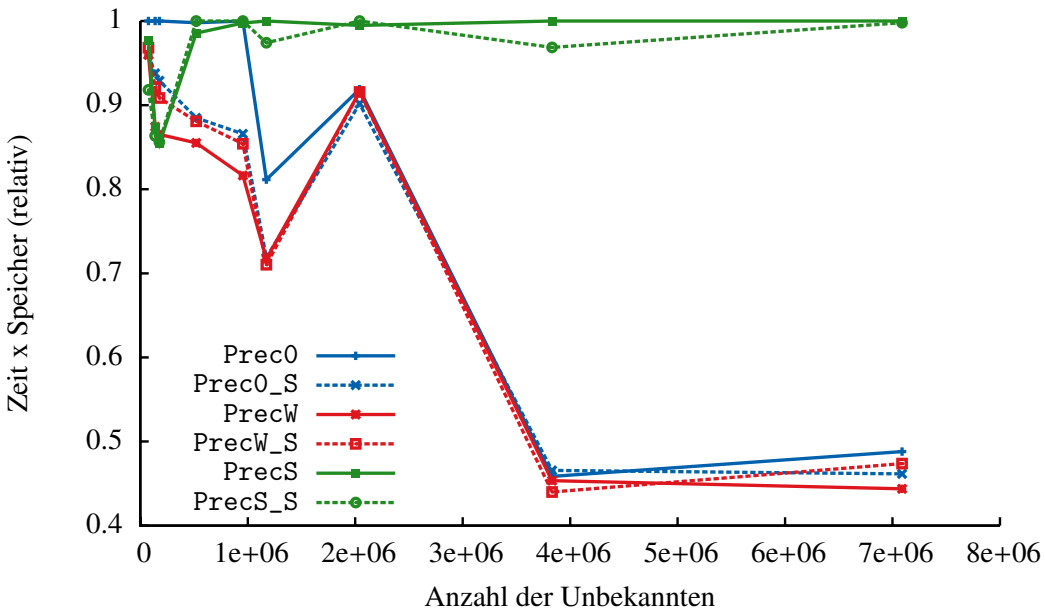


Abbildung 14: Das Produkt aus Laufzeit und Speicherbedarf ist relativ zum besten Vorkonditionierer dargestellt. Je besser der Vorkonditionierer ist, desto näher liegt dessen Wert bei Eins.

5 Berechnung minimaler Energiepfade der mikromagnetischen Energie[5]

Ferromagnetische Materialien besitzen in der Regel zwei stabile Zustände. Es ist möglich, mit Hilfe eines magnetischen Feldes zwischen diesen beiden zu wechseln [ACDP04, HS09]. Somit eignen sich solche Materialien für die permanente Speicherung von Informationen.

Aufgrund der zunehmenden Speicherdichte in magnetischen Datenträgern wird es immer wichtiger, ungewollte Schaltvorgänge zu vermeiden. Die Kenntnis von minimale Energiepfade kann dabei hilfreich sein, diese zu analysieren, siehe [ERVE03].

5.1 Das Landau-Lifschitz-Modell

Eine zentrale Bedeutung bei der Simulation von stationären mikromagnetischen Phänomenen hat die Landau-Lifschitz-Energie [LL35, HS98], im Folgenden meist als mikromagnetische Energie bezeichnet. Diese wird eindeutig durch die Magnetisierung bestimmt, welche punktweise innerhalb eines Gebietes Ω auf die Länge 1 eingeschränkt wird.

5.1.1 Betrachtung des Energiefunktionals

Sei $\Omega \subset \mathbb{R}^3$ eine beschränktes Lipschitz-Gebiet. Dann ist der Raum der möglichen Konfigurationen der Magnetisierung $\mathcal{A}$ gegeben durch

$$\mathcal{A} := \{\boldsymbol{m} \in H^1(\Omega; \mathbb{R}^3) : \|\boldsymbol{m}\|_2 = 1 \text{ fast überall in } \Omega\}. \tag{5.1}$$

Außerhalb des Gebietes sei $\boldsymbol{m}(x) = 0$ für $x \notin \Omega$.

Für ein $\boldsymbol{m} \in \mathcal{A}$ ist die mikromagnetische Energie $E : \mathcal{A} \to \mathbb{R}$ gegeben durch die Summe der folgenden Teil-Energien

$$E(\boldsymbol{m}) := E_E(\boldsymbol{m}) + E_A(\boldsymbol{m}) + E_Z(\boldsymbol{m}) + E_S(\boldsymbol{m}). \tag{5.2}$$

Der die Energieminimierung später bestimmende Term ist die Austauschenergie

$$E_E(\boldsymbol{m}) := \frac{1}{2} \int_\Omega \|D\boldsymbol{m}\|_F^2 \,\mathrm{d}x.$$

Dieser sorgt für eine parallele Ausrichtung der Magnetisierungsvektoren.

Hinzu kommen die anisotrope Energie und die Zeeman-Energie

$$E_A(\boldsymbol{m}) := \int_\Omega \phi(\boldsymbol{m}) \,\mathrm{d}x \quad \text{und} \quad E_Z(\boldsymbol{m}) := -\int_\Omega \boldsymbol{f} \cdot \boldsymbol{m} \,\mathrm{d}x,$$

mit $\boldsymbol{f} \in \mathbb{R}^3$. Die anisotrope Energie E_A berücksichtigt mit $\phi(\boldsymbol{m}) := 1 - (\boldsymbol{e} \cdot \boldsymbol{m})^2$ die Ausrichtung

[5] Teile dieses Kapitels werden veröffentlicht in [BBBa].

der kristallinen Achse $\boldsymbol{e} \in \mathbb{R}^3$ wobei $\|\boldsymbol{e}\|_2 = 1$. Hingegen modelliert die Zeeman-Energie E_Z die Auswirkungen eines externen Magnetfeldes $\boldsymbol{f}$.

Der numerisch am aufwendigsten zu berechnende Term in (5.2) ist die Streufeldenergie

$$E_S(\boldsymbol{m}) := \frac{\mu_0}{2} \int_{\mathbb{R}^3} \|H\|_2^2 \,\mathrm{d}x, \quad \mu_0 := 4\pi \cdot 10^{-7}.$$

Dabei bezeichnet H die magnetische Feldstärke (oder das Streufeld) bzgl. der Magnetisierung $\boldsymbol{m}$.

Das Streufeld H und die magnetische Induktion B sind gekoppelt über die Gleichung $B = \mu_0(H+\boldsymbol{m})$. Weiterhin vereinfachen sich die Maxwell-Gleichungen, siehe [Jac75], unter Abwesenheit von elektrischer Ladung und Spannung zu

$$\operatorname{div} B = 0, \tag{5.3a}$$
$$\operatorname{rot} H = 0. \tag{5.3b}$$

Es folgt aus (5.3b) die Existenz eines Potentials (magnetostatisches Potential) $u : \mathbb{R}^3 \to \mathbb{R}$ mit $H = -\nabla u$. Somit erhält man aus (5.3a)

$$\Delta u = \operatorname{div} \boldsymbol{m}$$

und die entsprechende schwache Formulierung

$$\int_{\mathbb{R}^3} \nabla u \cdot \nabla w = \int_\Omega \boldsymbol{m} \cdot \nabla w \quad \text{für alle } w \in H^1(\mathbb{R}^3). \tag{5.4}$$

Für die Streufeldenergie folgt mit der speziellen Wahl der Testfunktion $w = u$, dass

$$E_S(\boldsymbol{m}) = \frac{\mu_0}{2} \int_\Omega \boldsymbol{m} \cdot \nabla u \,\mathrm{d}x.$$

Weiterhin ergibt sich aus (5.4) mit Hilfe der Cauchy-Schwarzschen Ungleichung

$$\|\nabla u\|_{L^2(\mathbb{R}^3)} \leq \|\boldsymbol{m}\|_{L^2(\Omega)}. \tag{5.5}$$

Zur späteren Energieminimierung von $E(\boldsymbol{m})$ benötigen wir deren Ableitung. Falls $\boldsymbol{v}(x) \cdot \boldsymbol{m}(x) = 0$, für fast alle $x \in \Omega$, dann gilt

$$E'(\boldsymbol{m})(\boldsymbol{v}) = \int_\Omega \operatorname{spur} (D\boldsymbol{m})^T (D\boldsymbol{v}) \,\mathrm{d}x - 2 \int_\Omega (\boldsymbol{e} \cdot \boldsymbol{m})(\boldsymbol{e} \cdot \boldsymbol{v}) \,\mathrm{d}x - \int_\Omega f \cdot \boldsymbol{v} \,\mathrm{d}x + E_S'[\boldsymbol{m}](\boldsymbol{v}),$$

wobei

$$\begin{aligned} E_S'(\boldsymbol{m})(\boldsymbol{v}) &= \partial_{\boldsymbol{m}} \left(\frac{\mu_0}{2} \int_{\mathbb{R}^3} H(\boldsymbol{m}) \cdot H(\boldsymbol{m}) \,\mathrm{d}x \right) = \mu_0 \int_{\mathbb{R}^3} H(\boldsymbol{m}) \cdot H(\boldsymbol{v}) \,\mathrm{d}x \\ &= -\mu_0 \int_{\mathbb{R}^3} \nabla u \cdot H(\boldsymbol{v}) \,\mathrm{d}x = \mu_0 \int_\Omega \boldsymbol{v} \cdot \nabla u \,\mathrm{d}x \end{aligned}$$

aus (5.4) folgt.

5.1.2 Alternative Formulierungen der Streufeldenergie

Zur Berechnung der mikromagnetischen Energie benötigen wir einen geeigneten Ausdruck für das magnetostatische Potential u. Wie sich im folgenden Lemma zeigt, ergibt sich aus dem klassischen Randwerteproblem

$$\begin{aligned} \Delta u &= \begin{cases} \operatorname{div} \boldsymbol{m}, & \text{in } \Omega, \\ 0, & \text{in } \Omega^c := \mathbb{R}^3 \backslash \overline{\Omega}, \end{cases} \\ [u] &= 0 \quad \text{auf } \partial\Omega, \\ [\partial_\nu u] &= -\boldsymbol{m} \cdot \nu \quad \text{auf } \partial\Omega. \end{aligned} \tag{5.6}$$

die schwache Formulierung (5.4). Hierbei sei der Sprung $[v]$ am Rand von Ω definiert durch

$$[v] = \lim_{\substack{y \to x \\ y \in \overline{\Omega}^c}} v(y) - \lim_{\substack{y \to x \\ y \in \Omega}} v(y).$$

Lemma 5.1. *Aus dem Randwerteproblem* (5.6) *folgt die schwache Formulierung* (5.4)*.*

Beweis. Wenden wir das L^2-Skalarprodukt mit einer Testfunktion $w \in H^1(\mathbb{R}^3)$ auf (5.6) an, so erhalten wir

$$\int_\Omega w \Delta u \, \mathrm{d}x + \underbrace{\int_{\overline{\Omega}^c} w \Delta u \, \mathrm{d}x}_{=0} = \int_\Omega w \operatorname{div} \boldsymbol{m} \, \mathrm{d}x + \underbrace{\int_{\overline{\Omega}^c} w \operatorname{div} \boldsymbol{m} \, \mathrm{d}x}_{=0}.$$

Durch Anwendung mehrdimensionaler partieller Integration auf beide Seiten der Gleichung ergibt sich

$$\int_\Omega \boldsymbol{m} \cdot \nabla w \, \mathrm{d}x - \int_{\partial\Omega} w \, \boldsymbol{m} \cdot \nu \, \mathrm{d}s_x = \int_\Omega \nabla u \cdot \nabla w \, \mathrm{d}x - \int_{\partial\Omega} w \nabla u \cdot \nu \, \mathrm{d}s_x \tag{5.7}$$

$$0 = \int_{\overline{\Omega}^c} \nabla u \cdot \nabla w \, \mathrm{d}x - \int_{\partial\overline{\Omega}^c} w \nabla u \cdot \nu \, \mathrm{d}s_x. \tag{5.8}$$

Durch Addition der Gleichungen (5.7) und (5.8) erhalten wir

$$\int_\Omega \boldsymbol{m} \cdot \nabla w \, \mathrm{d}x - \int_{\partial\Omega} w \, \boldsymbol{m} \cdot \nu \, \mathrm{d}s_x = \int_{\mathbb{R}^3} \nabla u \cdot \nabla w \, \mathrm{d}\boldsymbol{x} - \int_{\partial\Omega} w \nabla u \cdot \nu \, \mathrm{d}s_x - \int_{\partial\overline{\Omega}^c} w \nabla u \cdot \nu \, \mathrm{d}s_x.$$

Die zweite Sprungbedingung aus (5.6) liefert sofort

$$\int_\Omega \boldsymbol{m} \cdot \nabla w \, \mathrm{d}x = \int_{\mathbb{R}^3} \nabla u \cdot \nabla w \, \mathrm{d}x.$$

Die erste Bedingung aus (5.6) erlaubt eine Einschränkung des Potentials auf $u \in H^1(\mathbb{R}^3)$. □

Das obige Randwertproblem (5.6) besitzt die Lösung

$$u(x) = -\int_\Omega \nabla \mathfrak{S}(x,y) \cdot \boldsymbol{m}(y)\,\mathrm{d}y,$$

siehe [Jac75], wobei

$$\mathfrak{S}(x,y) = -\frac{1}{4\pi}\frac{1}{|x-y|}$$

die Singularitätenfunktion des Laplace-Operators ist. Dieser Ansatz zur Bestimmung des Potentials u wurde in Verbindung mit $\mathcal{H}$-Matrizen bereits in [PP05] untersucht.

