Zweistufige multiplikative überlappende Gebietszerlegungsverfahren

Masterarbeit

im 1-Fach Masterstudiengang Mathematik
der Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Margaretha Franck

Erstgutachter: Prof. Dr. Steffen Börm
Zweitgutachter: Prof. Dr. Malte Braack

Kiel im Mai 2013

Inhaltsverzeichnis

1 Einleitung

Das numerische Lösen von partiellen Differentialgleichungen mit der Finite-Elemente-Methode führt auf große lineare Gleichungssysteme, die effizient gelöst werden sollen. Aufgrund der Größe und der im Allgemeinen schlechten Kondition dieser Gleichungssysteme verursacht die Verwendung von direkten Lösern und auch von einfachen iterativen Verfahren einen hohen Rechenaufwand. Um dieses Problem zu beheben, kann man das Gebiet, auf dem die Differentialgleichung definiert ist, zerlegen und auf Grundlage der entstandenen Teilgebiete Unterräume definieren. Auf diesen Unterräumen können nun aus dem ursprünglichen Problem gewonnene kleinere Gleichungssysteme gelöst werden, wodurch man ein lineares Iterationsverfahren zur Lösung des Problems erhält. Man spricht hierbei auch von einem Unterraumkorrekturverfahren. Diese Arbeit beschäftigt sich mit einer bestimmten Klasse dieses Verfahrens, den sogenannten zweistufigen multiplikativen überlappenden Gebietszerlegungsverfahren. Sie können als eigenständige Iterationsverfahren sowie als Vorkonditionierer für das cg-Verfahren verwendet werden. Es werden ein zweistufiges multiplikatives überlappendes Gebietszerlegungsverfahren und dessen symmetrische Version betrachtet. Ziel ist es, für eine konkrete Gebietszerlegung die Konvergenz nachzuweisen und zu zeigen, dass die Konvergenzrate unabhängig von den Diskretisierungsparametern ist. Für das symmetrische Verfahren wird zudem gezeigt, dass es als Vorkonditionierer für das cg-Verfahren geeignet ist. Die Verfahren werden auch implementiert und die aus einem Testbeispiel gewonnenen numerischen Resultate diskutiert.
Als Beispiel für partielle Differentialgleichungen betrachten wir im Verlauf der Arbeit die Poissongleichung

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega \\ u &= g \text{ auf } \partial\Omega. \end{aligned}$$

Hierbei bezeichnet Ω ein Gebiet im $\mathbb{R}^d$, $d \in \{1, 2, 3\}$ und $\partial\Omega$ dessen Rand, $f : \Omega \to \mathbb{R}$,

$g : \partial\Omega \to \mathbb{R}$ sind gegebene Funktionen und gesucht ist eine Funktion $u : \Omega \to \mathbb{R}$, die die obige Gleichung erfüllt.
Die Arbeit ist wie folgt aufgebaut: In Kapitel 2 werden benötigte Begriffe eingeführt und ein kurzer Überblick über die Finite-Elemente-Methode sowie über die Theorie von iterativen Lösern für lineare Gleichungssysteme gegeben. Das Verfahren der Unterraumkorrektur wird in Kapitel 3 wird vorgestellt und es werden ein allgemeiner Konvergenzbeweis sowie Schranken für das Spektrum des vorkonditionierten cg-Verfahrens geliefert. Die Grundlage dieses Kapitels bildet ein Übersichtsartikel von Yserentant [Yse93]. Kapitel 4 beschäftigt sich mit Gebietszerlegungsverfahren als Unterraumkorrekturverfahren. Es wird eine spezielle Gebietszerlegung vorgestellt und für diese werden die Annahmen, die dem allgemeinen Konvergenzbeweis zugrunde liegen, nachgewiesen. Dieses Kapitel orientiert sich an der Darstellung in der Monographie von Toselli und Widlund [TW04] sowie in der von Smith, Bjørstad und Gropp [SBG96]. Die Implementierung der Verfahren wird in Kaptiel 5 beschrieben. Dieses beinhaltet auch die gewonnenen numerischen Resultate und ihre Diskussion.

2 Grundlagen

In diesem Kapitel werden Begriffe vorgestellt, die im weiteren Verlauf der Arbeit wichtig sind. Außerdem wird ein Überblick über die Grundlagen der Finite-Elemente-Methode und iterativen Lösungsverfahren für lineare Gleichungssysteme gegeben, die im weiteren Verlauf der Arbeit benötigt werden. Wir verzichten hierbei weitestgehend auf Beweise und verweisen an einigen Stelle auf entsprechende Literatur.
In dieser Arbeit sei $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, immer ein beschränktes, polygonalberandetes Gebiet.

2.1 Grundlegende Begriffe und Notationen

In diesem Abschnitt bezeichne V einen endlichdimensionalen $\mathbb{R}$-Vektorraum und $(\cdot, \cdot)$ ein Skalarprodukt auf diesem. Die euklidische Norm im $\mathbb{R}^d$ bezeichenen wir mit $\|\cdot\|_2$.

Definition 2.1. (Durchmesser) *Wir definieren durch* $\operatorname{diam}(\Omega) := \sup\{\|x - y\|_2 : x, y \in \Omega\}$ *den Durchmesser von* Ω.

Definition 2.2. (Abstand) *Für* $x \in \mathbb{R}^d$ *bezeichne* $\operatorname{dist}(x, \Omega) := \inf\{\|x - y\|_2 : y \in \Omega\}$ *den* Abstand *von* x *zu* Ω.
Allgemein sei der Abstand zweier Mengen $N, M \subset \mathbb{R}^d$ *definiert durch* $\operatorname{dist}(M, N) := \inf\{\|x - y\|_2 : x \in M, y \in N\}$.

Lemma 2.3. *Für* $x, y \in \mathbb{R}^d$ *gilt* $|\operatorname{dist}(x, \Omega) - \operatorname{dist}(y, \Omega)| \leq \|x - y\|_2$.

Beweis. Für alle $z \in \Omega$ gilt aufgrund der Dreiecksungleichung

$$\operatorname{dist}(y, \Omega) \leq \|y - z\|_2 = \|y - x + x - z\|_2 \leq \|y - x\|_2 + \|x - z\|_2.$$

Sei nun $\varepsilon > 0$. Dann existiert ein $z \in \Omega$ mit $\|x - z\|_2 \leq \operatorname{dist}(x, \Omega) + \varepsilon$. Zusammen mit

der vorigen Überlegung erhalten wir

$$\operatorname{dist}(y,\Omega) \leq \|y-x\|_2 + \|x-z\|_2 \leq \|y-x\|_2 + \operatorname{dist}(x,\Omega) + \varepsilon,$$

also $\operatorname{dist}(y,\Omega) - \operatorname{dist}(x,\Omega) \leq \|y-x\|_2 + \varepsilon$. Da ε beliebig gewählt war, gilt also $\operatorname{dist}(y,\Omega) - \operatorname{dist}(x,\Omega) \leq \|x-y\|_2$. Ebenso zeigt man $\operatorname{dist}(x,\Omega) - \operatorname{dist}(y,\Omega) \leq \|x-y\|_2$ und erhält daraus die Behauptung. □

Definition 2.4. (Eigenwert) *Für eine lineare Abbildung $A : V \to V$ heißt $\lambda \in \mathbb{C}$* Eigenwert von A, *falls $x \in V\backslash\{0\}$ existiert mit $Ax = \lambda x$. Der Vektor x wird dann als zugehöriger* Eigenvektor *bezeichnet.*

Definition und Lemma 2.5. (Selbstadjungiert) *Eine lineare Abbildung $A : V \to V$ heißt* selbstadjungiert, *falls $(Av, w) = (v, Aw)$ für alle $v, w \in V$ gilt. Für eine darstellende Matrix $\mathbf{A}$ der Abbildung A gilt dann $\mathbf{A} = \mathbf{A}^T$ und sie wird als* symmetrische *Matrix bezeichnet.*

Definition und Satz 2.6. (Positiv definit) *Eine selbstadjungierte lineare Abbildung $A : V \to V$ nennen wir* positiv definit, *falls $(Ax, x) > 0$ für alle $x \in V\backslash\{0\}$ gilt. Als Abkürzung verwenden wir auch die Schreibweise spd für* symmetrisch positiv definit.
A ist genau dann positiv definit, wenn alle Eigenwerte von A reell und positiv sind.

Lemma 2.7. (Energienorm) *Durch eine positiv definite lineare Abbildung A auf V wird durch $(v, w)_A := (Av, w)$ für alle $v, w \in V$ ein Skalarprodukt induziert, das sogenannte* Energieskalarprodukt. *Dieses erzeugt eine Norm auf V, die wir als* Energienorm *bezeichnen: $\|\cdot\|_A := \sqrt{(\cdot,\cdot)_A}$.*

Bemerkung 2.8. (Cauchy-Schwarz-Ungleichung) *Selbstverständlich gilt für das Energieskalarprodukt und die dadurch induzierte Energienorm für alle $x, y \in V$ auch die Cauchy-Schwarz-Ungleichung:*

$$(Ax, y) = (x, y)_A \leq \|x\|_A \|y\|_A = \sqrt{(Ax, x)}\sqrt{(Ay, y)}.$$

Definition 2.9. (Träger) *Für eine Funktion $f : \Omega \to \mathbb{R}$ bezeichnet*

$$\operatorname{supp}(f) := \overline{\{x \in \Omega : f(x) \neq 0\}}$$

den Träger *von f.*

2.2 Finite-Elemente-Methode

Dieser Abschnitt gibt eine Zusammenfassung zur Finite-Elemente-Methode (FEM), die hauptsächlich an [Bra07] und [Ste03] angelehnt ist. Zusätzlich wurde [Bra12] verwendet. Es soll die partielle Differentialgleichung

$$\begin{aligned} Au &= f \text{ in } \Omega \\ u &= 0 \text{ auf } \partial\Omega \end{aligned}$$

gelöst werden, wobei A ein linearer Differentialoperator zweiter Ordnung, das heißt ein linearer Differentialoperator, der nur partielle Ableitungen bis zur zweiten Ordnung enthält, und $f : \Omega \to \mathbb{R}$ eine Funktion ist. Zur Vereinfachung werden Nullrandwerte betrachtet. Dies ist möglich, da ein Problem mit anderen Randwerten in ein Problem mit Nullrandwerten überführt werden kann. Gesucht ist eine hinreichend oft differenzierbare Funktion $u : \Omega \to \mathbb{R}$, die die Differentialgleichung erfüllt. Sie wird auch als *starke* oder *klassische Lösung* bezeichnet.

2.2.1 Schwache Ableitungen und Sobolevräume

Definition und Satz 2.10. (L_2-Raum) *Definiere*

$$L_2(\Omega) := \{f : \Omega \to \mathbb{R} : f \textit{ ist messbar und } f^2 \textit{ ist Lebesque-integrierbar }\}.$$

Zusammen mit dem Skalarprodukt

$$(f, g)_{L_2(\Omega)} := \int_\Omega f(x)g(x)dx$$

bildet $L_2(\Omega)$ einen Hilbertraum. Durch das Skalarprodukt wird die L_2-Norm

$$\|f\|_{L_2(\Omega)} := \left(\int_\Omega f(x)^2 dx\right)^{1/2}$$

induziert.

Hierbei beachte man, dass $L_2(\Omega)$ streng genommen keine Funktionen enthält, sondern Äquivalenzklassen von Funktionen, die sich nur auf einer Nullmenge unterscheiden. Wie üblich und auch schon in der Definition verwendet, werden wir stets einen Repräsentanten

der Äquivalenzklasse betrachten. Allerdings müssen wir beachten, dass keine Punktauswertungen erklärt sind.

Bemerkung 2.11. (Multiindex) *Als* Multiindex *bezeichnen wir ein* $\alpha \in \mathbb{N}_0^d$. *Den Betrag eines Multiindex definieren wir durch* $|\alpha| := \|\alpha\|_1 := \sum_{k=1}^d |\alpha_k|$. *Die Ableitung einer hinreichend oft differenzierbaren Funktion* $\phi : \Omega \to \mathbb{R}$ *bezüglich* α *ist definiert durch* $\partial^\alpha \phi := \sum_{k=1}^d \frac{\partial^{\alpha_k}}{\partial x_k^{\alpha_k}} \phi$.

Definition 2.12. (Schwache Ableitung) *Es sei* $D(\Omega) := \{\phi \in C^\infty(\Omega) : \mathrm{supp}(\phi)$ *kompakt*$\}$ *und* $\alpha \in \mathbb{N}_0^d$ *ein Multiindex. Existiert für ein* $u \in L_2(\Omega)$ *ein* $v \in L_2(\Omega)$ *mit*

$$\int_\Omega v\,\phi = (-1)^{|\alpha|} \int_\Omega u\,\partial^\alpha \phi \text{ für alle } \phi \in D(\Omega),$$

so nennen wir v *die* (schwache) Ableitung *von* u *bezüglich* α.

Wenn eine Funktion aus $L_2(\Omega)$ klassisch differenzierbar ist, so ist sie auch schwach differenzierbar und die Ableitungen stimmen überein, was die Schreibweise $v =: \partial^\alpha u$ für die schwache Ableitung einer Funktion u rechtfertigt.

Definition und Satz 2.13. (Sobolevraum) *Sei* $m \in \mathbb{N}_0$. *Dann ist der* Sobolevraum $H^m(\Omega)$ *definiert durch*

$$H^m(\Omega) := \{u \in L_2(\Omega) : \forall \alpha \in \mathbb{N}_0^d \text{ mit } |\alpha| \le m \text{ existiert die schwache Ableitung } \partial^\alpha u\}.$$

Durch

$$(u, v)_{H^m(\Omega)} := \sum_{|\alpha| \le m} (\partial^\alpha u, \partial^\alpha v)_{L_2(\Omega)}$$

wird ein Skalarprodukt auf $H^m(\Omega)$ *definiert. Dieses induziert die Sobolevnorm*

$$\|u\|_{H^m(\Omega)} := (u, u)_{H^m(\Omega)}^{1/2} = \left(\sum_{|\alpha| \le m} (\partial^\alpha u, \partial^\alpha u)_{L_2(\Omega)} \right)^{1/2}.$$

Durch

$$|u|_{H^m(\Omega)} := \left(\sum_{|\alpha|=m} \|\partial^\alpha u\|^2_{L_2(\Omega)} \right)^{1/2}$$

wird eine Halbnorm auf $H^m(\Omega)$ definiert.
$H^m(\Omega)$ bildet mit dem Skalarprodukt $(\cdot,\cdot)_{H^m(\Omega)}$ und der dadurch induzierten Norm einen Hilbertraum.

Definition 2.14. *($H_0^m(\Omega)$) Mit $H_0^m(\Omega)$ bezeichnen wir die Vervollständigung von $D(\Omega)$ bezüglich der Norm $\|\cdot\|_{H^m(\Omega)}$.*

Satz 2.15. *Auf $H_0^m(\Omega)$ sind die Norm $\|\cdot\|_{H^m(\Omega)}$ und die Halbnorm $|\cdot|_{H^m(\Omega)}$ äquivalent.*

Beweis. Siehe [Bra07] Kap. II, Satz 1.7. □

2.2.2 Variationsformulierung

Sei H ein Hilbertraum mit Skalarprodukt $(\cdot,\cdot)_H$ und der dadurch induzierten Norm $\|\cdot\|_H = \sqrt{(\cdot,\cdot)_H}$. $H' := \{f : H \to \mathbb{R} : f \text{ stetig und linear}\}$ bezeichne den Dualraum von H und $\langle\cdot,\cdot\rangle$ sei das Dualitätsprodukt, also $\langle f, v\rangle = f(v)$ für alle $f \in H', v \in H$. Dann ist durch

$$\|f\|_{H'} := \sup_{v \in H\setminus\{0\}} \frac{|\langle f, v\rangle|}{\|v\|_H} \tag{2.1}$$

eine Norm auf dem Dualraum gegeben.

Satz 2.16. *(*Darstellungssatz von Fréchet-Riesz*) Zu jedem $f \in H'$ existiert genau ein $u \in H$ mit $f(v) = (v,u)_H$ für alle $x \in H$ und es gilt $\|f\|_{H'} = \|u\|_H$.*

Beweis. Siehe [Wer07, Theorem V.3.6]. □

Aufgrund dieser Isomorphie sind Ausdrücke wie $\langle u, f\rangle$ für $u \in H, f \in H'$ erklärt.

Definition 2.17. *Ein Operator $A : H \to H'$ heißt*

- *H-*elliptisch*, falls es eine Konstante $c_1 > 0$ gibt mit $\langle Au, u\rangle \geq c_1\|u\|_H^2$ für alle $u \in H$.*

- beschränkt, *falls es eine Konstante* $c_2 > 0$ *gibt mit* $\|Au\|_{H'} \leq c_2\|u\|_H$ *für alle* $u \in H$.
- selbstadjungiert, *falls* $\langle Au, v\rangle = \langle u, Av\rangle$ *für alle* $u, v \in H$ *gilt.*

Durch $a(u, v) := \langle Au, v\rangle$ definiert A eine Bilinearform auf H. Falls H endlichdimensional ist, entspricht diese Bilinearform dem durch die darstellende Matrix von A erzeugten Energieskalarprodukt.
Die folgende Gleichung wird als *Operatorgleichung* bezeichnet:

$$Au = f, \tag{2.2}$$

wobei $f \in H'$ gegeben ist und $u \in H$ gesucht wird.
Diese Gleichung kann man in ein *variationelles Problem* überführen, indem man mit sogenannten *Testfunktionen* $v \in H$ multipliziert. Man erhält

$$\langle Au, v\rangle = \langle f, v\rangle \text{ für alle } v \in H. \tag{2.3}$$

Lemma 2.18. *Für* $u \in H$ *ist äquivalent: u ist Lösung der Operatorgleichung* (2.2) *und u ist Lösung des variationellen Problems* (2.3).

Beweis. Für einen Beweis siehe [Ste03] Abschnitt 3.1. □

Der folgende Satz trifft eine Aussage über die Lösbarkeit des Problems.

Satz 2.19. (Lax-Milgram) *Sei* $A : H \to H'$ *ein beschränkter, H-elliptischer Operator. Dann besitzt die Operatorgleichung für jedes* $f \in H'$ *eine eindeutig bestimmte Lösung* $u \in H$ *und es gilt*

$$\|u\|_H \leq \frac{1}{c_1}\|f\|_{H'}.$$

Beweis. Siehe [Ste03] Satz 3.2. □

2.2.3 Galerkin-Methode

Wir überführen nun unser ursprüngliches Problem in die variationelle Formulierung, suchen jedoch nur noch eine im schwachen Sinn hinreichend oft differenzierbare Funktion, erhalten also ein Problem im Hilbertraum $H_0^1(\Omega)$. Im Folgenden sei A als beschränkt,

elliptisch und selbstadjungiert angenommen. Mit $a(\cdot,\cdot)$ bezeichnen wir die vom Differentialoperator A definierte Bilinearform auf $H_0^1(\Omega)$. Aufgrund unserer Nullrandbedingung lautet das variationelle Problem:

$$\text{Finde } u \in V \text{ mit } a(u,\phi) = \langle f, \phi\rangle \text{ für alle } \phi \in H_0^1(\Omega).$$

Die Idee des Galerkin-Verfahrens ist es, dieses nicht im unendlichdimensionalen Raum $H_0^1(\Omega)$, sondern in einem endlichdimensionalen Teilraum $V \subset H_0^1(\Omega)$ zu lösen. Dieser ist als endlichdimensionaler Teilraum eines Hilbertraumes selber ein Hilbertraum. Sei $\{\phi_i : i \in I\}$ eine Basis von V, wobei I eine Indexmenge sei mit $|I| = \dim V$. Ebenso wie im unendlichdimensionalen Fall kann man das variationelle Problem formulieren. Dieses lautet:

$$\text{Finde } u_h \in V \text{ mit } a(u_h,\phi_i) = \langle f, \phi_i\rangle \text{ für alle } i \in I. \tag{2.4}$$

Wir bemerken, dass nicht mehr mit allen Funktionen aus V, sondern nur noch mit den Basisfunktionen getestet werden muss – wie schon in (2.4) genutzt – da jede Funktion als Linearkombination dieser dargestellt werden kann. Indem man auch die gesuchte Funktion $u_h \in V$ als Linearkombination der Basisfunktionen, d.h. $u_h = \sum_{j\in I} u_j\,\phi_j$ mit $u_j \in \mathbb{R}$, darstellt, erhält man ein lineares Gleichungssystem

$$\sum_{j\in I} u_j a(\phi_j,\phi_i) = \langle f, \phi_i\rangle \text{ für alle } i \in I$$

oder in Matrixform $\mathbf{Au} = \mathbf{f}$ mit $\mathbf{u} := (u_j)_{j\in I}$, $\mathbf{f} := (\langle f, \phi_i\rangle)_{i\in I}$ und $\mathbf{A} := (a(\phi_j,\phi_i))_{j,i\in I}$. Die Matrix $\mathbf{A}$ nennen wir auch *Steifigkeitsmatrix*.

