Konvergenzanalyse des Verfahrens der hierarchischen Basen

Masterarbeit

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

vorgelegt von

Markus Pfeil

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

Kiel im März 2014

Inhaltsverzeichnis

Abbildungsverzeichnis

Liste der Algorithmen

1 Einleitung

Anhand partieller Differentialgleichungen können viele physikalische Prozesse beschrieben werden. Da sich zumeist keine analytische Lösung einer partiellen Differentialgleichung angeben lässt, sind wir an der numerischen Berechnung einer approximativen Lösung interessiert. Die Finite Elemente Methode ist hierfür ein Verfahren, das die Berechnung einer solchen approximativen Lösung auf das Lösen eines linearen Gleichungssystems zurückführt. Weil die Anzahl der Unbekannten von einer Zerlegung des zugrundeliegenden Gebiets, auf dem die partielle Differentialgleichung definiert ist, abhängt, sind mehrere Tausend Unbekannte im zweidimensionalen und mehrere Millionen Unbekannte im dreidimensionalen Raum für eine hinreichend gute Approximation der Lösung nicht selten. Aufgrund der vielen Unbekannten und der im Allgemeinen schlechten Kondition dieser linearen Gleichungssysteme, ist der Einsatz von effizienten Lösungsverfahren von Interesse, da der Rechenaufwand von direkten Lösungsverfahren und klassischen Iterationsverfahren sehr hoch ist.

In dieser Arbeit behandeln wir das *Verfahren der hierarchischen Basen*, das ein effizientes Verfahren zum Lösen des linearen Gleichungssystems, welches aus der Diskretisierung einer partiellen Differentialgleichung im zweidimensionalen Raum durch die Methode der Finiten Elemente entsteht, darstellt und dessen Entwicklung auf [Yse86b] und [BDY88] beruht. Hierbei betrachten wir nur elliptische partielle Differentialgleichungen. Zudem stellen wir das Verfahren der hierarchischen Basen als Vorkonditionierer für das Verfahren der konjugierten Gradienten vor.

Das Verfahren der hierarchischen Basen stellen wir, wie in [Xu92] und [Yse93] beschrieben, als Unterraumkorrekturverfahren dar, wobei sich klassische lineare Iterationsverfahren wie das Jacobi-Verfahren, Gauß-Seidel-Verfahren, Mehrgitterverfahren oder Gebietszerlegungsverfahren ebenfalls als Unterraumkorrekturverfahren auffassen lassen. Diese Verfahren unterscheiden sich unter anderem durch die Wahl der jeweiligen Unterräume. Für das Verfahren der hierarchischen Basen verwenden wir Unterräume, die sich mithilfe der hierarchischen Basis ergeben. Um die Konvergenz dieses Verfahrens zu zeigen, entwickeln wir eine allgemeine Konvergenztheorie für Unterraumkorrekturverfahren, die auf drei Annahmen beruht.

Das Kapitel 2 enthält einerseits die Darstellung von verwendeten Grundlagen und andererseits eine kurze Einführung linearer Iterationsverfahren, um lineare Gleichungssysteme approximativ zu lösen. Die Methode der Finiten Elemente zur Berechnung einer Näherungslösung einer elliptischen partiellen Differentialgleichung wird in Kapitel 3 beschrieben. In Kapitel 4 befassen wir uns mit der allgemeinen Theorie der Unterraumkorrekturverfahren und stellen das multiplikative sowie das additive Unterraumkorrekturverfahren als zwei Varianten der Unterraumkorrekturverfahren vor. Die Konvergenztheorie für Unterraumkorrekturverfahren, mit der wir uns in Kapitel 5 beschäftigen, beruht im Wesentlichen auf drei Annahmen, die wir in Kapitel 5 vorstellen. Die Darstellung des Verfahrens der hierarchischen Basen, das ein Unterraumkorrekturverfahren ist, befindet sich mit dem Konvergenzbeweis für dieses Verfahren in Kapitel 6. Der Beweis der Konvergenz folgt, indem wir die drei Annahmen der Konvergenztheorie für Unterraumkorrekturverfahren zeigen. Eine Implementierung des Verfahrens mit den numerischen Resultaten zeigen wir in Kapitel 7.

2 Grundlagen

In diesem Kapitel stellen wir in Abschnitt 2.1 grundlegende Definitionen und Aussagen vor, die wir im weiteren Verlauf dieser Arbeit benötigen. Zudem geben wir in Kapitel 2.2 eine kurze Einführung in lineare Iterationsverfahren, um eine Approximation der Lösung eines linearen Gleichungssystems zu berechnen.

2.1 Theoretische Grundlagen

In diesem Kapitel sei X ein normierter Raum über $\mathbb{R}$ mit der Norm $\|\cdot\|_X$. Zudem sei $d \in \{1, 2, 3\}$. Die euklidische Norm im $\mathbb{R}^d$ bezeichnen wir mit $\|\cdot\|_2$.

Definition 2.1 (Volumen)**.** *Für eine messbare Menge $\Omega \subset \mathbb{R}^d$ ist das* Volumen *von Ω definiert durch*

$$|\Omega| := \int_\Omega dx.$$

Definition 2.2 (Durchmesser)**.** *Für $\Omega \subset \mathbb{R}^d$ ist der* Durchmesser *von Ω definiert als*

$$\operatorname{diam}(\Omega) := \sup\left\{\|x - y\|_2 : x, y \in \Omega\right\}.$$

Definition 2.3 (Träger)**.** *Sei $\Omega \subset \mathbb{R}^d$. Für eine Abbildung $f : \Omega \to \mathbb{R}$ heißt*

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

der Träger *von f.*

Definition 2.4 (Eigenwert)**.** *Sei V ein $\mathbb{R}$-Vektorraum und $f : V \to V$ eine lineare Abbildung. Dann heißt $\lambda \in \mathbb{C}$* Eigenwert *von f, wenn $v \in V\backslash\{0\}$ mit $f(v) = \lambda v$ existiert. Solch ein Vektor v wird als* Eigenvektor *zum Eigenwert λ bezeichnet. Die Menge $\sigma(f) := \{\lambda \in \mathbb{C} : \exists v \in V\backslash\{0\} : f(v) = \lambda v\}$ heißt* Spektrum *von f und $\varrho(f) := \max\{|\lambda| : \lambda \in \sigma(f)\}$* Spektralradius *von f.*

Definition 2.5 (Dualraum)**.** *Für zwei normierte Räume X und Y über $\mathbb{R}$ bezeichnet $\mathcal{L}(X,Y)$ den Raum der stetigen linearen Abbildungen zwischen X und Y. Für einen normierten Raum $(X, \|\cdot\|_X)$ über $\mathbb{R}$ heißt der Raum $\mathcal{L}(X,\mathbb{R})$ der* Dualraum *von X und wird mit X' bezeichnet. Durch*

$$\|\cdot\|_{X'} : X' \to \mathbb{R}, f \mapsto \sup_{x \in X\setminus\{0\}} \frac{|f(x)|}{\|x\|_X}$$

ist eine Norm auf X' definiert.

Auf dem normierten Raum X heißt die Bilinearform

$$(\cdot,\cdot) : X' \times X \to \mathbb{R}, (f,x) \mapsto f(x)$$

das *Dualitätsprodukt.*

Definition 2.6 (Selbstadjungiert)**.** *Ein stetiger linearer Operator $A : X \to X'$ heißt* selbstadjungiert*, falls $(Au, v) = (Av, u)$ für alle $u, v \in X$ gilt.*

Im Folgenden verwenden wir $(u, Av) := (Av, u)$ für alle $u, v \in X$ und einem stetigen linearen Operator $A : X \to X'$.

Definition 2.7 (Positiv definit)**.** *Ein selbstadjungierter Operator $A : X \to X'$ heißt* positiv definit*, falls $(Au, u) > 0$ für alle $u \in X\setminus\{0\}$ gilt.*

Definition 2.8 (Energienorm)**.** *Sei $A : X \to X'$ ein positiv definiter Operator. Das durch*

$$(u,v)_A := (Au, v)$$

für alle $u, v \in X$ definierte Skalarprodukt heißt das zu A gehörende Energieskalarprodukt*. Dieses Skalarprodukt erzeugt eine Norm $\|\cdot\|_A := (\cdot,\cdot)_A^{1/2}$ auf X, die wir als* Energienorm *bezeichnen.*

Lemma 2.9 (Cauchy-Schwarz-Ungleichung)**.** *Sei $A : X \to X'$ ein positiv definiter Operator. Dann gilt für alle $u, v \in X$ die Cauchy-Schwarz-Ungleichung*

$$(u,v)_A \leq \|u\|_A \cdot \|v\|_A .$$

Nach der Einführung dieser grundlegenden Begriffe und Notationen sind die folgenden drei Aussagen in dieser Arbeit von Interesse.