Wir wollen einen anderen Ansatz verwenden, welcher in [GC07] vorgeschlagen wurde. Dabei wird u in eine Summe $u = u_1 + u_2$ aufgespalten. Diese Summanden sind Lösungen der folgenden Randwertprobleme

$$\begin{aligned} \Delta u_1 &= \mathrm{div}\boldsymbol{m} \quad \text{in } \Omega, \\ u_1 &= 0 \quad \text{auf } \partial\Omega \end{aligned} \tag{5.9}$$

und

$$\begin{aligned} \Delta u_2 &= 0 \quad \text{in } \Omega \cup \Omega^c, \\ [u_2] &= 0 \quad \text{auf } \partial\Omega, \\ [\partial_\nu u_2] &= g \quad \text{auf } \partial\Omega, \end{aligned} \tag{5.10}$$

mit $g := (\nabla u_1 - \boldsymbol{m}) \cdot \nu$. Die Lösung des homogenen Problems (5.10) ist

$$u_2(x) = \int_{\partial\Omega} \mathfrak{S}(x,y) g(y)\,\mathrm{d}s_y,$$

siehe [GC07]. Somit muss zuerst u_1 berechnet werden, um das Potential u_2 innerhalb des Gebietes bestimmen zu können.

Dieser Ansatz hat den Vorteil, dass das Potential u in einen lokalen und nicht-lokalen Anteil zerfällt. Dabei beinhaltet der nicht-lokale Teil nur die Koppelung von Oberfläche mit Volumen.

Für weitere Umformungen benötigen wir den Begriff des Cauchyschen Hauptwerts.

Definition 5.2 (Cauchyscher Hauptwert)**.** *Ist das Integral $\int_\Omega f(x)\,\mathrm{d}x$ in $y \in \Omega$ uneigentlich, so bezeichnet*

$$\mathrm{CH}\int_\Omega f(x)\,\mathrm{d}x := \lim_{\varepsilon\to 0}\int_{\Omega\setminus B_\varepsilon(y)} f(x)\,\mathrm{d}x$$

den Cauchyschen Hauptwert.

Wendet man den Gradienten auf das Einfachschichtpotential an, so lassen sich Integral und Gradient vertauschen

$$\nabla u_2(x) = \nabla \int_{\partial\Omega} \mathfrak{S}(x,y) g(y)\,\mathrm{d}s_y = \mathrm{CH}\int_{\partial\Omega} \nabla\mathfrak{S}(x,y) g(y)\,\mathrm{d}s_y,$$

für einen Beweis siehe [Sch08].

Somit ergibt sich für die Streufeldenergie durch den Ansatz $u = u_1 + u_2$, dass

$$\begin{aligned} E_S(\boldsymbol{m}) &= \frac{\mu_0}{2} \int_\Omega \boldsymbol{m} \cdot \nabla u \, \mathrm{d}x \\ &= \frac{\mu_0}{2} \int_\Omega \boldsymbol{m} \cdot \nabla u_1 \, \mathrm{d}x + \frac{\mu_0}{2} \mathrm{CH} \int_\Omega \int_{\partial\Omega} \boldsymbol{m}(x) \cdot \nabla \mathfrak{S}(x,y) g(y) \, \mathrm{d}s_y \, \mathrm{d}x. \end{aligned}$$

Der Cauchysche Hauptwert auf der rechten Seite wird im nächsten Lemma zu schwach singulären Integralen umgeformt.

Lemma 5.3. *Sei $\boldsymbol{m} \in H^1(\Omega;\mathbb{R}^3)$, so ergibt sich folgende Umformung*

$$\mathrm{CH} \int_\Omega \boldsymbol{m}(x) \cdot \nabla \mathfrak{S}(x,y) \, \mathrm{d}x = \int_{\partial\Omega} \mathfrak{S}(x,y) \boldsymbol{m}(x) \cdot \nu_x \, \mathrm{d}s_x - \int_\Omega \mathfrak{S}(x,y) \mathrm{div} \boldsymbol{m} \, \mathrm{d}x.$$

Beweis. Durch partielle Integration erhält man

$$\begin{aligned} \mathrm{CH} \int_\Omega \boldsymbol{m}(x) \cdot \nabla \mathfrak{S}(x,y) \, \mathrm{d}x &= \lim_{\varepsilon \to 0} \int_{\Omega \setminus B_\varepsilon(y)} \boldsymbol{m}(x) \cdot \nabla \mathfrak{S}(x,y) \, \mathrm{d}x \\ &= \lim_{\varepsilon \to 0} \int_{\partial\Omega \cup \partial B_\varepsilon(y)} \mathfrak{S}(x,y) \boldsymbol{m}(x) \cdot \nu \, \mathrm{d}s_x \\ &\quad - \lim_{\varepsilon \to 0} \int_{\Omega \setminus B_\varepsilon(y)} \mathfrak{S}(x,y) \mathrm{div} \boldsymbol{m} \, \mathrm{d}x. \end{aligned}$$

Wir erhalten weiterhin durch Transformation auf Kugelkoordinaten und Grenzwertbildung

$$\begin{aligned} \lim_{\varepsilon \to 0} \int_{B_\varepsilon(y)} \mathfrak{S}(x,y) \mathrm{div} \boldsymbol{m}(x) \, \mathrm{d}x &= -\frac{1}{4\pi} \lim_{\varepsilon \to 0} \int_{B_\varepsilon(y)} \frac{1}{|x-y|} \mathrm{div} \boldsymbol{m}(x) \, \mathrm{d}x \\ &= -\frac{1}{4\pi} \lim_{\varepsilon \to 0} \int_{B_\varepsilon(0)} \frac{1}{|\tilde{x}|} \mathrm{div} \boldsymbol{m}(\tilde{x}, y) \, \mathrm{d}\tilde{x} \\ &= -\frac{1}{4\pi} \lim_{\varepsilon \to 0} \int_{B_\varepsilon(0)} \frac{1}{r} r^2 \sin\psi \, \mathrm{div} \boldsymbol{m}(r,\varphi,\psi,y) \, \mathrm{d}r \, \mathrm{d}\varphi \, \mathrm{d}\psi = 0 \end{aligned}$$

und

$$\begin{aligned}\lim_{\varepsilon\to 0}\int\limits_{\partial B_\varepsilon(y)} \mathfrak{S}(x,y)\boldsymbol{m}(x)\cdot\nu(x)\,\mathrm{d}s_x &= -\frac{1}{4\pi}\lim_{\varepsilon\to 0}\int\limits_{\partial B_\varepsilon(y)} \frac{1}{|x-y|}\boldsymbol{m}(x)\cdot\nu(x)\,\mathrm{d}s_x\\ &= -\frac{1}{4\pi}\lim_{\varepsilon\to 0}\int\limits_{\partial B_\varepsilon(0)} \frac{1}{|\tilde{x}|}\boldsymbol{m}(\tilde{x},y)\cdot\nu(\tilde{x},y)\,\mathrm{d}s_{\tilde{x}}\\ &= -\frac{1}{4\pi}\lim_{\varepsilon\to 0}\int\limits_{\partial B_\varepsilon(0)} \frac{1}{\varepsilon}\varepsilon^2\sin\varphi\,\boldsymbol{m}(\varphi,\psi,y)\cdot\nu(\varphi,\psi,y)\,\mathrm{d}\varphi\,\mathrm{d}\psi = 0.\end{aligned}$$

Hieraus folgt die Behauptung. □

Mit Hilfe von Lemma 5.3 ergibt sich, dass

$$\begin{aligned}E_S(\boldsymbol{m}) = \frac{\mu_0}{2}\int_\Omega \boldsymbol{m}\cdot\nabla u_1\,\mathrm{d}x - \frac{\mu_0}{2}\int_\Omega\int_{\partial\Omega}\mathfrak{S}(x,y)(\operatorname{div}\boldsymbol{m})g(y)\,\mathrm{d}s_y\,\mathrm{d}x\\ + \frac{\mu_0}{2}\int_{\partial\Omega}\int_{\partial\Omega}\mathfrak{S}(x,y)\nu_x\cdot\boldsymbol{m}(x)g(y)\,\mathrm{d}s_y\,\mathrm{d}s_x.\end{aligned} \tag{5.11}$$

Im folgenden Abschnitt werden wir zeigen, wie (5.11) genutzt werden kann, um die Streufeldenergie effizient numerisch zu berechnen.

5.2 Effiziente numerische Berechnung der einzelnen Teil-Energien

Nun betrachten wir die Diskretisierung der verschiedenen Teil-Energien des Landau-Lifschitz-Modells. Dabei ist der numerisch anspruchsvollste Teil die Berechnung der Streufeldenergie durch den nichtlokalen Anteil u_2. Hierfür benötigen wir die Duffy-Transformationen zur Berechnung von singulären Integralen.

5.2.1 Diskretisierung

Im Folgenden bezeichnet $\boldsymbol{m}_h \in [\mathcal{M}^1]^3$ die Magnetisierung $\boldsymbol{m}$ welche mittels linearer Ansatzfunktionen diskretisiert wurde, so dass

$$\boldsymbol{m}_h = \sum_{i\in I}\boldsymbol{\alpha}_i\varphi_i$$

gilt. Somit erhält man für die Ableitung $D\boldsymbol{m}_h = \sum_{i\in I}\boldsymbol{\alpha}_i(\nabla\varphi_i)^T$. Die Austauschenergie aus (5.2) kann berechnet werden durch

$$\begin{aligned}\int_\Omega \|D\boldsymbol{m}_h\|_F^2\,\mathrm{d}x &= \int_\Omega \operatorname{spur}(D\boldsymbol{m}_h)^T(D\boldsymbol{m}_h)\,\mathrm{d}x = \sum_{i,j\in I}\int_\Omega \operatorname{spur}\nabla\varphi_i\boldsymbol{\alpha}_i^T\boldsymbol{\alpha}_j(\nabla\varphi_j)^T\,\mathrm{d}x\\ &= \sum_{i,j\in I}\boldsymbol{\alpha}_i\cdot\boldsymbol{\alpha}_j\int_\Omega \operatorname{spur}\nabla\varphi_i(\nabla\varphi_j)^T\,\mathrm{d}x = \sum_{i,j\in I}\boldsymbol{\alpha}_i\cdot\boldsymbol{\alpha}_j\int_\Omega\nabla\varphi_i\cdot\nabla\varphi_j\,\mathrm{d}x.\end{aligned}$$

Für die anisotrope Energie und die Zeeman-Energie ergeben sich

$$\int_\Omega \phi(\boldsymbol{m}_h)\,\mathrm{d}x = \int_\Omega 1-(\boldsymbol{e}\cdot\boldsymbol{m}_h)^2\,\mathrm{d}x = |\Omega| - \sum_{i,j\in I}(\boldsymbol{e}\cdot\boldsymbol{\alpha}_i)(\boldsymbol{e}\cdot\boldsymbol{\alpha}_j)\int_\Omega \varphi_i\varphi_j\,\mathrm{d}x$$

und

$$\int_\Omega \boldsymbol{f}\cdot\boldsymbol{m}_h\,\mathrm{d}x = \sum_{i\in I}\boldsymbol{f}\cdot\boldsymbol{\alpha}_i\int_\Omega\varphi_i\,\mathrm{d}x.$$

Da u_1 auf dem Rand $\partial\Omega$ verschwindet, benötigen wir für die Diskretisierung von u_1 nur die inneren Freiheitsgrade, d.h.

$$u_1 = \sum_{j\in I_{in}}\beta_j\varphi_j$$

mit $I_{\text{in}} := I\backslash I_{\text{bd}}$, wobei I_{bd} die Indizes auf dem Rand darstellen. Sei $I_{\text{lay}} \subset I_{\text{in}}$ die Menge der Knoten mit einem Nachbarn in der Menge der Rand-Indizes I_{bd}. Somit folgt für die Einschränkung von ∇u_1 auf den Rand $\partial\Omega$, dass $(\nabla u_1)|_{\partial\Omega} = \sum_{i\in I_{lay}}\beta_i\nabla\varphi_i$ und es ist

$$g = \sum_{j\in I_{lay}}\beta_j\nu\cdot\nabla\varphi_j - \sum_{j\in I_{bd}}\nu\cdot\boldsymbol{\alpha}_j\varphi_j.$$

Die Streufeldenergie lässt sich umformen zu

$$\begin{aligned}
E_S[\boldsymbol{m}_h] &= \frac{\mu_0}{2}\sum_{i\in I}\left(\sum_{j\in I_{in}}\beta_j\boldsymbol{\alpha}_i\cdot\int_\Omega\varphi_i\nabla\varphi_j\,\mathrm{d}x - \boldsymbol{\alpha}_i\cdot\int_\Omega\int_{\partial\Omega}\nabla\varphi_i(x)\mathfrak{S}(x,y)g(y)\,\mathrm{d}s_y\,\mathrm{d}x\right)\\
&\quad+\frac{\mu_0}{2}\sum_{i\in I_{bd}}\boldsymbol{\alpha}_i\cdot\int_{\partial\Omega}\int_{\partial\Omega}\nu_x\varphi_i(x)\mathfrak{S}(x,y)g(y)\,\mathrm{d}s_y\,\mathrm{d}x\\
&= \frac{\mu_0}{2}\sum_{i\in I}\left(\sum_{j\in I_{in}}\beta_j\boldsymbol{\alpha}_i\cdot\int_\Omega\varphi_i\nabla\varphi_j\,\mathrm{d}x - \sum_{j\in I_{lay}}\beta_j\boldsymbol{\alpha}_i\cdot\boldsymbol{a}_{ij} + \sum_{j\in I_{bd}}\boldsymbol{\alpha}_i\cdot(B_{ij}\boldsymbol{\alpha}_j)\right)\\
&\quad+\frac{\mu_0}{2}\sum_{i\in I_{bd}}\left(\sum_{j\in I_{lay}}\beta_j\boldsymbol{\alpha}_i\cdot\boldsymbol{c}_{ij} - \sum_{j\in I_{bd}}\boldsymbol{\alpha}_i\cdot(D_{ij}\boldsymbol{\alpha}_j)\right)
\end{aligned}$$

wobei

$$\boldsymbol{a}_{ij} = \int_\Omega\int_{\partial\Omega}\nabla\varphi_i(x)\mathfrak{S}(x,y)\nu_y\cdot\nabla\varphi_j(y)\,\mathrm{d}s_y\,\mathrm{d}x, \tag{5.12a}$$

$$B_{ij} = \int_\Omega\int_{\partial\Omega}\nabla\varphi_i(x)\mathfrak{S}(x,y)\varphi_j(y)\nu_y^T\,\mathrm{d}s_y\,\mathrm{d}x, \tag{5.12b}$$

$$\boldsymbol{c}_{ij} = \int_{\partial\Omega}\int_{\partial\Omega}\nu_x\varphi_i(x)\mathfrak{S}(x,y)\nu_y\cdot\nabla\varphi_j(y)\,\mathrm{d}s_y\,\mathrm{d}s_x, \tag{5.12c}$$

$$D_{ij} = \int_{\partial\Omega}\int_{\partial\Omega}\nu_x\varphi_i(x)\mathfrak{S}(x,y)\varphi_j(y)\nu_y^T\,\mathrm{d}s_y\,\mathrm{d}s_x. \tag{5.12d}$$

Im nächsten Abschnitt gehen wir näher auf die Berechnung dieser singulären Integrale ein.