Der folgende Satz liefert eine Bestapproximationsaussage für die Güte der Näherung von u durch das so bestimmte u_h.

Satz 2.20. (Céa's Lemma) *Für $H-$elliptisches, beschränktes A gelten folgende Aussagen über Stabilität und Fehler*

$$\|u_h\|_{H_0^1(\Omega)} \le \frac{1}{c_1}\|f\|_{H_0^1(\Omega)'}$$
$$\|u - u_h\|_{H_0^1(\Omega)} \le \frac{c_2^2}{c_1}\inf_{v_h\in V}\|u - v_h\|_{H_0^1(\Omega)}.$$

Dabei sind die Konstanten c_1, c_2 die Konstanten aus der $H-$Elliptizität bzw. der Beschränktheit der Bilinearform wie in Definition 2.17.

Beweis. Siehe [Ste03] Satz 8.1. □

Finite-Elemente-Räume

Als endlichdimensionale Teilräume V haben sich die sogenannten Finite-Elemente-Räume als nützlich erwiesen. Zu ihrer Konstruktion wird das der Operatorgleichung zugrundeliegende Gebiet Ω in einfache geometrische Strukturen – wir werden Simplizes betrachten – partitioniert. Der Finite-Elemente-Raum V besteht dann aus stetigen Funktionen, die auf den einzelnen Objekten Polynomfunktionen sind.

Definition 2.21. (Triangulierung) *Eine* Triangulierung *von Ω ist eine Zerlegung $\mathcal{T} = \{\tau_1, \ldots, \tau_m\}$ in Simplizes, d.h. Intervalle ($d = 1$), Dreiecke ($d = 2$) oder Tetraeder ($d = 3$). Mit $\mathcal{N}$ bezeichnen wir die Menge der* Ecken *und mit $\mathcal{E}$ die der* Kanten *einer Triangulierung. Für ein Element $\tau \in \mathcal{T}$ bezeichne ρ_τ den* Innenkreisradius *und h_τ den* Außenkreisradius*; $h := \max_{\tau \in \mathcal{T}} h_\tau$ heißt auch* Maschenweite *von $\mathcal{T}$. Für eine Triangulierung $\mathcal{T}$ mit Maschenweite h schreiben wir an einigen Stellen auch $\mathcal{T}_h$.*

- *Eine Triangulierung $\mathcal{T}$ heißt* zulässig*, falls zwei verschiedene Elemente entweder leeren Schnitt haben oder ihr Schnitt nur einen Knoten ($d = 1, 2, 3$), eine Kante ($d = 2, 3$) oder ein Dreieck ($d = 3$) enthält.*

- *Eine Folge von Triangulierungen $\{\mathcal{T}_h\}_h$ heißt* quasi-uniform*, wenn ein $\kappa > 0$ existiert, so dass*

$$h_\tau \leq \kappa \rho_\tau \text{ für alle } \tau \in \mathcal{T}_h, \text{ für alle } \mathcal{T}_h \in \{\mathcal{T}_h\}_h$$

  gilt.

- *Wir bezeichnen eine zulässige Triangulierung $\mathcal{T}_h$ als* Verfeinerung *einer ebenfalls zulässigen Triangulierung $\mathcal{T}_H$, falls sie durch Unterteilung der Elemente in $\mathcal{T}_H$ entstanden ist.*

Anstelle von Triangulierung sprechen wir auch von Gittern. Im Folgenden betrachten wir nur quasi-uniforme Folgen von zulässigen Triangulierungen.

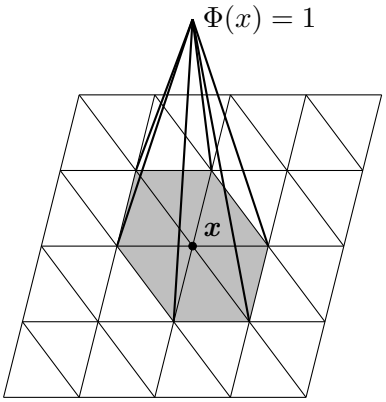

Abbildung 2.1: Hutfunktion Φ zum Knoten x und ihr Träger in 2D.

In dieser Arbeit beschränken wir uns auf die Behandlung von linearen Finiten-Elementen, die im Folgenden definiert werden. Auf gleiche Weise kann man auch quadratische, kubische etc. Finite-Elemente definieren.

Definition 2.22. (Lineare Finite-Elemente) *Zu einer Triangulierung $\mathcal{T}_h$ sei* $V_h := \{u \in \mathcal{C}(\overline{\Omega}) : u\big|_\tau \in P_1(\tau)$ *für alle* $\tau \in \mathcal{T}_h\}$.

Hierbei bezeichnet $P_1(\tau)$ den Polynomraum der linearen Polynome auf τ.
Bei linearen Finiten-Elementen ist eine Funktion bereits durch ihre Funktionswerte an den Eckpunkten der Triangulierungselemente bestimmt, wodurch die Wahl der folgenden Basis naheliegt.

Definition und Satz 2.23. (Knotenbasis) *Es sei V_h ein Finite-Elemente-Raum und $\mathcal{N}\backslash\partial\Omega = \{x_1, \ldots, x_n\}$ die Menge der* inneren Knoten *der zugehörigen Triangulierung. Für $i \in \{1, \ldots, n\}$ definieren wir $\Phi_i \in V_h$ durch $\Phi_i(x_j) = \delta_{ij}$ für alle $x_j \in \mathcal{N}\backslash\partial\Omega$. Dann bildet $\{\Phi_1, \ldots, \Phi_n\}$ eine Basis von V_h, die wir als* Knotenbasis *bezeichnen.*

Beweis. Vergleiche [Bra12] Abschnitt 5.7. □

Die Elemente der Knotenbasis haben nur lokalen Träger – diejenigen Elemente der Triangulierung, die den entsprechenden Knoten enthalten – und werden aufgrund ihres Aussehens auch „Hutfunktionen“ (vgl. Abb. 2.1) genannt. Wegen der lokalen Träger ist die entstehende Steifigkeitsmatrix $\mathbf{A}$ schwachbesetzt. Für das Poissonproblem ergibt sich

die Steifigkeitsmatrix $\mathbf{A}_P$ durch

$$(\mathbf{A}_P)_{i,j} = \int_\Omega \nabla\Phi_i(x)\nabla\Phi_j(x)\,dx.$$

Auf einem gleichmäßigen Gitter, d.h. jedes Element hat den gleichen Umkreisradius h, ist ihre Konditionszahl ist gegeben durch $\kappa(\mathbf{A}_P) = \mathcal{O}(h^{-2})$.

Abschätzungen für lineare Finite-Elemente

Lemma 2.24. (Lokale inverse Ungleichung) *Für $v \in V_h$ und $\tau \in \mathcal{T}_h$ gilt die* lokale inverse Ungleichung

$$|v|_{H^1(\tau)} \leq ch_\tau^{-1}\|v\|_{L_2(\tau)}$$

mit einer von h unabhängigen, positiven Konstanten c.

Beweis. Siehe [Ste03] Lemma 9.4. □

Definition 2.25. (L_2-Projektion) *Die L_2-*Projektion *$Q_h : L_2(\Omega) \to V_h$, $u \mapsto Q_h u$ ist definiert als eindeutige Lösung von*

$$(Q_h u, v)_{L_2(\Omega)} = (u, v)_{L_2(\Omega)} \text{ für alle } v \in V_h.$$

Satz 2.26. (H^1-Stabilität der L_2-Projektion) *Bei einer quasi-uniformen Triangulierung $\mathcal{T}_h$ und einem linearen Finite-Elemente-Raum gilt*

$$|Q_h v|_{H^1(\Omega)} \leq c|v|_{H^1(\Omega)} \text{ für } v \in H_0^1(\Omega)$$
$$\|v - Q_h v\|_{L_2(\Omega)} \leq c'h|v|_{H^1(\Omega)} \text{ für } v \in H_0^1(\Omega)$$

mit von h unabhängigen Konstanten c, c'.

Beweis. Siehe [BX91] Theorem 3.2 und Theorem 3.4. □

Bemerkung 2.27. *Aus dem vorherigen Satz folgt sofort durch Anwendung der Dreiecksungleichung, dass es eine Konstante c gibt mit $|v - Q_h v|_{H^1(\Omega)} \leq c|v|_{H^1(\Omega)}$ für alle $v \in H_0^1(\Omega)$.*

Definition 2.28. (Lineare Interpolierende) *Für eine stetige Funktion $v \in \mathcal{C}(\Omega)$ auf einem Gitter $\mathcal{T}$ mit n inneren Knoten $\{x_1, \ldots, x_n\}$ wird durch*

$$I_h[v](x) := \sum_{k=1}^{n} v(x_k)\Phi_k(x)$$

die Interpolierende im Raum der stetigen, stückweise linearen Funktionen definiert, wobei $\{\Phi_k : k = 1, \ldots, n\}$ die Knotenbasis sei.

Lemma 2.29. *Sei u_h eine stetige, stückweise quadratische Funktion, die auf dem Gitter $\mathcal{T}$ definiert ist. Dann gibt es eine von h unabhängige Konstante C, so dass*

$$|I_h[u_h]|_{H^1(\tau)} \leq C|u_h|_{H^1(\tau)} \text{ für alle } \tau \in \mathcal{T}$$

gilt.

Beweis. Siehe [TW04] Lemma 3.9. □

2.3 Iterative Löser für lineare Gleichungssysteme

Dieser Abschnitt behandelt allgemeine lineare Iterationsverfahren sowie das cg-Verfahren zur Lösung eines linearen Gleichungssystems $Au = f$ mit einer regulären Matrix $A \in \mathbb{R}^{I\times I}$, $u, f \in \mathbb{R}^I$, wobei I eine endliche Indexmenge sei. Wir orientieren uns bei der Darstellung an [Bör11]. Es bezeichne $(\cdot,\cdot)_2$ das euklidische Skalarprodukt auf $\mathbb{R}^I$ und $\|\cdot\|_A$ die Energienorm zu A.

2.3.1 Lineare Iterationsverfahren

Bei den zu betrachtenden Gebietszerlegungsverfahren handelt es sich um lineare Iterationsverfahren, weshalb im Folgenden eine kurze Zusammenfassung über diese Verfahren gegeben wird.

Definition 2.30. (Iterationsverfahren) *Eine Abbildung $\Phi : \mathbb{R}^I \times \mathbb{R}^I \to \mathbb{R}^I$, die im ersten Argument stetig ist, heißt* Iterationsverfahren. *Für gegebenes $f, u^{(0)} \in \mathbb{R}^I$ heißt $(u^{(m)})_{m\in\mathbb{N}_0}$ mit $u^{(m)} := \Phi(u^{(m-1)}, f)$ für alle $m \in \mathbb{N}$ die Folge der Iterierten.*

Definition und Lemma 2.31. (Konvergenz und Konsistenz) *Ein Iterationsverfahren Φ heißt*

- konvergent, *falls für alle* $f \in \mathbb{R}^I$ *ein* $\hat{u} \in \mathbb{R}^I$ *existiert, so dass für alle* $u^{(0)} \in \mathbb{R}^I$ *für die Folge der Iterierten* $(u^{(m)})_{m\in\mathbb{N}_0}$ *gilt* $\lim_{m\to\infty} u^{(m)} = \hat{u}$.
- konsistent, *falls die Lösung* u^* *von* $Au = f$ *für jedes* $f \in \mathbb{R}^I$ *ein Fixpunkt zu* Φ *und* f *ist, d.h.* $u^* = \Phi(u^*, f)$.

Für ein konsistentes und konvergentes Iterationsverfahren Φ *und eine rechte Seite* f *konvergiert die Folge der Iterierten zu einem beliebigen Startwert gegen die Lösung des linearen Gleichungssystems.*

Definition 2.32. (Lineares Iterationsverfahren) *Ein Iterationsverfahren* Φ *heißt* linear, *falls Matrizen* $M, B \in \mathbb{R}^{I\times I}$ *mit* $\Phi(u, f) = Mu + Bf$ *für alle* $u, f \in \mathbb{R}^I$ *existieren.*

Lemma 2.33. *Ein lineares Iterationsverfahren ist genau dann konsistent, wenn* $M = I - BA$ *ist. Dann hat das Iterationsverfahren die Form* $\Phi(u, f) = u + B(f - Au)$.

Satz 2.34. *Ein lineares Iterationsverfahren, das durch die Matrizen* M, B *gegeben ist, ist genau dann konvergent, falls für den Spektralradius von* M *gilt* $\rho(M) < 1$.

Beweis. Siehe [Bör11] Satz 2.24. □

2.3.2 cg-Verfahren

Das Verfahren der *konjugierten Gradienten* (engl.: *c*onjugated *g*radients), kurz cg-Verfahren, ist ein semiiteratives Verfahren zur Lösung von positiv definiten Gleichungssystemen. Es gehört zur Klasse der Krylow-Unterraum-Verfahren. Wir stellen es hier vor, da wir ein Gebietszerlegungsverfahren als Vorkonditionierer für dieses verwenden.

Definition 2.35. (cg-Verfahren) *Sei* $A \in \mathbb{R}^{I\times I}$ *positiv definit und* $u^{(0)}, f \in \mathbb{R}^I$. *Die für alle* $m \in \mathbb{N}_0$ *durch*

$$
\begin{aligned}
r^{(m)} &:= f - Au^{(m)} \\
p^{(m)} &:= \begin{cases} r^{(0)} & \text{falls } m = 0 \\ r^{(m)} - \frac{(r^{(m)}, Ap^{(m-1)})_2}{(p^{(m-1)}, Ap^{(m-1)})_2} p^{(m-1)} & \text{falls } m > 0 \text{ und } p^{(m-1)} \neq 0 \\ 0 & \text{sonst} \end{cases} \\
u^{(m+1)} &:= u^{(m)} + \frac{(p^{(m)}, r^{(m)})_2}{(p^{(m)}, Ap^{(m)})_2} p^{(m)}
\end{aligned}
$$

definierte Folge $(u^{(m)})_{m\in\mathbb{N}_0}$ *bezeichnen wir als die Folge der Semiiterierten des cg-Verfahrens.*

Satz 2.36. (Konvergenz) *Sei* $A \in \mathbb{R}^{I\times I}$ *positiv definit,* $\alpha, \beta \in \mathbb{R}_{>0}$ *mit* $\sigma(A) \subset [\alpha, \beta]$, $f \in \mathbb{R}^I$, $u^* = A^{-1}f$. *Die Folge der Semiiterierten des cg-Verfahrens zum Startvektor* $u^{(0)} \in \mathbb{R}^I$ *sei gegeben durch* $(u^{(m)})_{m\in\mathbb{N}_0}$. *Dann gilt*

$$\|u^{(m)} - u^*\|_A \leq \frac{2c^m}{1+c^{2m}}\|u^{(0)} - u^*\|_A$$

mit $c := \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$, *wobei die Konstante* κ *gegeben ist durch* $\kappa := \frac{\beta}{\alpha}$.

Beweis. Vergleiche [Bör11] Satz 3.26. □

Durch Vorkonditionierung mit einer positiv definiten Matrix B kann das cg-Verfahren beschleunigt werden.

Definition 2.37. (Vorkonditioniertes cg-Verfahren) *Seien* $A, B \in \mathbb{R}^{I\times I}$ *positiv definit und* $u^{(0)}, f \in \mathbb{R}^I$. *Die für alle* $m \in \mathbb{N}_0$ *durch*

$$\begin{aligned}
r^{(m)} &:= f - Au^{(m)} \\
q^{(m)} &:= Br^{(m)} \\
p^{(m)} &:= \begin{cases} q^{(0)} & \text{falls } m = 0 \\ q^{(m)} - \frac{(q^{(m)}, Ap^{(m-1)})_2}{(p^{(m-1)}, Ap^{(m-1)})_2} p^{(m-1)} & \text{falls } m > 0 \text{ und } p^{(m-1)} \neq 0 \\ 0 & \text{sonst} \end{cases} \\
u^{(m+1)} &:= u^{(m)} + \frac{(p^{(m)}, r^{(m)})_2}{(p^{(m)}, Ap^{(m)})_2} p^{(m)}
\end{aligned}$$

definierte Folge $(u^{(m)})_{m\in\mathbb{N}_0}$ *bezeichnen wir als die Folge der Semiiterierten des vorkonditionierten cg-Verfahrens.*

Satz 2.38. (Konvergenz) *Seien* $A, B \in \mathbb{R}^{I\times I}$ *positiv definit* $\alpha, \beta \in \mathbb{R}_{>0}$ *mit* $\sigma(BA) \subset [\alpha, \beta]$ *und* $f \in \mathbb{R}^I$, $u^* = A^{-1}f$. *Die Folge der Semiiterierten des cg-Verfahrens zum Startvektor* $u^{(0)} \in \mathbb{R}^I$ *sei gegeben durch* $(u^{(m)})_{m\in\mathbb{N}_0}$. *Dann gilt*

$$\|u^{(m)} - u^*\|_A \leq \frac{2c^m}{1+c^{2m}}\|u^{(0)} - u^*\|_A$$

mit $c := \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$*, wobei die Konstante* κ *durch* $\kappa := \frac{\beta}{\alpha}$ *gegeben ist.*

Beweis. Siehe [Bör11] Satz 3.28. □

Falls also die Konstante κ der vorkonditionierten Matrix BA kleiner ist als die Konstante κ der Matrix A selber, ist die Konvergenzgeschwindigkeit des vorkonditionierten Verfahrens größer als die des urspünglichen cg-Verfahrens.

3 Konvergenztheorie

Mit Hilfe der Finite-Elemente-Methode können wir das Lösen einer partiellen Differentialgleichung – zumindest approximativ – auf das Lösen eines linearen Gleichungssystems zurückführen. Wir möchten also das Problem

$$Au = f \text{ mit } u \in V,\ f \in V' \tag{3.1}$$

lösen, wobei V ein endlichdimensionaler Funktionenraum, V' sein Dualraum und $A : V \to V'$ ein selbstadjungierter, elliptischer linearer Operator sei. Mit $(\cdot, \cdot)$ bezeichnen wir das Dualitätsprodukt. Das Energieskalarprodukt zu A ist durch die Bilinearform $a(\cdot, \cdot)$ gegeben, die dadurch induzierte Norm sei $\|\cdot\|$.

Bemerkung 3.1. *(*Adjungierte*) Für eine lineare Abbildung $X : V \to V$ bezeichne X^* die Adjungierte zu X bezüglich des Energieskalarprodukts. Das heißt, es gilt $a(Xv, w) = a(v, X^*w)$ für alle $v, w \in V$.*

In unserer Anwendung ist $V = V_h$ versehen mit dem L_2-Skalarprodukt.

3.1 Unterraumkorrekturverfahren

Seien $W_l \subset V$, $l \in \{0, \ldots, J\}$, Untervektorräume mit $V = \sum_{l=0}^{J} W_l$, wobei die Summe keine direkte Summe sein muss. Die Idee von Unterraumkorrekturverfahren ist es, eine approximative Lösung $\tilde{u}$ von (3.1) zu betrachten und den Fehler $e = u^* - \tilde{u}$ zwischen exakter Lösung u^* und approximierter Lösung auf einem der Unterräume W_l zu bestimmen, d.h. einen Fehler e_l zu ermitteln. Da $u^* = \tilde{u} + e$ gilt, besteht die Hoffnung, dass $\tilde{u}_{neu} := \tilde{u} + e_l$ eine bessere Approximation der Lösung darstellt als $\tilde{u}$.
Um diese Idee exakter zu fassen, definieren wir zu jedem Untervektorraum orthogonale Projektionen bezüglich des Energieskalarprodukts:

Definition 3.2. *Definiere für alle* $l \in \{0, \ldots, J\}$

$$P_l : V \to W_l \text{ durch } a(P_l u, w_l) = a(u, w_l) \text{ für alle } u \in V, w_l \in W_l \tag{3.2}$$

a-orthogonale Projektionen auf die Unterräume.