Satz 2.10. *Für jede Matrix $M \in \mathbb{R}^{n\times n}$ mit $n \in \mathbb{N}$ und jeder induzierten Matrixnorm $\|\cdot\|$ gilt die Ungleichung $\varrho(M) \leq \|M\|$.*

Beweis. Siehe [Hac93, Lemma 2.9.1]. □

Satz 2.11 (Parallelogrammgleichung)**.** *Sei V ein Prähilbertraum. Für alle $u, v \in V$ gilt die Parallelogrammgleichung*

$$\|u+v\|_V^2 + \|u-v\|_V^2 = 2\left(\|u\|_V^2 + \|v\|_V^2\right). \tag{2.1}$$

Beweis. Siehe [Wer07, Satz V.1.7]. □

Satz 2.12 (Greensche Formel)**.** *Es sei $\Omega \subset \mathbb{R}^d$ beschränkt. Dann gilt für Funktionen $u, v \in C^2(\Omega)$ die Greensche Formel*

$$\int_\Omega \langle \nabla u, \nabla v\rangle_2 \, dx = -\int_\Omega u\, \Delta v \, dx + \int_{\partial\Omega} u \, \langle n, \nabla v\rangle_2 \, dx, \tag{2.2}$$

wobei n der äußere Normaleneinheitsvektor zu $\partial\Omega$ ist.

Beweis. Siehe [DW11, Satz A.5]. □

Die folgenden Aussagen über Dreiecke benötigen wir für Triangulierungen eines Gebiets $\Omega \subset \mathbb{R}^2$, die bei der Finite Elemente Methode auftreten.

Definition 2.13. *$A, B \subset \mathbb{R}^d$ heißen* kongruent*, falls eine Isometrie $f : \mathbb{R}^d \to \mathbb{R}^d$, also eine Abbildung mit $\|x-y\|_2 = \|f(x)-f(y)\|_2$ für alle $x, y \in A$, so existiert, dass $f(A) = B$ gilt.*

Zwei Mengen $A, B \subset \mathbb{R}^d$ heißen ähnlich*, wenn eine Abbildung $f : \mathbb{R}^d \to \mathbb{R}^d$ mit $f(A) = B$ existiert, die Hintereinanderausführung von Isometrien und Streckungen ist.*

Satz 2.14 (Kongruenzsatz für Dreiecke)**.** *Zwei Dreiecke sind kongruent, falls in einer der folgenden Aussagen die Größen beider Dreiecke übereinstimmen:*

(1) Die Längen ihrer drei Seiten.

(2) Die Längen zweier Seiten und der von ihnen eingeschlossene Winkel.

(3) Die Längen zweier Seiten und der der längeren Seite gegenüberliegende Winkel.

(4) Die Länge einer Seite und die beiden daran anliegenden Winkel.

Beweis. Siehe [AF11, Satz 15]. □

Satz 2.15 (Ähnlichkeitssatz für Dreiecke)**.** *Zwei Dreiecke sind ähnlich, falls in einer der folgenden Aussagen die Größen beider Dreiecke übereinstimmen:*

(1) Die Verhältnisse der Längen ihrer Seiten.

(2) Die Verhältnisse der Längen zweier Seiten und der von ihnen eingeschlossene Winkel.

(3) Die Verhältnisse der Längen zweier Seiten und die der größeren Seite gegenüberliegenden Winkel.

(4) Zwei Winkel.

Beweis. Siehe [AF11, Satz 17]. □

Satz 2.16. *Der Radius ρ_T des Inkreises eines Dreiecks T mit den Seiten a, b und c ist*

$$\rho_T = \frac{2 \cdot |T|}{a + b + c}. \tag{2.3}$$

Beweis. Siehe [AF11, Satz 27]. □

2.2 Lineare Iterationsverfahren

In diesem Kapitel sei $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, $\mathcal{J}$ eine endliche Indexmenge und $A \in \mathbb{K}^{\mathcal{J} \times \mathcal{J}}$ eine reguläre Matrix. Wir definieren allgemeine lineare Iterationsverfahren, um eine beliebig genaue Approximation der Lösung $x \in \mathbb{K}^{\mathcal{J}}$ des linearen Gleichungssystems

$$Ax = b \tag{2.4}$$

mit einer rechten Seite $b \in \mathbb{K}^{\mathcal{J}}$ zu berechnen. Die Einführung dieser Verfahren orientiert sich an [Hac93].

Definition 2.17 (Iterationsverfahren)**.** *Eine Abbildung*

$$\Phi : \mathbb{K}^{\mathcal{J}} \times \mathbb{K}^{\mathcal{J}} \to \mathbb{K}^{\mathcal{J}},$$

die im ersten Argument stetig ist, heißt Iterationsverfahren*. Für $x^{(0)}, b \in \mathbb{K}^{\mathcal{J}}$ bezeichnet die Folge $\left(x^{(m)}\right)_{m \in \mathbb{N}_0}$, die durch*

$$x^{(m)} := \Phi\left(x^{(m-1)}, b\right)$$

für alle $m \in \mathbb{N}$ definiert ist, die Folge der Iterierten *des Iterationsverfahrens.*

Ein Iterationsverfahren Φ berechnet aus einer gegebenen Iterierten und der rechten Seite b eine neue Iterierte. Falls wir ein Iterationsverfahren mit der Lösung des Gleichungssystems (2.4) ausführen, soll das Ergebnis wieder diese Lösung sein. Mithin soll die Lösung ein Fixpunkt des Iterationsverfahrens sein.

Definition 2.18 (Fixpunkt)**.** *Ein Vektor $x^* \in \mathbb{K}^{\mathcal{J}}$ heißt* Fixpunkt *eines Iterationsverfahrens Φ zu einem Vektor $b \in \mathbb{K}^{\mathcal{J}}$, falls $\Phi(x^*, b) = x^*$ gilt.*

Definition 2.19 (Konsistenz)**.** *Ein Iterationsverfahren Φ heißt* konsistent*, falls die Lösung x^* des Gleichungssystems* (2.4) *ein Fixpunkt von Φ zu b ist, also die Gleichung $\Phi(x^*, b) = x^*$ erfüllt.*

Um mit einem Iterationsverfahren eine beliebig genaue Approximation der Lösung des Gleichungssystems zu berechnen, muss die Folge der Iterierten konvergieren.

Definition 2.20 (Konvergenz)**.** *Ein Iterationsverfahren Φ heißt* konvergent*, falls für alle $b \in \mathbb{K}^{\mathcal{J}}$ ein $x^* \in \mathbb{K}^{\mathcal{J}}$ so existiert, dass für jeden Startvektor $x^{(0)} \in \mathbb{K}^{\mathcal{J}}$ die Folge $\left(x^{(m)}\right)_{m \in \mathbb{N}_0}$ der Iterierten gegen den Grenzwert x^* konvergiert.*

Lemma 2.21. *Sei Φ ein konsistentes und konvergentes Iterationsverfahren und $b \in \mathbb{K}^{\mathcal{J}}$. Dann konvergiert die Folge $\left(x^{(m)}\right)_{m \in \mathbb{N}_0}$ der Iterierten zu jedem beliebigen Startvektor $x^{(0)} \in \mathbb{K}^{\mathcal{J}}$ gegen die Lösung $x^* \in \mathbb{K}^{\mathcal{J}}$ des Gleichungssystems* (2.4)*.*

Beweis. Siehe [Hac93, Zusatz 3.2.8]. □

Eine besondere Klasse von Iterationsverfahren sind lineare Iterationsverfahren, die zum Lösen linearer Gleichungssysteme verwendet werden.

Definition 2.22 (Lineares Iterationsverfahren)**.** *Ein Iterationsverfahren Φ heißt* linear*, falls es Matrizen $M, N \in \mathbb{K}^{\mathcal{J} \times \mathcal{J}}$ so gibt, dass*

$$\Phi(x, b) = Mx + Nb$$

für alle $x, b \in \mathbb{K}^{\mathcal{J}}$ gilt. Die Matrix M heißt Iterationsmatrix *und die Darstellung des Iterationsverfahren* erste Normalform*.*

Für lineare Iterationsverfahren geben wir im Folgenden jeweils ein Kriterium an, um die Konsistenz und Konvergenz dieser Verfahren zu charakterisieren.

Lemma 2.23. *Sei Φ ein lineares Iterationsverfahren und $M, N \in \mathbb{K}^{\mathcal{J}\times\mathcal{J}}$ die Matrizen der ersten Normalform. Φ ist genau dann konsistent, wenn $M = I - NA$ gilt. In diesem Fall lässt sich das Iterationsverfahren in der* zweiten Normalform

$$\Phi(x, b) = x - N(Ax - b)$$

für alle $x, b \in \mathbb{K}^{\mathcal{J}}$ darstellen.

Beweis. Siehe [Hac93, Satz 3.2.2]. □

Satz 2.24. *Sei Φ ein lineares Iterationsverfahren mit der Iterationsmatrix $M \in \mathbb{K}^{\mathcal{J}\times\mathcal{J}}$. Φ ist genau dann konvergent, wenn $\varrho(M) < 1$ für den Spektralradius von M gilt.*

Beweis. Siehe [Hac93, Satz 3.2.7]. □

2.2.1 Jacobi-Verfahren

Als erstes lineares Iterationsverfahren führen wir die Jacobi-Iteration ein, wobei die Inverse der Diagonalen der Matrix A als Matrix N der ersten Normalform verwendet wird. Dazu setzen wir voraus, dass alle Diagonaleinträge der Matrix A von Null verschieden sind.