5.2.2 Duffy-Transformation

Die effiziente numerische Berechnung der singulären Integrale (5.12a), (5.12b), (5.12c) und (5.12d) ist eine anspruchsvolle Aufgabe. Eine Möglichkeit der Berechnung besteht in der Verwendung von Kugelkoordinaten. Dabei hat man jedoch den Nachteil von komplizierten Integralgrenzen.

Ein weiterer Ansatz ist die Verwendung der Duffy-Transformation, siehe [Duf82]. Dabei wird im Wesentlichen ein Dreieck auf ein Viereck transformiert, um die Singularitäten zu eliminieren. Anhand des folgenden Beispiels wird dieses Prinzip demonstriert. Mit den Transformationen $x = \xi$ und $y = \xi\eta$ erhält man

$$\int_0^1 \int_0^x \frac{1}{x+y} \,\mathrm{d}y \,\mathrm{d}x = \int_0^1 \int_0^1 \frac{1}{1+\eta} \,\mathrm{d}\eta \,\mathrm{d}\xi,$$

so dass die Integration mit gewöhnlichen Methoden durchgeführt werden kann.

Dieses Prinzip kann auf die Integration über zwei Dreiecke verallgemeinert werden, siehe [SS11]. Dabei unterscheidet man die folgenden drei Fälle: gleiches Dreieck, gleiche Kante und gleicher Knoten. Mit Hilfe dieser Transfomationen können die Integrale (5.12c) und (5.12d) berechnet werden. Weitere Arbeiten zu diesem Thema, welche sich mit exponentieller Konvergenz beschäftigen, sind [SW92, CPS11].

Dieses Verfahren kann auf die Kombination von Referenz-Dreieck mit Referenz-Tetraeder übertragen werden. Sei eine Kernfunktion $\kappa : \mathbb{R}^3 \times \mathbb{R}^2 \to \mathbb{R}$ gegeben und das zu berechnende Integral hat die Form

$$\mathcal{I} := \int_0^1 \int_0^{1-x_1} \int_0^{1-x_1-x_2} \int_0^1 \int_0^{1-y_1} \kappa(\boldsymbol{x}, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} \,\mathrm{d}\boldsymbol{x},$$

mit $\boldsymbol{x} := (x_1, x_2, x_3)$ und $\boldsymbol{y} := (y_1, y_2)$. Durch Relativkoordinaten wird $\mathcal{I}$ transformiert zu

$$\mathcal{I} = \int_0^1 \int_0^{\tilde{x}_1} \int_0^{\tilde{x}_1-\tilde{x}_2} \int_{-\tilde{x}_1}^{1-\tilde{x}_1} \int_{-\tilde{x}_2}^{\tilde{y}_1+\tilde{x}_1-\tilde{x}_2} \kappa\left(\begin{pmatrix} 1-\tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{pmatrix}, \begin{pmatrix} 1-\tilde{y}_1-\tilde{x}_1 \\ \tilde{y}_2+\tilde{x}_2 \end{pmatrix}\right) \,\mathrm{d}\tilde{\boldsymbol{y}} \,\mathrm{d}\tilde{\boldsymbol{x}}.$$

Nur an einem einzelnen Punkt $\tilde{\boldsymbol{y}} = 0$, $\tilde{x}_3 = 0$ existiert eine Singularität welche mit Hilfe der Duffy-Transformation eliminiert werden kann.

Ähnlich zu dem Ansatz in [SS11] spalten wir das Integrationsgebiet in die folgenden Teile auf.

$$\left\{\begin{pmatrix} -1 \le \tilde{y}_1 \le 0 \\ -1 \le \tilde{y}_2 \le \tilde{y}_1 \\ -\tilde{y}_2 \le \tilde{x}_1 \le 1 \\ -\tilde{y}_2 \le \tilde{x}_2 \le \tilde{x}_1 \\ 0 \le \tilde{x}_3 \le \tilde{x}_1 - \tilde{x}_2 \end{pmatrix}\right\} \cup \left\{\begin{pmatrix} -1 \le \tilde{y}_1 \le 0 \\ \tilde{y}_1 \le \tilde{y}_2 \le 0 \\ -\tilde{y}_1 \le \tilde{x}_1 \le 1 \\ -\tilde{y}_2 \le \tilde{x}_2 \le \tilde{x}_1 + \tilde{y}_1 - \tilde{y}_2 \\ 0 \le \tilde{x}_3 \le \tilde{x}_1 - \tilde{x}_2 \end{pmatrix}\right\} \cup \left\{\begin{pmatrix} -1 \le z_1 \le 0 \\ 0 \le \tilde{y}_2 \le 1 + \tilde{y}_1 \\ \tilde{y}_2 - \tilde{y}_1 \le \tilde{x}_1 \le 1 \\ 0 \le \tilde{x}_2 \le \tilde{x}_1 + \tilde{y}_1 - \tilde{y}_2 \\ 0 \le \tilde{x}_3 \le \tilde{x}_1 - \tilde{x}_2 \end{pmatrix}\right\}$$
$$\cup \left\{\begin{pmatrix} 0 \le \tilde{y}_1 \le 1 \\ -1 + \tilde{y}_1 \le \tilde{y}_2 \le 0 \\ -\tilde{y}_2 \le \tilde{x}_1 \le 1 - \tilde{y}_1 \\ -\tilde{y}_2 \le \tilde{x}_2 \le \tilde{x}_1 \\ 0 \le \tilde{x}_3 \le \tilde{x}_1 - \tilde{x}_2 \end{pmatrix}\right\} \cup \left\{\begin{pmatrix} 0 \le \tilde{y}_1 \le 1 \\ 0 \le \tilde{y}_2 \le \tilde{y}_1 \\ 0 \le \tilde{x}_1 \le 1 - \tilde{y}_1 \\ 0 \le \tilde{x}_2 \le \tilde{x}_1 \\ 0 \le \tilde{x}_3 \le \tilde{x}_1 - \tilde{x}_2 \end{pmatrix}\right\} \cup \left\{\begin{pmatrix} 0 \le \tilde{y}_1 \le 1 \\ \tilde{y}_1 \le \tilde{y}_2 \le 1 \\ \tilde{y}_2 - \tilde{y}_1 \le \tilde{x}_1 \le 1 - \tilde{y}_1 \\ 0 \le \tilde{x}_2 \le \tilde{y}_1 - \tilde{y}_2 + \tilde{x}_1 \\ 0 \le \tilde{x}_3 \le \tilde{x}_1 - \tilde{x}_2 \end{pmatrix}\right\}.$$

Die entsprechenden Integrale $\mathcal{I} := \mathcal{I}_1 + \ldots + \mathcal{I}_6$, können jeweils auf $(0,1)^5$ transformiert werden. Somit ergeben sich die folgenden Terme.

- $\mathcal{I}_1$:

$$\mathcal{I}_1 = \int\limits_{(0,1)^5} p_1 \kappa \left(\begin{pmatrix} 1-\eta_5 \\ \eta_5(1-\eta_1+\eta_1\eta_2) \\ \eta_1\eta_4\eta_5(1-\eta_2) \end{pmatrix}, \begin{pmatrix} 1-\eta_5+\eta_1\eta_2\eta_3\eta_5 \\ \eta_5(1-\eta_1) \end{pmatrix} \right) \mathrm{d}\boldsymbol{\eta},$$
$$p_1 := \eta_1^3\eta_2\eta_5^4(1-\eta_2)$$

- $\mathcal{I}_2$:

$$\mathcal{I}_2 = \int\limits_{(0,1)^5} p_2 \kappa \left(\begin{pmatrix} 1-\eta_5 \\ \eta_1\eta_5(1-\eta_2+\eta_2\eta_3) \\ \eta_4\eta_5(1-\eta_1+\eta_1\eta_2-\eta_1\eta_2\eta_3) \end{pmatrix}, \begin{pmatrix} 1-\eta_5+\eta_1\eta_2\eta_5 \\ \eta_1\eta_5(1-\eta_2) \end{pmatrix} \right) \mathrm{d}\boldsymbol{\eta},$$
$$p_2 := \eta_1^2\eta_2\eta_5^4(1-\eta_1+\eta_1\eta_2-\eta_1\eta_2\eta_3)$$

- $\mathcal{I}_3$:

$$\mathcal{I}_3 = \int\limits_{(0,1)^5} p_3 \kappa \left(\begin{pmatrix} 1-\eta_5 \\ \eta_1\eta_5(1-\eta_2) \\ \eta_4\eta_5(1-\eta_1+\eta_1\eta_2) \end{pmatrix}, \begin{pmatrix} 1-\eta_5+\eta_1\eta_2\eta_3\eta_5 \\ \eta_1\eta_5(1-\eta_2\eta_3) \end{pmatrix} \right) \mathrm{d}\boldsymbol{\eta},$$
$$p_3 := \eta_1^2\eta_2\eta_5^4(1-\eta_1+\eta_1\eta_2),$$

- $\mathcal{I}_4$:

$$\mathcal{I}_4 = \int\limits_{(0,1)^5} p_4 \kappa \left(\begin{pmatrix} 1-\eta_5+\eta_1\eta_2\eta_3\eta_5 \\ \eta_1\eta_5(1-\eta_2\eta_3) \\ \eta_4\eta_5(1-\eta_1) \end{pmatrix}, \begin{pmatrix} 1-\eta_5 \\ \eta_1\eta_5(1-\eta_2) \end{pmatrix} \right) \mathrm{d}\boldsymbol{\eta},$$
$$p_4 := \eta_1^2\eta_2\eta_5^4(1-\eta_1)$$

- $\mathcal{I}_5$:

$$\mathcal{I}_5 = \int\limits_{(0,1)^5} p_5 \kappa \left(\begin{pmatrix} 1-\eta_5+\eta_1\eta_2\eta_5 \\ \eta_1\eta_5(1-\eta_2) \\ \eta_4\eta_5(1-\eta_1) \end{pmatrix}, \begin{pmatrix} 1-\eta_5 \\ \eta_1\eta_5(1-\eta_2+\eta_2\eta_3) \end{pmatrix} \right) \mathrm{d}\boldsymbol{\eta},$$
$$p_5 := \eta_1^2\eta_2\eta_5^4(1-\eta_1)$$

- $\mathcal{I}_6$:

$$\mathcal{I}_6 = \int\limits_{(0,1)^5} p_6 \kappa \left(\begin{pmatrix} 1-\eta_5+\eta_1\eta_2\eta_3\eta_5 \\ \eta_5(1-\eta_1) \\ \eta_1\eta_4\eta_5(1-\eta_2\eta_3) \end{pmatrix}, \begin{pmatrix} 1-\eta_5 \\ \eta_5(1-\eta_1+\eta_1\eta_2) \end{pmatrix} \right) \mathrm{d}\boldsymbol{\eta},$$
$$p_6 := \eta_1^3\eta_2\eta_5^4(1-\eta_2\eta_3)$$

Die Integrale $\mathcal{I}_1, \ldots, \mathcal{I}_6$ können nun effizient mit Standardverfahren, wie der Gauß-Quadratur, berechnet werden.