Außerdem sei die Einbettung des Dualraums von V in den Dualraum eines Untervektorraums W_l gegeben durch:

Definition 3.3.

$$Q_l : V' \to W_l' \text{ durch } (Q_l u, w_l) = (u, w_l) \text{ für alle } u \in V', w_l \in W_l.$$

Zudem sei eine Einschränkung des Operators A auf W_l gegeben durch

Definition 3.4.

$$A_l : W_l \to W_l', u_l \mapsto Q_l A u_l.$$

Dann gilt insbesondere $(A_l u_l, w_l) = (Q_l A u_l, w_l) = (A u_l, w_l)$ für alle $u_l, w_l \in W_l$. Da A positiv definit ist, ist auch A_l positiv definit.
Wir bemerken, dass $A_l P_l = Q_l A$ für alle $l \in \{0, \ldots, J\}$ ist, denn für alle $u \in V, w_l \in W_l$ gilt: $(A_l P_l u, w_l) = (A P_l u, w_l) = a(P_l u, w_l) = a(u, w_l) = (Au, w_l) = (Q_l A u, w_l)$.
Sei nun $u^{(0)} \in V$ eine approximative Lösung von (3.1) und u^* die exakte Lösung. In einem ersten Teilschritt projizieren wir den Fehler $u^* - u^{(0)}$ bezüglich des Energieskalarprodukts orthogonal auf W_0 und addieren diesen Fehleranteil zu $u^{(0)}$. Dadurch erhalten wir die erste von $J+1$ Teiliterierten $u^{(1/(J+1))} = u^{(0)} + P_0(u^* - u^{(0)})$. Die nächste Teiliterierte erhalten wir, indem wir den neuen Fehler auf W_1 anstatt auf W_0 projizieren. Die erste Iterierte $u^{(1)}$ ergibt sich dadurch, dass der Teilschritt auf allen $J + 1$ Teilgebieten durchgeführt wird. Das Vorgehen ist für zwei Unterräume auch in Abbildung 3.1 dargestellt. Es ergibt sich folgendes Schema:

Definition 3.5. (Exaktes Unterraumkorrekturverfahren) *Es sei ein Startwert $u^{(0)} \in V$ gegeben. Für $n \in \mathbb{N}_0$ berechnet sich die nächste Iterierte durch*

$$\begin{aligned} u^{(n+1/(J+1))} &= u^{(n)} + P_0(u^* - u^{(n)}) \\ u^{(n+2/(J+1))} &= u^{(n+1/(J+1))} + P_1(u^* - u^{(n+1/(J+1))}) \\ &\vdots \\ u^{(n+1)} &= u^{(n+J/(J+1))} + P_J(u^* - u^{(n+J/(J+1))}). \end{aligned}$$

Da die Unterraumkorrekturen nacheinander immer mit den aktuellen Werten aus dem vorangegangenen Iterationsteilschritt durchgeführt werden, spricht man von multiplikativen Unterraumkorrekturverfahren. Im Gegensatz dazu gibt es auch additive Unterraumkorrekturverfahren, bei denen die Korrekturen auf jedem Unterraum gleichzeitig bestimmt weden. Im Rahmen dieser Arbeit wird jedoch nur der multiplikative Fall betrachtet.

Indem alle Gleichungen eines Iterationsschritts ineinander eingesetzt werden, erhält man:

$$u^{(n+1)} = u^{(n)} + (I - (I - P_J) \cdots (I - P_1)(I - P_0))(u^* - u^{(n)}).$$

Bei diesem Verfahren besteht nun aber das Problem, dass das unbekannte u^* verwendet wird. Mit Hilfe der Gleichung $u^* = A^{-1}f$ formulieren wir das Verfahren um und sehen uns dazu wiederum einen Einzelschritt an:

$$\begin{aligned} u^{(n+(l+1)/(J+1))} &= u^{(n+l/(J+1))} + P_l(u^* - u^{(n+l/(J+1))}) \\ &= u^{(n+l/(J+1))} + P_l A^{-1}(f - Au^{(n+l/(J+1))}) \\ &= u^{(n+l/(J+1))} + A_l^{-1} Q_l(f - Au^{(n+l/(J+1))}). \end{aligned}$$

In einem Teilschritt berechnen wir also das aktuelle Residuum $f - Au^{(n+l/(J+1))}$, projizieren es orthogonal auf den Raum W_l, lösen dort ein kleineres lineares Gleichungssystem und addieren diesen Korrekturterm dann zu der aktuellen Iterierten. Dadurch, dass das zu lösende Gleichungssystem kleiner geworden ist, sind wir in der Lage, einen Teilschritt tatsächlich durchzuführen. Bei diesem Vorgehen sprechen wir auch von einem exakten Unterraumkorrekturverfahren.

Die Vorgehensweise ist allerdings problematisch, falls die Unterräume W_l noch recht groß

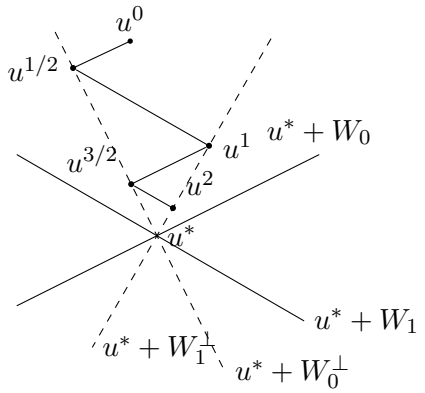

Abbildung 3.1: Unterraumkorrekturverfahren im Fall $J = 1$. Anstelle von a-Orthogonalität wird hier die Orthogonalität bezüglich des euklidischen Skalarprodukts verwendet (nach [DW11])

sind oder das exakte Lösen aus anderen Gründen aufwändig ist. Um dies zu umgehen, kann man stattdessen approximative Löser B_l^{-1} auf den Unterräumen betrachten. Hierbei seien die $B_l : W_l \to W_l'$ positiv definite Operatoren, mit denen lineare Gleichungssysteme einfach zu lösen sind. Das Unterraumkorrekturverfahren hat dann die Form:

Definition 3.6. (Approximatives Unterraumkorrekturverfahren) *Es sei ein Startwert $u^{(0)} \in V$ gegeben. Für $n \in \mathbb{N}_0$ berechnet sich die nächste Iterierte durch*

$$\begin{aligned} u^{(n+1/(J+1))} &= u^{(n)} + B_0^{-1} Q_0 (f - A u^{(n)}) \\ u^{(n+2/(J+1))} &= u^{(n+1/(J+1))} + B_1^{-1} Q_1 (f - A u^{(n+1/(J+1))}) \\ &\vdots \\ u^{(n+1)} &= u^{(n+J/(J+1))} + B_J^{-1} Q_J (f - A u^{(n+J/(J+1))}). \end{aligned}$$

Wie oben möchten wir auch das approximative Unterraumkorrekturverfahren kompakter schreiben. Dafür definiere $T_k := B_k^{-1} Q_k A = B_k^{-1} A_k P_k$, im Falle eines exakten Unterraumkorrekturverfahrens ist also $T_k = P_k$. Durch Ineinandereinsetzen erhält man

$$\begin{aligned} u^{(n+1)} &= u^{(n)} + (I - (I - T_J) \cdots (I - T_0))(u^* - u^{(n)}) \\ &= u^{(n)} + (I - (I - T_J) \cdots (I - T_0)) A^{-1} (f - A u^{(n)}) =: u^{(n)} + B_{mult} (f - A u^{(n)}). \end{aligned}$$

Wir haben also das approximative Unterraumkorrekturverfahren als konsistentes lineares Iterationsverfahren mit den linearen Abbildungen $B = B_{mult}$ und damit $M = I - B_{mult}A$ dargestellt.

Bemerkung 3.7. (Eigenschaften der T_k) *Die Operatoren T_k sind selbstadjungiert bezüglich $a(\cdot,\cdot)$, da B_k und damit B_k^{-1} selbstadjungiert bezüglich $(\cdot,\cdot)$ ist.*
Außerdem gilt $a(T_k v, v) = a(B_k^{-1}A_k P_k v, v) = (B_k^{-1}A_k P_k v, A_k P_k v) \geq 0$ für alle $v \in V$.
Also erzeugen die T_k positiv-semidefinite Bilinearformen auf V durch $a(T_k v, w)$ für alle $v, w \in V$. Für diese Bilinearform gilt auch die Cauchy-Schwarz-Ungleichung.

Für den Fehler im n-ten Iterationsschritt $e^{(n)} := u^* - u^{(n)}$ gilt dann $e^{(n)} = (I - T_J)\cdots(I - T_0)e^{(n-1)}$. Um die Konvergenz des Verfahrens nachzuweisen, reicht es also, den Fehlerfortpflanzungsoperator $E_{mult} := (I - T_J)\cdots(I - T_0)$ zu untersuchen und nachzuweisen, dass dessen Norm echt kleiner als eins ist.

Die Unterraumkorrekturverfahren können als eigenständige Lösungsverfahren verwendet werden. Sie können aber auch als Vorkonditionierer für andere Lösungsverfahren – wie zum Beispiel das cg-Verfahren – genutzt werden. Durch die Vorkonditionierung wollen wir das cg-Verfahren, dessen Konvergenzgeschwindigkeit vom Spektrum des Operators A abhängt, beschleunigen, indem wir das lineare Gleichungssystem durch einen Vorkonditionierer umformen und so ein kleineres Intervall erhalten, in dem das Spektrum liegt. Als Vorkonditionierer benötigen wir jedoch einen positiv definiten, also insbesondere einen selbstadjungierten Operator, was B_{mult} im Allgemeinen nicht ist. Um ein symmetrisches Verfahren zu konstruieren, wenden wir die Unterraumkorrekturen nochmals in umgekehrter Reihenfolge an, erhalten also

$$\begin{aligned} u^{(n+1)} &= u^{(n)} + (I - (I - T_0)\cdots(I - T_J)(I - T_J)\cdots(I - T_0))(u^* - u^{(n)}) \\ &= u^{(n)} + (I - (I - T_0)\cdots(I - T_J)(I - T_J)\cdots(I - T_0))A^{-1}(f - Au^{(n)}) \\ &=: u^{(n)} + B_{m,s}(f - Au^{(n)}). \end{aligned}$$

Da die T_k selbstadjungiert bezüglich des Energieskalarprodukts sind, ist $B_{m,s} = (I - (I - T_0)\cdots(I - T_J)(I - T_J)\cdots(I - T_0))A^{-1} = A^{-1}(I - (I - T_0)\cdots(I - T_J)(I - T_J)\cdots(I - T_0))$ und somit $B_{m,s}$ selbstadjungiert bezüglich des Dualitätsprodukts. Die positive Definitheit von $B_{m,s}$ werden wir unter gewissen Voraussetzungen in Abschnitt 3.3 nachweisen.

3.2 Konvergenz multiplikativer Unterraumkorrekturverfahren

In diesem Abschnitt soll die Konvergenz der multiplikativen Unterraumkorrekturverfahren nachgewiesen werden, wobei wir uns an [Yse93] orientieren.

3.2.1 Annahmen

Wir treffen zunächst einige Annahmen für den Rest des Kapitels.

Annahme 1. (Stabilität der Zerlegung) *Es gebe Unterräume $V_k \subset W_k$, $k \in \{0, \dots, J\}$, mit $V = \sum_{k=0}^{J} V_k$. Weiter gebe es eine Konstante K_1 so, dass für jedes $v \in V$ eine Zerlegung $v = \sum_{k=0}^{J} v_k$, $v_k \in V_k$, besteht mit*

$$\sum_{k=0}^{J} (B_k v_k, v_k) \leq K_1 \left\| \sum_{k=0}^{J} v_k \right\|^2 = K_1 \|v\|^2. \tag{3.3}$$

Im Fall von direkten Lösern lautet Annahme 1: $\sum_{k=0}^{J} \|v_k\|^2 \leq K_1 \left\| \sum_{k=0}^{J} v_k \right\|^2$. Sie besagt also, dass die Energienorm der einzelnen Teile von v sich – bis auf eine Konstante – durch die Energienorm von v selbst beschränken lässt, sie also nicht zu groß wird.

Annahme 2. *Es gibt eine Konstante K_2 so, dass für jede Teilmenge $S \subset \{0, \dots, J\} \times \{0, \dots, J\}$ und alle $w_k \in W_k, v_l \in W_l$*

$$\sum_{(k,l) \in S} a(w_k, v_l) \leq K_2 \left(\sum_{k=0}^{J} (B_k w_k, w_k) \right)^{1/2} \left(\sum_{l=0}^{J} (B_l v_l, v_l) \right)^{1/2}$$

gilt.

Annahme 2 impliziert insbesondere, dass für $v = \sum_{k=0}^{J} v_k$ wie in Annahme 1 gilt $\|v\|^2 = \sum_{k,l=0}^{J} a(v_k, v_l) \leq K_2 \sum_{k=0}^{J} (B_k v_k, v_k)$. Im Fall von direkten Lösern wird also die Energienorm der Teilfunktionen nach unten beschränkt. Es wird also sichergestellt, dass diese nicht zu klein ist.

Annahme 3. *Es gibt eine Konstante $0 < \omega < 2$ so, dass für alle $l \in \{1, \dots, J\}$ $(A_l w_l, w_l) \leq \omega (B_l w_l, w_l)$ für alle $w_l \in W_l$ gilt.*

Diese Annahme ist äquivalent zur Konvergenz der dem approximativen Unterraumkorrekturverfahren zugrundeliegenden Lösungsverfahren. Da A_l und B_l positiv definit sind,

folgt für alle $v_l \in W_l$

$$0 < \frac{(A_l v_l, v_l)}{(B_l v_l, v_l)} \leq \omega < 2.$$

Weiter folgt mit Hilfe des Rayleigh-Quotienten, da $I - B_l^{-1} A_l$ selbstadjungiert bezüglich des von B_l erzeugten Energieskalarprodukts ist,

$$\rho(I - B_l^{-1} A_l) = \max_{v_l \in W_l \setminus \{0\}} \left| \frac{(B_l(I - B_l^{-1} A_l) v_l, v_l)}{(B_l v_l, v_l)} \right| = \max_{v_l \in W_l \setminus \{0\}} \left| 1 - \frac{(A_l v_l, v_l)}{(B_l v_l, v_l)} \right|.$$

Also gilt $\rho(I - B_l^{-1} A_l) < 1$. Mit Satz 2.34 folgt die Konvergenz des verwendeten approximativen Lösers. Die andere Richtung folgt aus den gleichen Überlegungen.

3.2.2 Konvergenzbeweis für das multiplikative Unterraumkorrekturverfahren

Mit Hilfe der drei Annahmen aus dem vorangegangenen Abschnitt beweisen wir folgenden Satz:

Satz 3.8. (Konvergenz) *Jeder volle Schritt des Unterraumkorrekturverfahrens verringert den Fehler mindestens um den Faktor $\|E_{mult}\|$. Es ist*

$$\|E_{mult}\|^2 \leq 1 - \frac{2 - \omega}{K_1 (1 + K_2)^2}. \tag{3.4}$$

Da $K_1, K_2 > 0$ und $\omega < 2$ sind, ist $\frac{2-\omega}{K_1(1+K_2)^2} > 0$, also $\|E_{mult}\| < 1$. Es ist wünschenswert, dass die Konstanten unabhängig von der Anzahl der Unterräume J sind, um eine von J unabhängige Konvergenzrate zu erhalten.
Bevor wir zum Beweis des Satzes kommen, zeigen wir drei Hilfssätze.

Lemma 3.9. *Für alle $v \in V$ mit der Darstellung $v = \sum_{k=0}^{J} v_k$ aus Annahme 1 und alle $u_k \in V$ gilt*

$$\sum_{k=0}^{J} a(v_k, u_k) \leq \sqrt{K_1} \left\| \sum_{k=0}^{J} v_k \right\| \left(\sum_{k=0}^{J} a(T_k u_k, u_k) \right)^{1/2}.$$

Beweis. Mit Hilfe der Cauchy-Schwarz-Ungleichung für das von B_k^{-1} erzeugte Energie-

skalarprodukt und Annahme 1 erhalten wir

$$\begin{aligned}
\sum_{k=0}^{J} a(v_k,u_k) &= \sum_{k=0}^{J} a(v_k,P_k u_k) = \sum_{k=0}^{J} (v_k, A_k P_k u_k) \\
&= \sum_{k=0}^{J} (B_k^{-1} B_k v_k, A_k P_k u_k) = \sum_{k=0}^{J} (B_k v_k, B_k^{-1} A_k P_k u_k) \\
&\overset{\text{C.S}}{\leq} \sum_{k=0}^{J} (v_k, B_k v_k)^{1/2} (B_k^{-1} A_k P_k u_k, A_k P_k u_k)^{1/2} \\
&\overset{\text{C.S.}}{\leq} \left(\sum_{k=0}^{J} (B_k v_k, v_k) \right)^{1/2} \left(\sum_{k=0}^{J} (B_k^{-1} A_k P_k u_k, A_k P_k u_k) \right)^{1/2} \\
&= \left(\sum_{k=0}^{J} (B_k v_k, v_k) \right)^{1/2} \left(\sum_{k=0}^{J} (T_k u_k, A_k P_k u_k) \right)^{1/2} \\
&= \left(\sum_{k=0}^{J} (B_k v_k, v_k) \right)^{1/2} \left(\sum_{k=0}^{J} a(T_k u_k, P_k u_k) \right)^{1/2} \\
&= \left(\sum_{k=0}^{J} (B_k v_k, v_k) \right)^{1/2} \left(\sum_{k=0}^{J} a(T_k u_k, u_k) \right)^{1/2} \\
&\overset{\text{Ann. 1}}{\leq} \sqrt{K_1} \left\| \sum_{k=0}^{J} v_k \right\| \left(\sum_{k=0}^{J} a(T_k u_k, u_k) \right)^{1/2}.
\end{aligned}$$

□

Lemma 3.10. *Für alle $u \in V$, $k \in \{0,\dots,J\}$ gilt $\|T_k u\|^2 \leq \omega a(T_k u, u)$.*

Beweis. Wir wenden Annahme 3 sowie die Definition der T_k an und erhalten

$$\begin{aligned}
\|T_k u\|^2 = (T_k u, A_k T_k u) &\overset{\text{Ann. 3}}{\leq} \omega (T_k u, B_k T_k u) \\
&= \omega (T_k u, B_k B_k^{-1} A_k P_k u) \\
&= \omega a(T_k u, P_k u) = \omega a(T_k u, u).
\end{aligned}$$

□

Da ein Schritt des multiplikativen Unterraumkorrekturverfahrens aus mehreren, hintereinander ausgeführten Einzelschritten besteht, ist es sinnvoll, den Fehler in jedem Einzelschritt zu betrachten, was folgende Definition motiviert:

Definition 3.11. (Abgeschnittener Fehlerfortpflanzungsoperator) *Wir definieren den* abgeschnittenen Fehlerfortpflanzungsoperator *durch* $E_k := (I - T_k)\dots(I - T_0)$ *für* $k \in \{0, \dots, J\}$ *und* $E_{-1} := I$. *Man beachte, dass* $E_{mult} = E_J$ *ist.*

Lemma 3.12. *Für die abgeschnittenen Fehlerfortpflanzungsoperatoren gilt die Gleichung* $I - E_{l-1} = \sum_{k=0}^{l-1} T_k E_{k-1}$ *für alle* $l \in \{0, \dots, J\}$.

Beweis. Wir führen den Beweis per Induktion. Für $l = 0$ gilt

$$I - E_{-1} = I - I = 0 = \sum_{k=0}^{-1} T_k E_{k-1}.$$

Sei $l \in \{1, \dots, J-1\}$ und es gelte $I - E_{l-1} = \sum_{k=0}^{l-1} T_k E_{k-1}$. Dann gilt auch

$$I - E_l = I - (I - T_l)E_{l-1} = I - E_{l-1} + T_l E_{l-1} = \sum_{k=0}^{l-1} T_k E_{k-1} + T_l E_{l-1} = \sum_{k=0}^{l} T_k E_{k-1}.$$

□

Nun kommen wir zum Beweis des Konvergenzsatzes:

Beweis. (Satz 3.8) Der erste Teil des Satzes folgt sofort, da für den Fehler gilt

$$\|e^{(n)}\| = \|E_{mult} e^{(n-1)}\| \leq \|E_{mult}\| \|e^{(n-1)}\|.$$

Wir werden die Operatornorm des Fehlerfortpflanzungsoperators nach oben beschränken. Dazu formulieren wir die zu zeigende Aussage zunächst um, wodurch sich

$$\begin{aligned}
&\|E_{mult}\|^2 \leq 1 - \frac{2-\omega}{K_1(1+K_2)^2} \\
\iff &\frac{\|E_{mult} v\|^2}{\|v\|^2} \leq 1 - \frac{2-\omega}{K_1(1+K_2)^2} && \forall v \in V \setminus \{0\} \\
\iff &\frac{2-\omega}{K_1(1+K_2)^2} \leq \frac{\|v\|^2 - \|E_{mult} v\|^2}{\|v\|^2} && \forall v \in V \setminus \{0\} \\
\iff &(2-\omega)\|v\|^2 \leq K_1(1+K_2)^2(\|v\|^2 - \|E_{mult} v\|^2) && \forall v \in V \setminus \{0\} \qquad (3.5)
\end{aligned}$$

ergibt.