Definition 2.25 (Gedämpfte Jacobi-Iteration)**.** *Sei $\theta \in \mathbb{K}$ und $D \in \mathbb{K}^{\mathcal{J}\times\mathcal{J}}$ die Diagonale von A, die durch $D := \operatorname{diag}(A)$ definiert ist, und sei D invertierbar, also jedes Diagonalelement von A von Null verschieden. Das durch*

$$\Phi_{Jac,\theta}(x, b) := x - \theta D^{-1}(Ax - b)$$

für alle $x, b \in \mathbb{K}^{\mathcal{J}\times\mathcal{J}}$ gegebene lineare Iterationsverfahren heißt gedämpfte Jacobi-Iteration *mit Dämpfungsparameter θ.*

Weil die Jacobi-Iteration in zweiter Normalform gegeben ist, ist dieses Verfahren nach Lemma 2.23 konsistent. Für $\theta = 1$ erhalten wir die ungedämpfte Jacobi-Iteration. Die Konvergenz der gedämpften Jacobi-Iteration hängt von der Wahl des Dämpfungsparameters θ ab.

Lemma 2.26 (Konvergenz)**.** *Sei A positiv definit. Falls $\frac{2}{\theta}D - A$ für ein $\theta \in \mathbb{R}_{>0}$ positiv definit ist, ist die gedämpfte Jacobi-Iteration $\Phi_{Jac,\theta}$ konvergent. Es gibt ein $\theta_{\max} \in \mathbb{R}_{>0}$, sodass die gedämpfte Jacobi-Iteration für alle $\theta \in (0, \theta_{\max})$ konvergent ist.*

Beweis. Siehe [Hac93, Satz 4.4.14]. □

2.2.2 Gauß-Seidel-Verfahren

Als zweites lineares Iterationsverfahren führen wir die Gauß-Seidel-Iteration ein, bei der die Inverse der unteren Dreiecksmatrix von A als Matrix N der ersten Normalform verwendet wird. Dazu sei $\iota : \mathcal{J} \to \{1, \ldots, n\}$ eine Bijektion mit $n := |\mathcal{J}|$. Die Matrix A zerlegen wir in die Diagonalmatrix $D \in \mathbb{K}^{\mathcal{J}\times\mathcal{J}}$, die strikte untere Dreiecksmatrix $E \in \mathbb{K}^{\mathcal{J}\times\mathcal{J}}$ und die strikte obere Dreiecksmatrix $F \in \mathbb{K}^{\mathcal{J}\times\mathcal{J}}$, sodass

$$A = D - E - F \tag{2.5}$$

mit

$$D_{ij} := \begin{cases} A_{ii}, & i = j \\ 0, & \text{sonst} \end{cases}$$

$$E_{ij} := \begin{cases} -A_{ij}, & \iota(i) > \iota(j), \\ 0, & \text{sonst} \end{cases}$$

$$F_{ij} := \begin{cases} -A_{ij}, & \iota(i) < \iota(j), \\ 0, & \text{sonst} \end{cases}$$

für alle $i, j \in \mathcal{J}$ gilt.

Definition 2.27 (Gauß-Seidel-Iteration)**.** *Seien $D, E, F \in \mathbb{K}^{\mathcal{J}\times\mathcal{J}}$ mit $A = D - E - F$ wie in* (2.5) *gegeben und sei D invertierbar, das heißt alle Diagonalelemente von A sind von Null verschieden. Das durch*

$$\Phi_{GS}(x, b) := x - (D - E)^{-1}(Ax - b)$$

gegebene lineare Iterationsverfahren heißt Gauß-Seidel-Iteration*.*

Mit den Matrizen $M_{GS} := I - (D - E)^{-1} A$ und $N_{GS} := (D - E)^{-1}$ der ersten Normalform der Gauß-Seidel-Iteration ist dieses Verfahren nach Lemma 2.23 konsistent.

Lemma 2.28 (Konvergenz)**.** *Sei A positiv definit. Dann ist die Gauß-Seidel-Iteration konvergent.*

Beweis. Siehe [Hac93, Satz 4.4.18]. □

Die Gauß-Seidel-Iteration konvergiert im Gegensatz zur Jacobi-Iteration für jede positiv definite Matrix und es ist nicht erforderlich, einen Dämpfungsparameter zu verwenden.

2.2.3 Verfahren der konjugierten Gradienten

Das *Verfahren der konjugierten Gradienten* oder kurz *cg-Verfahren* ist ein semiiteratives Verfahren zur Lösung linearer Gleichungssysteme mit einer positiv definiten Matrix $A \in \mathbb{K}^{\mathcal{J}\times\mathcal{J}}$, wobei ein semiiteratives Verfahren eine Abbildung

$$\Sigma : \left(\bigcup_{m\in\mathbb{N}_0} \left(\mathbb{K}^{\mathcal{J}}\right)^{m+1} \right) \to \mathbb{K}^{\mathcal{J}}$$

ist. Das heißt, dass bei der Berechnung einer neuen Iterierten des semiiterativen Verfahrens die $m+1$ vorherigen Iterierten anstatt wie bei linearen Iterationsverfahren nur die vorherige Iterierte eingehen. Für eine genaue Einführung semiiterativer Verfahren sei auf [Hac93, Kapitel 7] verwiesen. Weil $x \in \mathbb{K}^{\mathcal{J}}$ genau dann eine Lösung des linearen Gleichungssystems (2.4) ist, wenn x die Funktion

$$f : \mathbb{K}^{\mathcal{J}} \to \mathbb{R}, x \mapsto \frac{1}{2}\langle Ax, x\rangle_2 - \langle b, x\rangle_2$$

minimiert (siehe [Hac93, Lemma 9.1.1]), bietet sich als Suchrichtung $p \in \mathbb{K}^{\mathcal{J}}\backslash\{0\}$ das Residuum $r := b - Ax \in \mathbb{K}^{\mathcal{J}}$ mit der optimalen Schrittweite $\lambda_{\text{opt}} = \frac{\langle r,p\rangle_2}{\langle Ap,p\rangle_2}$ (siehe [Hac93, Lemma 9.1.2]) an. Da die Optimalität bezüglich einer Suchrichtung im folgenden Schritt wieder verloren gehen kann, empfiehlt es sich, die neue Suchrichtung auf allen vorherigen Suchrichtungen bezüglich des Energieskalarprodukts zu orthogonalisieren, um die Optimalität bezüglich der vorherigen Suchrichtungen zu erhalten. Damit ergibt sich insgesamt das folgende Verfahren der konjugierten Gradienten.

Definition 2.29 (cg-Verfahren)**.** *Es sei $A \in \mathbb{K}^{\mathcal{J}\times\mathcal{J}}$ positiv definit und $x^{(0)}, b \in \mathbb{K}^{\mathcal{J}}$ gegeben. Die für alle $m \in \mathbb{N}_0$ durch*

$$\begin{aligned}
r^{(m)} &:= b - Ax^{(m)}, \\
p^{(m)} &:= \begin{cases} r^{(0)}, & \text{falls } m = 0, \\ r^{(m)} - \frac{\langle r^{(m)}, Ap^{(m-1)}\rangle_2}{\langle p^{(m-1)}, Ap^{(m-1)}\rangle_2} p^{(m-1)}, & \text{falls } m > 0 \text{ und } p^{(m-1)} \neq 0, \\ 0, & \text{sonst} \end{cases} \\
x^{(m+1)} &:= x^{(m)} + \frac{\langle p^{(m)}, r^{(m)}\rangle_2}{\langle p^{(m)}, Ap^{(m)}\rangle_2} p^{(m)}
\end{aligned}$$

definierte Folge $\left(x^{(m)}\right)_{m\in\mathbb{N}_0}$ bezeichnen wir als Folge der Semiiterierten des Verfahrens der konjugierten Gradienten.

Weil die Schrittweiten für jede Suchrichtung optimal gewählt werden und alle Suchrichtungen orthogonal zu einander sind, berechnet das Verfahren der konjugierten Gradienten spätestens nach $m_0 := \dim\left(\mathbb{K}^{\mathcal{J}}\right)$ Schritten die exakte Lösung, wie in [Hac93, Satz 9.4.2] gezeigt wurde.

Satz 2.30 (Konvergenz)**.** *Es sei $A \in \mathbb{K}^{\mathcal{J}\times\mathcal{J}}$ positiv definit und $\alpha, \beta \in \mathbb{R}_{>0}$ mit $\sigma(A) \subseteq [\alpha, \beta]$. Weiter sei $\left(x^{(m)}\right)_{m\in\mathbb{N}_0}$ die Folge der Semiiterierten des Verfahrens der konjugierten Gradienten zu einem Startvektor $x^{(0)} \in \mathbb{K}^{\mathcal{J}}$ und $b \in \mathbb{K}^{\mathcal{J}}$. Mit $x^* := A^{-1}b$ gilt für alle $m \in \mathbb{N}_0$*

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

mit $c := \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ und $\kappa := \frac{\beta}{\alpha}$.