Durch folgende Beispiele erkennt man die Effektivität der obigen Transformation. Zum einen wurden die Integrale durch Gauß-Quadratur und zum anderen durch Gauß-Quadratur mit vorheriger Duffy-Transformation berechnet. Zur Bestimmung der Referenzlösung der Integrale wurde eine adaptive Simpson-Quadratur, siehe [HB06], implementiert. Die Werte in runden Klammern geben die relative Abweichung zur Referenzlösung an.

$$\int_0^1 \int_0^{1-y_1} \int_0^1 \int_0^{1-x_1} \int_0^{1-x_1-x_2} \frac{1}{\left| \begin{pmatrix} x_1-y_1 \\ x_2-y_2 \\ x_3 \end{pmatrix} \right|} \mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} = 0{,}2150, \tag{5.13a}$$

Gauß-Quad	Duffy+Gauß-Quad
0,2074 (-3,5e-02)	0,2152 (9,3e-04)

$$\int_0^1 \int_0^{1-y_1} \int_0^1 \int_0^{1-x_1} \int_0^{1-x_1-x_2} \frac{1-y_1-y_2}{\left| \begin{pmatrix} x_1-y_1 \\ x_2-y_2 \\ x_3 \end{pmatrix} \right|} \mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} = 0{,}0764368, \tag{5.13b}$$

Gauß-Quad	Duffy+Gauß-Quad
0,07626 (-2,3e-03)	0,07649 (6,9e-04)

$$\int_0^1 \int_0^{1-y_1} \int_0^1 \int_0^{1-x_1} \frac{1-y_1-y_2}{|\boldsymbol{x}-\boldsymbol{y}|} \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} = 0{,}341632, \tag{5.13c}$$

Gauß-Quad	Duffy+Gauß-Quad
n.a.	0,34164 (2,3e-05)

$$\int_0^1 \int_0^{1-y_1} \int_0^1 \int_0^{1-x_1} \frac{(1-y_1-y_2)(1-x_1-x_2)}{|\boldsymbol{x}-\boldsymbol{y}|} \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} = 0{,}136653, \tag{5.13d}$$

Gauß-Quad	Duffy+Gauß-Quad
n.a.	0,136657 (2,9e-05)

Es ist zu erkennen, dass die Duffy-Transformation die Integrale (5.13a) und (5.13b) glättet und somit eine genauere Berechnung ermöglicht. Für (5.13c) und (5.13d) ist es nur mit Hilfe der Transformation möglich, die Integrale zu berechnen, da sonst bei der verwendeten Implementierung der Gauß-Quadratur auf der Singularität ausgewertet wird.

5.3 Energieminimierung

Die wesentliche Schwierigkeit bei der Minimierung des Energiefunktionals (5.2) liegt in der Beachtung der punktweisen Nebenbedingung $\|\boldsymbol{m}\|_2 = 1$ für die Magnetisierung $\boldsymbol{m}$. Wir werden uns in diesem Abschnitt damit auseinandersetzen, wie diese numerisch effizient und stabil umgesetzt werden kann.

5.3.1 Das Minimierungsverfahren

Alouges stellte in [Alo97, Alo01] ein Verfahren zur Minimierung des Energiefunktionals (5.2) vor. Wir untersuchen die Konvergenz der diskretisierten Methode und verwenden Ideen aus [Alo97, ACDP04, Bar05].

Betrachtung eines vereinfachten Modells: Zunächst beschränken wir uns auf das vereinfachte Funktional

$$E(\boldsymbol{m}) = \Theta(\boldsymbol{m}) + \frac{1}{2}\int_\Omega \|D\boldsymbol{m}\|_F^2 \,\mathrm{d}x \tag{5.14}$$

mit einer hinreichend glatten Funktion $\Theta : H^1(\Omega;\mathbb{R}^3) \to \mathbb{R}$. Anschließend werden wir unsere Aussagen auf das mikromagnetische Energie-Funktional (5.2) übertragen.

Algorithmus (Minimierungsverfahren nach Alouges): Sei $\boldsymbol{m}_h^0 \in [\mathcal{M}^1]^3$, so dass $\|\boldsymbol{m}_h^0(z)\|_2 = 1$ für alle Knoten der Diskretisierung $z \in \mathcal{N}_h$. Man führe für $n = 0, 1, 2, ..$ die folgenden Iterationen aus:

(1) Berechne $\boldsymbol{w}_h^n \in [\mathcal{M}^1]^3$ mit $\int_\Omega \boldsymbol{w}_h^n \,\mathrm{d}x = 0$, so dass $\boldsymbol{w}_h^n(z) \cdot \boldsymbol{m}_h^n(z) = 0$ für alle $z \in \mathcal{N}_h$ und

$$\Theta'(\boldsymbol{m}_h^n)[\boldsymbol{v}_h] + \int_\Omega \text{spur}\,[(D(\boldsymbol{m}_h^n + \boldsymbol{w}_h^n))^T (D\boldsymbol{v}_h)]\,\mathrm{d}x = 0$$

für alle $\boldsymbol{v}_h \in [\mathcal{M}^1]^3$ mit $\int_\Omega \boldsymbol{v}_h \,\mathrm{d}x = 0$.

(2) Sei $\boldsymbol{m}_h^{n+1} \in [\mathcal{M}^1]^3$ definiert durch

$$\boldsymbol{m}_h^{n+1}(z) := \frac{\boldsymbol{m}_h^n(z) + \alpha \boldsymbol{w}_h^n(z)}{\|\boldsymbol{m}_h^n(z) + \alpha \boldsymbol{w}_h^n(z)\|_2}$$

für alle $z \in \mathcal{N}_h$ mit einem angemessenen $\alpha > 0$.

Um die Konvergenz des Verfahrens zu zeigen werden wir ähnlich wie in [Alo97, ACDP04, Bar05] vorgehen. Zunächst betrachten wir folgendes Lemma für eine gegebene, schwach spitzwinklige Triangulierung (für Details siehe [Bar05]).

Lemma 5.4. *Seien $\boldsymbol{m}_h, \boldsymbol{w}_h \in [\mathcal{M}^1]^3$, so dass $\|\boldsymbol{m}_h(z)\|_2 = 1$ und $\boldsymbol{m}_h(z) \cdot \boldsymbol{w}_h(z) = 0$ für alle $z \in \mathcal{N}_h$. Dann gilt*

$$\int_\Omega \left\| D\left(\frac{\boldsymbol{m}_h + \alpha \boldsymbol{w}_h}{\|\boldsymbol{m}_h + \alpha \boldsymbol{w}_h\|_2} \right) \right\|_F^2 \mathrm{d}x \leq \int_\Omega \|D(\boldsymbol{m}_h + \alpha \boldsymbol{w}_h)\|_F^2 \,\mathrm{d}x$$

für alle $\alpha \in \mathbb{R}$.

Beweis. Für einen Beweis siehe [Bar05]. □

Seien $\boldsymbol{m}_h^n$ und $\boldsymbol{w}_h^n$ wie im obigen Algorithmus. Wir setzen $\boldsymbol{v}_h = \boldsymbol{w}_h^n$, so dass

$$\int_\Omega \text{spur}[(D\boldsymbol{m}_h^n)^T (D\boldsymbol{w}_h^n)]\,\mathrm{d}x = -\Theta'(\boldsymbol{m}_h^n)[\boldsymbol{w}_h^n] - \int_\Omega \|D\boldsymbol{w}_h^n\|_F^2 \,\mathrm{d}x.$$

Es folgt mit Lemma 5.4, dass

$$\begin{aligned}
\frac{1}{2}\int_\Omega \|D\boldsymbol{m}_h^{n+1}\|_F^2 - \|D\boldsymbol{m}_h^n\|_F^2 \,\mathrm{d}x &\leq \frac{1}{2}\int_\Omega \|D(\boldsymbol{m}_h^n + \alpha \boldsymbol{w}_h^n)\|_F^2 - \|D\boldsymbol{m}_h^n\|_F^2 \,\mathrm{d}x \\
&= \alpha \int_\Omega \text{spur}[(D\boldsymbol{m}_h^n)^T (D\boldsymbol{w}_h^n)] + \frac{\alpha}{2}\|D\boldsymbol{w}_h^n\|_F^2 \,\mathrm{d}x \\
&= -\alpha \Theta'(\boldsymbol{m}_h^n)[\boldsymbol{w}_h^n] + (\alpha^2/2 - \alpha)\int_\Omega \|D\boldsymbol{w}_h^n\|_F^2 \,\mathrm{d}x.
\end{aligned}$$

Somit erhält man

$$\begin{aligned}
E(\boldsymbol{m}_h^{n+1}) - E(\boldsymbol{m}_h^n) &= \Theta(\boldsymbol{m}_h^{n+1}) - \Theta(\boldsymbol{m}_h^n) + \frac{1}{2}\int_\Omega \|D\boldsymbol{m}_h^{n+1}\|_F^2 - \|D\boldsymbol{m}_h^n\|_F^2 \,\mathrm{d}x \\
&= \Theta(\boldsymbol{m}_h^{n+1}) - \Theta(\boldsymbol{m}_h^n) - \alpha\Theta'(\boldsymbol{m}_h^n)[\boldsymbol{w}_h^n] + \\
&\quad + (\alpha^2/2 - \alpha)\int_\Omega \|D\boldsymbol{w}_h^n\|_F^2 \,\mathrm{d}x.
\end{aligned}$$

Um die Konvergenz des Verfahrens zu beweisen, nehmen wir folgende Glattheitsannahmen für Θ an

$$\Theta(\boldsymbol{m}_h^{n+1}) - \Theta(\boldsymbol{m}_h^n + \alpha \boldsymbol{w}_h^n) \leq C_1 \|\boldsymbol{m}_h^{n+1} - (\boldsymbol{m}_h^n + \alpha \boldsymbol{w}_h^n)\|_{L^2(\Omega)}, \tag{5.15a}$$

$$\Theta(\boldsymbol{m}_h^n + \alpha \boldsymbol{w}_h^n) - \Theta(\boldsymbol{m}_h^n) - \alpha \Theta'(\boldsymbol{m}_h^n)[\boldsymbol{w}_h^n] \leq C_2 \alpha^2 \|\boldsymbol{w}_h^n\|_{L^2(\Omega)}^2, \tag{5.15b}$$

mit Konstanten $C_1, C_2 \geq 0$ unabhängig von $\boldsymbol{m}_h^n$, $\boldsymbol{m}_h^{n+1}$, $\boldsymbol{w}_h^n$ und α. Weiterhin erhält man für alle $z \in \mathcal{N}_h$

$$\begin{aligned}
\|\boldsymbol{m}_h^{n+1}(z) - \boldsymbol{m}_h^n(z) - \alpha \boldsymbol{w}_h^n(z)\|_2 &= \left\| \frac{\boldsymbol{m}_h^n(z) + \alpha \boldsymbol{w}_h^n(z)}{\|\boldsymbol{m}_h^n(z) + \alpha \boldsymbol{w}_h^n(z)\|_2} - \boldsymbol{m}_h^n(z) - \alpha \boldsymbol{w}_h^n(z) \right\|_2 \\
&= \|\boldsymbol{m}_h^n(z) + \alpha \boldsymbol{w}_h^n(z)\|_2 - 1 \\
&= \left(1 + \alpha^2 \|\boldsymbol{w}_h^n(z)\|_2^2\right)^{1/2} - 1 \\
&\leq \frac{\alpha^2}{2} \|\boldsymbol{w}_h^n(z)\|_2^2.
\end{aligned}$$

Somit ergibt sich mit Hilfe der stetigen Einbettung von $H^1(\Omega) \hookrightarrow L^4(\Omega)$, vgl. Satz 1.6, und der Poincaré-Ungleichung, die folgende Abschätzung

$$\begin{aligned}
E(\boldsymbol{m}_h^{n+1}) - E(\boldsymbol{m}_h^n) &\leq (-\alpha + \alpha^2/2) \int_\Omega \|D\boldsymbol{w}_h^n\|_F^2 \,\mathrm{d}x + C_3 \alpha^2 \|\boldsymbol{w}_h^n\|_{L^4(\Omega)}^2 \\
&\leq -\alpha(1 - \alpha/2 - C_F C_3 \alpha) \int_\Omega \|D\boldsymbol{w}_h^n\|_F^2 \,\mathrm{d}x,
\end{aligned}$$

wobei $C_3 := C_1/2 + C_2$.

Sei α hinreichend klein, so dass $(1 - \alpha/2 - C_F C_3 \alpha) \geq 1/2$, dann erhalten wir

$$E(\boldsymbol{m}_h^{n+1}) - E(\boldsymbol{m}_h^n) \leq -\alpha/2 \int_\Omega \|D\boldsymbol{w}_h^n\|_F^2 \,\mathrm{d}x.$$

Durch Summation über $n = 0, 1, ..., N$ ergibt sich

$$E(\boldsymbol{m}_h^{N+1}) + (\alpha/2) \sum_{n=0}^{N} \int_\Omega \|D\boldsymbol{w}_h^n\|_F^2 \,\mathrm{d}x \leq E(\boldsymbol{m}_h^0).$$

Somit ist die Konvergenz der numerischen Methode gezeigt, da die Energie nach unten beschränkt ist.

Anwendung auf die mikromagnetische Energie: Das Funktional Θ in (5.14) ist im Falle der mikromagnetischen Energie (5.2) gegeben durch

$$\Theta(\boldsymbol{m}) = \int_\Omega 1 - (\boldsymbol{e} \cdot \boldsymbol{m})^2 - \boldsymbol{f} \cdot \boldsymbol{m} + \frac{\mu_0}{2} \boldsymbol{m} \cdot \nabla u_{\boldsymbol{m}} \,\mathrm{d}x.$$

Für die Konvergenz des obigen Minimierungsverfahrens bleibt zu zeigen, dass Θ die Glattheitsannahmen (5.15) erfüllt.