Nun beweisen wir die letzte Ungleichung (3.5). Die abgeschnittenen Fehlerfortpflanzungsoperatoren erfüllen für alle $k \in \{0, \ldots, J\}$ und $v \in V$ die Gleichung

$$\begin{aligned}\|E_{k-1}v\|^2 - \|E_k v\|^2 &= \|E_{k-1}v\|^2 - \|(I - T_k)E_{k-1}v\|^2 \\ &= \|E_{k-1}v\|^2 - a((I - T_k)E_{k-1}v, (I - T_k)E_{k-1}v) \\ &= \|E_{k-1}v\|^2 - \|E_{k-1}v\|^2 - \|T_k E_{k-1}v\|^2 + 2a(T_k E_{k-1}v, E_{k-1}v) \\ &= 2a(T_k E_{k-1}v, E_{k-1}v) - \|T_k E_{k-1}v\|^2.\end{aligned}$$

Lemma 3.10 auf diese Gleichung angewendet liefert

$$\begin{aligned}\|E_{k-1}v\|^2 - \|E_k v\|^2 &= 2a(T_k E_{k-1}v, E_{k-1}v) - \|T_k E_{k-1}v\|^2 \\ &\geq 2a(T_k E_{k-1}v, E_{k-1}v) - \omega a(T_k E_{k-1}v, E_{k-1}v) \\ &= (2 - \omega)a(T_k E_{k-1}v, E_{k-1}v).\end{aligned}$$

Daraus erhalten wir

$$\|v\|^2 - \|E_{mult}v\|^2 = \sum_{k=0}^{J} \|E_{k-1}v\|^2 - \|E_k v\|^2 \geq (2 - \omega) \sum_{k=0}^{J} a(T_k E_{k-1}v, E_{k-1}v). \quad (3.6)$$

Wenn wir nun zeigen, dass für alle $v \in V$

$$\|v\|^2 \leq K_1(1 + K_2)^2 \sum_{k=0}^{J} a(T_k E_{k-1}v, E_{k-1}v) \quad (3.7)$$

gilt, erhalten wir durch Kombination von (3.7) und (3.6) die gewünschte Ungleichung (3.5).

Dazu verwenden wir die Darstellung von $v = \sum_{l=0}^{J} v_l$ gemäß Annahme 1. Dann ist

$$\begin{aligned}\|v\|^2 &= \sum_{l=0}^{J} a(v, v_l) = \sum_{l=0}^{J} a(E_{l-1}v, v_l) + \sum_{l=0}^{J} a((I - E_{l-1})v, v_l) \\ &= \sum_{l=0}^{J} a(E_{l-1}v, v_l) + \sum_{l=1}^{J} a((I - E_{l-1})v, v_l). \quad (3.8)\end{aligned}$$

Die Terme im rechten Teil der Gleichung werden nun einzeln betrachtet. Für den ersten

Term liefert Lemma 3.9

$$\sum_{l=0}^{J} a(v_l, E_{l-1}v) \leq \sqrt{K_1}\|v\| \left(\sum_{l=0}^{J} a(T_l E_{l-1}v, E_{l-1}v) \right)^{1/2}. \tag{3.9}$$

Für den zweiten Term nutzen wir Lemma 3.12 aus. Damit folgt

$$\begin{aligned}
\sum_{l=1}^{J} a((I - E_{l-1})v, v_l) =& \sum_{l=1}^{J} \sum_{k=0}^{l-1} a(T_k E_{k-1}v, v_l) \\
\overset{\text{Ann. 2}}{\leq}& K_2 \left(\sum_{k=0}^{J} (B_k T_k E_{k-1}v, T_k E_{k-1}v) \right)^{1/2} \left(\sum_{l=0}^{J} (B_l v_l, v_l) \right)^{1/2} \\
=& K_2 \left(\sum_{k=0}^{J} a(E_{k-1}v, T_k E_{k-1}v) \right)^{1/2} \left(\sum_{l=0}^{J} (B_l v_l, v_l) \right)^{1/2} \\
\overset{\text{Ann. 1}}{\leq}& K_2 \sqrt{K_1}\|v\| \left(\sum_{k=0}^{J} a(T_k E_{k-1}v, E_{k-1}v) \right)^{1/2}.
\end{aligned} \tag{3.10}$$

Zusammen erhalten wir also aus (3.8), (3.9) und (3.10)

$$\begin{aligned}
\|v\|^2 &= \sum_{l=0}^{J} a(E_{l-1}v, v_l) + \sum_{l=1}^{J} (a(I - E_{l-1})v, v_l) \\
&\leq \sqrt{K_1}\|v\| \left(\sum_{l=0}^{J} a(T_l E_{l-1}v, E_{l-1}v) \right)^{1/2} \\
&\quad + K_2 \sqrt{K_1}\|v\| \left(\sum_{k=0}^{J} a(T_k E_{k-1}v, E_{k-1}v) \right)^{1/2} \\
&= \sqrt{K_1}(1 + K_2)\|v\| \left(\sum_{k=0}^{J} a(T_k E_{k-1}v, E_{k-1}v) \right)^{1/2}.
\end{aligned}$$

Dies entspricht (3.6) und damit ist die Behauptung gezeigt. □

Für das symmetrische multiplikative Verfahren ist der Fehlerfortpflanzungsoperator gegeben durch $E_{m,s} = E_J^T E_J$. Daher folgt aus der Konvergenz des multiplikativen Verfahrens die Konvergenz des symmetrischen multiplikativen Verfahrens.

3.3 Schranken für das Spektrum des vorkonditionierten Operators

Für die Konvergenz des vorkonditionierten cg-Verfahrens ist das Spektrum des vorkonditionierten Operators $B_{m,s}A$ ausschlaggebend. In diesem Abschnitt möchten wir daher den minimalen und maximalen Eigenwert von $B_{m,s}A$ nach unten bzw. oben beschränken, wobei wir der Argumentation in [SBG96] folgen. Außerdem zeigen wir, dass der Vorkonditionierer, also das multiplikative, symmetrische Unterraumkorrekturverfahren $B_{m,s}$, positiv definit ist.
In diesem Abschnitt gelten die Annahmen 1, 2 und 3 aus dem vorangegangen Abschnitt. Auch die abgeschnittenen Fehlerfortpflanzungsoperatoren werden wieder verwendet.
Mit Hilfe dieser lässt sich der Vorkonditionierer schreiben als $B_{m,s} = (I-(I-T_0)\dots(I-T_J)(I-T_J)\dots(I-T_0))A^{-1} = (I-E_J^*E_J)A^{-1}$

Lemma 3.13. *$B_{m,s}$ ist positiv definit.*

Beweis. In Satz 3.8 wurde gezeigt, dass $\|E_J\| < 1$ gilt, was äquivalent zu $a(E_Jv, E_Jv) < a(v,v)$ für alle $v \in V\backslash\{0\}$ ist. Dann gilt für jedes $v \in V\backslash\{0\}$ auch

$$\begin{aligned} 0 < a(v,v) - a(v, E_J^*E_Jv) &= a(v, (I - E_J^TE_J)v) \\ &= (Av, (I - E_J^*E_J)A^{-1}Av) = (Av, B_{m,s}Av). \end{aligned}$$

Da A regulär ist, ist $B_{m,s}$ demnach positiv definit. □

Satz 3.14. *Für den maximalen Eigenwert von $B_{m,s}A$ gilt: $\lambda_{max}(B_{m,s}A) \leq 1$.*

Beweis. Für alle $u \in V\backslash\{0\}$ betrachten wir den Rayleigh-Quotienten von $B_{m,s}A$ im Energieskalarprodukt, wodurch sich

$$\begin{aligned} \frac{a(B_{m,s}Au,u)}{a(u,u)} = \frac{a((I-E_J^*E_J)u,u)}{a(u,u)} &= \frac{a(u,u)}{a(u,u)} - \frac{a(E_Ju,E_Ju)}{a(u,u)} \\ &= 1 - \underbrace{\frac{a(E_Ju,E_Ju)}{a(u,u)}}_{\geq 0, \text{ da } a \text{ pos. def.}} \leq 1 \end{aligned}$$

ergibt. Da der maximale Eigenwert vom Rayleigh-Quotienten zu einem Eigenvektor dieses Eigenwerts angenommen wird, gilt $\lambda_{max}(B_{m,s}A) \leq \max_{u\in V\backslash\{0\}} \left(\frac{a(B_{m,s}Au,u)}{a(u,u)}\right)$, also auch $\lambda_{max}(B_{m,s}A) \leq 1$. □

Für die untere Schranke des minimalen Eigenwerts von $B_{m,s}A$ ist etwas mehr Arbeit erforderlich. Wir stellen den Satz voran, zeigen aber vor dem eigentlichen Beweis einige Lemmata.

Satz 3.15. *Für den minimalen Eigenwert von $B_{m,s}A$ gilt*

$$\lambda_{min}(B_{m,s}A) \geq \frac{2-\omega}{2K_1(K_2\omega + (1+K_2)^2)}.$$

Lemma 3.16. *Es ist $a(\sum_{k=0}^{J} T_k v, v) \leq K_2 a(v,v)$ für alle $v \in V$.*

Beweis. Wir wollen also den maximalen Eigenwert von $\sum_{k=0}^{J} T_k$ durch K_2 beschränken. Dazu nutzen wir den Rayleigh-Quotienten. Sei $v \in V$. Dann ist

$$\begin{aligned} a\left(\sum_{k=0}^{J} T_k v, \sum_{l=0}^{J} T_l v\right) &= \sum_{k,l=0}^{J} a(T_k v, T_l v) \overset{\text{Ann. 2}}{\leq} K_2 \sum_{k=0}^{J} (B_k T_k v, T_k v) \\ &= K_2 \sum_{k=0}^{J} (B_k B_k^{-1} A_k P_k v, T_k v) = K_2 \sum_{k=0}^{J} a(v, T_k v) \\ &= K_2\, a(v, \sum_{k=0}^{J} T_k v) \overset{\text{C.S.}}{\leq} K_2\, a(v,v)^{1/2}\, a\left(\sum_{k=0}^{J} T_k v, \sum_{k=0}^{J} T_k v\right)^{1/2}. \end{aligned}$$

Weiter bemerken wir, dass $\sum_{k=0}^{J} T_k$ als Summe selbstadjungierter Operatoren selbstadjungiert bezüglich des Energieskalarprodukts ist.
Zusammen ergibt sich $a\left(\left(\sum_{k=0}^{J} T_k\right)^2 v, v\right) = a\left(\sum_{k=0}^{J} T_k v, \sum_{k=0}^{J} T_k v\right) \leq K_2^2 a(v,v)$, also $\lambda_{max}\left(\left(\sum_{k=0}^{J} T_k\right)^2\right) \leq K_2^2$ und damit wie gewünscht $\lambda_{max}\left(\sum_{k=0}^{J} T_k\right) \leq K_2$. □

Lemma 3.17. *Es ist $a\left(\sum_{k=0}^{J} T_k v, v\right) \leq 2\left(K_2\omega + (1+K_2)^2\right) \sum_{k=0}^{J} a(E_{k-1}v, T_k E_{k-1} v)$ für alle $v \in V$.*

Beweis. Durch Umstellen folgt aus Lemma 3.12 $I = T_0 + E_{k-1} + \sum_{l=1}^{k-1} T_l E_{l-1}$ für alle $k \in \{0, \ldots, J\}$ und wir erhalten

$$a\left(\sum_{k=0}^{J} T_k v, v\right) = \sum_{k=0}^{J} a(T_k v, v)$$

$$
\begin{aligned}
&= \sum_{k=0}^{J} a(T_k v, T_0 v) + a(T_k v, E_{k-1} v) + a(T_k v, \sum_{l=1}^{k-1} T_l E_{l-1} v) \\
&\overset{\text{C.S.}}{\leq} \sum_{k=0}^{J} a(T_k v, v)^{1/2} a(T_k T_0 v, T_0 v)^{1/2} \\
&\quad + \sum_{k=0}^{J} a(T_k v, v)^{1/2} a(T_k E_{k-1} v, E_{k-1} v)^{1/2} + \sum_{k=0}^{J} \sum_{l=1}^{k-1} a(T_k v, T_l E_{l-1} v) \\
&\overset{\text{Ann. 2}}{\leq} \sum_{k=0}^{J} a(T_k v, v)^{1/2} a(T_k T_0 v, T_0 v)^{1/2} \\
&\quad + \sum_{k=0}^{J} a(T_k v, v)^{1/2} a(T_k E_{k-1} v, E_{k-1} v)^{1/2} \\
&\quad + K_2 \left(\sum_{k=0}^{J} (B_k T_k v, T_k v) \right)^{1/2} \left(\sum_{k=0}^{J} (B_k T_k E_{k-1} v, T_k E_{k-1} v) \right)^{1/2} \\
&\overset{\text{Def } T_k}{=} \sum_{k=0}^{J} a(T_k v, v)^{1/2} a(T_k T_0 v, T_0 v)^{1/2} \\
&\quad + \sum_{k=0}^{J} a(T_k v, v)^{1/2} a(T_k E_{k-1} v, E_{k-1} v)^{1/2} \\
&\quad + K_2 \left(\sum_{k=0}^{J} a(v, T_k v) \right)^{1/2} \left(\sum_{k=0}^{J} (B_k T_k E_{k-1} v, T_k E_{k-1} v) \right)^{1/2} \\
&\overset{\text{C.S.}}{\leq} \left(\sum_{k=0}^{J} a(T_k v, v) \right)^{1/2} \left[\left(\sum_{k=0}^{J} a(T_k T_0 v, T_0 v) \right)^{1/2} \right. \\
&\quad \left. + (1 + K_2) \left(\sum_{k=0}^{J} a(T_k E_{k-1} v, E_{k-1} v) \right)^{1/2} \right].
\end{aligned}
$$

Die erste der folgenden Ungleichungen erhalten wir, indem wir in obigem Resultat kürzen und quadrieren, die zweite folgt, weil $(x+y)^2 \leq 2(x^2+y^2)$ für alle $x, y \in \mathbb{R}$ gilt:

$$
a\left(\sum_{k=0}^{J} T_k v, v \right) \leq \left[\left(\sum_{k=0}^{J} a(T_k T_0 v, T_0 v) \right)^{1/2} + (1 + K_2) \left(\sum_{k=0}^{J} a(T_k E_{k-1} v, E_{k-1} v) \right)^{1/2} \right]^2
$$

$$\leq 2\left[\sum_{k=0}^{J} a(T_k T_0 v, T_0 v) + (1+K_2)^2 \sum_{k=0}^{J} a(T_k E_{k-1} v, E_{k-1} v)\right].$$

Auf die erste Summe kann nun Lemma 3.16 angewendet werden

$$\begin{aligned}\sum_{k=0}^{J} a(T_k T_0 v, T_0 v) &\overset{\text{Lemma 3.16}}{\leq} K_2\, a(T_0 v, T_0 v) = K_2 (A_0 T_0 v, T_0 v) \\ &\overset{\text{Ann. 3}}{\leq} K_2 \omega (B_0 B_0^{-1} A_0 P_0 v, T_0 v) = K_2 \omega\, a(P_0 v, T_0 E_{-1} v) \\ &= K_2 \omega\, a(E_{-1} v, T_0 E_{-1} v).\end{aligned}$$

Dann ergibt sich, da $a(T_k E_{k-1} v, E_{k-1} v) \geq 0$ für alle $k \in \{1, \ldots, J\}$ und $v \in V$ gilt,

$$\begin{aligned}a\left(\sum_{k=0}^{J} T_k v, v\right) &\leq 2\left(K_2 \omega a(E_{-1} v, T_0 E_{-1} v) + (1+K_2)^2 \sum_{k=0}^{J} a(T_k E_{k-1} v, E_{k-1} v)\right) \\ &\leq 2\left(K_2 \omega + (1+K_2)^2\right) \sum_{k=0}^{J} a(T_k E_{k-1} v, E_{k-1} v).\end{aligned}$$

□

Mit Hilfe dieser Vorüberlegungen können wir nun Satz 3.15 beweisen.

Beweis. (Satz 3.15) Aus den Eigenschaften des Rayleigh-Quotienten folgt, dass

$$\lambda_{min}(B_{m,s} A) \geq \min_{v \in V \setminus \{0\}} \frac{a(B_{m,s} A v, v)}{a(v, v)}$$

ist. Deshalb genügt es zu zeigen, dass für alle $v \in V \setminus \{0\}$

$$\frac{a(B_{m,s} A v, v)}{a(v, v)} \geq \frac{2 - \omega}{K_1 2 (K_2 \omega + (1+K_2)^2)}$$

gilt. Dazu betrachten wir zunächst Zähler und Nenner getrennt.
Wie im Beweis von Satz 3.14 benutzen wir, dass $B_{m,s} A = I - E_J^T E_J$ ist. Dann erhalten wir für den Zähler

$$a(B_{m,s} A v, v) = a(v, v) - a(E_J v, E_J v) = \sum_{k=0}^{J} a(E_{k-1} v, E_{k-1} v) - a(E_k v, E_k v)$$

$$
\begin{aligned}
&= \sum_{k=0}^{J} a(E_{k-1}v, E_{k-1}v) - a((E_{k-1} - T_kE_{k-1})v, (E_{k-1} - T_kE_{k-1})v) \\
&= \sum_{k=0}^{J} 2a(E_{k-1}v, T_kE_{k-1}v) - a(T_kE_{k-1}v, T_kE_{k-1}v) \\
&= \sum_{k=0}^{J} 2a(E_{k-1}v, T_kE_{k-1}v) - (A_kT_kE_{k-1}v, T_kE_{k-1}v) \\
&\overset{\text{Ann 3}}{\geq} \sum_{k=0}^{J} 2a(E_{k-1}v, T_kE_{k-1}v) - \omega(B_kT_kE_{k-1}v, T_kE_{k-1}v) \\
&= \sum_{k=0}^{J} 2a(E_{k-1}v, T_kE_{k-1}v) - \omega(B_kB_k^{-1}A_kP_kE_{k-1}v, T_kE_{k-1}v) \\
&= \sum_{k=0}^{J} 2a(E_{k-1}v, T_kE_{k-1}v) - \omega a(E_{k-1}v, T_kE_{k-1}v) \\
&= (2-\omega)\sum_{k=0}^{J} a(E_{k-1}v, T_kE_{k-1}v).
\end{aligned}
$$

Für den Nenner gilt mit der Zerlegung $v = \sum_{k=0}^{J} v_k$ gemäß Annahme 1 und Lemma 3.9 angewendet mit $u_k = v$ für alle $k \in 0, \ldots, J$

$$
a(v,v) = a\left(\sum_{k=0}^{J} v_k, v\right) \leq \sqrt{K_1}a(v,v)^{1/2}\left(\sum_{k=0}^{J} a(T_kv, v)\right)^{1/2}.
$$

Daraus erhalten wir durch Kürzen, Quadrieren und mit Hilfe des Lemmas 3.17

$$
a(v,v) \leq K_1 \sum_{k=0}^{J} a(T_kv, v) \leq 2K_1\left(K_2\omega + (1+K_2)^2\right)\sum_{l=0}^{J} a(E_{l-1}v, T_lE_{l-1}v).
$$

Zusammen ergibt sich also wie gewünscht

$$
\begin{aligned}
\frac{a(B_{m,s}Av, v)}{a(v,v)} &\geq \frac{(2-\omega)\sum_{k=0}^{J} a(E_{k-1}v, T_kE_{k-1}v)}{2K_1(K_2\omega + (1+K_2)^2)\sum_{k=0}^{J} a(E_{k-1}v, T_kE_{k-1}v)} \\
&= \frac{(2-\omega)}{2K_1(K_2\omega + (1+K_2)^2)}.
\end{aligned}
$$

□

Aus Lemma 3.14 und Satz 3.15 folgt sofort

Korollar 3.18. *Für das Spektrum der vorkonditionierten Matrix gilt*

$$\sigma(B_{m,s}A) \subset \left[\frac{(2-\omega)}{2K_1(K_2\omega + (1+K_2)^2)}, 1\right].$$

Das heißt für die Konstante aus dem vorkonditionierten cg-Verfahren gilt

$$\kappa = 2K_1(K_2\omega + (1+K_2)^2)(2-\omega).$$

Sie ist unabhängig von der Problemgröße, sofern die Konstanten K_1, K_2 und ω es sind.

4 Überlappende Gebietszerlegungsverfahren

In diesem Kapitel betrachten wir überlappende Gebietszerlegungsverfahren im Rahmen der Theorie von Unterraumkorrekturverfahren. Zunächst untersuchen wir den allgemeinen Fall einer räumlichen Zerlegung des Gebiets, auf dem die Differentialgleichung definiert ist und konstruieren daraus die benötigten Unterräume. Dann gehen wir auf eine konkrete Möglichkeit der Gebietszerlegung ein und beweisen für diese die Annahmen aus dem Kapitel 3.