Beweis. Siehe [Hac93, Satz 9.4.12]. □

Das Verfahren der konjugierten Gradienten kann beschleunigt werden, indem das lineare Gleichungssystem mit einer positiv definiten Matrix derart vorkonditioniert wird, dass die vorkonditionierte Matrix eine kleinere Konditionszahl besitzt, womit sich die Anzahl der benötigten Iterationen reduziert.

Definition 2.31 (Vorkonditioniertes cg-Verfahren)**.** *Es seien $A, N \in \mathbb{K}^{\mathcal{J}\times\mathcal{J}}$ positiv definit und $x^{(0)}, b \in \mathbb{K}^{\mathcal{J}}$ gegeben. Die für alle $m \in \mathbb{N}_0$ durch*

$$\begin{aligned}
r^{(m)} &:= b - Ax^{(m)},\\
q^{(m)} &:= Nr^{(m)},\\
p^{(m)} &:= \begin{cases} q^{(0)}, & \text{falls } m = 0,\\ q^{(m)} - \frac{\langle q^{(m)}, Ap^{(m-1)}\rangle_2}{\langle p^{(m-1)}, Ap^{(m-1)}\rangle_2} p^{(m-1)}, & \text{falls } m > 0 \text{ und } p^{(m-1)} \neq 0,\\ 0, & \text{sonst} \end{cases}\\
x^{(m+1)} &:= x^{(m)} + \frac{\langle p^{(m)}, r^{(m)}\rangle_2}{\langle p^{(m)}, Ap^{(m)}\rangle_2} p^{(m)}
\end{aligned}$$

definierte Folge $\left(x^{(m)}\right)_{m\in\mathbb{N}_0}$ bezeichnen wir als die Folge der Semiiterierten des vorkonditionierten Verfahrens der konjugierten Gradienten*. Die Matrix N heißt in diesem Kontext* Vorkonditionierer*.*

Satz 2.32 (Konvergenz)**.** *Es seien $A, N \in \mathbb{K}^{\mathcal{J}\times\mathcal{J}}$ positiv definit und $\alpha, \beta \in \mathbb{R}_{>0}$ mit $\sigma(NA) \subseteq [\alpha, \beta]$. Weiter sei $\left(x^{(m)}\right)_{m\in\mathbb{N}_0}$ die Folge der Semiiterierten des vorkonditio-*

nierten Verfahrens der konjugierten Gradienten zu einem Startvektor $x^{(0)} \in \mathbb{K}^{\mathcal{J}}$ *und* $b \in \mathbb{K}^{\mathcal{J}}$*. Mit* $x^* := A^{-1}b$ *gilt für alle* $m \in \mathbb{N}_0$

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

mit $c := \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ *und* $\kappa := \frac{\beta}{\alpha}$*.*

Beweis. Siehe [Hac93, Satz 9.4.14]. □

2.2.4 Mehrgitterverfahren

Die Idee der Mehrgitterverfahren besteht aus der Kombination zweier Verfahren, um sowohl die hochfrequenten als auch die niedrigfrequenten Anteile des Fehlers bei der Lösung eines Gleichungssystems zu reduzieren. Wie in [DW11, Kapitel 5.4] dargestellt ist, glätten klassische Iterationsverfahren den Gesamtfehler, das heißt, dass hochfrequente Anteile des Fehlers schnell reduziert werden. Deshalb bieten sich als *Glättungsverfahren* beispielsweise das Jacobi-Verfahren oder das Gauß-Seidel-Verfahren an. Für die Reduktion der niedrigfrequenten Fehleranteile besteht die Möglichkeit, diese auf einem gröberen Gitter zu approximieren. Dazu sei eine Hierarchie von Indexmengen $\mathcal{J}_\ell$ und Gleichungssystemen mit regulären Matrizen $A_\ell \in \mathbb{K}^{\mathcal{J}_\ell \times \mathcal{J}_\ell}$ und rechten Seiten $b_\ell \in \mathbb{K}^{\mathcal{J}_\ell}$ für $\ell \in \mathbb{N}_0$ gegeben. Die Gleichungssysteme für $\ell \in \mathbb{N}_0$ gehören zu verschiedenen feinen Auflösungen derselben zugrundeliegenden partiellen Differentialgleichung, wobei das gegebene Gleichungssystem $Ax = b$ einer Gitterstufe $m \in \mathbb{N}_0$ entspricht. Um dieses Gleichungssystem $A_m x_m = b_m$ mithilfe der Gitterstufen $\ell \in \mathbb{N}_0$ mit $\ell < m$ zu lösen, benötigen wir Transferoperatoren zwischen den einzelnen Gittern.

Definition 2.33 (Gittertransfer)**.** *Sei* $\ell \in \mathbb{N}$*. Eine injektive Matrix* $p_\ell \in \mathbb{K}^{\mathcal{J}_\ell \times \mathcal{J}_{\ell-1}}$ *bezeichnen wir als* Prolongation *und eine surjektive Matrix* $r_\ell \in \mathbb{K}^{\mathcal{J}_{\ell-1} \times \mathcal{J}_\ell}$ *als* Restriktion*.*

Definition 2.34 (Galerkin-Eigenschaft)**.** *Falls für alle* $\ell \in \mathbb{N}$ *die Gleichung*

$$A_{\ell-1} = r_\ell A_\ell p_\ell$$

gilt, besitzt die Hierarchie der Gleichungssysteme die Galerkin-Eigenschaft*.*

Um das Gleichungssystem auf einer Stufe $\ell \in \mathbb{N}$ zu lösen, führen wir einige Schritte eines Glättungsverfahrens mit dem Ergebnis $\bar{x}_\ell \in \mathbb{K}^{\mathcal{J}_\ell}$ durch, um hochfrequente Fehleranteile zu reduzieren. Aus dem zugehörigen Fehler $f_\ell := \bar{x}_\ell - x_\ell$ ergibt sich durch

Algorithmus 1 : Mehrgitterverfahren

Eingabe : ℓ, b_ℓ, x_ℓ, ν_1, ν_2, γ
Ausgabe : Näherungslösung x_ℓ
if $\ell > 0$ **then**
  for $i = 1$ **to** ν_1 **do**
    $x_\ell \leftarrow \Phi_{\mathrm{Gl},\ell}(x_\ell, b_\ell)$
  $d_\ell \leftarrow A_\ell x_\ell - b_\ell$
  $b_{\ell-1} \leftarrow r_\ell d_\ell$
  $x_{\ell-1} \leftarrow 0$
  for $i = 1$ **to** γ **do**
    $x_{\ell-1} \leftarrow$ Mehrgitterverfahren $(\ell - 1, b_{\ell-1}, x_{\ell-1}, \nu_1, \nu_2, \gamma)$
  $x_\ell \leftarrow x_\ell - p_\ell x_{\ell-1}$
  for $i = 1$ **to** ν_2 **do**
    $x_\ell \leftarrow \Phi_{\mathrm{Gl},\ell}(x_\ell, b_\ell)$
else
  $x_\ell \leftarrow A_\ell^{-1} b_\ell$

$x_\ell = \bar{x}_\ell - f_\ell$ die exakte Lösung, wobei f_ℓ die Gleichung

$$A_\ell f_\ell = A_\ell(\bar{x}_\ell - x_\ell) = A_\ell \bar{x}_\ell - A_\ell x_\ell = A_\ell \bar{x}_\ell - b_\ell = d_\ell$$

mit dem Defekt $d_\ell := A_\ell \bar{x}_\ell - b_\ell$ erfüllt. Weil der Fehler f_ℓ nach den Glättungsschritten hinreichend glatt ist, approximieren wir f_ℓ durch $p_\ell f_{\ell-1}$, wobei $f_{\ell-1}$ die Lösung des Gleichungssystems $A_{\ell-1} f_{\ell-1} = d_{\ell-1}$ mit der rechten Seite $d_{\ell-1} := r_\ell d_\ell$ ist. Insgesamt erhalten wir somit das folgende Verfahren zur Reduktion niedrigfrequenter Fehleranteile.

Definition 2.35 (Grobgitterkorrektur)**.** *Es sei $\ell \in \mathbb{N}$. Das für alle $x_\ell, b_\ell \in \mathbb{K}^{\mathcal{J}_\ell}$ durch*

$$\Phi_{GGK,\ell}(x_\ell, b_\ell) = x_\ell - p_\ell A_{\ell-1}^{-1} r_\ell (A_\ell x_\ell - b_\ell)$$

definierte lineare Iterationsverfahren heißt Grobgitterkorrektur *auf der Gitterstufe ℓ.*

Da die Grobgitterkorrektur, wie in [Hac93, Bemerkung 10.1.5] gezeigt, zwar konsistent aber nicht konvergent ist, erhalten wir erst durch Kombination der Grobgitterkorrektur mit einem Glättungsverfahren ein effizientes Verfahren.