Beschränken wir uns zuerst auf (5.15a). Für $\boldsymbol{m}^1, \boldsymbol{m}^2 \in \mathcal{A}$ folgt, dass

$$\Theta(\boldsymbol{m}^1) - \Theta(\boldsymbol{m}^2) \leq \int_\Omega (\boldsymbol{e}\cdot\boldsymbol{m}^1)^2 - (\boldsymbol{e}\cdot\boldsymbol{m}^2)^2 \,\mathrm{d}x + \|\boldsymbol{f}\|_{L^2(\Omega)}\|\boldsymbol{m}^1 - \boldsymbol{m}^2\|_{L^2(\Omega)} + \frac{\mu_0}{2}\int_\Omega \nabla u_{\boldsymbol{m}^1}\cdot\boldsymbol{m}^1 - \nabla u_{\boldsymbol{m}^2}\cdot\boldsymbol{m}^2 \,\mathrm{d}x.$$

Mit Hilfe der Cauchy-Schwarzschen Ungleichung kann man den ersten Term auf der rechten Seite beschränken durch

$$\begin{aligned}\int_\Omega (\boldsymbol{e}\cdot\boldsymbol{m}^1)^2 - (\boldsymbol{e}\cdot\boldsymbol{m}^2)^2 \,\mathrm{d}x &\leq \int_\Omega \|\boldsymbol{m}^1\|_2^2 - \|\boldsymbol{m}^2\|_2^2 \,\mathrm{d}x = \int_\Omega (\boldsymbol{m}^1 - \boldsymbol{m}^2, \boldsymbol{m}^1 + \boldsymbol{m}^2)\,\mathrm{d}x\\ &\leq \int_\Omega \|\boldsymbol{m}^1 - \boldsymbol{m}^2\|_2\|\boldsymbol{m}^1 + \boldsymbol{m}^2\|_2 \,\mathrm{d}x\\ &\leq \|\boldsymbol{m}^1 - \boldsymbol{m}^2\|_{L^2(\Omega)}\|\boldsymbol{m}^1 + \boldsymbol{m}^2\|_{L^2(\Omega)}.\end{aligned}$$

In ähnlicher Weise erhält man mit (5.5), dass

$$\begin{aligned}\int_\Omega \nabla u_{\boldsymbol{m}^1}\cdot\boldsymbol{m}^1 - \nabla u_{\boldsymbol{m}^2}\cdot\boldsymbol{m}^2 \,\mathrm{d}x &= \int_{\mathbb{R}^3} \|\nabla u_{\boldsymbol{m}^1}\|_2^2 - \|\nabla u_{\boldsymbol{m}^2}\|_2^2 \,\mathrm{d}x\\ &\leq \|\nabla(u_{\boldsymbol{m}^1} - u_{\boldsymbol{m}^2})\|_{L^2(\mathbb{R}^3)}\|\nabla(u_{\boldsymbol{m}^1} + u_{\boldsymbol{m}^2})\|_{L^2(\mathbb{R}^3)}\\ &\leq \|\boldsymbol{m}^1 - \boldsymbol{m}^2\|_{L^2(\Omega)}(\|\boldsymbol{m}^1\|_{L^2(\Omega)} + \|\boldsymbol{m}^2\|_{L^2(\Omega)}).\end{aligned}$$

Somit ist die erste Annahme (5.15a) erfüllt.

Die zweite Annahme (5.15b) lässt sich umformen zu

$$\begin{aligned}&\Theta(\boldsymbol{m}+\alpha\boldsymbol{w}) - \Theta(\boldsymbol{m}) - \alpha\Theta'(\boldsymbol{m})[\boldsymbol{w}]\\ &= \int_\Omega (\boldsymbol{e}\cdot(\boldsymbol{m}+\alpha\boldsymbol{w}))^2 - (\boldsymbol{e}\cdot\boldsymbol{m})^2 - 2\alpha(\boldsymbol{e}\cdot\boldsymbol{m})(\boldsymbol{e}\cdot\boldsymbol{w})\,\mathrm{d}x\\ &\quad + \frac{\mu_0}{2}\int_\Omega \nabla u_{\boldsymbol{m}+\alpha\boldsymbol{w}}\cdot(\boldsymbol{m}+\alpha\boldsymbol{w}) - \nabla u_{\boldsymbol{m}}\cdot\boldsymbol{m} - 2\alpha\nabla u_{\boldsymbol{m}}\cdot\boldsymbol{w}\,\mathrm{d}x\\ &= \alpha^2\int_\Omega (\boldsymbol{e}\cdot\boldsymbol{w})^2\,\mathrm{d}x + \frac{\mu_0}{2}\int_\Omega \nabla u_{\boldsymbol{m}+\alpha\boldsymbol{w}}\cdot(\boldsymbol{m}+\alpha\boldsymbol{w}) - \nabla u_{\boldsymbol{m}}\cdot\boldsymbol{m} - 2\alpha\nabla u_{\boldsymbol{m}}\cdot\boldsymbol{w}\,\mathrm{d}x\\ &\leq \alpha^2\|\boldsymbol{w}\|_{L^2(\Omega)}^2 + \frac{\mu_0}{2}\int_\Omega \nabla u_{\boldsymbol{m}+\alpha\boldsymbol{w}}\cdot(\boldsymbol{m}+\alpha\boldsymbol{w}) - \nabla u_{\boldsymbol{m}}\cdot\boldsymbol{m} - 2\alpha\nabla u_{\boldsymbol{m}}\cdot\boldsymbol{w}\,\mathrm{d}x.\end{aligned}$$

Aufgrund von$\nabla u_{\boldsymbol{m}+\alpha\boldsymbol{w}} = \nabla u_{\boldsymbol{m}} + \alpha\nabla u_{\boldsymbol{w}}$ und (5.4), erhält man

$$\begin{aligned}&\int_{\mathbb{R}^3} \|\nabla u_{\boldsymbol{m}+\alpha\boldsymbol{w}}\|_2^2 - \|\nabla u_{\boldsymbol{m}}\|_2^2\,\mathrm{d}x - 2\alpha\int_\Omega \nabla u_{\boldsymbol{m}}\cdot\boldsymbol{w}\,\mathrm{d}x\\ &= \int_{\mathbb{R}^3} \|\nabla u_{\boldsymbol{m}}\|_2^2 + 2\alpha\nabla u_{\boldsymbol{m}}\cdot\nabla u_{\boldsymbol{w}} + \|\nabla u_{\boldsymbol{w}}\|_2^2 - \|\nabla u_{\boldsymbol{m}}\|_2^2\,\mathrm{d}x - 2\alpha\int_\Omega \nabla u_{\boldsymbol{m}}\cdot\boldsymbol{w}\,\mathrm{d}x\\ &= \int_{\mathbb{R}^3} \|\nabla u_{\alpha\boldsymbol{w}}\|_2^2\,\mathrm{d}x \leq \alpha^2\int_\Omega \|\boldsymbol{w}\|_2^2\,\mathrm{d}x.\end{aligned}$$

Somit erfüllt die mikromagnetische Energie unsere Glattheitsforderungen (5.15) und der Minimierungs-Algorithmus konvergiert.

5.3.2 String-Methode

Unser Ziel ist es, einen Pfad zu berechnen, dessen maximale Energie minimal ist für alle Kurven, die zwei bestimmte Zustände der Magnetisierung verbinden. In unserem Fall sind diese beiden Zustände lokale Minima und wir nehmen deshalb an, dass unser Energiefunktional $E : \mathcal{A} \to \mathbb{R}$, siehe (5.2), mindestens zwei lokale Minima $\boldsymbol{m}_0, \boldsymbol{m}_1 \in \mathcal{A}$ besitzt. Weiterhin sei eine stetige Abbildung

$$\varphi(\cdot, t) : [0, 1] \to \mathcal{A}, \quad t \geq 0$$

gegeben, wobei $\varphi(0, t) = \boldsymbol{m}_0$ und $\varphi(1, t) = \boldsymbol{m}_1$.

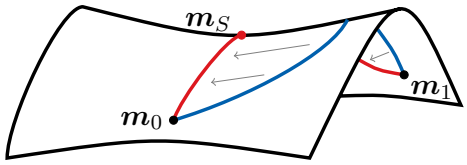


Abbildung 15: In der schematischen Skizze ist eine Energielandschaft abgebildet. Dabei repräsentiert der blaue Pfad einen Initialisierungspfad und der rote Pfad stellt den MEP dar.

Ein minimaler Energiepfad (MEP) bezeichnet einen Pfad zwischen $\boldsymbol{m}_0, \boldsymbol{m}_1 \in \mathcal{A}$, wobei für jeden Zustand die Komponente von ∇E normal zu φ verschwindet, d.h.

$$(\nabla E)^{\perp}(\varphi) = 0 \tag{5.16}$$

wobei

$$(\nabla E)^{\perp}(\varphi) = \nabla E(\gamma) - (\nabla E(\gamma), \tau)\tau.$$

Hierbei ist τ die Einheitstangente der Kurve φ und $(\cdot, \cdot)$ das Euklidische Produkt.

Zur numerischen Berechnung eines minimalen Energiepfades verwenden wir die String-Methode, welche in [ERVE02] vorgestellt und in [ERVE07] modifiziert wurde. Die modifizierte Variante zeichnet sich vor allem durch ihre einfache Implementierbarkeit aus. Ein weiteres Verfahren zur Berechnung eines minimalen Energiepfades ist die nudged elastic band (NEB)-Methode, siehe [JMJ98].

Algorithmus (vereinfachte String-Methode): Seien zwei lokale Energieminima $\boldsymbol{m}_0, \boldsymbol{m}_1 \in \mathcal{A}$ gegeben. Man bestimme einen Pfad $\varphi^0 : \{0, \ldots, N\} \to \mathcal{A}$ mit $\varphi^0(0) = \boldsymbol{m}_0$ und $\varphi^0(N) = \boldsymbol{m}_1$. Die Zwischenstellen $\varphi^0(i)$, $i = 1, \ldots, N-1$ werden mittels Interpolation berechnet. Nun führe man für $n = 0, 1, 2, ..$ die folgenden Iterationen aus:

(1) Seien $\varphi^n_*(0) = \varphi^n(0)$ und $\varphi^n_*(N) = \varphi^n(N)$. Man berechne jeweils für die Konfigurationen $\varphi^n(i)$ mit $i = 1, \ldots, N-1$ einen Minimierungsschritt des Verfahrens nach Alouges und bezeichne die Resultate mit $\varphi^n_*(i)$.

(2) Man bestimme mittels Interpolation aus $\left(\varphi^n_*(i)\right)_{i=0,\ldots,N}$ den neuen Pfad $\left(\varphi^{n+1}(i)\right)_{i=0,\ldots,N}$.

Für eine schematische Skizze zur Beschreibung der vereinfachten String-Methode siehe auch Abbildung 15. Hierbei sind $\boldsymbol{m}_0$, $\boldsymbol{m}_1$ die lokalen Energieminima und $\boldsymbol{m}_S$ stellt das Maximum des minimalen Energiepfades dar.

Die Interpolationen, die während der vereinfachten String-Methode durchgeführt werden, können in beliebiger Weise erfolgen. In unserem Fall interpolieren wir, so dass sich die jeweiligen Konfigurationen äquidistant auf der Kugeloberfläche befinden.

Der Vorteil dieser Methode ist, dass sich aufgrund des Alouges-Verfahrens die punktweise Nebenbedingung $\|\boldsymbol{m}(x)\|_2 = 1$, $x \in \Omega$ auf die jeweiligen Konfigurationen übertragen.

5.4 Numerische Ergebnisse

Um verlässliche numerische Ergebnisse zu erzeugen, werden wir zuerst die implementierten Routinen überprüfen. Anschließend untersuchen wir das asymptotische Verhalten der präsentierten numerischen Methode in Bezug auf Rechenzeit und Speicherbedarf. Im Anschluss berechnen wir einen minimalen Energiepfad zwischen dem Vortex- und dem Flower-Zustand in einem Würfel.

5.4.1 Verifizierung der Implementierung

Ein Referenzbeispiel zur Überprüfung der implementierten Routinen zur Berechnung der mikromagnetischen Energie wurde von Alex Hubert, Universität Erlangen-Nürnberg, vorgeschlagen. Hierbei bestimmt man die Kantenlänge eines Würfels, bei der die Energieminimas Flower- und Vortex-Zustand gleiche Energie besitzen. Der Test ist bekannt als μ-mag Standard-Problem #3, siehe [McM08].

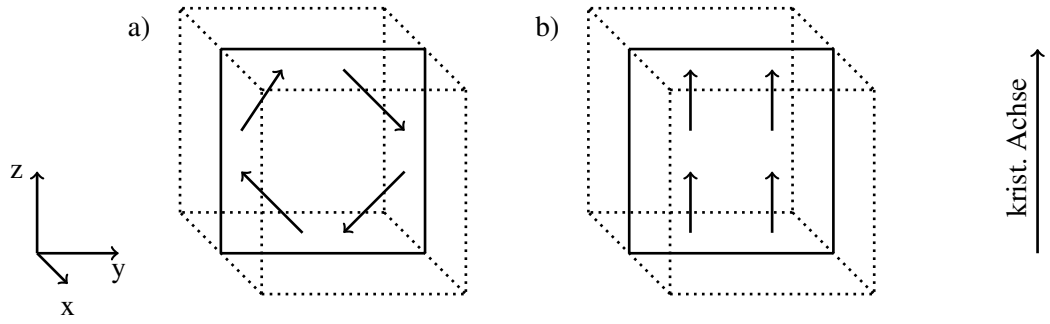


Abbildung 16: Initialisierungssetting nach [McM08] für den a) Vortex- und b) Flower-Zustand.