4.1 Multiplikatives Schwarz-Verfahren

Die Idee für Gebietszerlegungsverfahren stammt von Hermann Schwarz (1843-1921) (vgl. [Sch70]). Dieser unterteilte ein Gebiet Ω in zwei sich überlappende Teilgebiete Ω_1, Ω_2 und führte über ein alternierendes Verfahren einen konstruktiven Existenzbeweis für harmonische Funktionen. Ihm zu Ehren nennt man überlappende Gebietszerlegungsverfahren auch Schwarz-Verfahren. Wir folgen bei diesen der Idee von Schwarz und unterteilen das Gebiet Ω in mehrere Teilgebiete Ω_k, $k \in \{1, \dots, J\}$, so dass $\overline{\Omega} = \bigcup_{k=1}^{J} \overline{\Omega}_k$. Dabei fordern wir keine Partition von Ω, sondern lassen Überschneidungen ausdrücklich zu.
Auf diesen Teilgebieten definieren wir nun folgendermaßen die Unterräume unseres Finite-Elemente-Raums V_h:

$$W_k := \{\varphi \in V : \operatorname{supp}(\varphi) \subset \overline{\Omega}_k\}.$$

Es gilt $V = \sum_{k=1}^{J} W_k$.
Bei Anwendung der Unterraumkorrekturmethode auf diese Zerlegung findet in jedem Iterationsschritt nur ein Datenaustausch zwischen den sich überlappenden Teilgebieten statt, was dazu führt, dass Informationen aus einem Teilgebiet in allen anderen Teilgebie-

ten vorhanden sind. Dies kann die Konvergenzgeschwindigkeit beeinträchtigen. Um einen globalen Datenaustausch zu gewährleisten, führen wir zusätzlich einen Grobgitterraum W_0 ein, der auf ganz Ω definiert ist. Hierbei bietet sich zum Beispiel der Finite-Elemente-Raum V_H zu einer Triangulierung $\mathcal{T}_H$, aus der $\mathcal{T}_h$ durch Verfeinerung hervorgegangen ist, an. Da hierbei zwei verschiedene Gitter eingehen, spricht man auch von zweistufigen Verfahren im Gegensatz zu einstufigen.

4.2 Konstruktion einer überlappenden Gebietszerlegung

Nun konstruieren wir eine konkrete überlappende Gebietszerlegung für Ω. Sei dazu $\{\mathcal{T}_h\}_h$ eine quasi-uniforme Familie von zulässigen Triangulierungen von Ω. Es gebe eine Triangulierung $\mathcal{T}_H = \{\hat{\Omega}_1, \dots, \hat{\Omega}_J\} \in \{\mathcal{T}_h\}_h$ mit Maschenweite H, die zusätzlich $h_{\hat{\Omega}_k} = H$ für alle $k \in \{1, \dots, J\}$ erfüllt.[1] Diese bezeichnen wir im Folgenden auch als grobe Triangulierung, da H relativ groß sein soll. Sei weiterhin $\mathcal{T}_h$ eine zulässige Triangulierung mit Maschenweite h, die aus $\mathcal{T}_H$ durch Verfeinerung entstanden ist. Sie wird auch als feines Gitter bezeichnet. Wir konstruieren die überlappenden Gebiete folgendermaßen, wobei die Überlappung von einem Parameter $\delta \in \mathbb{R}_{>0}$ abhängt. Mit $B_\delta(x)$ bezeichnen wir die offene Kugel um $x \in \mathbb{R}^d$ mit Radius δ.

Definition 4.1. (überlappende Gebiete) *Definiere $\Omega_k \subset \Omega$ für alle $k \in \{1, \dots, J\}$ durch*

- $\hat{\Omega}_k \subset \Omega_k$,
- $\Omega_k \cap \tau \in \{\emptyset, \tau\}$ *für alle* $\tau \in \mathcal{T}_h$,
- $B_\delta(x) \cap \Omega \subset \Omega_k$ *für alle* $x \in \hat{\Omega}_k$,
- *für alle* $\tau \in \mathcal{T}_h$ *mit* $\tau \subset \Omega_k$ *ist* $\operatorname{dist}(\tau, \hat{\Omega}_k) < \delta$.

Das heißt, wir fügen zu einem Element der groben Triangulierung $\hat{\Omega}_k$ Elemente der feineren Triangulierung hinzu, sodass der Abstand von $\hat{\Omega}_k$ zum Rand des neuen Gebiets Ω_k mindestens δ beträgt. Die Konstruktion ist für den zweidimensionalen Fall in Abb. 4.1 dargestellt. Die letzte Bedingung soll sicherstellen, dass wir nur so viele Elemente der feinen Triangulierung hinzunehmen, wie unbedingt nötig ist, um einen Überlappungsbereich mit einer Breite von mindestens δ zu erhalten.

[1] Es ist auch möglich, ein Gitter mit verschiedenen Umkreisradien zu betrachten, siehe hierzu [TW04].

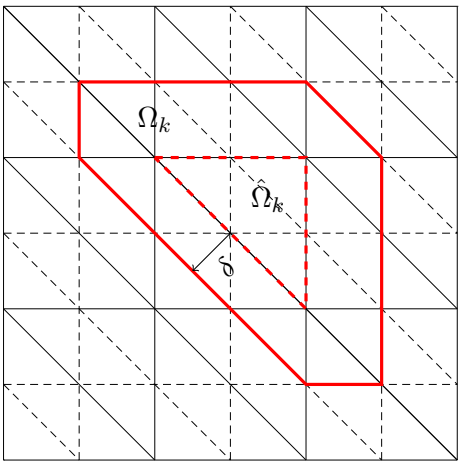

Abbildung 4.1: Konstruktion einer überlappenden Gebietszerlegung in 2D; grobes Gitter (schwarz), feines Gitter (schwarz gestrichelt), Ursprungselement $\hat{\Omega}_k$ (rot gestrichelt), daraus entstandenes Gebiet Ω_k (rot) und der Überlappungsparameter δ.

Weiterhin wählen wir δ so klein, dass für alle $l \in \{1, \ldots, J\} \backslash \{k\}$ gilt $\hat{\Omega}_l \not\subseteq \Omega_k$. In einem der überlappenden Teilgebiete ist damit genau ein Gebiet aus der groben Triangulierung komplett enthalten.

4.3 Beweis der Annahmen für die überlappende Gebietszerlegung

Für die im vorangegangenen Abschnitt beschriebene Gebietszerlegung beweisen wir die Annahmen aus Kapitel 3, die dem Konvergenzbeweis zugrunde liegen. Wir arbeiten mit exakten Lösern auf den Unterräumen, d.h. $B_k = A_k$. Außerdem beschränken wir uns auf das Poissonproblem, dann ist $a(u, v) = \int_\Omega \nabla u(x) \nabla v(x) dx$ für alle $u, v \in V$. Dies ist gerechtfertigt, weil für ein anderes beschränktes, elliptisches Problem die aus der zugehörigen Bilinearform entstehende Norm äquivalent zur H^1-Halbnorm ist, was der durch die Poissongleichung induzierten Norm entspricht (vgl. [Bra07, S. 37]).

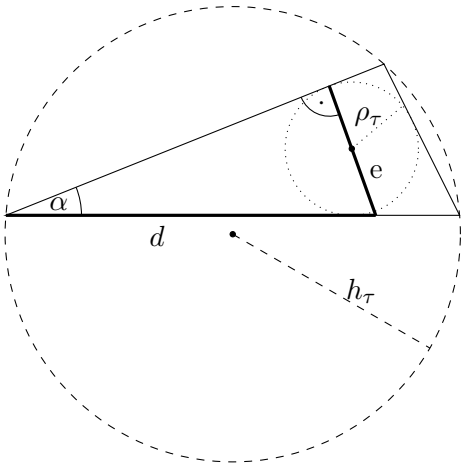

Abbildung 4.2: Dreieck τ mit Winkel α, Inkreis mit Radius ρ_τ, Umkreis mit Radius h_τ und Hilfsgrößen (fette Strecken) d, e.

4.3.1 Technische Hilfsmittel

Definition 4.2. (Nachbarschaft) *Für $k \in \{1, \ldots, J\}$ definiere die Nachbarschaftsmenge* $N_k := \{l \in \{1, \ldots, J\} : \Omega_l \cap \Omega_k \neq \emptyset\}$.
Sei $n := max_{k\in\{1,\ldots,J\}}|N_k|$.

Die Nachbarschaftsmenge N_k enthält also die Indizes der Gebiete, die mit dem Gebiet Ω_k überlappen. Ω_k hat dann $|N_k|-1$ viele Nachbarn und die maximale Anzahl von Nachbarn eines Gebiets ist beschränkt durch $n-1$.

Lemma 4.3. *Sei $d = 2$ und $\{\mathcal{T}_h\}_h$ eine quasi-uniforme Familie von Triangulierungen. Die Innenwinkel eines Dreiecks $\tau \in \mathcal{T}_h$, $\mathcal{T}_h \in \{\mathcal{T}_h\}_h$ seien gegeben durch $\alpha_\tau, \beta_\tau, \gamma_\tau$. Dann gibt es ein $\alpha_0 > 0$ mit*

$$\alpha_0 \leq \min_{\substack{\tau\in\mathcal{T}_h \\ \mathcal{T}_h\in\{\mathcal{T}_h\}_h}} \{\alpha_\tau, \beta_\tau, \gamma_\tau\}.$$

Beweis. Da $\{\mathcal{T}_h\}_h$ quasi-uniform ist, existiert eine Konstante $\kappa > 0$ mit $\frac{h_\tau}{\rho_\tau} < \kappa$ für alle $\tau \in \mathcal{T}_h$, $\mathcal{T}_h \in \{\mathcal{T}_h\}_h$. Sei $\mathcal{T}_h \in \{\mathcal{T}_h\}_h$ und $\tau \in \mathcal{T}_h$. Es sei o.B.d.A. α der kleinste Innenwinkel von τ, die Strecken d, e seien wie in Abb. 4.2 gewählt. Offenbar ist $d \leq 2h_\tau$ und $e = 2\rho_\tau$ also $\sin(\alpha) = \frac{e}{d} \geq \frac{\rho_\tau}{h_\tau} \geq \frac{1}{\kappa} > 0$. Da $\alpha \in (0, \frac{\pi}{2})$ liegt und der Sinus auf diesem Intervall monoton wächst, gibt es eine untere Schranke $\alpha_0 := \arcsin(\frac{1}{\kappa})$ für den kleinsten

Innenwinkel. □

Im zweidimensionalen Fall ist also bei quasi-uniformen Gittern der minimale Innenwinkel der Dreiecke beschränkt. Da diese Schranke $\alpha_0 \in (0, \pi)$ ist, können an einem Knoten höchstens $\lceil 2\pi/\alpha_0 \rceil$ Dreiecke aufeinander treffen. Aus diesem Grund ist n unabhängig von der Anzahl der Teilgebiete J. Auch im dreidimensionalen Fall führt die Quasi-Uniformität zu einer Beschränkung der Anzahl der Nachbarn.

Lemma 4.4. *Ein Element der feinen Triangulierung $\tau \in \mathcal{T}_h$ liegt in höchstens n der überlappenden Teilgebiete Ω_k.*

Beweis. Sei $\tau \in \mathcal{T}_h$. Dann gibt es ein $k \in \{1, \dots, J\}$ mit $\tau \subset \Omega_k$. Für jedes weitere $l \in \{1, \dots, J\} \backslash \{k\}$ mit $\tau \subset \Omega_l$ gilt $\Omega_k \cap \Omega_l \neq \emptyset$, also sind Ω_k und Ω_l benachbart. Da Ω_k höchstens $n-1$ viele Nachbarn hat, kann τ in höchstens n Teilgebieten enthalten sein. □

Lemma 4.5. (Partition der Eins) *Zu der überlappenden Zerlegung Ω_k, $k \in \{1, \dots, J\}$, gibt es eine Familie von schwach differenzierbaren Funktionen $\{\theta_k\}$, $\theta_k : \overline{\Omega} \to \mathbb{R}$ mit*

$$0 \leq \theta_k(x) \leq 1 \text{ für alle } x \in \overline{\Omega} \tag{4.1}$$

$$\operatorname{supp}(\theta_k) \subset \overline{\Omega}_k \tag{4.2}$$

$$\sum_{k=1}^{J} \theta_k(x) = 1 \text{ für alle } x \in \overline{\Omega} \tag{4.3}$$

$$\|\nabla \theta_k\|_{L^\infty(\Omega)} \leq \frac{C}{\delta} \text{ mit einer von } \delta \text{ und } H \text{ unabhängigen Konstanten } C. \tag{4.4}$$

Für den eindimensionalen Fall ist eine solche Familie von Funktionen in Abbildung 4.3 dargestellt.
Im Beweis folgen wir weitestgehend demjenigen von [DW11, Lemma 7.13].

Beweis. Wir definieren für $k \in \{1, \dots, J\}$ zuerst eine Hilfsfunktion

$$\hat{\theta}_k(x) := \max \left\{ 0, 1 - \frac{\operatorname{dist}(x, \hat{\Omega}_k)}{\delta} \right\}$$

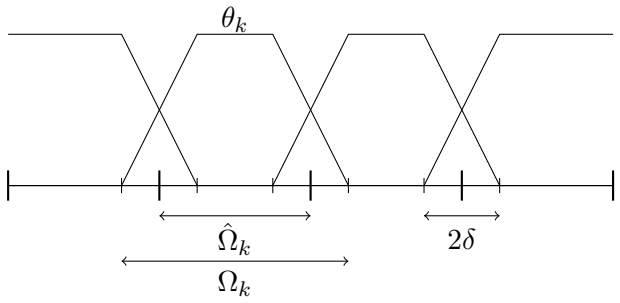

Abbildung 4.3: Partition der Eins im eindimensionalen Fall auf einem Intervall Ω, das in vier Teilintervalle unterteilt ist.

und damit

$$\theta_k := \frac{\hat{\theta}_k}{\sum_{l=1}^{J} \hat{\theta}_l}.$$

Für $x \in \overline{\Omega}$ gibt es ein $k \in \{1, \ldots, J\}$ mit $x \in \overline{\hat{\Omega}}_k$, also $\hat{\theta}_k(x) = 1$ und damit $\sum_{l=1}^{J} \hat{\theta}_l \geq 1 > 0$. Also ist θ_k wohldefiniert. Außerdem ist θ_k mit Ausnahme der Ränder der überlappenden Teilgebiete, die eine Nullmenge bilden, stetig differenzierbar und damit auf ganz Ω schwach differenzierbar.
Nun zu den geforderten Eigenschaften:

- Die erste Eigenschaft (4.1) ist offenbar erfüllt, da der Zähler zwischen 0 und 1 liegt und der Nenner größer gleich 1 ist.

- Zur zweiten Eigenschaft (4.2): Sei $x \in \Omega$ mit $\theta_k(x) \neq 0$, dann ist $\hat{\theta}_k(x) \neq 0$, also $\operatorname{dist}(x, \hat{\Omega}_k) \leq \delta$ und damit $x \in \overline{\Omega}_k$.

- Die dritte Eigenschaft (4.3) ist ebenfalls erfüllt, da für $x \in \overline{\Omega}$ gilt $\sum_{k=1}^{J} \theta_k(x) = \sum_{k=1}^{J} \frac{\hat{\theta}_k(x)}{\sum_{l=1}^{J} \hat{\theta}_l(x)} = \frac{\sum_{k=1}^{J} \hat{\theta}_k(x)}{\sum_{l=1}^{J} \hat{\theta}_l(x)} = 1$.

Um die vierte Eigenschaft (4.4) nachzuweisen, zeigen wir, dass θ_k lokal Lipschitz-stetig mit Lipschitz-Konstante $(n+1)/\delta$ ist. Dazu weisen wir zuerst die Lipschitz-Stetigkeit von $\hat{\theta}_k$ mit Lipschitz-Konstante $1/\delta$ nach. Hierbei benutzen wir die Ungleichung aus Lemma 2.3.

- 1. Fall $x, y \in \overline{\Omega}_k$: Dann ist

$$\begin{aligned}|\hat{\theta}_k(x) - \hat{\theta}_k(y)| &= |(1 - \frac{\operatorname{dist}(x, \hat{\Omega}_k)}{\delta}) - (1 - \frac{\operatorname{dist}(y, \hat{\Omega}_k)}{\delta})| \\ &= \frac{1}{\delta}|\operatorname{dist}(x, \hat{\Omega}_k) - \operatorname{dist}(y, \hat{\Omega}_k)| \leq \frac{1}{\delta}\|x - y\|_2.\end{aligned}$$

- 2. Fall $x \in \overline{\Omega}_k, y \notin \overline{\Omega}_k$: Hier gilt

$$\begin{aligned}|\hat{\theta}_k(x) - \hat{\theta}_k(y)| = |1 - \frac{\operatorname{dist}(x, \hat{\Omega}_k)}{\delta}| &= \frac{1}{\delta}(\delta - \operatorname{dist}(x, \hat{\Omega}_k)) \\ &\leq \frac{1}{\delta}(\operatorname{dist}(y, \hat{\Omega}_k) - \operatorname{dist}(x, \hat{\Omega}_k)) = \frac{1}{\delta}|(\operatorname{dist}(y, \hat{\Omega}_k) - \operatorname{dist}(x, \hat{\Omega}_k))| \\ &\leq \frac{1}{\delta}\|x - y\|_2.\end{aligned}$$

- 3. Fall $x \notin \overline{\Omega}_k, y \in \overline{\Omega}_k$: Analog zum 2. Fall.

- 4. Fall $x, y \notin \overline{\Omega}_k$: Es folgt $|\hat{\theta}_k(x) - \hat{\theta}_k(y)| = 0 \leq \frac{1}{\delta}\|x - y\|_2$.

Nun zur lokalen Lipschitz-Stetigkeit der θ_k. Sei $x \in \overline{\hat{\Omega}_i}$. Dann gilt für alle $y \in B_\delta(x) \subset \Omega_i$

$$\begin{aligned}|\theta_k(x) - \theta_k(y)| &= \left|\frac{\hat{\theta}_k(x)}{\sum_{l=1}^J \hat{\theta}_l(x)} - \frac{\hat{\theta}_k(y)}{\sum_{l=1}^J \hat{\theta}_l(y)}\right| \\ &= \left|\frac{\hat{\theta}_k(x)}{\sum_{l=1}^J \hat{\theta}_l(x)} - \frac{\hat{\theta}_k(y)}{\sum_{l=1}^J \hat{\theta}_l(x)} + \frac{\hat{\theta}_k(y)}{\sum_{l=1}^J \hat{\theta}_l(x)} - \frac{\hat{\theta}_k(y)}{\sum_{l=1}^J \hat{\theta}_l(y)}\right| \\ &\leq \frac{|\hat{\theta}_k(x) - \hat{\theta}_k(y)|}{\sum_{l=1}^J \hat{\theta}_l(x)} + \left|\frac{\hat{\theta}_k(y)}{\sum_{l=1}^J \hat{\theta}_l(x)} - \frac{\hat{\theta}_k(y)}{\sum_{l=1}^J \hat{\theta}_l(y)}\right| \\ &\overset{\text{Lip.Stet.}\hat{\theta}_k}{\leq} \frac{\frac{1}{\delta}\|x - y\|_2}{\sum_{l=1}^J \hat{\theta}_l(x)} + \left|\frac{\hat{\theta}_k(y)}{\sum_{l=1}^J \hat{\theta}_l(x)} - \frac{\hat{\theta}_k(y)}{\sum_{l=1}^J \hat{\theta}_l(y)}\right| \\ &= \frac{\frac{1}{\delta}\|x - y\|_2}{\sum_{l=1}^J \hat{\theta}_l(x)} + \hat{\theta}_k(y)\frac{|\sum_{l=1}^J \hat{\theta}_l(y) - \sum_{l=1}^J \hat{\theta}_l(x)|}{\sum_{l=1}^J \hat{\theta}_l(y) \sum_{l=1}^J \hat{\theta}_l(x)} \\ &= \frac{\frac{1}{\delta}\|x - y\|_2}{\sum_{l=1}^J \hat{\theta}_l(x)} + \theta_k(y)\frac{|\sum_{l=1}^J \hat{\theta}_l(y) - \sum_{l=1}^J \hat{\theta}_l(x)|}{\sum_{l=1}^J \hat{\theta}_l(x)} \\ &\overset{\theta_k \leq 1}{\leq} \frac{\frac{1}{\delta}\|x - y\|_2}{\sum_{l=1}^J \hat{\theta}_l(x)} + \frac{|\sum_{l=1}^J \hat{\theta}_l(y) - \hat{\theta}_l(x)|}{\sum_{l=1}^J \hat{\theta}_l(x)}\end{aligned}$$

$$\overset{\sum_{l=1}^{J}\hat{\theta}_l(x)\geq 1}{\leq} \frac{1}{\delta}\|x-y\|_2 + \sum_{l=1}^{J}|\hat{\theta}_l(y)-\hat{\theta}_l(x)|.$$

Falls $\hat{\Omega}_i$ und $\hat{\Omega}_l$ nicht benachbart sind, ist – da der Überlappungsparameter hinreichend klein gewählt ist – $\theta_l(x) = 0 = \theta_l(y)$, das heißt, in der Summe bleiben höchstens n Summanden übrig, auf die die Lipschitz-Stetigkeit angewendet wird. Dadurch erhalten wir schließlich $|\theta_k(x) - \theta_k(y)| \leq \frac{n+1}{\delta}\|x-y\|_2$. Aufgrund dieser Eigenschaft ist auch $\|\nabla\theta_k\|_{L^\infty(\Omega)} \leq {}^{(n+1)}/_{\delta}$. □

Zum Beweis der Annahmen benötigen wir auch eine Abschätzung der L_2-Norm auf einem Randstreifen der Elemente der groben Triangulierung $\hat{\Omega}_k$. Dazu definiere $\Gamma_{k,\delta} := \{\tau \in \mathcal{T}_h : \tau \subset \hat{\Omega}_k, \mathrm{dist}(\tau, \partial\hat{\Omega}_k) < \delta\}$. Dies entspricht dem Überlappungsbereich von $\hat{\Omega}_k$. Wir erhalten damit folgendes Lemma:

Lemma 4.6. *Es existiert eine Konstante $\hat{C}$, so dass für alle $u \in H^1(\hat{\Omega}_k)$*

$$\|u\|^2_{L_2(\Gamma_{k,\delta})} \leq \hat{C}\,\delta^2\left(\left(1+\frac{H}{\delta}\right)|u|^2_{H^1(\hat{\Omega}_k))} + \frac{1}{H\delta}\|u\|^2_{L_2(\hat{\Omega}_k)}\right).$$

gilt.