Definition 2.36 (Zweigitterverfahren)**.** *Es sei $\ell \in \mathbb{N}$ und $\Phi_{Gl,\ell}$ ein lineares Iterationsverfahren für die Matrix A_ℓ. Das für alle $x_\ell, b_\ell \in \mathbb{K}^{\mathcal{J}_\ell}$ durch*

$$\Phi_{ZGV,\ell}(x_\ell, b_\ell) = \Phi_{GGK,\ell}(\Phi_{Gl,\ell}(x_\ell, b_\ell), b_\ell)$$

definierte lineare Iterationsverfahren heißt Zweigitterverfahren *auf der Gitterstufe ℓ mit dem Glättungsverfahren (oder kurz Glätter) $\Phi_{Gl,\ell}$.*

Das Zweigitterverfahren ist nach [Hac93, Lemma 10.2.1] konsistent. Weil das Glättungsverfahren den Fehler in einem Schritt nur um einen gewissen Faktor reduziert, genügt es auf dem groben Gitter nur eine Approximation der Lösung zu berechnen. Dazu verwenden wir als Iteration das Zweigitterverfahren auf den Gitterstufen $\ell - 1$ und $\ell - 2$. Wenn wir in dieser Weise rekursiv vorgehen, bis die Gitterstufe 0 erreicht ist, erhalten wir das in Algorithmus 1 angegebene Mehrgitterverfahren mit $\nu_1 = 1$, $\nu_2 = 0$ und $\gamma = 1$. Die Parameter $\nu_1, \nu_2 \in \mathbb{N}_0$ geben die Anzahl der *Vor-* und *Nachglättungsschritte* an. Zudem gibt der Parameter $\gamma \in \mathbb{N}$ die Anzahl der zur Approximation der Grobgitterlösung verwendeten Mehrgitterschritte an. Das Mehrgitterverfahren mit $\gamma = 1$ heißt *V-Zyklus* und jenes mit $\gamma = 2$ wird *W-Zyklus* genannt. Für die Konvergenztheorie der Mehrgitterverfahren sei auf [Hac93, Kapitel 10.6] verwiesen. Falls die Anzahl der Freiheitsgrade auf dem groben Gitter jeweils viel kleiner als auf dem feinen Gitter ist, ergibt sich, wie in [Hac93, Satz 10.4.2] gezeigt, ein Rechenaufwand für das Mehrgitterverfahren, der sich proportional zu der Anzahl der Freiheitsgrade verhält.

3 Finite Elemente Methode

Wir stellen in diesem Kapitel das Verfahren der Finiten Elemente vor, das benutzt werden kann, um partielle Differentialgleichungen zu lösen. Hierbei beschränken wir uns auf elliptische partielle Differentialgleichungen. Die Einführung der Finite Elemente Methode orientiert sich an [Bra07], [Hac96] und [Eva10].

Im Folgenden sei $d \in \{1, 2, 3\}$ und $\Omega \subset \mathbb{R}^d$ ein Gebiet mit stückweise glattem Rand. Wir suchen eine Lösung $u \in C^2(\Omega) \cap C\left(\bar{\Omega}\right)$ des Randwertproblems

$$\begin{aligned} Lu &= f && \text{in } \Omega \\ u &= g && \text{auf } \partial\Omega, \end{aligned} \tag{3.1}$$

wobei $f : \Omega \to \mathbb{R}$ und $g : \partial\Omega \to \mathbb{R}$ Funktionen sind und L ein Differentialoperator zweiter Ordnung mit Divergenzstruktur ist. Das heißt, dass der Differentialoperator L die Form

$$Lu = -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_j}\left(\alpha_{ij}\frac{\partial}{\partial x_i}u\right) + \sum_{i=1}^{d} \beta_i \frac{\partial}{\partial x_i}u + \gamma u \tag{3.2}$$

mit beschränkten und messbaren Koeffizientenfunktionen $\alpha_{ij}, \beta_i, \gamma : \Omega \to \mathbb{R}$ für alle $i, j \in \{1, \ldots, d\}$ und $u \in C^2(\Omega)$ besitzt. Die Lösung u, die die Differentialgleichung in jedem Punkt von Ω erfüllt und die Randwerte g stetig annimmt, heißt *klassische Lösung*.

Die Differentialgleichung überführen wir im Folgenden in eine Variationsformulierung, indem wir sie mit Testfunktionen aus einem geeigneten Funktionenraum multiplizieren und über das Gebiet Ω integrieren. Folglich reicht es aus, die Differentialgleichung im Integralmittel statt punktweise zu erfüllen. Danach ersetzen wir den Funktionenraum durch einen endlichdimensionalen Teilraum, in dem die Variationsformulierung einem linearen Gleichungssystem entspricht. Die Lösung des Gleichungssystems stellt eine Approximation der Lösung der Variationsformulierung und daher der ursprünglichen Differentialgleichung dar. Die Idee des Finite Elemente Verfahrens besteht darin, bei der Wahl der endlichdimensionalen Teilräume das Gebiet Ω in kleine, einfach strukturierte Teilgebiete zu zerlegen und auf diesen Teilgebieten stückweise polynomiale Ansatzfunktionen, eben die Finiten Elemente, zu verwenden.

3.1 Sobolev-Räume

In diesem Kapitel führen wir die Funktionenräume ein, in denen wir die Variationsformulierung verfassen werden.

Definition 3.1 ($L^2(\Omega)$)**.** *Der Raum $L^2(\Omega)$ besteht aus allen messbaren Funktionen $u : \Omega \to \mathbb{R}$, sodass $|u|^2$ Lebesgue-integrierbar auf Ω ist, das heißt*

$$L^2(\Omega) := \left\{ u : \Omega \to \mathbb{R} : u \text{ ist Lebesgue-messbar}, |u|^2 \text{ ist Lebesgue-integrierbar} \right\}.$$

Der Raum $L^2(\Omega)$ enthält die Äquivalenzklassen der Äquivalenzrelation $\sim$, die für $u, v \in L^2(\Omega)$ definiert ist durch

$$u \sim v \ :\Leftrightarrow \text{ es existiert eine Nullmenge } N \subset \Omega \text{ mit } u(x) = v(x) \text{ für alle } x \in \Omega \backslash N.$$

Das bedeutet, wir identifizieren zwei Funktionen $u, v \in L^2(\Omega)$ miteinander, wenn u und v überall bis auf einer Nullmenge übereinstimmen.

Lemma 3.2. *Der Raum $L^2(\Omega)$ ist mit dem durch*

$$\langle u, v \rangle_{L^2(\Omega)} := \int_\Omega u(x) v(x) \, dx$$

für $u, v \in L^2(\Omega)$ definierten Skalarprodukt ein Hilbertraum. Dieses Skalarprodukt induziert für $u \in L^2(\Omega)$ die Norm

$$\|u\|_{L^2(\Omega)} := \sqrt{\langle u, u \rangle_{L^2(\Omega)}} = \left(\int_\Omega |u(x)|^2 \right)^{1/2}.$$

Beweis. Siehe [Dob10, Korollar 4.18]. □

Definition 3.3 ($L^\infty(\Omega)$)**.** *$L^\infty(\Omega)$ ist der Raum aller messbarer Funktionen $u : \Omega \to \mathbb{R}$, für die eine Nullmenge $N \subset \Omega$ mit*

$$\sup_{x \in \Omega \backslash N} |u(x)| < \infty$$

existiert. Die Norm $\|\cdot\|_{L^\infty(\Omega)}$ auf $L^\infty(\Omega)$ ist für $u \in L^\infty(\Omega)$ definiert durch

$$\|u\|_{L^\infty(\Omega)} := \inf_{\substack{N \subset \Omega \\ N \text{ Nullmenge}}} \sup_{x \in \Omega \backslash N} |u(x)|.$$

Definition 3.4 (Multiindex)**.** *Ein* Multiindex *ist ein Vektor* $\alpha \in \mathbb{N}_0^d$*. Für einen Multiindex* $\alpha \in \mathbb{N}_0^d$ *bezeichnet* $|\alpha| := \sum_{k=1}^d \alpha_k$ *den* Betrag *und*

$$D^\alpha := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$$

den zugehörigen partiellen Ableitungsoperator.

Es sei

$$C_0^\infty(\Omega) := \{u \in C^\infty(\Omega) : \operatorname{supp}(u) \text{ ist kompakt}\}.$$

Definition 3.5 (Schwache Ableitung)**.** *Es sei* $\alpha \in \mathbb{N}_0^d$*. Eine Funktion* $u \in L^2(\Omega)$ *besitzt eine* schwache Ableitung*, wenn eine Funktion* $u_\alpha \in L^2(\Omega)$ *existiert mit*

$$\int_\Omega u(x) D^\alpha \phi(x)\, dx = (-1)^{|\alpha|} \int_\Omega u_\alpha(x)\phi(x)\, dx \qquad \text{für alle } \phi \in C_0^\infty(\Omega).$$

Lemma 3.6. *Falls die schwache Ableitung einer Funktion* $u \in L^2(\Omega)$ *existiert, ist diese eindeutig. Wenn eine Funktion* $u \in L^2(\Omega)$ *klassisch differenzierbar ist, so ist sie auch schwach differenzierbar und beide Ableitungen stimmen überein.*

Beweis. Siehe [Dob10, Lemma 5.4]. □

Wegen dieses Lemmas schreiben wir für $\alpha \in \mathbb{N}_0^d$ und einer Funktion $u \in L^2(\Omega)$ auch $D^\alpha u$ anstelle von u_α für die schwache Ableitung und unterscheiden nicht zwischen der klassischen und der schwachen Ableitung.