Für unsere Tests wurde der Würfel in 24576 Tetraeder und 3072 Oberflächendreiecke unterteilt. Die Magnetisierung wurde entsprechend wie in Abbildung 16 initialisiert. Anschließend wurden die Energieminima jeweils mit dem oben beschriebenen Minimierungsverfahren nach Alouges bestimmt. Die Ergebnisse für die verschiedenen Kantenlängen sind in Abbildung 17 zu sehen. Der theoretisch vorhergesagt Wert von ungefähr 8 stimmt mit unserem von $8{,}23$ überein.

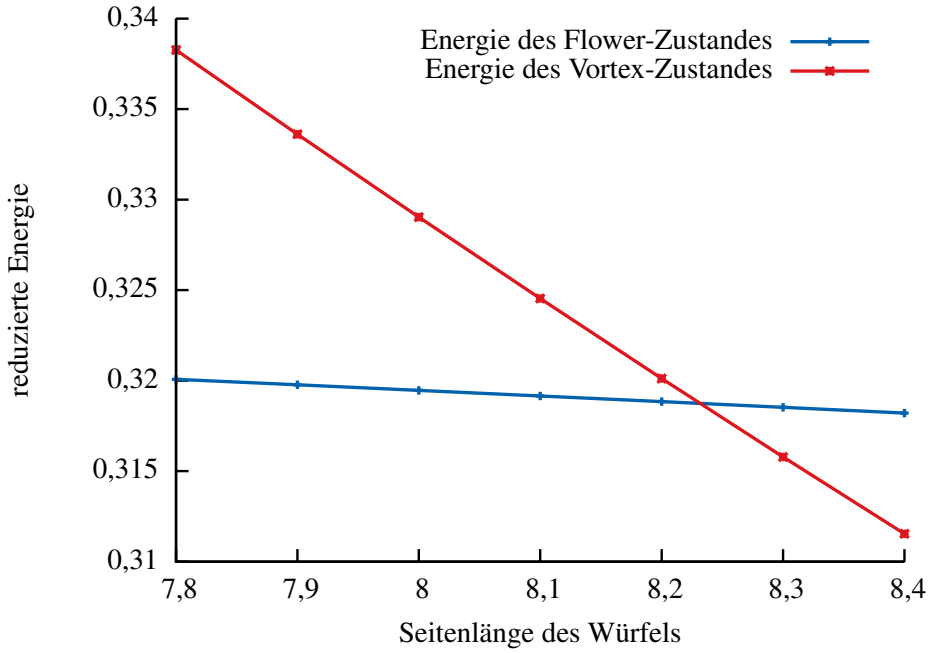


Abbildung 17: Testergebnisse für das μ-mag Standard-Problem #3.

5.4.2 Speicher- und Rechenzeitbedarf

Um das quasi-lineare Verhalten der vorgestellten numerischen Methode zu demonstrieren, führen wir Tests mit verschiedenen Diskretisierungen des Einheitswürfels durch. Dabei wurden sämtliche Verfeinerungen mit Hilfe von netgen[6] erstellt.

Zuerst betrachten wir die Erstellung der Matrizen (5.12a), (5.12b), (5.12c) und (5.12d). Ein wesentlicher Vorteil hierbei ist, dass diese Matrizen für eine bestimmte Geometrie und Approximationsgenauigkeit nur einmal erstellt werden müssen. Wie in Abbildung 18 zu sehen ist, verhält sich die Zeit und der Speicher, welcher zur Erstellung der Matrizen (5.12a) und (5.12b) verwendet wurde, linear, bis auf logarithmische Terme zur Anzahl der Tetraeder. Weiterhin ist in Abbildung 19 zu sehen, dass die Erstellung der Matrizen (5.12c) und (5.12d) quasi-optimal in Bezug auf die Anzahl der Oberflächendreiecke ist.

Bei der String-Methode hat der Minimierungsalgorithmus aus Abschnitt 5.3.1 den entscheidenden Anteil an der Gesamtberechnungszeit. Aus diesem Grund werden in Abbildung 20 einzelne Minimierungsschritte dargestellt. Das zu beobachtende asymptotische Verhalten ist quasi-linear in Bezug auf die Anzahl der Tetraeder. Wie in Abbildung 20 weiterhin zu erkennen ist, beträgt die Rechenzeit für einen Minimierungsschritt selbst bei einem relativ kleinen Beispiel, ca. 200000 Tetraeder, bereits mehr als eine Minute. Somit ist es entscheidend, ein Verfahren zu verwenden, welches quasi-optimale Komplexität besitzt, um wesentlich größere Beispiele in vertretbarer Zeit zu rechnen.

[6] `http://www.hpfem.jku.at/netgen/`

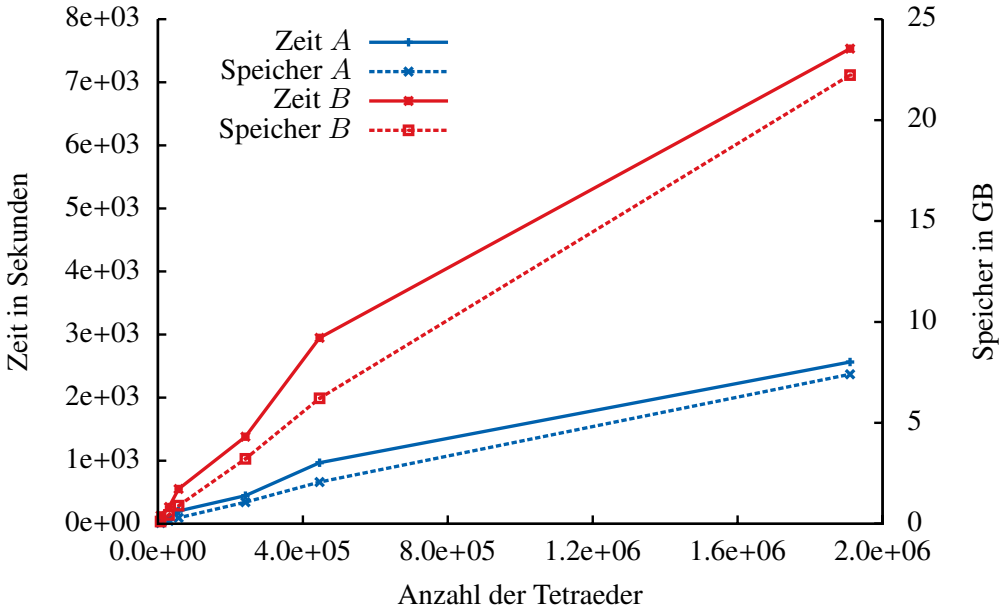


Abbildung 18: Zeit- und Speicherverbrauch für die Matrizen A und B, vgl. (5.12a) und (5.12b).

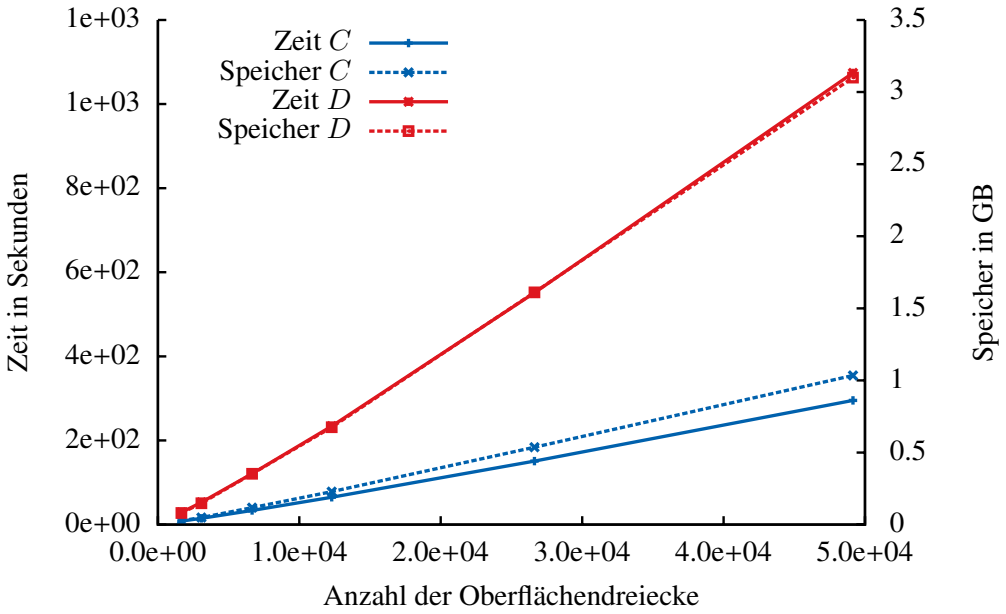


Abbildung 19: Zeit- und Speicherverbrauch für die Matrizen C und D, vgl. (5.12c) und (5.12d).

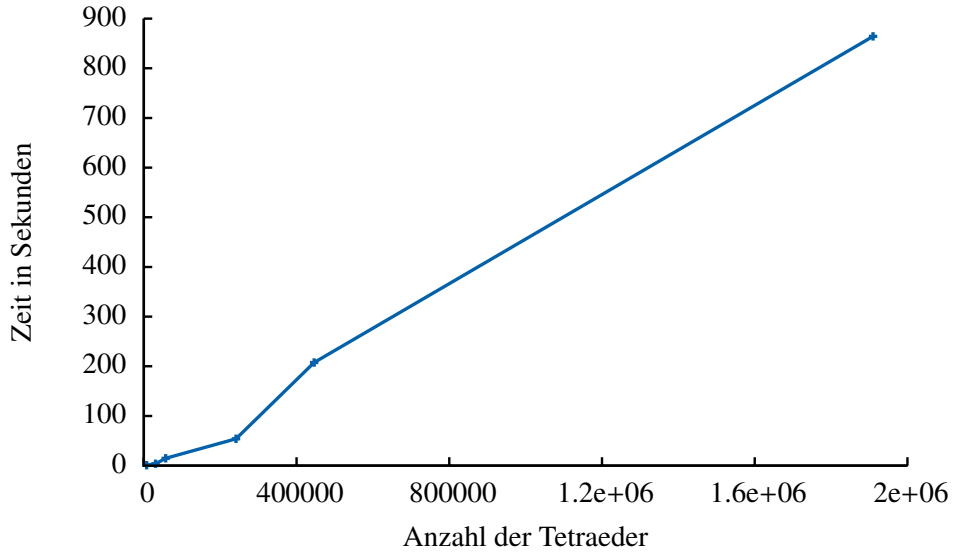


Abbildung 20: Benötigte Zeit für einen einzelnen Minimierungsschritt.

5.4.3 Minimaler Energiepfad

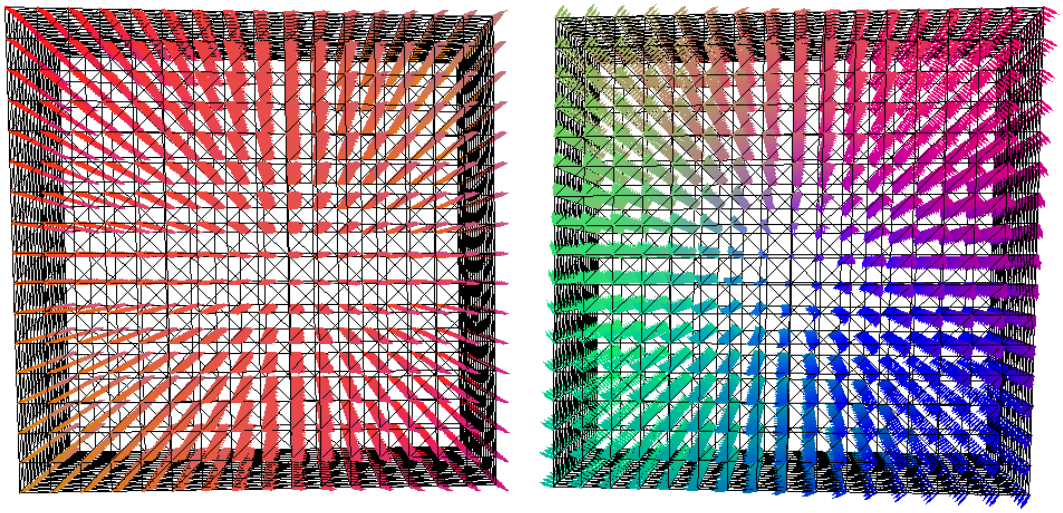
Wir berechnen mit der in Abschnitt 5.3.2 beschriebenen String-Methode einen minimalen Energiepfad zwischen dem Vortex- und dem Flower-Zustand. Als Setting wurde ein Würfel mit der Parameterkonfiguration aus dem μ-mag Standard-Problem #3 gewählt, wobei die Kantenlänge $8{,}2$ beträgt. Für die Diskretisierung wurde der Würfel in 24576 Tetraeder und 3072 Oberflächendreiecke unterteilt. Die Anzahl der einzelnen Zustände des Pfades zwischen den Minima beträgt 41.

In Abbildung 21 werden die beiden Energieminima (Vortex- und Flower-Zustand) gezeigt, welche als Anfangs- und Endpunkt für den MEP dienen. Dabei repräsentieren unterschiedliche Farben der Pfeile unterschiedliche Richtungen im Raum. Die Abbildung wurde mit Hilfe der Programmierschnittstelle OpenGL[7] erstellt.