Beweis. Siehe [TW04, Lemma 3.10]. □

4.3.2 Beweis der Annahmen für das konstruierte multiplikative Schwarz-Verfahren

Wir zeigen in diesem Abschnitt, dass das vorgestellte Gebietszerlegungsverfahren die getroffenen Annahmen 1, 2 und 3 aus Kapitel 3 erfüllt. Damit ist die Konvergenztheorie aus diesem anwendbar. Die Annahmen eins und zwei werden in jeweils einem Unterabschnitt behandelt. Für die dritte Annahme ergibt sich

Bemerkung 4.7. *Annahme 3 ist wegen $B_k = A_k$ trivialerweise mit $\omega = 1$ erfüllt.*

Annahme 1

Nun kommen wir zur ersten Annahme, der Stabilitätsbedingung. Beim Nachweis dieser orientieren wir uns an [SBG96].

Wir machen einige Vorüberlegungen zur Wahl der V_k. Eine erste Idee – wie in Abb. 4.4 a zu sehen – könnte darin bestehen, nur den Anteil einer Funktion v auf $\hat{\Omega}_k$ zu betrachten und v_k außerhalb auf 0 zu setzen. Das Problem hierbei ist, dass die resultierenden Funktionen nicht stetig sind, also die V_k keine Unterräume von V bilden würden. Um diesem abzuhelfen, kann man die abgeschnittene Funktion interpolieren und so eine stetige Funktion erhalten. Bei diesem Vorgehen kann es jedoch – wie in Abb. 4.4 b dargestellt – zu sehr steilen Funktionsanteilen kommen. Hierdurch kann die zu betrachtende H^1-Halbnorm sehr groß werden, sie verhält sich wie $1/h$. Um diese großen Ableitungen zu verhindern, nutzen wir die in Abschnitt 4.3.1 vorgestellte Partition der Eins (vgl. Abb. 4.4 c). Durch die Multiplikation der abgeschnittenen Funktion mit dieser und anschließender Interpolation erhalten wir einen flacheren Anstieg im Überlappungsbereich. Der Anstieg verhält sich hier wie $1/\delta$. Problematisch werden können nun noch – wie in Abb. 4.4 d angedeutet – sehr glatte Funktionen, z.B. konstante, sein. Mittels dieser können trotz der Verwendung der Partition der Eins beliebig große Steigungen erzeugt werden. Allerdings können diese Funktionen bereits auf dem groben Gitter gut dargestellt werden, weswegen wir den glatten Anteil einer Funktion durch eine L_2-Projektion auf das grobe Gitter, d.h. auf die bestmögliche Darstellung der Funktion auf dem groben Gitter bezüglich der L_2-Norm, abbilden, von der eigentlichen Funktion abziehen und auf den übrig bleibenden Funktionsanteil die vorangegangenen Überlegungen anwenden.

Definition 4.8. *Definiere $V_0 := \{Q_H u : u \in V\}$, wobei Q_H die L_2-Projektion auf das grobe Gitter sei (vgl. Definition 2.25). Für $k \in \{1, \dots, J\}$ sei $V_k := \{I^h[\theta_k(u - Q_H u)] : u \in V\}$ mit der Knoteninterpolierenden I^h aus Definition 2.28.*

Lemma 4.9. *Für alle $k \in \{0, \dots, J\}$ ist $V_k \subset W_k$ und $\sum_{k=0}^{J} V_k = V$.*

Beweis. Nach Definition ist $V_0 \subset W_0$. Für $k \in \{1, \dots, J\}$, $v \in V$ ist $\operatorname{supp}(\theta_k v) \subset \overline{\Omega}_k$, also ist bei linearer Interpolation auch $\operatorname{supp}(I^h[\theta_k v]) \subset \overline{\Omega}_k$ und damit $V_k \subset W_k$.
Weiter gilt für $v \in V$ und $v_0 := Q_H v \in V_0$, $v_k := I^h[\theta_k(v - Q_H v)] \in V_k$ für alle $k \in \{1, \dots, J\}$ wegen der Linearität des Interpolationsoperators:

$$\begin{aligned}\sum_{k=0}^{J} v_k &= Q_H v + \sum_{k=1}^{J} I^h[\theta_k(v - Q_H v)] = Q_H v + I^h\left[\sum_{k=1}^{J} \theta_k(v - Q_H v)\right] \\ &= Q_H v + I^h[v - Q_H v] = Q_H v + v - Q_H v = v.\end{aligned}$$

□

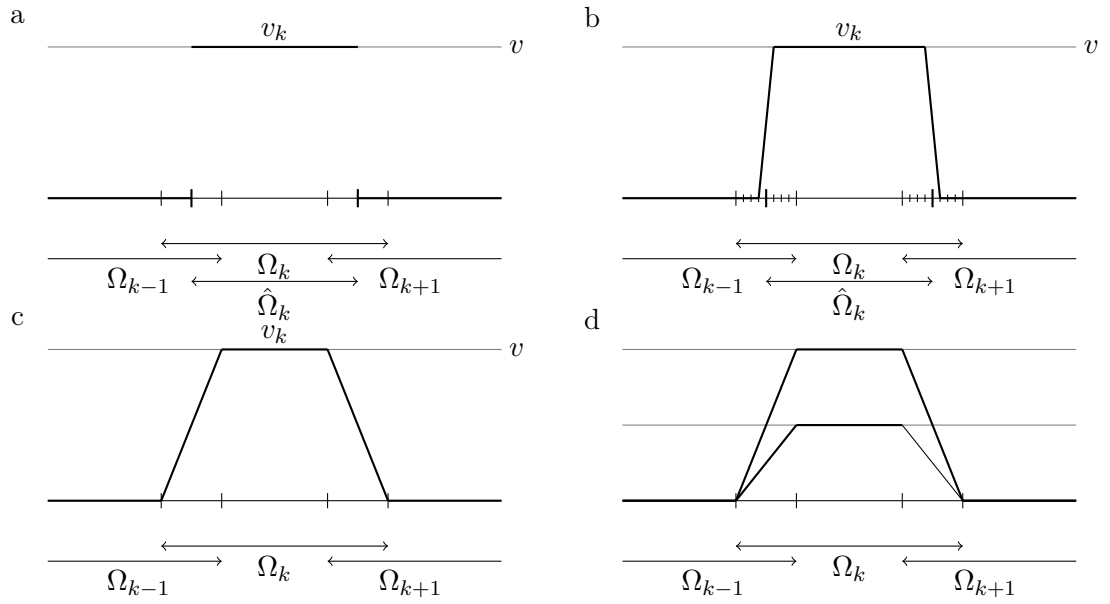

Abbildung 4.4: Zur Wahl der Unterräume V_k. a) Funktion abschneiden, b) Interpolation der abgeschnittenen Funktion, c) Hinzunehmen der Partition der Eins, d) Probleme bei glatten Funktionen.

Satz 4.10. *Die V_k mit der Zerlegung $v = \sum_{k=0}^{J} v_k$ wie im vorigen Beweis erfüllen Annahme 1.*

Der Beweis ist eine Ausarbeitung der Beweisskizze von Theorem 2 aus [SBG96].

Beweis. Wir bemerken, dass $(A_k v_k, v_k) = a(v_k, v_k) = |v_k|^2_{H^1(\Omega)}$ für alle $k \in \{0, \ldots, J\}$ ist. In einem ersten Schritt betrachten wir jeden Summanden einzeln, wobei der Grobgitterraum $k = 0$ einen Sonderfall darstellt. In diesem Fall ist nämlich wegen der H^1-Stabilität der L_2-Projektion (Satz 2.26) mit der Konstanten $\bar{c}$

$$a(v_0, v_0) = |Q_H v|^2_{H^1(\Omega)} \leq \bar{c}|v|^2_{H^1(\Omega)} = \bar{c}\, a(v, v). \tag{4.5}$$

Zur Abkürzung führen wir $w := v - v_0$ ein. Sei nun $k \in \{1, \ldots, J\}$ und sei $\tau \in \mathcal{T}_h$ ein einzelnes Element der feinen Triangulierung. Es sei $\bar{\theta}_{k,\tau} := \frac{1}{\int_\tau dy} \int_\tau \theta_k(y)\, dy$ der Mittelwert von θ_k auf τ. Aufgrund der Lipschitz-Eigenschaft der Partition der Eins und wegen

$\operatorname{diam}(\tau) \leq 2h_\tau$ ist

$$\begin{aligned}
\|\theta_k - \bar{\theta}_{k,\tau}\|_{L^\infty(\tau)} &= \sup_{x\in\tau} \left|\theta_k(x) - \frac{1}{\int_\tau dy} \int_\tau \theta_k(y)\, dy\right| \\
&= \sup_{x\in\tau} \left|\frac{\int_\tau \theta_k(x) dy}{\int_\tau dy} - \frac{1}{\int_\tau dy} \int_\tau \theta_k(y)\, dy\right| \\
&\leq \sup_{x\in\tau} \frac{1}{\int_\tau dy} \int_\tau |\theta_k(x) - \theta_k(y)|\, dy \\
&\leq \sup_{x\in\tau} \frac{1}{\int_\tau dy} \int_\tau \frac{n+1}{\delta} \|x-y\|_2\, dy \qquad (4.6) \\
&\leq \sup_{x\in\tau} \frac{1}{\int_\tau dy} \int_\tau \frac{(n+1)2h_\tau}{\delta}\, dy \\
&= \frac{2(n+1)h_\tau}{\delta}.
\end{aligned}$$

Nun betrachten wir die H^1-Halbnorm von v_k auf τ und erhalten dafür

$$\begin{aligned}
|v_k|^2_{H^1(\tau)} = &\quad \left|I^h[\theta_k w]\right|^2_{H^1(\tau)} = \left|I^h[\bar{\theta}_{k,\tau} w + (\theta_k - \bar{\theta}_{k,\tau})w]\right|^2_{H^1(\tau)} \\
\leq &\quad 2\left|I^h[\bar{\theta}_{k,\tau} w]\right|^2_{H^1(\tau)} + 2\left|I^h[(\theta_k - \bar{\theta}_{k,\tau})w]\right|^2_{H^1(\tau)} \\
\overset{\bar{\theta}_{k,\tau} w \in V}{=} &\quad 2\left|\bar{\theta}_{k,\tau} w\right|^2_{H^1(\tau)} + 2\left|I^h[(\theta_k - \bar{\theta}_{k,\tau})w]\right|^2_{H^1(\tau)}.
\end{aligned}$$

Wir wenden jetzt auf den zweiten Term der rechten Seite zuerst die Interpolationsabschätzung (Lemma 2.29) und danach die inverse Ungleichung (Lemma 2.24) mit den Konstanten $C^{1/2}$ bzw. $c^{1/2}$ an und nutzen für den ersten aus, dass $|\bar{\theta}_{k,\tau}| \leq 1$ ist. So erhalten wir

$$|v_k|^2_{H^1(\tau)} \leq 2|w|^2_{H^1(\tau)} + 2C|(\theta_k - \bar{\theta}_{k,\tau})w|^2_{H^1(\tau)} \leq 2|w|^2_{H^1(\tau)} + \frac{2Cc}{h_\tau^2}\|(\theta_k - \bar{\theta}_{k,\tau})w\|^2_{L_2(\tau)}.$$

Wir summieren nun über alle Indizes k und beachten, dass v_k konstant Null auf τ ist, falls $\tau \not\subseteq \Omega_k$. Außerdem liegt nach Lemma 4.4 jedes τ in höchstens n Teilgebieten Ω_k, wodurch wir

$$\sum_{k=1}^{J} |v_k|^2_{H^1(\tau)} = \sum_{\substack{k\in\{1,\dots,J\} \\ \tau\subset\Omega_k}} |v_k|^2_{H^1(\tau)} \leq \sum_{\substack{k\in\{1,\dots,J\} \\ \tau\subset\Omega_k}} 2|w|^2_{H^1(\tau)} + \frac{2Cc}{h_\tau^2}\|(\theta_k - \bar{\theta}_{k,\tau})w\|^2_{L_2(\tau)}$$

$$\leq 2n|w|^2_{H^1(\tau)} + \frac{2Cc}{h_\tau^2} \sum_{\substack{k\in\{1,\dots,J\}\\ \tau\subset\Omega_k}} \|(\theta_k - \bar{\theta}_{k,\tau})w\|^2_{L_2(\tau)}$$

erhalten. Wir bemerken, dass für ein $x \in \tau \subset \Omega_k$ nur $\theta_k(x) - \bar{\theta}_{k,\tau} \neq 0$ gilt, falls τ im Bereich von Ω_k liegt, der nicht überlappt wird. Sonst wäre nämlich θ_k konstant 1 auf τ. Das heißt, falls $\theta_k(x) - \bar{\theta}_{k,\tau} \neq 0$ für $x \in \tau \subset \Omega_k$ gilt, folgt $\tau \subset \Gamma_{k,\delta}$. Dies nutzen wir bei der folgenden Summation über die Elemente der feinen Triangulierung:

$$\begin{aligned}
\sum_{k=1}^{J} |v_k|^2_{H^1(\Omega)} = &\sum_{k=1}^{J}\sum_{\tau\in\mathcal{T}_h} |v_k|^2_{H^1(\tau)} = \sum_{\tau\in\mathcal{T}_h}\sum_{k=1}^{J} |v_k|^2_{H^1(\tau)} \\
\leq &\sum_{\tau\in\mathcal{T}_h} 2n|w|^2_{H^1(\tau)} + \sum_{\tau\in\mathcal{T}_h} \frac{2Cc}{h_\tau^2} \sum_{\substack{k\in\{1,\dots,J\}\\ \tau\subset\Omega_k}} \|(\theta_k - \bar{\theta}_{k,\tau})w\|^2_{L_2(\tau)} \\
\leq &\; 2n|w|^2_{H^1(\Omega)} + 2Cc \sum_{\tau\in\mathcal{T}_h} \frac{1}{h_\tau^2} \sum_{\substack{k\in\{1,\dots,J\}\\ \tau\subset\Omega_k}} \|(\theta_k - \bar{\theta}_{k,\tau})\|^2_{L^\infty(\tau)} \|w\|^2_{L_2(\tau)} \\
= &\; 2n|w|^2_{H^1(\Omega)} + 2Cc \sum_{\tau\in\mathcal{T}_h} \frac{1}{h_\tau^2} \sum_{\substack{k\in\{1,\dots,J\}\\ \tau\subset\Gamma_{k,\delta}}} \|(\theta_k - \bar{\theta}_{k,\tau})\|^2_{L^\infty(\tau)} \|w\|^2_{L_2(\tau)} \\
\overset{(4.6)}{\leq} &\; 2n|w|^2_{H^1(\Omega)} + 2Cc \sum_{\tau\in\mathcal{T}_h} \frac{1}{h_\tau^2} \sum_{\substack{k\in\{1,\dots,J\}\\ \tau\subset\Gamma_{k,\delta}}} \frac{4(n+1)^2h_\tau^2}{\delta^2} \|w\|^2_{L_2(\tau)} \\
= &\; 2n|w|^2_{H^1(\Omega)} + \frac{8(n+1)^2Cc}{\delta^2} \sum_{k=1}^{J} \sum_{\substack{\tau\in\mathcal{T}_h\\ \tau\subset\Gamma_{k,\delta}}} \|w\|^2_{L_2(\tau)} \\
= &\; 2n|w|^2_{H^1(\Omega)} + \frac{8(n+1)^2Cc}{\delta^2} \sum_{k=1}^{J} \|w\|^2_{L_2(\Gamma_{k,\delta})} \\
\overset{\text{Lemma 4.6}}{\leq} &\; 2n|w|^2_{H^1(\Omega)} \\
&\quad + \frac{8(n+1)^2Cc}{\delta^2} \sum_{k=1}^{J} \hat{c}\,\delta^2 \left[\left(1+\frac{H}{\delta}\right)|w|^2_{H^1(\hat{\Omega}_k)} + \frac{1}{H\delta}\|w\|^2_{L_2(\hat{\Omega}_k)}\right] \\
= &\; 2n|w|^2_{H^1(\Omega)} + 8(n+1)^2Cc\hat{c}\left[\left(1+\frac{H}{\delta}\right)|w|^2_{H^1(\Omega)} + \frac{1}{H\delta}\|w\|^2_{L_2(\Omega)}\right] \\
= &\; 2n|w|^2_{H^1(\Omega)}
\end{aligned}$$

$$+ 8(n+1)^2 Cc\hat{c}\left[\left(1+\frac{H}{\delta}\right)|v-v_0|^2_{H^1(\Omega)} + \frac{1}{H\delta}\|v-v_0\|^2_{L_2(\Omega)}\right].$$

Wir nehmen o.B.d.A. $Cc\hat{c} \geq 1$ an und fassen $8(n+1)^2 Cc\hat{c}$ in einer Konstanten $\hat{C}$ zusammen. Nun wenden wir noch die H^1-Stabilität der L_2-Projektion (Satz 2.26) mit den Konstanten c, c' an und erhalten

$$\begin{aligned}\sum_{k=1}^{J}|v_k|^2_{H^1(\Omega)} \leq& \hat{C}\left[2\left(1+\frac{H}{\delta}\right)|v-v_0|^2_{H^1(\Omega)} + \frac{1}{H\delta}\|v-v_0\|^2_{L_2(\Omega)}\right]\\ \leq& \hat{C}\left[2\left(1+\frac{H}{\delta}\right)c^2|v|^2_{H^1(\Omega)} + \frac{1}{H\delta}c'^2H^2|v|^2_{H^1(\Omega)}\right].\end{aligned}$$

Wenn wir nun noch die Abschätzung für v_0 hinzunehmen, ergibt sich

$$\begin{aligned}\sum_{k=0}^{J}|v_k|^2_{H^1(\Omega)} \leq& \hat{C}\left[2\left(1+\frac{H}{\delta}\right)c^2|v|^2_{H^1(\Omega)} + \frac{1}{H\delta}c'^2H^2|v|^2_{H^1(\Omega)}\right] + \bar{c}|v|^2_{\Omega}\\ \leq& \tilde{K}_1(1+\frac{H}{\delta})|v|^2_{H^1(\Omega).}\end{aligned}$$

Also erhalten wir unsere Behauptung mit einer Konstanten K_1, die sich linear zu $1+\frac{H}{\delta}$ verhält. □

Annahme 2

Für den Nachweis von Annahme 2 beginnen wir mit einigen Vorbetrachtungen, die wir allgemein auch für nicht exakte Löser auf den Unterräumen durchführen. Hierbei folgen wir der Argumentation in [Xu92].

Definition 4.11. (Ungleichung vom Cauchy-Schwarz-Typ) *Sei* $\mathcal{E} = (\varepsilon_{kl}) \in \mathbb{R}^{J+1\times J+1}$ *mit* ε_{kl} *der kleinsten Konstanten mit* $a(w_k, v_l) \leq \varepsilon_{kl}\omega(B_k w_k, w_k)^{1/2}(B_l v_l, v_l)^{1/2}$ *für alle* $w_k \in W_k, v_l \in W_l$.

Dann ist $\mathcal{E}$ symmetrisch und wegen Annahme 3 und der Cauchy-Schwarz-Ungleichung gilt $0 \leq \varepsilon_{kl} \leq 1$.

Lemma 4.12. *Sei* $\mathcal{J}_0 \subset \{0,\ldots,J\}$, $\mathcal{J}_0^c := \{0,\ldots,J\}\backslash\mathcal{J}_0$ *das Komplement und* $\gamma_0 := |\mathcal{J}_0|$ *die Mächtigkeit von* $\mathcal{J}_0$. *Weiter sei* $\sigma_0 := \max_{k\notin\mathcal{J}_0}\sum_{l\notin\mathcal{J}_0}\varepsilon_{lk}$. *Dann ist* $K_2 \leq 2\omega(\gamma_0+\sigma_0)$.