Definition 3.7 (Sobolev-Raum)**.** *Sei* $m \in \mathbb{N}_0$*. Der Raum*

$$H^m(\Omega) := \left\{u \in L^2(\Omega) : \forall \alpha \in \mathbb{N}_0^d : |\alpha| \le m\ \exists D^\alpha u \in L^2(\Omega)\right\}$$

heißt Sobolev-Raum *m-ter Ordnung.*

Lemma 3.8. *Sei* $m \in \mathbb{N}_0$*. Der Sobolev-Raum* $H^m(\Omega)$ *ist mit dem durch*

$$\langle u, v\rangle_{H^m(\Omega)} := \sum_{\substack{\alpha \in \mathbb{N}_0^d \\ |\alpha| \le m}} \langle D^\alpha u, D^\alpha v\rangle_{L^2(\Omega)}$$

für $u, v \in H^m(\Omega)$ *definierten Skalarprodukt ein Hilbertraum. Dieses Skalarprodukt indu-*

ziert für $u \in H^m(\Omega)$ die Norm

$$\|u\|_{H^m(\Omega)} := \sqrt{\langle u, u\rangle_{H^m(\Omega)}} = \left(\sum_{\substack{\alpha \in \mathbb{N}_0^d \\ |\alpha| \leq m}} \|D^\alpha u\|^2_{L^2(\Omega)} \right)^{1/2}.$$

Beweis. Siehe [Dob10, Korollar 5.11]. □

Sei $m \in \mathbb{N}_0$. Wir lassen im Folgenden als Schreibweise sowohl bei der Norm $\|\cdot\|_{H^m(\Omega)}$ als auch bei dem Skalarprodukt $\langle \cdot, \cdot \rangle_{H^m(\Omega)}$ des Sobolev-Raums $H^m(\Omega)$ bei der Summe $\alpha \in \mathbb{N}_0^d$ weg. Also schreiben wir für $u, v \in H^m(\Omega)$

$$\|u\|_{H^m(\Omega)} = \left(\sum_{|\alpha| \leq m} \|D^\alpha u\|^2_{L^2(\Omega)} \right)^{1/2} \quad \text{und} \quad \langle u, v\rangle_{H^m(\Omega)} = \sum_{|\alpha| \leq m} \langle D^\alpha u, D^\alpha v\rangle_{L^2(\Omega)}.$$

Neben der Norm $\|\cdot\|_{H^m(\Omega)}$ definieren wir durch

$$|u|_{H^m(\Omega)} := \left(\sum_{\substack{\alpha \in \mathbb{N}_0^d \\ |\alpha| = m}} \|D^\alpha u\|^2_{L^2(\Omega)} \right)^{1/2}$$

für alle $u \in H^m(\Omega)$ eine Halbnorm auf $H^m(\Omega)$. Wie für die Norm schreiben wir im Folgenden die Halbnorm $|\cdot|_{H^m(\Omega)}$ ohne $\alpha \in \mathbb{N}_0^d$. Mithilfe dieser Halbnorm können wir die Norm für $u \in H^m(\Omega)$ als

$$\|u\|_{H^m(\Omega)} = \left(\sum_{i=0}^{m} |u|^2_{H^i(\Omega)} \right)^{1/2}$$

darstellen.

Satz 3.9 (Meyers und Serrin)**.** *Sei $m \in \mathbb{N}_0$. Dann ist $C^\infty(\Omega) \cap H^m(\Omega)$ dicht in $H^m(\Omega)$.*

Beweis. Siehe [Dob10, Satz 5.16]. □

Nach dem Satz 3.9 ist $H^m(\Omega)$ die Vervollständigung von $C^\infty(\Omega) \cap H^m(\Omega)$. Auf diese Weise verallgemeinern wir Funktionen mit Nullrandbedingungen.

Definition 3.10 ($H_0^m(\Omega)$)**.** *Sei $m \in \mathbb{N}_0$. Der* Sobolev-Raum *m-ter Ordnung mit Dirichlet-Randwerten $H_0^m(\Omega)$ ist die Vervollständigung von $C_0^\infty(\Omega)$ im Raum $H^m(\Omega)$ bezüglich*

der Norm $\|\cdot\|_{H^m(\Omega)}$*. Das heißt*

$$H_0^m(\Omega) = \{u \in H^m(\Omega) : u \text{ ist Grenzwert einer Folge in } C_0^\infty(\Omega)\}\,.$$

Die Halbnorm $|\cdot|_{H^m(\Omega)}$ ist eine Norm auf $H_0^m(\Omega)$ für $m \in \mathbb{N}_0$. Zudem ist $H_0^m(\Omega)$ ein Hilbertraum, da $H_0^m(\Omega)$ ein abgeschlossener Unterraum des Hilbertraums $H^m(\Omega)$ ist. Es gilt insbesondere $L^2(\Omega) = H^0(\Omega) = H_0^0(\Omega)$.

Lemma 3.11. *Sei* $m \in \mathbb{N}_0$*. Wenn* Ω *beschränkt ist, sind die Normen* $\|\cdot\|_{H^m(\Omega)}$ *und* $|\cdot|_{H^m(\Omega)}$ *in* $H_0^m(\Omega)$ *äquivalent.*

Beweis. Siehe [Hac96, Lemma 6.2.11]. □

Die im nächsten Satz eingeführte Poincaré-Ungleichung liefert einen Teil des Beweises der Normäquivalenz und stellt eine Abschätzung der Norm $\|\cdot\|_{L^2(\Omega)}$ dar.

Satz 3.12 (Poincaré-Ungleichung)**.** *Sei* $\Omega \subset \mathbb{R}^d$ *konvex und in einer Kugel vom Durchmesser* $r \in \mathbb{R}_{\geq 0}$ *enthalten. Dann gilt für alle* $u \in H^1(\Omega)$ *mit* $\int_\Omega u\,dx = 0$

$$\|u\|_{L^2(\Omega)} \leq 2^{d/2} r\,|u|_{H^1(\Omega)}\,. \tag{3.3}$$

Beweis. Siehe [Dob10, Satz 6.23]. □

3.2 Variationsformulierung

Im Folgenden nehmen wir $\alpha_{ij}, \beta_i, \gamma \in L^\infty(\Omega)$ für alle $i, j \in \{1, \ldots, d\}$ für die Koeffizientenfunktionen des Differentialoperators L aus (3.2) und $g = 0$ für die Randwerte des Randwertproblems (3.1) an. Diese Randbedingung heißt *homogene Dirichlet-Randbedingung*. Zudem beschränken wir uns auf elliptische Differentialoperatoren, die folgendermaßen definiert sind.

Definition 3.13. *Der Differentialoperator* L *aus* (3.2) *heißt* gleichmäßig elliptisch*, wenn eine Konstante* $c_e \in \mathbb{R}_{>0}$ *mit*

$$\sum_{i,j=1}^{d} \alpha_{ij}(x)\xi_i\xi_j \geq c_e \|\xi\|_2^2$$

für alle $x \in \Omega$ *und* $\xi \in \mathbb{R}^d$ *existiert.*

Da nicht immer eine klassische Lösung des Randwertproblems (3.1) existiert, verallgemeinern wir die Lösung dadurch, dass wir eine schwache Lösung suchen. Dazu überführen

wir das Randwertproblem in die Variationsformulierung, indem wir mit einer *Testfunktion* $v \in C_0^1(\Omega)$ multiplizieren und anschließend über das Gebiet Ω integrieren. Somit ergibt sich mittels partieller Integration insgesamt die Variationsformulierung

$$\int_\Omega \sum_{i,j=1}^{d} \alpha_{ij} \frac{\partial}{\partial x_i} u \ \frac{\partial}{\partial x_j} v + \sum_{i=1}^{d} \beta_i \frac{\partial}{\partial x_i} u \ v + \gamma uv \ dx = \int_\Omega fv \ dx, \tag{3.4}$$

weil die Randterme wegen $v|_{\partial\Omega} = 0$ verschwinden. In der Variationsformulierung braucht nur noch das Produkt der Ableitungen der Funktionen u und v integrierbar sein, sodass wir den Sobolev-Raum $H_0^1(\Omega)$ mit den schwachen Ableitungen als Verallgemeinerung der klassischen Ableitungen verwenden. Mithilfe der durch

$$a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}, (u, v) \mapsto \int_\Omega \sum_{i,j=1}^{d} \alpha_{ij} D_i u \ D_j v + \sum_{i=1}^{d} \beta_i D_i u \ v + \gamma uv \ dx$$

definierten Bilinearform erhalten wir die *Variationsformulierung*: Gesucht ist eine Funktion $u \in H_0^1(\Omega)$ mit

$$a(u, v) = (f, v) \qquad \text{für alle } v \in H_0^1(\Omega). \tag{3.5}$$

Dabei ist D_i für $i \in \{1, \dots, d\}$ eine Schreibweise für D^α mit dem Multiindex $\alpha \in \mathbb{N}_0^d$ mit $\alpha_i = 1$ und $|\alpha| = 1$, das heißt nur an der i-ten Stelle steht eine 1, und $(f, \cdot) : H_0^1(\Omega) \to \mathbb{R}, v \mapsto \int_\Omega v(x) f(x) \ dx$ ist das Dualitätsprodukt, das dem Skalarprodukt $\langle \cdot, \cdot \rangle_{L^2(\Omega)}$ entspricht. Eine Funktion $u \in H_0^1(\Omega)$, welche die Variationsformulierung (3.5) erfüllt, heißt *schwache Lösung* des Randwertproblems (3.1). Eine klassische Lösung des Randwertproblems ist nach Konstruktion der Variationsformulierung auch eine schwache Lösung. Falls die schwache Lösung eindeutig ist und eine klassische Lösung existiert, stimmen diese überein.