Die Abbildung 22 vergleicht die reduzierte Energie des initialisierten Pfades mit dem des MEP. Anschließend werden in Tabelle 6 ausgewählte Zwischenzustände dargestellt, wobei Zustand 25 die Energie maximiert. Bemerkenswert am Vortex-Zustand ist (Zustand 40), dass auf der Vorderseite ein sich öffnender und auf der Rückseite ein sich schließender Wirbel zu sehen ist.

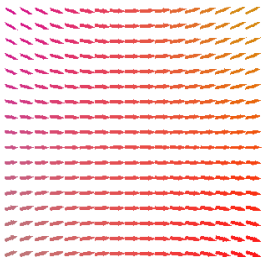
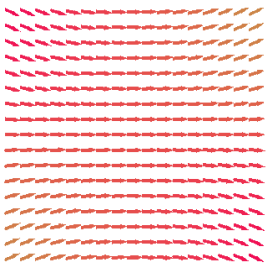
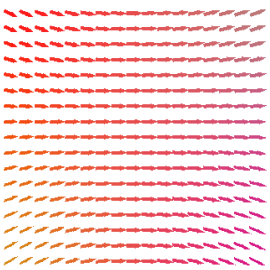
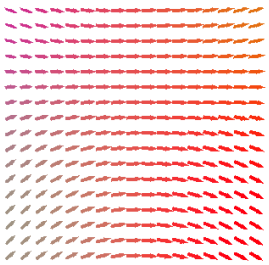
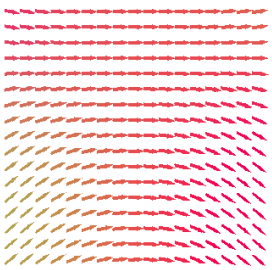
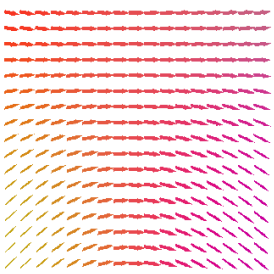
Der limitierende Faktor bei den durchgeführten numerischen Tests ist die Laufzeit. So beträgt diese bei dem gewählten Beispiel ca. 2 Tage. Aus diesem Grund ist es wichtig, schnelle Methoden zu verwenden, wie bereits in Abschnitt 5.4.2 erläutert wurde.

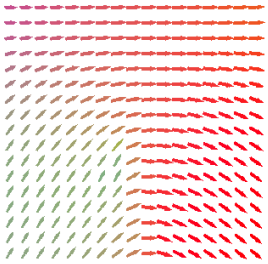
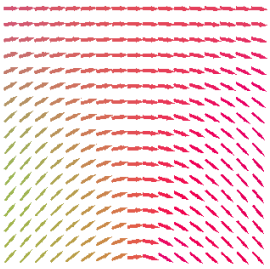
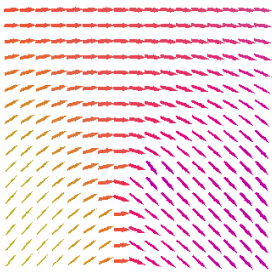
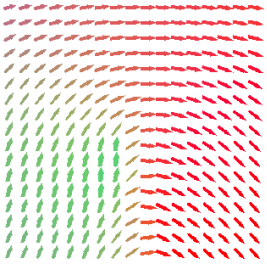
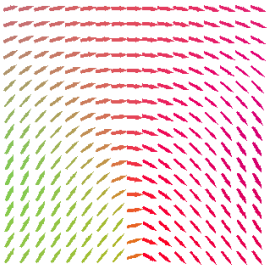
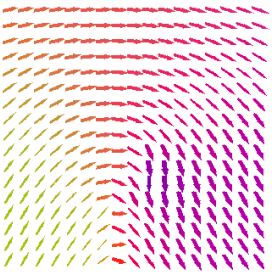
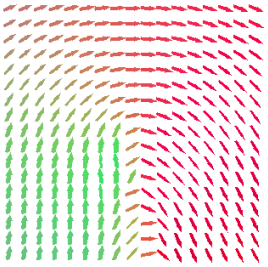
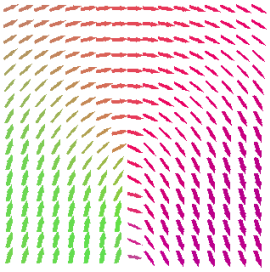
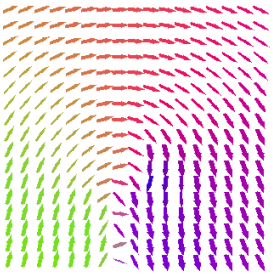
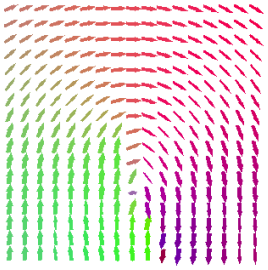
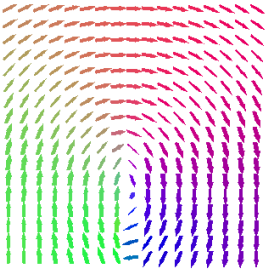
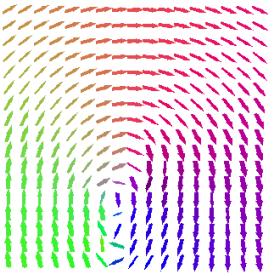
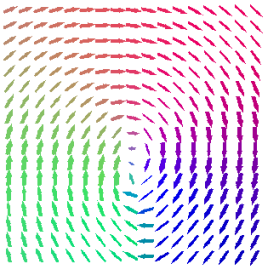
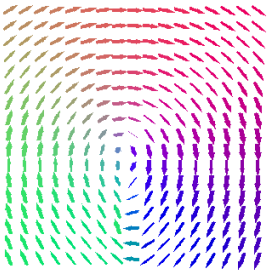
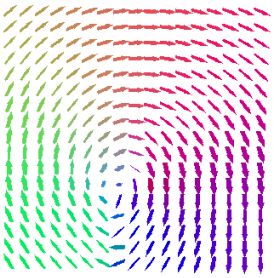
[7] `http://www.opengl.org/`



(a) Flower-Zustand (b) Vortex-Zustand

Abbildung 21: Die zur Initialisierung des MEP verwendeten Energieminima.

Zustand	Magnetisierung des Würfels		
	Vorderseite	Mitte	Rückseite
0			
5			

Zustand	Magnetisierung des Würfels		
	Vorderseite	Mitte	Rückseite
10			
15			
20			
25			
30			

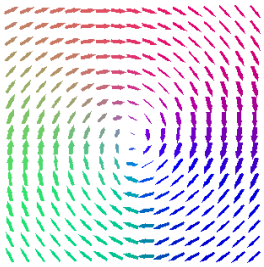
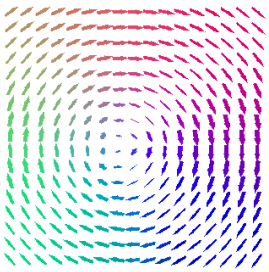
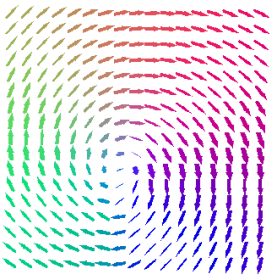
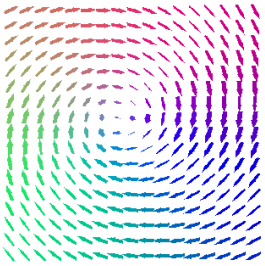
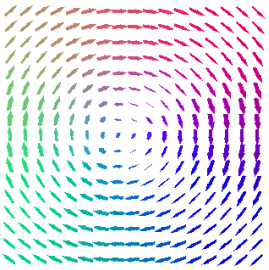
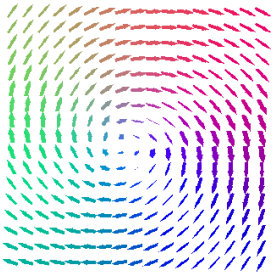
Zustand	Magnetisierung des Würfels		
	Vorderseite	Mitte	Rückseite
35			
40			

Tabelle 6: Zwischenzustände des MEP vom Vortex- zum Flower-Zustand.

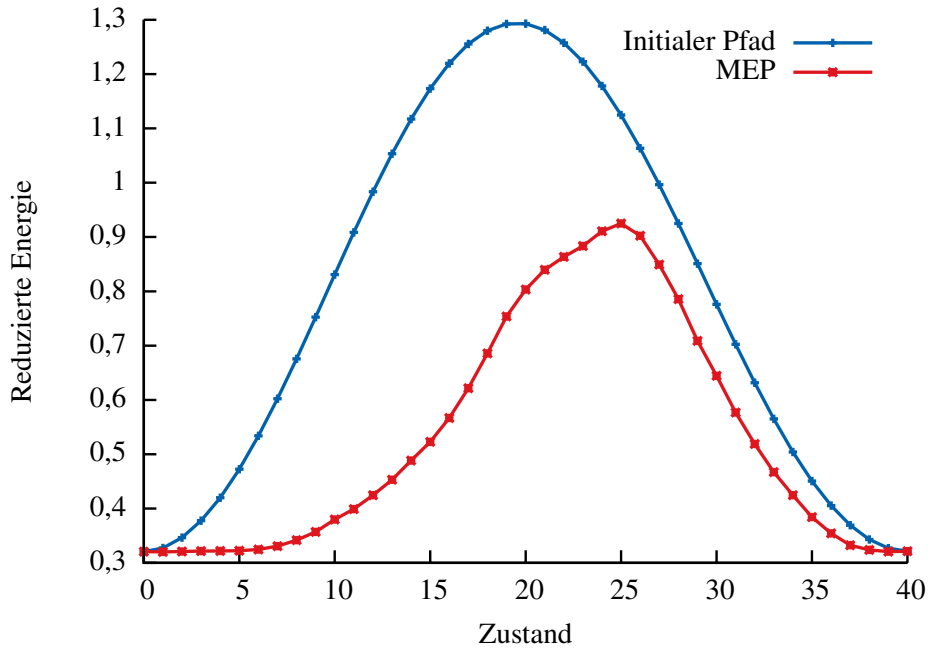


Abbildung 22: Energie der Zwischenkonfigurationen des initialen Pfads und des MEPs vom Flower- (links) zum Vortex-Zustand (rechts).

6 Ausblick

Die vorgestellte Arbeit lässt sich in verschiedene Richtungen erweitern bzw. auf andere Bereiche übertragen. So bietet es sich zum Beispiel an, Teile von Beweisideen auf andere Problemstellungen zu übertragen oder vorgestellte Techniken wie den Erhalt von Nebenbedingungen für $\mathcal{H}$-Matrizen bei anderen Verfahren zu nutzen.

Die Darstellung des Schurkomplements mit Hilfe des Poincaré-Steklov-Operators in Abschnitt 3.2.1 kann genutzt werden, um die blockweise Approximierbarkeit der hierarchischen LU-Zerlegung einer FE-Steifigkeitsmatrix zu beweisen. Damit wird die asymptotische Entwicklung des blockweisen Ranges in Abschnitt 4.2.2 genauer bestimmt. Zwar existieren bereits Abschätzungen in [Beb08, Theorem 4.31], die die Approximierbarkeit beweisen, jedoch kann davon ausgegangen werden, dass sie nicht scharf sind. Im Wesentlichen beruhen die Beweise auf Abschätzungen der Inversen der Steifigkeitsmatrix und erschweren genauere Aussagen über die Güte der LU-Zerlegung.

Eine Möglichkeit, die Technik des in Kapitel 4 vorgestellten Erhalts von Vektoren auf andere Probleme zu übertragen, ist die Anwendung auf Maxwell-Gleichungen. So ist man zum Beispiel daran interessiert, für eine gegebene Stromquelle j_0 das Vektor-Potential $u \in \left\{ v \in \left[L^2(\Omega)\right]^d : \operatorname{rot} v \in \left[L^2(\Omega)\right]^d \right\}$ für das Problem

$$\begin{aligned} \operatorname{rot} \frac{1}{\mu} \operatorname{rot} u &= j_0 \qquad \text{in } \Omega, \\ u \times \nu &= 0 \qquad \text{auf } \partial\Omega, \end{aligned}$$

zu berechnen, wobei $\mu \in L^\infty(\Omega)$ die magnetische Permeabilität und ν die äußere Normale auf Ω bezeichnet. Experimente zeigen, dass ein naiver Erhalt der blockweisen Zeilen- und Spaltensumme, wie beim Laplace-Problem, zu keiner Verbesserung der Iterationszahlen führt. Es muss also untersucht werden, inwiefern sich die Norm des entsprechenden Schurkomplements verhält und welche Basis die Inverse der Systemmatrix gut approximiert.

Es besteht weiterhin die Möglichkeit, gewisse Ergebnisse des theoretischen Teils dieser Arbeit zu verbessern. So kann zum Beispiel versucht werden, durch veränderte Beweistechniken in Abschnitt 3.4 die Annahme (3.14) an die Greensche Funktion zu umgehen und auf bekannte Abschätzungen für dessen Abklingverhalten wie aus [GW82, DM95, HK07] zurückzugreifen. Desweiterhin erscheint es sinnvoll, den Raum

$$H^1(\Omega, \mathcal{L}) := \{ u \in H^1(\Omega) : \mathcal{L}u \in L^2(\Omega) \}$$

für die Theorie in Kapitel 1.2.2 zu verwenden, so dass man nicht auf H^2-Regularität angewiesen ist. Wie in [Cos88] gezeigt wurde, übertragen sich die meisten unserer Aussagen aus diesem Abschnitt direkt. Leider sind dem Autor jedoch keine Existenzaussagen für die Lösung des Randwertproblems (1.2) bekannt.