Beweis. Sei $S \subset \{0,\dots,J\} \times \{0,\dots,J\}$. Definiere

$$S_{11} := S \cap (\mathcal{J}_0 \times \mathcal{J}_0) \qquad S_{12} := S \cap (\mathcal{J}_0 \times \mathcal{J}_0^c)$$
$$S_{21} := S \cap (\mathcal{J}_0^c \times \mathcal{J}_0) \qquad S_{22} := S \cap (\mathcal{J}_0^c \times \mathcal{J}_0^c)$$

Dann ist $S = S_{11} \dot\cup S_{21} \dot\cup S_{12} \dot\cup S_{22}$.
Für $v_l \in W_l, w_k \in W_k$ betrachten wir nun Summen über diese Teilmengen von S. Für die Elemente aus S_{11} erhalten wir:

$$\begin{aligned}
\left(\sum_{(k,l)\in S_{11}} a(w_k,v_l)\right)^2 &\le \left(\sum_{(k,l)\in S_{11}} \varepsilon_{kl}\omega (B_k w_k,w_k)^{1/2}(B_l v_l,v_l)^{1/2}\right)^2 \\
&\le \omega^2 \left(\sum_{k\in\mathcal{J}_0}\sum_{l\in\mathcal{J}_0} \varepsilon_{kl}(B_k w_k,w_k)^{1/2}(B_l v_l,v_l)^{1/2}\right)^2 \\
&\overset{\varepsilon_{kl}\le 1}{\le} \omega^2 \left(\sum_{k\in\mathcal{J}_0} (B_k w_k,w_k)^{1/2} \sum_{l\in\mathcal{J}_0} (B_l v_l,v_l)^{1/2}\right)^2 \\
&\overset{\text{C.S.}}{\le} \omega^2\gamma_0^2 \sum_{k\in\mathcal{J}_0} (B_k w_k,w_k) \sum_{l\in\mathcal{J}_0} (B_l v_l,v_l) \\
&\overset{B_k \text{ spd}}{\le} \omega^2\gamma_0^2 \sum_{k=0}^{J} (B_k w_k,w_k) \sum_{l=0}^{J} (B_l v_l,v_l)
\end{aligned}$$

Bei den Elementen aus S_{22} ergibt sich:

$$\begin{aligned}
\left(\sum_{(k,l)\in S_{22}} a(w_k,v_l)\right)^2 &\le \omega^2 \left(\sum_{k\in\mathcal{J}_0^c}\sum_{l\in\mathcal{J}_0^c} \varepsilon_{kl}(B_k w_k,w_k)^{1/2}(B_l v_l,v_l)^{1/2}\right)^2 \\
&\le \omega^2 \left(\sum_{k\in\mathcal{J}_0^c} \max_{l\in\mathcal{J}_0^c}\varepsilon_{kl}(B_k w_k,w_k)^{1/2} \sum_{l\in\mathcal{J}_0^c} (B_l v_l,v_l)^{1/2}\right)^2 \\
&\overset{\varepsilon_{kl}=\varepsilon_{lk}}{=} \omega^2 \left(\sum_{k\in\mathcal{J}_0^c} (\max_{l\in\mathcal{J}_0^c}\varepsilon_{kl})^{1/2}(B_k w_k,w_k)^{1/2} \right. \\
&\qquad \left. \sum_{l\in\mathcal{J}_0^c} (\max_{k\in\mathcal{J}_0^c}\varepsilon_{kl})^{1/2}(B_l v_l,v_l)^{1/2}\right)^2
\end{aligned}$$

$$\overset{\text{C.S}}{\leq} \omega^2\sigma_0^2 \sum_{k\in\mathcal{J}_0^c} (B_k w_k, w_k) \sum_{l\in\mathcal{J}_0^c} (B_l v_l, v_l)$$
$$\overset{B_k \text{ spd}}{\leq} \omega^2\sigma_0^2 \sum_{k=0}^{J} (B_k w_k, w_k) \sum_{l=0}^{J} (B_l v_l, v_l)$$

Die Analyse der Summation über die Elemente von S_{12} ist etwas aufwändiger. Für $k \in \mathcal{J}_0$ sei $\mathcal{J}_0^c(k) := \{l \in \mathcal{J}_0^c : (k,l) \in S_{12}\}$. Damit erhalten wir

$$\begin{aligned}
\left(\sum_{(k,l)\in S_{12}} a(w_k, v_l)\right)^2 &= \left(\sum_{k\in\mathcal{J}_0} a(w_k, \sum_{l\in\mathcal{J}_0^c(k)} v_l)\right)^2 \\
&\overset{\text{C.S.}}{\leq} \left(\sum_{k\in\mathcal{J}_0} a(w_k,w_k)^{1/2}\, a\left(\sum_{l\in\mathcal{J}_0^c(k)} v_l, \sum_{l\in\mathcal{J}_0^c(k)} v_l\right)^{1/2}\right)^2 \\
&\leq \gamma_0 \sum_{k\in\mathcal{J}_0} a(w_k,w_k)\, a\left(\sum_{l\in\mathcal{J}_0^c(k)} v_l, \sum_{l\in\mathcal{J}_0^c(k)} v_l\right) \\
&\overset{\text{Ann. 3}}{\leq} \gamma_0\omega \sum_{k\in\mathcal{J}_0} (B_k w_k, w_k)\, a\left(\sum_{l\in\mathcal{J}_0^c(k)} v_l, \sum_{l\in\mathcal{J}_0^c(k)} v_l\right).
\end{aligned}$$

Für jedes $m \in \mathcal{J}_0$ gilt

$$\begin{aligned}
a\left(\sum_{l\in\mathcal{J}_0^c(m)} v_l, \sum_{l\in\mathcal{J}_0^c(m)} v_l\right) &= \sum_{k\in\mathcal{J}_0^c(m)} \sum_{l\in\mathcal{J}_0^c(m)} a(v_k, v_l) \\
&\leq \sum_{k\in\mathcal{J}_0^c(m)} \sum_{l\in\mathcal{J}_0^c(m)} \varepsilon_{kl}\omega (B_k v_k, v_k)^{1/2} (B_l v_l, v_l)^{1/2} \\
&\overset{B_k \text{ spd}}{\leq} \sum_{k\in\mathcal{J}_0^c} \sum_{l\in\mathcal{J}_0^c} \varepsilon_{kl}\omega (B_k v_k, v_k)^{1/2} (B_l v_l, v_l)^{1/2} \\
&\leq \sigma_0\omega \sum_{k\in\mathcal{J}_0^c} (B_k v_k, v_k).
\end{aligned}$$

Also

$$\left(\sum_{(k,l)\in S_{12}} a(w_k, v_l)\right)^2 \leq \gamma_0\sigma_0\omega^2 \sum_{k\in\mathcal{J}_0} (B_k w_k, w_k) \sum_{l\in\mathcal{J}_0^c} (B_l v_l, v_l)$$
$$\overset{B_k \text{ spd}}{\leq} \omega^2\gamma_0\sigma_0 \sum_{k=0}^{J} (B_k w_k, w_k) \sum_{l=0}^{J} (B_l v_l, v_l).$$

Analog ergibt sich

$$\left(\sum_{(k,l)\in S_{21}} a(w_k, v_l)\right)^2 \leq \omega^2\gamma_0\sigma_0 \sum_{k=0}^{J} (B_k w_k, w_k) \sum_{l=0}^{J} (B_l v_l, v_l).$$

Nun fügen wir die eben erhaltenen Ungleichungen zusammen und nutzen, dass für alle $w, x, y, z \in \mathbb{R}$ gilt $(w+x+y+z)^2 \leq 4(w^2+x^2+y^2+z^2)$. Dann erhalten wir

$$\begin{aligned}
&\left(\sum_{(k,l)\in S} a(w_k, v_l)\right)^2 \\
&= \left(\sum_{(k,l)\in S_{11}} a(w_k, v_l) + \sum_{(k,l)\in S_{12}} a(w_k, v_l) + \sum_{(k,l)\in S_{21}} a(w_k, v_l) + \sum_{(k,l)\in S_{22}} a(w_k, v_l)\right)^2 \\
&\leq 4\left[\left(\sum_{(k,l)\in S_{11}} a(w_k, v_l)\right)^2 + \left(\sum_{(k,l)\in S_{12}} a(w_k, v_l)\right)^2\right. \\
&\qquad \left. + \left(\sum_{(k,l)\in S_{21}} a(w_k, v_l)\right)^2 + \left(\sum_{(k,l)\in S_{22}} a(w_k, v_l)\right)^2\right] \\
&\leq 4\left[(\omega^2\gamma_0^2 + 2\omega^2\gamma_0\sigma_0 + \omega^2\sigma_0^2) \sum_{k=0}^{J} (B_k w_k, w_k) \sum_{l=0}^{J} (B_l v_l, v_l)\right].
\end{aligned}$$

Damit ist

$$\sum_{(k,l)\in S} a(w_k, v_l) \leq 2\omega(\gamma_0+\sigma_0) \left(\sum_{k=0}^{J} (B_k w_k, w_k)\right)^{1/2} \left(\sum_{l=0}^{J} (B_l v_l, v_l)\right)^{1/2}$$

und damit $K_2 \leq 2\omega(\gamma_0 + \sigma_0)$. □

Wir wenden uns wieder dem Gebietszerlegungsverfahren mit exakten Lösern auf den Teilräumen zu.

Lemma 4.13. *Annahme 2 ist erfüllt mit* $K_2 \leq 2(n+1)$.

Beweis. Wir wenden Lemma 4.12 an mit $\mathcal{J}_0 = \{0\}$. Dann ist $\gamma_0 = 1$.
In der Ungleichung vom Cauchy-Schwarz-Typ (Definition 4.11) ist für alle $k, l \in \{1, \ldots, J\}$ $\varepsilon_{kl} = 0$, falls $\Omega_k \cap \Omega_l = \emptyset$, denn in diesem Fall haben $w_k \in V_k, v_l \in V_l$ disjunkte Träger, also $a(w_k, v_l) = \int_\Omega \nabla w_k \nabla v_l \, dx = 0$. Außerdem ist $\varepsilon_{kl} \leq 1$ und damit erhalten wir

$$\sigma_0 = \max_{k \in \{1,\ldots,J\}} \sum_{l=1}^{J} \varepsilon_{kl} = \max_{k \in \{1,\ldots,J\}} \sum_{\substack{l \in \{1,\ldots,J\} \\ \Omega_k \cap \Omega_l \neq \emptyset}} \varepsilon_{kl} \leq \max_{k \in \{1,\ldots,J\}} |N_k| = n.$$

Da $\omega = 1$ ist, ergibt sich die Behauptung. □

Aus Kapitel 3 wissen wir, dass für den Fehlerfortpflanzungsoperator gilt

$$\|E_{mult}\|^2 \leq 1 - \frac{2-\omega}{K_1(1+K_2)^2} = 1 - \frac{1}{C(1+\frac{H}{\delta}) + c}.$$

Dieser verhält sich also wie $1 - \frac{1}{H/\delta}$ und bleibt immer kleiner als 1.
Für das Spektrum von $B_{m,s}A$ gilt $\sigma(B_{m,s}A) \subset [\frac{2-\omega}{2K_1(K_2\omega+(1+K_2)^2)}, 1]$. Mit den in diesem Kapitel gewonnenen Schranken gilt also mit einer Konstanten C, die von der Anzahl der Nachbarn n abhängt, $\sigma(B_{m,s}A) \subset [(C(1+H/\delta))^{-1}, 1]$. Das Intervall wird also größer, das heißt, die Konstante κ aus Satz 2.38 und damit die Konvergenzrate wird schlechter, je größer $\frac{H}{\delta}$ ist.

5 Implementierung

Im Rahmen dieser Arbeit wurde das vorgestellte multiplikative Schwarzverfahren für den zweidimensionalen Fall implementiert. Dieses Kapitel beschäftigt sich mit der Implementierung des Verfahrens und den an einem Modellbeispiel durchgeführten Tests. Die daraus resultierenden Ergebnisse werden ausgewertet und diskutiert.

5.1 Algorithmus

Im Folgenden stellen wir die implementierten Algorithmen vor. Zunächst erklären wir das Erstellen der überlappenden Gebietszerlegung sowie die Konstruktion der Restriktionen und Teilmatrizen aus dieser. Dann gehen wir auf das multiplikative Gebietszerlegungsverfahren als eigenständiges Verfahren ein und stellen zum Schluss die Anwendung des symmetrischen Verfahrens als Vorkonditionierer für das cg-Verfahren vor.
Bei der Implementierung wurden Routinen aus der Programmbibliothek SimpleFEM der Arbeitsgruppe Scientific Computing an der CAU Kiel verwendet, die Anwendungen für die Lösung von partiellen Differentialgleichungen mit der Finite-Elemente-Methode bereithält.

5.1.1 Erzeugung einer überlappenden Gebietszerlegung

Gegeben seien ein grobes Gitter $\mathcal{T}_H$, die Maschenweite für das feine Gitter h und der Überlappungsparameter δ. In einem Vorbereitungsschritt für die Anwendung des Schwarzverfahrens sollen die überlappende Gebietszerlegung, die dazugehörigen Teilmatrizen und die Restriktionen auf die Unterräume erstellt werden.
Die Funktion refine_with_origin (Algorithmus 1) erzeugt aus $\mathcal{T}_H$ mit Hilfe der Routine refine_tri2d aus der SimpleFEM-Bibliothek ein feines Gitter der Maschenweite h und speichert für jedes entstehende kleine Dreieck, aus welchem großen es entstanden ist. Refine_tri2d verfeinert mit der Strategie der Rotverfeinerung, welche eine Familie von

Algorithmus 1: refine_with_origin

Input : grobe Triangulierung $\mathcal{T}_H$, feine Maschenweite h
Output : feines Gitter $\mathcal{T}_h$, Prolongation P_0, Ursprungsinformationen origin

```
P_0 ← I
for τ ∈ T_H do
    origin(τ) ← τ
end
origin_help ← origin
while mesh_width(T_H > h) do
    T_h ← refine_tri2d(T_H, t2ref)
    P̂ ← build_prolongation_tri2dp1(T_h, T_H)
    P_0 ← P̂ P_0
    for τ ∈ T_h do
        origin(τ) ← origin_help(t2ref(τ))
    end
    origin_help ← origin
    T_H ← T_h
end
```

quasi-uniformen Gittern erzeugt. Wir nutzen aus, dass es mit refine_tri2d möglich ist zu speichern, aus welchem Dreieck ein um einen Schritt feineres Dreieck entstanden ist, um die Ursprungsinformationen zu erhalten. Zusätzlich wird in jedem Verfeinerungsschritt mit build_prolongation_tri2dp1 aus der SimpleFEM-Bibliothek eine Projektionsmatrix vom aktuellen auf das verfeinerte Gitter erstellt. Durch Hintereinanderausführung der Projektionsmatrizen aus jedem Schritt erhält man die Projektionsmatrix vom groben auf das feine Gitter.
Zur Bestimmung der Maschenweite wird eine Funktion mesh_width genutzt, die für jedes Dreieck τ mit den Seitenlängen a_τ, b_τ, c_τ einer gegebenen Triangulierung den Umkreisradius

$$h_\tau = \frac{a_\tau b_\tau c_\tau}{\sqrt{(a_\tau + b_\tau + c_\tau)(b_\tau + c_\tau - a_\tau)(a_\tau - b_\tau + c_\tau)(a_\tau + b_\tau - c_\tau)}}$$

mit Hilfe der Formel von Heron für den Flächeninhalt (vgl. [KK07]) bestimmt und daraus das Maximum ermittelt.
Zum entstandenen feinen Gitter kann nun die zur Problemstellung gehörige Steifigkeitsmatrix A erstellt werden; dies geschieht durch Routinen aus der SimpleFEM-Bibliothek.

Mehr Informationen hierzu gibt es in Abschnitt 5.2.

Um aus der Steifigkeitsmatrix die Teilmatrizen zu den überlappenden Gebieten zu erhalten, benötigen wir zunächst die Restriktionen auf die Teilgebiete. Die Funktion NewRestriction (Algorithmus 2), ermittelt, welche Dreiecke der feinen Triangulierung zu den

Algorithmus 2: NewRestriction

```
Input  : feines Gitter T_h, Ursprungsinformationen origin, Überlappungsparameter δ
Output : Dreiecke in jeweiligem Teilgebiet R = (R̂_1, ..., R̂_J)
for i = 1 to J do
    l ← 0
    for τ ∈ T_h do
        if origin(τ) = Ω̂_i then
            R̂_i[l] ← τ
            l++
        end
        else if dist(τ, Ω̂_i) < δ then
            R̂_i[l] ← τ
            l++
        end
    end
end
```

jeweiligen überlappenden Gebieten gehören und speichert diese jeweils in $\hat{R}_i$. Die Funktion Restriction (Algorithmus 3) erstellt die eigentlichen Restriktionen für die Gebiete. Hierbei werden aber für jedes Teilgebiet nur die zugehörigen Knotennummern in R_i gespeichert. Die Knotennummern werden für jedes Dreieck durch die SimpleFEM-Prozedur getvertices_tri2d ermittelt und in einem Hilfsvektor v gespeichert.

Mit Hilfe der vorgestellten Routinen kann nun der eigentliche Vorbereitungsschritt in DomainDecomposition_init (Algorithmus 4) durchgeführt werden. Mittels der Information aus den Restriktionen werden durch GetSubmatrices die Teilmatrizen aus der großen Matrix A ausgelesen, anschließend wird eine LR-Zerlegung für diese Matrizen berechnet, damit direktes Lösen auf den Teilräumen möglich ist. Die Matrix zum Grobgitterraum wird mit Hilfe der Prolongation vom groben auf das feine Gitter P_0 bestimmt; auch für sie wird eine LR-Zerlegung bestimmt.

Algorithmus 3: Restriction

Input : feines Gitter $\mathcal{T}_h$, Dreiecksinformationen R, Gebietsnummer i
Output : Knoten zum i-ten Teilgebiet R_i

```
l ← 0;
for j ∈ R̂_i do
    getvertices_tri2d(T_h, j, v)
    for k=1 to 3 do
        if v[k] ∉ R_i then
            R_i[l] ← v[k]
            l++
        end
    end
end
```

Algorithmus 4: DomainDecomposition_init

Input : Ursprungsinformationen origin, feines Gitter $\mathcal{T}_h$, Überlappungsparameter δ, Steifigkeitsmatrix A
Output : LR-Zerlegung der Teilmatrizen $A_0, \ldots, A_J$, Prolongation P_0, Restriktionen $R_1, \ldots R_J$

```
*Konstruktion der überlappenden Teilgebiete*
R ← NewRestriction(origin, T_h, T_H, δ)
for i = 1 to J do
    R_i ← Restriction(T_h, R, i)
end
*Erstellen der Teilmatrizen*
A_0 ← P_0^T A_0 P_0
LR(A_0)
for i = 1 to J do
    A_i ← GetSubmatrices(A, R_i)
    LR(A_i)
end
```

5.1.2 Einsatz als direktes Lösungsverfahren

In der Routine DomainDecomposition_step (Algorithmus 5) wird ein voller Schritt des Unterraumkorrekturverfahrens durchgeführt.

Algorithmus 5: DomainDecomposition_step

Input : Systemmatrix A, rechte Seite b, Iterationsvektor x, Teilmatrizen $A_0, \ldots, A_J$, Projektion P_0, Restriktionen $R_1, \ldots, R_J$
Output : aktueller Iterationsvektor x, Residuum r

$r \leftarrow b - Ax$
Grobgitterkorrektur $\hat{r} \leftarrow P_0^T r$
Löse $A_0 z = \hat{r}$
$x \leftarrow x + P_0 z$
$r \leftarrow b - Ax$
Unterraumkorrekturen **for** *i=1* **to** J **do**
  $\hat{r} \leftarrow R_i r$
  Löse $A_i z = \hat{r}$
  $x \leftarrow x + R_i^T z$
  $r \leftarrow b - Ax$
end

Das Lösen der linearen Gleichungssysteme geschieht auf Grundlage der in DomainDecomposition_init berechneten LR-Zerlegungen durch Vorwärts- und Rückwärtseinsetzen. Bei den überlappenden Gebieten werden die Matrix-Vektor-Multiplikationen $R_i r$ und $R_i^T z$ nicht tatsächlich ausgeführt. Vielmehr sind in R_i die zum i-ten Teilgebiet gehörigen Knotennummern abgespeichert. Die passenden Einträge werden dann aus r ausgelesen und im kleineren Vektor $\hat{r}$ gespeichert, bzw. die entsprechenden Komponenten von z werden an die passenden Stellen eines Nullvektors geschrieben. Im Hauptprogramm wird diese Routine solange aufgerufen, bis der Fehler klein genug ist – beispielsweise bis das relative Residuum eine vorgegebene Schranke unterschreitet.

5.1.3 Einsatz als Vorkonditionierer für das cg-Verfahren

Das symmetrische Unterraumkorrekturverfahren wird durch die Routinen DomainDecomposition_init aus Abschnitt 5.1.1 und DomainDecompositionSymm_step realisiert. DomainDecompositionSymm_step arbeitet fast genauso wie die aus Abschnitt 5.1.2 bekannte Prozedur DomainDecomposition_step. Der einzige Unterschied zwischen den beiden

Funktionen besteht darin, dass die Korrekturen nochmals in umgekehrter Reihenfolge durchgeführt werden. Wegen dieser Ähnlichkeit verzichten wir auf eine Darstellung in Pseudocode.