Satz 3.14 (Rieszscher Darstellungssatz)**.** *Es sei H ein reeller Hilbertraum.*

1. *Zu jedem stetigen linearen Funktional $f \in H'$ gibt es genau ein $u \in H$ mit*

$$\langle u, v \rangle_H = f(v) \qquad \textit{für alle } v \in H \tag{3.6}$$

   und $\|f\|_{H'} = \|u\|_H$.

2. *Das eindeutig bestimmte $u \in H$ in* (3.6) *ist auch die eindeutig bestimmte Lösung*

des Minimierungsproblems: Finde $u' \in H$ mit

$$F(u') \leq F(v) \qquad \text{für alle } v \in H \tag{3.7}$$

mit $F : H \to \mathbb{R}, v \mapsto \frac{1}{2}\langle v, v\rangle_H - f(v)$.

Beweis. Siehe [Dob10, Satz 2.25]. □

Der Satz liefert die Existenz und Eindeutigkeit einer schwachen Lösung, falls die Bilinearform $a(\cdot,\cdot)$ das Skalarprodukt eines Hilbertraums darstellt. Zudem folgt aus diesem Satz, dass wir die Variationsformulierung auch als Minimierungsproblem betrachten können.

Um in weiteren Fällen eine eindeutige Lösung zu erhalten, verallgemeinern wir den Rieszschen Darstellungssatz, indem wir Einschränkungen an der Bilinearform $a(\cdot,\cdot)$ vornehmen.

Definition 3.15 (Beschränktheit und Elliptizität)**.** *Eine Bilinearform $b : H \times H \to \mathbb{R}$ auf einem reellen Hilbertraum H heißt* beschränkt*, wenn eine Konstante $\alpha_1 \in \mathbb{R}_{\geq 0}$ so existiert, dass*

$$|b(u,v)| \leq \alpha_1 \|u\|_H \|v\|_H \tag{3.8}$$

für alle $u, v \in H$ gilt. Außerdem heißt die Bilinearform elliptisch*, wenn sie beschränkt ist und eine Konstante $\alpha_2 \in \mathbb{R}_{>0}$ so existiert, dass*

$$b(u,u) \geq \alpha_2 \|u\|_H^2 \tag{3.9}$$

für alle $u \in H$ gilt.

Jede elliptische Bilinearform $b(\cdot,\cdot)$ auf einem reellen Hilbertraum H induziert durch $\|\cdot\|_b := \sqrt{b(\cdot,\cdot)}$ eine Norm, die zur Norm des Hilbertraums äquivalent ist. Der folgende Satz von Lax-Milgram liefert die Existenz und Eindeutigkeit von schwachen Lösungen, wenn die zugehörige Bilinearform $a(\cdot,\cdot)$ elliptisch auf dem Hilbertraum $H_0^1(\Omega)$ ist. Insbesondere liefert dieser Satz eine eindeutige Lösung für unsymmetrische Bilinearformen.

Satz 3.16 (Lax-Milgram)**.** *Sei H ein reeller Hilbertraum und $b : H \times H \to \mathbb{R}$ eine elliptische Bilinearform. Für jedes stetige lineare Funktional $f \in H'$ existiert genau eine Lösung $u \in H$ des Variationsproblems*

$$b(u,v) = (f,v) \qquad \text{für alle } v \in H.$$

Diese Lösung erfüllt die Stabilitätsaussage

$$\|u\|_H \leq \frac{1}{\alpha_1} \|f\|_{H'} .$$

Beweis. Siehe [Alt12, Folgerung 4.3]. □

Die Variationsformulierung des elliptischen Randwertproblems (3.1) besitzt nach dem Satz 3.16 eine eindeutige Lösung, falls die Bilinearform $a(\cdot,\cdot)$ aus (3.5) elliptisch ist. Falls

$$d \cdot c_P \max_{i \leq d} \|\beta_i\|_{L^\infty(\Omega)} + c_P^2 \|\gamma\|_{L^\infty(\Omega)} < \alpha_2$$

mit der Konstanten $c_P \in \mathbb{R}_{\geq 0}$ aus der Poincaré-Ungleichung gilt, ist die Bilinearform $a(\cdot,\cdot)$ elliptisch, wie in [Dob10, Kapitel 7.2] gezeigt wurde. Die elliptische Bilinearform $a(\cdot,\cdot)$ induziert einen positiv definiten Operator $A : H_0^1(\Omega) \to H_0^1(\Omega)'$ durch

$$(Au, v) = a(u, v) \qquad \text{für alle } u, v \in H_0^1(\Omega).$$

Wenn das Randwertproblem (3.1) eine *inhomogene Dirichlet-Randbedingung* besitzt, das heißt es gilt $g \neq 0$ für die Funktion $g : \partial\Omega \to \mathbb{R}$, können wir die entsprechende Variationsformulierung in eine Variationsformulierung mit homogenen Dirichlet-Randbedingungen transformieren. Dazu sei $u_0 \in H^1(\Omega)$ mit $u_0|_{\partial\Omega} = g$. Dann ist $u \in H^1(\Omega)$ genau dann eine Lösung der ursprünglichen Variationsformulierung, wenn $w := u - u_0 \in H_0^1(\Omega)$ eine Lösung der Variationsformulierung

$$a(w, v) = (f, v) - a(u_0, v) \qquad \text{für alle } v \in H_0^1(\Omega)$$

ist. Daher stellt es keine Einschränkung dar, im Folgenden nur Randwertprobleme mit homogenen Dirichlet-Randbedingungen zu betrachten.

3.3 Galerkin-Verfahren

Die eindeutige Lösung $u \in V$ mit $V := H_0^1(\Omega)$ der aus dem Randwertproblem (3.1) entstehenden Variationsformulierung

$$a(u, v) = (f, v) \qquad \text{für alle } v \in V \tag{3.10}$$

mit einer elliptischen Bilinearform $a(\cdot,\cdot)$ lässt sich mithilfe des *Galerkin-Verfahrens* approximativ berechnen. Die Idee des Galerkin-Verfahrens besteht darin, den Raum V

durch einen endlichdimensionalen Raum $V_h \subset V$ zu ersetzen und als Approximation der Lösung der ursprünglichen Variationsformulierung (3.10) die Lösung der folgenden Variationsformulierung zu verwenden: Gesucht ist eine Funktion $u_h \in V_h$, die

$$a(u_h, v_h) = (f, v_h) \qquad \text{für alle } v_h \in V_h \tag{3.11}$$

erfüllt. Dabei steht h für einen Diskretisierungsparameter, der die Feinheit der Diskretisierung beschreibt. Weil der Raum V_h als endlichdimensionaler Teilraum von V abgeschlossen in V ist (vergleiche [Alt12, Lemma 2.9]), ist V_h selbst ein Hilbertraum mit dem Skalarprodukt $a(\cdot, \cdot)$. Somit existiert nach dem Satz 3.16 eine eindeutige Lösung des auf den Raum V_h eingeschränkten Variationsproblems (3.11).

Um eine Lösung der Variationsformulierung (3.11) zu berechnen, sei $(\phi_i)_{i \in \mathcal{J}}$ eine Basis des Raums V_h, wobei $\mathcal{J}$ eine Indexmenge der Kardinalität der Dimension des Raums V_h sei. Aufgrund der Linearität von $a(u_h, \cdot)$ und $(f, \cdot)$ ist (3.11) äquivalent zu

$$a(u_h, \phi_i) = (f, \phi_i) \qquad \text{für alle } i \in \mathcal{J}. \tag{3.12}$$

Indem wir die Lösung $u_h \in V_h$ anhand der Basis als $u_h = \sum_{j \in \mathcal{J}} x_j \phi_j$ mit dem Koeffizientenvektor $x \in \mathbb{R}^{\mathcal{J}}$ darstellen, erhalten wir x als Lösung des linearen Gleichungssystems

$$\sum_{j \in \mathcal{J}} x_j a(\phi_j, \phi_i) = (f, \phi_i) \qquad \text{für alle } i \in \mathcal{J}.$$

Mit der durch $a_{ij} := a(\phi_j, \phi_i)$ für alle $i, j \in \mathcal{J}$ definierten Matrix $\mathbf{A} \in \mathbb{R}^{\mathcal{J} \times \mathcal{J}}$ und dem durch $b_i := (f, \phi_i)$ für alle $i \in \mathcal{J}$ definierten Vektor $b \in \mathbb{R}^{\mathcal{J}}$ können wir das lineare Gleichungssystem in Matrix-Vektor-Form als

$$\mathbf{A}x = b \tag{3.13}$$

schreiben. Die Matrix $\mathbf{A}$ nennen wir *Steifigkeitsmatrix* und den Vektor b *Lastvektor*. Falls die Bilinearform $a(\cdot, \cdot)$ symmetrisch ist, folgt direkt die Symmetrie der Steifigkeitsmatrix $\mathbf{A}$. Weiter impliziert die Elliptizität der Bilinearform $a(\cdot, \cdot)$, dass die Matrix $\mathbf{A}$ positiv definit, also insbesondere regulär ist. Daher besitzt das lineare Gleichungssystem (3.13) für elliptische Bilinearformen genau eine Lösung.