Die Betrachtungen aus Kapitel 5 über die mikromagnetische Energie können auf zeitabhängige Probleme übertragen werden. Hierfür verwendet man in der Regel die Landau-Lifschitz-Gilbert-Gleichungen (LLG), vgl. [HS98], in der Form

$$\partial_t \boldsymbol{m} + \lambda \boldsymbol{m} \times \partial_t \boldsymbol{m} = -\mu \boldsymbol{m} \times \frac{\partial E[\boldsymbol{m}]}{\partial \boldsymbol{m}}$$

mit $\lambda > 0$ und $\mu = 1 + \lambda^2$, wobei $E[\boldsymbol{m}]$ mit dem Energiefunktional (5.2) aus dem Landau-Lifschitz-Modell übereinstimmt. In diesem Fall können große Teile der Implementierung für den stationären Fall übernommen werden. Geeignete Minimierungsverfahren für die LLG-Gleichungen werden in [AJ06, BKP08] vorgestellt und analysiert.

Literatur

[ACDP04] Alouges, F. ; Conti, S. ; DeSimone, A. ; Pokern, Y.: Energetics and Switching of Quasi-Uniform States in Small Ferromagnetic Particles. In: *Mathematical Modelling and Numerical Analysis* 38 (2004), S. 235–248

[Ada75] Adams, R. A.: *Sobolev Spaces*. Academic Press, 1975

[AJ06] Alouges, F. ; Jaisson, P.: Convergence of a Finite Element Discretization for the Landau-Lifshitz Equations in Micromagnetism. In: *Mathematical Models and Methods in Applied Sciences* 2 (2006), S. 299–316

[Alo97] Alouges, F.: A new algorithm for computing liquid crystal stable configurations: the harmonic mapping case. In: *SIAM J. Numer. Anal.* 34 (1997), S. 1708–1726

[Alo01] Alouges, F.: Computation of the demagnetizing potential in micromagnetics using a coupled finite and infinite elements method. In: *ESAIM Control Optim. Calc. Var.* 6 (2001), S. 629–647

[Axe94] Axelsson, O.: *Iterative Solution Methods*. Cambridge University Press, 1994

[Bar05] Bartels, S.: Stability and convergence of finite-element approximation schemes for harmonic maps. In: *SIAM J. Numer. Anal.* 43 (2005), S. 220–238

[BBBa] Bartels, S. ; Bebendorf, M. ; Bratsch, M.: *A fast and accurate numerical method for the computation of unstable micromagnetic configurations*. – geplante Veröffentlichung in 2012

[BBBb] Bebendorf, M. ; Bollhöfer, M. ; Bratsch, M.: *On the Spectral Equivalence of Hierarchical Matrix Preconditioners for Elliptic Problems*. – geplante Veröffentlichung in 2012

[BBB11] Bebendorf, M. ; Bollhöfer, M. ; Bratsch, M.: Hierarchical Matrix Approximation with Blockwise Constraints / SFB 611. University of Bonn, April 2011 (494). – Forschungsbericht

[BC07] Bebendorf, M. ; Chen, Y.: Efficient solution of nonlinear elliptic problems using hierarchical matrices with Broyden updates. In: *Computing* 81 (2007), S. 239–257

[Beb07] Bebendorf, M.: Why finite element discretizations can be factored by triangular hierarchical matrices. In: *SIAM J. Num. Anal.* 45 (2007), Nr. 4, S. 1472–1494

[Beb08] Bebendorf, M.: *Lecture Notes in Computational Science and Engineering (LNCSE)*. Bd. 63: *Hierarchical Matrices: A Means to Efficiently Solve Elliptic Boundary Value Problems*. Springer-Verlag, 2008. – ISBN 978-3-540-77146-3

[BF11] Bebendorf, M. ; Fischer, T.: On the purely algebraic data-sparse approximation of the inverse and the triangular factors of sparse matrices. In: *Num. Lin. Alg. Appl.* 18 (2011), S. 105–122

[BH86] Barnes, J. ; Hut, P.: A hierarchical $\mathcal{O}(N \log N)$ force calculation algorithm. In: *Nature* 324 (1986), S. 446–449

[BH03] Bebendorf, M. ; Hackbusch, W.: Existence of H-matrix approximants to the inverse FE-matrix of elliptic operators with L-coefficients. In: *Numer. Math.* 95 (2003), S. 1–28

[BKP08] Bartels, S. ; Ko, J. ; Prohl, A.: Numerical analysis of an explicit approximation scheme for the Landau-Lifshitz-Gilbert equation. In: *Math. Comp.* 77 (2008), S. 773–788

[Bra07] Braess, D.: *Finite Elemente*. Springer, 2007

[Bre99] Brenner, S. C.: The condition number of the Schur complement in domain decomposition. In: *Numer. Math. 83* 2 (1999), S. 187–203

[BS02] Brenner, S. C. ; Scott, L. R.: *The Mathematical Theory of Finite Element Methods*. Springer, 2002

[BVV12] Brezina, M. ; Vaněk, P. ; Vassilevski, P. S.: An improved convergence analysis of smoothed aggregation algebraic multigrid. In: *Numer. Linear Algebra Appl.* 19 (2012), S. 441–469

[Cos88] Costabel, M.: Boundary integral operators on Lipschitz domains: elementary results. In: *SIAM J. Math. Anal.* 19 (1988), S. 613–626

[CPS11] Chernov, A. ; Petersdorff, T. von ; Schwab, C.: Exponential convergence of hp quadrature for integral operators with Gevrey kernels. In: *ESAIM Math. Model. Numer. Anal.* 45 (2011), S. 387–422

[DM95] Dolzmann, G. ; Müller, S.: Estimates for Green's matrices of elliptic systems by L^p theory. In: *manuscripta math.* 88 (1995), S. 261–273

[Duf82] Duffy, M.: Quadrature over a pyramid or cube of integrands with a singularity at a vertex. In: *SIAM J. Numer. Anal.* 19 (1982), S. 1260–1262

[ERVE02] E, W. ; Ren, W. ; Vanden-Eijnden, E.: String method for the study of rare events. In: *Physical Review B* 66 (2002), S. 052301 1–4

[ERVE03] E, W. ; Ren, W. ; Vanden-Eijnden, E.: Energy landscape and thermally activated switching of submicron-sized ferromagnetic elements. In: *Journal of Applied Physics* 93 (2003), S. 2275–2282

[ERVE07] E, W. ; Ren, W. ; Vanden-Eijnden, E.: Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. In: *The Journal of Chemical Physics* 126 (2007), S. 164103 1–8

[GC07] García-Cervera, J. M.: Numerical Micromagnetics: A Review. In: *Bol. Soc. Esp. Matem. Apl.* 39 (2007), S. 103–135

[GH03] Grasedyck, L. ; Hackbusch, W.: Construction and arithmetics of H-matrices. In: *Computing 70* 4 (2003), S. 295–334

[Gie01] Giebermann, K.: Multilevel Approximation of Boundary Integral Operators. In: *Computing* 67 (2001), S. 183–207

[GR87] Greengard, L. F. ; Rokhlin, V.: A fast algorithm for particle simulations. In: *J. Comput. Phys.* 2 (1987), S. 325–348

[GR97] Greengard, L. F. ; Rokhlin, V.: A new version of the fast multipole method for the Laplace equation in three dimensions. In: *Acta numerica, 1997*. Cambridge Univ. Press, 1997

[Gre97] Greenbaum, A.: *Iterative Methods for Solving Linear Systems*. Frontiers in Applied Mathematics, 17, SIAM, 1997

[GT83] Gilbarg, D. ; Trudinger, N. S.: *Elliptic Partial Differential Equations of Second Order*. Springer-Verlag, 1983

[Gus78] Gustafsson, I.: A Class of First Order Factorization Methods. In: *BIT* 18 (1978), S. 142–156

[GW82] Grüter, M. ; Widman, K.-O.: The Green Function for uniformly elliptic equations. In: *manuscripta math.* 37 (1982), S. 303–342

[Hac86] Hackbusch, W.: *Theorie und Numerik elliptischer Differentialgleichungen*. B.G. Teubner, 1986

[Hac99] Hackbusch, W.: A sparse matrix arithmetic based on $\mathcal{H}$-matrices. Part I: Introduction to $\mathcal{H}$-matrices. In: *Computing* 62 (1999), S. 89–108

[Hac09] Hackbusch, W.: *Hierarchische Matrizen: Algorithmen und Analysis*. Springer, 2009

[HB06] Hanke-Bourgeois, M.: *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. B. G. Teubner, 2006

[HK00] Hackbusch, W. ; Khoromskij, B. N.: A sparse $\mathcal{H}$-matrix arithmetic. Part II: Application to multi-dimensional problems. In: *Computing* 64 (2000), S. 21–47

[HK07] Hofmann, S. ; Kim, S.: The Green function estimates for strongly elliptic systems of second order. In: *manuscripta math.* 124 (2007), S. 139–172

[HN89] Hackbusch, W. ; Nowak, Z. P.: On the fast matrix multiplication in the boundary element method by panel clustering. In: *Numer. Math.* 54 (1989), S. 463–491

[HS52] Hestenes, M. R. ; Stiefel, E.: Methods of Conjugate Gradients for Solving Linear Systems. In: *Journal of Research of the National Bureau of Standards* 49 (1952), S. 409 – 436

[HS98] Huber, A. ; Schäfer, R.: *Magnetic Domains*. Springer, 1998

[HS09] Hubert, A. ; Schäfer, R.: *Magnetic Domains – The Analysis of Magnetic Microstructures*. Springer-Verlag, 2009

[Jac75] Jackson, J. D.: *Classical electrodynamics*. John Wiley & Sons Inc, New York, 1975

[JMJ98] Jónsson, H. ; Mills, G. ; Jacobsen, K. W.: *Classical and Quantum Dynamics in Condensed Phase Simulations*. World Scientific, Singapore, 1998

[KW04] Khoromskij, B. N. ; Wittum, G.: *Numerical Solution of Elliptic Differential Equations by Reduction to the Interface*. Springer-Verlag, 2004

[LL35] Landau, L. ; Lifshitz, E.: On the theory of the dispersion of magnetic permeability in ferromagnetic bodies. In: *Phys. Zeitsch. der Sow.* 8 (1935), S. 153–169

[McM08] McMichael, R.D.: Standard problem number 3 - problem specification and reported solutions. In: *http://www.ctcms.nist.gov/∼rdm/mumag.html* (2008)

[Meu99] Meurant, G.: *Computer Solution of Large Linear Systems*. Elsevier, 1999

[Mir60] Mirsky, L.: Symmetric Gauge Functions and Unitarily Invariant Norms. In: *Quart. J. Math. Oxford (2)* 11 (1960), S. 50–59

[MV77] Meijerink, J. A. ; Vorst, H. A. d.: An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix. In: *Math. Comp.* 31 (1977), S. 148–162

[Neč67] Nečas, J.: *Les méthodes directes en théorie des équations elliptiques*. Masson, Paris, 1967

[PP05] Popović, N. ; Praetorius, D.: Applications of H-matrix techniques in micromagnetics. In: *Computing 74* 3 (2005), S. 177–204

[QSS01] Quarteroni, A. ; Sacco, R. ; Saleri, F.: *Numerische Mathematik 1*. Springer, 2001

[Rit08] Ritz, W.: Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik. In: *Journal für die reine und angewandte Mathematik* 135 (1908), S. 1–61

[Saa03] Saad, Y.: *Iterative methods for sparse linear systems. Second Edition.* Society for Industrial and Applied Mathematics, Philadelphia, 2003

[Sch08] Schlömerkemper, A.: About solutions of Poisson's equation with transition condition in non-smooth domains. In: *Z. Anal. Anwend. 27* 3 (2008), S. 253–281

[SF73] Strang, G. ; Fix, G. J.: *An analysis of the finite element method*. Prentice-Hall, 1973

[Sha03] Shapira, Y.: *Matrix-Based Multigrid: Theory and Applications*. Kluwer Academic Publishers, 2003

[SS11] Sauter, S. ; Schwab, C.: *Boundary element methods*. Springer-Verlag, 2011

[Ste03] Steinbach, O.: *Numerische Näherungsverfahren für elliptische Randwertprobleme*. Teubner, 2003

[SW92] Schwab, C. ; Wendland, W. L.: On numerical cubatures of singular surface integrals in boundary element methods. In: *Numerische Mathematik* 62 (1992), S. 343–369

[TOS00] Trottenberg, U. ; Oosterlee, C.W. ; Schüller, A.: *Multigrid: Basics, Parallelism and Adaptivity*. Academic Press, 2000

[Tyr98] Tyrtyshnikov, E. E.: Mosaic-skeleton approximations. In: *Calcolo* 33 (1998), S. 47–57

[Van92] Vaněk, P.: Acceleration of convergence of a two level algorithm by smoothing transfer operator. In: *Appl. Math.* 37 (1992), S. 265–274

[Van95] Vaněk, P.: Fast multigrid solver. In: *Appl. Math.* 40 (1995), S. 1–20

[Var60] Varga, R.S.: Factorization and normalized iterative methods. In: *Boundary problems in differential equations*. Univ. of Wisconsin Press, Madison, 1960, S. 121–142

[VBM01] Vaněk, P. ; Brezina, M. ; Mandel, J.: Convergence of algebraic multigrid based on smoothed aggregation. In: *Numer. Math.* 88 (2001), S. 559–579

[Wlo82] Wloka, J.: *Partielle Differentialgleichungen*. B.G. Teubner, 1982