Das verwendete vorkonditionierte cg-Verfahren ist eine leicht veränderte Version der in der SimpleFEM-Bibliothek implementierten Routinen init_pcg und step_pcg (vgl. Algorithmen 6 und 7).

Hierbei ist zu beachten, dass im vorkonditionierten cg-Verfahren $q = B_{m,s}r$ berechnet werden muss. Dies entspricht $q = M \cdot 0 + B_{m,s}r$, wobei $M = I - B_{m,s}A$ die Matrix aus der ersten Normalform sei. Also können wir q durch einen Aufruf von DomainDecompositionSymm_step mit rechter Seite b und Startvektor 0 bestimmen.

Algorithmus 6: init_pcg_Schwarz

Input : Systemmatrix A, rechte Seite b, Iterationsvektor x, Teilmatrizen $A_0, \ldots, A_J$, Prolongation P_0, Restriktionen $R_1, \ldots, R_J$
Output : Residuum r, Suchrichtung p

$r \leftarrow b - Ax$
$q \leftarrow 0$
`DomainDecompositionSymm_step` $(A, r, q, A_0, \ldots, A_J, P_0, R_1, \ldots, R_J)$
$p \leftarrow q$

Algorithmus 7: step_pcg_Schwarz

Input : Systemmatrix A, Iterationsvektor x, Suchrichtung $p \neq 0$, Residuum r, Teilmatrizen $A_0, \ldots, A_J$, Prolongation P_0, Restriktionen $R_1, \ldots, R_J$
Output : neue Iterierte x, neue Suchrichtung p, neues Residuum r

$a \leftarrow Ap$
$\gamma \leftarrow (p, a)_2$
$\lambda \leftarrow \frac{(p,r)_2}{\gamma}$
$x \leftarrow x + \lambda p$
$r \leftarrow r - \lambda a$
$q \leftarrow 0$
`DomainDecompositionSymm_step` $(A, r, q, A_0, \ldots, A_J, P_0, R_1, \ldots, R_J)$
$p \leftarrow q - \frac{(q,a)_2}{\gamma} p$

5.2 Testbeispiel

Um die theoretisch getroffenen Aussagen zu überprüfen, werden numerische Experimente durchgeführt. Als Testbeispiel betrachten wir folgendes Problem auf dem Einheitsquadrat $\Omega = (-1,1) \times (-1,1)$:

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

Dieses Beispiel hat den Vorteil, dass der Fehler und das Residuum leicht zu bestimmen sind, da die exakte Lösung 0 ist. Das heißt, der Fehler ist die Iterierte selber und das Residuum der Operator A angewendet auf die Iterierte.
Das betrachtete Gebiet wird gleichmäßig trianguliert; die ersten Elemente der entstehenden Familie von Triangulierungen sind in Abb. 5.1 gezeigt. Diese Familie von Trian-

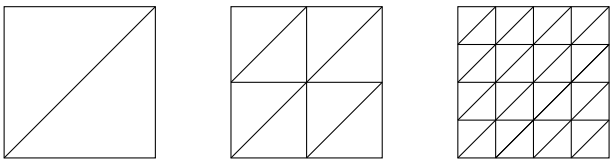

Abbildung 5.1: Erste Elemente einer quasi-uniformen Familie von Triangulierungen des Einheitsquadrats. Ausgangstriangulierung, eine und zwei Verfeinerungen.

gulierungen ist quasi-uniform. In unseren Experimenten wählen wir die grobe und die feine Triangulierung $\mathcal{T}_H, \mathcal{T}_h$ aus dieser Familie, indem wir die Anzahl der Verfeinerungen angeben, die wir für $\mathcal{T}_H$ mit M und für $\mathcal{T}_h$ mit m bezeichnen. Damit ergeben sich für die Maschenweiten die Werte $H = \frac{\sqrt{2}}{2^M}$ bzw. $h = \frac{\sqrt{2}}{2^m}$ und für die Anzahl der überlappenden Teilgebiete erhält man $J = 2 \cdot 4^M$.
Zum Erzeugen der Systemmatrix A und der rechten Seite b werden die Routinen assemble_poisson_tri2dp1 und functional_tri2dp1 aus der SimpleFEM-Bibliothek benutzt.
Es werden mehrere Testläufe mit zufällig generierten Startvektoren und für unterschiedliche Parametereinstellungen der Anzahl der Verfeinerungen für das grobe Gitter M, für das feine Gitter m und des Verhältnisses von H zu δ durchgeführt. Über H/δ kann auch gesteuert werden, wie viele „Dreiecksschichten“ bei der Erstellung der überlappenden Gebiete dazugenommen werden. Es wird jeweils die Anzahl der Iterationsschritte gezählt, die benötigt werden, bis der relative Fehler eine vorgegebene Schranke unterschreitet. In

den Tabellen 5.1 und 5.3 in den folgenden Abschnitten ist jeweils der auf eine ganze Zahl abgerundete Mittelwert angegeben.

5.2.1 Numerische Resultate für das eigenständige Verfahren

Die Ergebnisse der numerischen Experimente für das eigenständige Verfahren für eine Reduktion des relativen Fehlers auf höchstens 10^{-6} sind in Tabelle 5.1 abgebildet. In den Spalten ist die Anzahl der Verfeinerungen des groben Gitters angegeben; sie entsprechen 128, 32 und 8 überlappenden Gebieten. In den Zeilen ist die Anzahl der hinzugenommenen Dreiecksschichten dargestellt. Ein Block von jeweils vier Zeilen gehört zu einer der Verfeinerungsstufen m des feinen Gitters. Durch die Wahl von Zweierpotenzen in den Zeilen ist das Verhältnis von H zu δ auf den Diagonalen der einzelnen Blöcke konstant. Bei den mit ‚–‘ bezeichneten Einträgen in der Tabelle handelt es sich um die Fälle, in

	δ/h	$M=3$	$M=2$	$M=1$
$m=4$	1	6	7	6
	2	5	6	6
	4	–	4	4
	8	–	–	3
$m=5$	1	14	10	18
	2	12	8	12
	4	10	6	8
	8	–	6	5
$m=6$	1	16	16	29
	2	16	12	19
	4	13	11	13
	8	11	8	8
$m=7$	1	$*$	31	65
	2	$*$	23	44
	4	$*$	18	28
	8	$*$	15	17

Tabelle 5.1: Anzahl der Iterationsschritte zur Reduktion des relativen Fehlers auf höchstens 10^{-6} in Abhängigkeit von δ/h für $m = 4, 5, 6, 7$, $M = 1, 2, 3$.

denen die Überlappung so groß wäre, dass ein Teilgebiet nicht nur sein Ursprungsdreieck der groben Triangulierung, sondern auch noch weitere Dreiecke aus der groben Triangu-

lierung ganz enthalten würde. Dieser Fall wurde im Theorieteil ausgeschlossen und auch bei der Implementierung nicht umgesetzt.

Der Fall $m = 7$, $M = 3$ führt zu einem sehr hohen Speicher- und Rechenaufwand, weshalb entsprechende Tests nicht durchgeführt wurden und somit die zugehörigen Tabelleneinträge keine Werte enthalten. Diese Einträge sind in Tabelle 5.1 mit ‚$*$' gekennzeichnet.

In Kapitel 4 wurde gezeigt, dass $\|E_{mult}\|^2 \leq 1 - \frac{1}{C(1+\frac{H}{\delta})+c}$ ist.

In den Spalten eines Blocks von Tabelle 5.1 ist jeweils gut zu erkennen, dass je mehr Schichten hinzugenommen werden – das heißt je größer der Überlappungsparameter δ wird – die Anzahl der Iterationsschritte sinkt. Dies entspricht unseren Erwartungen, da die Konvergenzrate bei konstantem H kleiner wird, je größer δ wird.

In den Zeilen der Tabelle 5.1 wird bei konstantem Überlappungsparameter die Grobgittermaschenweite H größer; hier erwarten wir demnach ein schlechteres Konvergenzverhalten. Die erzielten Ergebnisse sind in diesem Punkt jedoch uneindeutig und lassen keine allgemeine Tendenz erkennen. Hier könnten weitere Tests mit mehr Verfeinerungen des groben Gitters unter Umständen eindeutigere Ergebnisse liefern; auf diese wurde jedoch verzichtet, weil die LR-Zerlegung in diesen Fällen nicht mehr effizient durchzuführen wäre.

Bei konstantem H/δ erwarten wir, dass die Anzahl der Iterationsschritte im Wesentlichen konstant bleibt. Wie oben bereits erwähnt, ist H/δ auf den Diagonalen der jeweiligen Blöcke in Tabelle 5.1 konstant. Um einen einfacheren Vergleich zu ermöglichen sind die Werte für konstantes H/δ in 5.2 nochmals anders zusammengestellt aufgetragen. Zusätzlich wurden noch sieben weitere Werte zur Vervollständigung der Tabelle berechnet, die beiden mit ‚$*$' bezeichneten Einträge konnten aufgrund des hohen Speicherbedarfs nicht bestimmt werden. Die kursiv dargestellten Werte entsprechen ihren Vorgängerwerten. Dies hat den Ursprung darin, dass nicht weniger als eine „Dreiecksschicht" hinzugenommen werden kann.

Vergleicht man die Werte für konstantes H/δ zunächst blockweise miteinander, ist zu erkennen, dass sie sich für $M = 1$ und $M = 2$ kaum unterscheiden, es aber ab $m = 5$ einen relativ großen Sprung zu $M = 3$ gibt. Diese Beobachtung legt nahe, dass für Werte von $M \in \{1, 2\}$ die obere Schranke des Fehlerfortpflanzungsoperators noch nicht erreicht wird. Eine denkbare Erklärung ist, dass bei $M = 1$ alle und bei $M = 2$ viele Dreiecke nicht die maximal mögliche Anzahl von Nachbarn besitzen. Da diese quadratisch in K_1 und linear in K_2 und damit in der vierten Potenz in unsere Abschätzung eingeht, kann sie

	H/δ	1	2	2^2	2^3	2^4	2^5	2^6
m=4	$M=1$	3	4	6	6	*6*	*6*	*6*
	$M=2$	4	6	7	*7*	*7*	*7*	*7*
	$M=3$	5	6	*6*	*6*	*6*	*6*	*6*
m=5	$M=1$	3	5	8	12	18	*18*	*18*
	$M=2$	6	6	8	10	*10*	*10*	*10*
	$M=3$	10	12	14	*14*	*14*	*14*	*14*
m=6	$M=1$	4	5	8	13	19	29	*29*
	$M=2$	7	8	11	12	16	*16*	*16*
	$M=3$	11	13	16	16	*16*	*16*	*16*
m=7	$M=1$	$*$	6	9	17	28	44	65
	$M=2$	$*$	10	15	18	23	31	*31*

Tabelle 5.2: Anzahl der Iterationsschritte zur Reduktion des relativen Fehlers auf höchstens 10^{-6} in Abhängigkeit von H/δ für $m = 4, 5, 6, 7$, $M = 1, 2, 3$.

bereits bei kleinen Werten einen relativ großen Effekt haben und damit die Konvergenz verschlechtern.

Beim Vergleich der konstanten H/δ-Bereiche zwischen den Blöcken stellen wir fest, dass die Werte für $M = 3$ relativ dicht zusammenliegen, ebenso diejenigen für $M \in \{1, 2\}$. Vergleicht man die Werte für gleiches H/δ und gleiches M, ist zu bemerken, dass diese für wachsendes m – das heißt kleineres h – steigen. Diese Steigerung fällt jedoch nicht besonders groß aus, was für eine von h unabhängige Konvergenzrate spricht. Weiterhin ist zu beobachten, dass die Anzahl der Iterationsschritte ansteigt, wenn H/δ wächst – das heißt, wenn wir die Spalten von Tabelle 5.1 miteinander vergleichen. Letzteres deckt sich wiederum mit unseren Erwartungen.

5.2.2 Numerische Resultate für den Einsatz als Vorkonditionierer für das cg-Verfahren

Auch für den Vorkonditionierer werden Tests durchgeführt. Dabei wird die Anzahl der Iterationsschritte, die für eine Verringerung des relativen Fehlers auf weniger als 10^{-3} benötigt wird, gemessen. Die daraus resultierenden Ergebnisse sind in Tabelle 5.3 abgebildet. Ihr Aufbau entspricht demjenigen aus dem vorangegangenen Abschnitt; zusätzlich ist die Anzahl der benötigten Schritte des cg-Verfahrens für die jeweilige feine Maschen-

weite zum Vergleich angegeben.

	δ/h	$M=3$	$M=2$	$M=1$
$m=4$	1	3	5	6
	2	3	4	5
	4	–	4	4
	8	–	–	2
	cg		27	
$m=5$	1	5	7	7
	2	4	6	6
	4	4	5	5
	8	–	5	3
	cg		53	
$m=6$	1	8	12	23
	2	7	9	21
	4	5	8	13
	8	5	6	9
	cg		96	
$m=7$	1	$*$	19	49
	2	$*$	14	48
	4	$*$	12	31
	8	$*$	10	17
	cg		174	

Tabelle 5.3: Anzahl der Iterationsschritte zur Reduktion des relativen Fehlers auf weniger als 10^{-3} mit dem vorkonditionierten cg-Verfahren in Abhängigkeit von δ/h für $m = 4, 5, 6, 7$, $M = 1, 2, 3$.

Wir betrachten die Diagonalen der Blöcke, auf denen H/δ konstant ist, also die Konvergenzrate dieselbe sein sollte. Dieses erwartete Verhalten ist in den einzelnen Blöcken gut zu erkennen. Die Sonderstellung von $M = 3$ aus dem vorangegangenen Abschnitt ist hier nicht sichtbar. Auch blockübergreifend, das heißt für unterschiedliches h, sind wenig Veränderungen an den Werten zu bemerken.

In den Spalten der Blöcke ist wiederum zu erkennen, dass bei größerer Überlappung die Konvergenzgeschwindigkeit wie erwartet ansteigt. In den Zeilen ist in diesem Fall zu sehen, dass die Konvergenzgeschwindigkeit für größeres H – wie in der Theorie gezeigt – geringer wird.

Beim Vergleich der Zahl der Iterationsschritte des vorkonditionierten und des normalen cg-Verfahrens ist zu erkennen, dass das vorkonditionierte Verfahren sehr viel weniger Schritte benötigt, um das gewünschte Resultat zu erzielen. Es ist also gelungen, das Spektrum der Systemmatrix geeignet zu transformieren. Bei der Verwendung des Vorkonditionierers ist allerdings zu beachten, dass der notwendige Vorbereitungsschritt Rechenzeit und Speicherplatz benötigt und dadurch der Einsatz des nicht vorkonditionierten Verfahrens an einigen Stellen sinnvoller ist.
Wie im vorangegangenen Abschnitt sind die Ergebnisse im Hinblick auf H/δ nochmals gesondert in einer weiteren Tabelle (vgl. Tabelle 5.4) susammengestellt.

	H/δ	1	2	2^2	2^3	2^4	2^5	2^6
m=4	$M=1$	2	4	5	6	*6*	*6*	*6*
	$M=2$	4	4	5	*5*	*5*	*5*	*5*
	$M=3$	3	3	*3*	*3*	*3*	*3*	*3*
m=5	$M=1$	2	3	5	6	7	*7*	*7*
	$M=2$	5	5	6	7	*7*	*7*	*7*
	$M=3$	4	4	5	*5*	*5*	*5*	*5*
m=6	$M=1$	2	2	9	13	21	23	*23*
	$M=2$	5	6	8	9	12	*12*	*12*
	$M=3$	5	5	7	8	*8*	*8*	*8*
m=7	$M=1$	$*$	6	10	17	31	48	49
	$M=2$	$*$	10	10	12	14	19	*19*

Tabelle 5.4: Anzahl der Iterationsschritte zur Reduktion des relativen Fehlers mit dem vorkonditionierten cg-Verfahren auf höchstens 10^{-3} in Abhängigkeit von H/δ für $m = 4, 5, 6, 7,\ M = 1, 2, 3$.

Beim Vergleich der Werte in den Spalten von Tabelle 5.4 sieht man in den einzelen m-Blöcken wenig Unterschiede bei der Anzahl der Iterationsschritte. Ein Sprung auf $M = 3$ wie beim eigenständigen Verfahren ist nicht zu erkennen. Auch blockübergreifend fallen die Unterschiede gering aus. Für steigendes H/δ ist zu erkennen, dass auch die Anzahl der Iterationsschritte steigt. Dies deckt sich mit unserer Erwartung einer schlechteren Vorkonditionierung.

6 Zusammenfassung und Ausblick

In dieser Arbeit wurden zweistufige multiplikative überlappende Gebietszerlegungsverfahren als eigenständige Verfahren und als Vorkonditionierer vorgestellt. In Kapitel 3 wurde unter gewissen Annahmen ein Konvergenzbeweis, sowie Schranken für das Spektrum des vorkonditionierten Operators gegeben. Es wurde gezeigt, dass die Konvergenzgeschwindigkeit bzw. die Schranken nicht von den Diskretisierungsparametern H, h abhängen, sondern nur vom Verhältnis der Maschenweite H zum Überlappungsparameter δ. Durch geeignete Wahl von δ kann also die Konvergenzrate beziehungsweise der Spektralbereich beliebig gewählt werden. In Kapitel 4 wurde eine konkrete Form der Gebietszerlegungsverfahren entwickelt. Für diese wurden die Annahmen unter Zuhilfenahme von Methoden aus dem Bereich der Finiten-Elemente nachgewiesen. Das entstandene Verfahren wurde implementiert und in numerischen Experimenten konnte das theoretisch vorhergesagte Verhalten wiedergefunden werden.

In diesen numerischen Experimenten bemerkte man schon bei den relativ kleinen Problemdimensionen den Aufwand von $\mathcal{O}(n^3)$ für eine LR-Zerlegung deutlich in der Laufzeit der Testprogramme. Deswegen wäre es interessant, die Annahmen nicht nur für direkte, sondern auch für approximative Löser auf den Unterräumen nachzuweisen und die Verfahren zu implementieren. Außerdem gibt es die Möglichkeit der Parallelisierung mit Hilfe von Färbungen, worauf aber in dieser Arbeit nicht eingegangen wurde.

Als weitere verwandte Verfahren können zudem noch mehrstufige Schwarz-Verfahren betrachtet werden, bei denen das Lösen auf den Teilräumen wiederum durch Unterraumkorrekturverfahren realisiert wird, oder auch additive überlappende Gebietszerlegungsverfahren untersucht werden.

Abbildungsverzeichnis

Tabellenverzeichnis

Liste der Algorithmen

Literaturverzeichnis

[Bör11] Börm, Steffen: *Vorlesungsskript - Iterative Verfahren für große Gleichungssysteme.* WS 2010 / 2011

[Bra07] Braess, Dietrich: *Finite Elemente.* Berlin: Springer-Verlag, 2007

[Bra12] Braack, Malte: *Vorlesungsskript - Finite Elemete Methode.* WS 2011 / 2012

[BX91] Bramble, James H. ; Xu, Jinchao: Some estimates for a weighted L^2 projection. In: *Mathematics of Computation* 56 (1991), Nr. 194, S. 463–476

[DW11] Deuflhard, Peter ; Weiser, Martin: *Numerische Mathematik 3. Adaptive Lösung partieller Differentialgleichungen.* Berlin: de Gruyter, 2011

[KK07] Koecher, Max ; Krieg, Aloys: *Ebene Geometrie.* Springer-Lehrbuch. Berlin: Springer., 2007

[SBG96] Smith, Barry ; Bjørstad, Peter ; Gropp, William: *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations.* Cambridge University Press, 1996

[Sch70] Schwarz, Hermann: Über einen Grenzübergang durch alternierendes Verfahren. In: *Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich* (1870), S. 272–286

[Ste03] Steinbach, Olaf: *Numerische Näherungsverfahren Für Elliptische Randwertprobleme: Finite Elemente und Randelemente.* Stuttgart: Vieweg+Teubner Verlag, 2003 (Advances in Numerical Mathematics)

[TW04] Toselli, Andrea ; Widlund, Olof: *Domain Decomposition Methods - Algorithms and Theory.* Bd. 34. Springer, 2004

[Wer07] Werner, Dirk: *Funktionalanalysis.* Springer, Limited, 2007 (Springer-Lehrbuch). – ISBN 9783540725367

[Xu92] Xu, Jinchao: Iterative methods by space decomposition and subspace correction. In: *SIAM Review* 34 (1992), December, Nr. 4, S. 581–613

[Yse93] Yserentant, Harry: Old and new convergence proofs for multigrid methods. In: *Acta Numerica* 2 (1993), S. 285–326

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe angefertigt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Weiterhin versichere ich, dass diese Arbeit noch nicht als Abschlussarbeit an anderer Stelle vorgelegen hat.

Kiel, den 31. Mai 2013