Abschließend stellt sich die Frage, wie sich der Diskretisierungsfehler zwischen der Lösung $u \in V$ der Variationsformulierung (3.10) zu der Lösung $u_h \in V_h$ der diskreten Variationsformulierung (3.11) verhält.

Satz 3.17 (Céas Lemma)**.** *Die Bilinearform $a : V \times V \to \mathbb{R}$ sei elliptisch und $f \in V'$*

ein stetiges Funktional. Weiter sei $u \in V$ die Lösung der Variationsformulierung (3.10) *in V und $u_h \in V_h$ die Lösung der Variationsformulierung* (3.11) *in V_h. Dann gilt*

$$\|u - u_h\|_V \leq \frac{\alpha_1}{\alpha_2} \inf_{v_h \in V_h} \|u - v_h\|_V .$$

Beweis. Siehe [Alt12, Lemma 7.25]. □

Der Satz 3.17 besagt, dass der Diskretisierungsfehler durch den Approximationsfehler, der durch die Wahl des Raums V_h entsteht, beschränkt ist. Deswegen hängt die Genauigkeit der approximativen Lösung $u_h \in V_h$ wesentlich von der Wahl des endlichdimensionalen Teilraums V_h ab, sodass die Lösung $u \in V$ gut approximiert werden kann. Für Abschätzungen des Approximationsfehlers sei auf [Bra07, Kapitel II.6] verwiesen.

3.4 Finite Elemente

Als Wahl der endlichdimensionalen Teilräume V_h eignen sich die sogenannten *Finite Elemente Räume*. Dazu zerlegen wir das Gebiet Ω in endlich viele Teilgebiete, die wir als *Elemente* bezeichnen. Als Funktionen betrachten wir stückweise Polynome, welche wir *Finite Elemente* nennen. Wir beschränken uns im Folgenden darauf, dass Ω ein polygonales Gebiet ist. Zudem verwenden wir als Elemente nur einfache geometrische Objekte, wobei diese im eindimensionalen Raum aus Intervallen, im zweidimensionalen Raum aus Dreiecken oder Vierecken und im dreidimensionalen Raum aus Tetraedern oder Quadern bestehen. Wir beschränken uns je nach Dimension auf Intervalle, Dreiecke beziehungsweise Tetraeder. Für andere Elemente mit zugehörigen Finiten Elementen verweisen wir auf [Bra07] oder [Hac96].

Definition 3.18 (Triangulierung)**.** *(1) Eine* Triangulierung *von Ω ist eine Zerlegung $\mathcal{T} := \{T_1, \ldots, T_t\}$ mit Elementen $T_i \subset \mathbb{R}^d$ für alle $i \in \{1, \ldots, t\}$ und der Eigenschaft*

$$\overline{\Omega} = \bigcup_{i=1}^{t} T_i.$$

(2) Eine Triangulierung $\mathcal{T} = \{T_1, \ldots, T_t\}$ von Ω heißt zulässig*, falls der Schnitt $T_i \cap T_j$ zweier Elemente $T_i, T_j \in \mathcal{T}$ mit $i, j \in \{1, \ldots, t\}$ und $i \neq j$ entweder leer ist oder nur aus einem gemeinsamen Eckpunkt, einer gemeinsamen Kante oder einem gemeinsamen Dreieck von T_i und T_j besteht.*

(3) Sei $\mathcal{T}$ eine zulässige Triangulierung von Ω. Für ein Element $T \in \mathcal{T}$ bezeichnet h_T

den Umkreisradius, ρ_T *den* Inkreisradius *und* $m_T := |T|^{1/d}$ *die* Maschenweite *von* T.

Wir nennen eine zulässige Triangulierung $\mathcal{T}$ von Ω auch *Gitter* und die Menge der Eckpunkte dieser Triangulierung bezeichnen wir als $\mathcal{N}$. Des Weiteren heißt $\mathcal{T}$ *Verfeinerung* einer zulässigen Triangulierung $\mathcal{T}_1$, falls $\mathcal{T}$ durch Unterteilung der Elemente von $\mathcal{T}_1$ entstanden ist. Mögliche Gitterverfeinerungen stellen wir in Kapitel 3.5 vor. Mittels solcher Verfeinerungen können wir aus einer initialen zulässigen Triangulierung $\mathcal{T}_0$ eine Folge von Triangulierungen $(\mathcal{T}_i)_{i\in\mathbb{N}_0}$ erzeugen. Wir betrachten im Folgenden nur zulässige Triangulierungen.

Definition 3.19. *Eine Folge von Triangulierungen* $(\mathcal{T}_i)_{i\in\mathbb{N}_0}$ *von* Ω *heißt* quasiuniform*, falls eine Konstante* $\kappa \in \mathbb{R}_{>0}$ *so existiert, dass für alle* $i \in \mathbb{N}_0$ *für jedes* $T \in \mathcal{T}_i$

$$h_T \leq \kappa \rho_T$$

gilt.

Wir definieren mit Multiindizes den Raum der d-dimensionalen Polynome, indem wir für einen Multiindex $\alpha \in \mathbb{N}_0^d$

$$x^\alpha := x_1^{\alpha_1} \cdot \ldots \cdot x_d^{\alpha_d}$$

für $x \in \mathbb{R}^d$ definieren. Damit definieren wir den Raum der d-dimensionalen Polynome vom Grad $k \in \mathbb{N}_0$ durch

$$\mathcal{P}_k^d := \operatorname{span}\left(\left\{x^\alpha : \alpha \in \mathbb{N}_0^d, |\alpha| \leq k\right\}\right).$$

Weil wir als Finite Elemente stückweise Polynome verwenden, können wir mithilfe des Raums $\mathcal{P}_k^d$ den Finite Elemente Raum V_h definieren. Die Funktionen aus V_h sind stetig, auf jedem Element einer Triangulierung $\mathcal{T}$ von Ω ein Polynom und erfüllen die Randbedingung. Das heißt der Raum V_h ist durch

$$V_h := \left\{u \in C\left(\bar{\Omega}\right) : u|_{\partial\Omega} = 0, u|_T \in \mathcal{P}_k^d \text{ für alle } T \in \mathcal{T}\right\} \tag{3.14}$$

definiert. Wir beschränken uns im weiteren Verlauf auf stückweise lineare Finite Elemente. Mithin betrachten wir den Raum V_h mit Polynomen vom Grad $k = 1$.

Definition 3.20. *Die Finite Elemente Methode mit dem Finite Elemente Raum* V_h *heißt* konform*, falls* $V_h \subset V$ *gilt. Andernfalls heißt die Finite Elemente Methode* nicht konform*.*

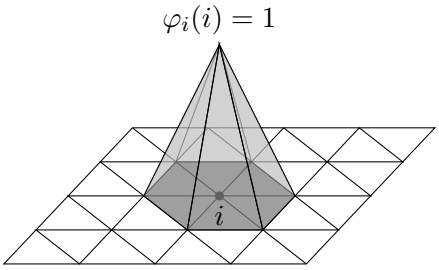


Abbildung 3.1: Basisfunktion φ_i der Knotenbasis für einen Knoten $i \in \mathcal{J}$ mit dem Träger $\operatorname{supp}(\varphi_i)$ im zweidimensionalen Fall.

Satz 3.21. *Sei $k \in \mathbb{N}$, Ω beschränkt und $\mathcal{T}$ eine Triangulierung von Ω. Für eine Funktion $v : \bar{\Omega} \to \mathbb{R}$, wobei $v|_T$ für alle $T \in \mathcal{T}$ beliebig oft differenzierbar sei, gilt genau dann $v \in H^k(\Omega)$, wenn $v \in C^{k-1}\left(\bar{\Omega}\right)$ gilt.*

Beweis. Siehe [Bra07, Kapitel II, Satz 5.2]. □

Aus diesem Satz folgt, dass die Methode der Finiten Elemente mit stückweise linearen Finiten Elementen konform ist, da $V_h \subset V$ gilt.

Um eine Lösung der Variationsformulierung (3.11) durch das Gleichungssystem (3.13) zu berechnen, benötigen wir eine Basis $(\phi_i)_{i\in\mathcal{J}}$ von V_h. Da jede Funktion $v_h \in V_h$ auf $T \in \mathcal{T}$ durch die Funktionswerte an den Eckpunkten von T festgelegt ist (siehe [Bra07, Kapitel II, Bemerkung 5.4]), ist v_h durch die Funktionswerte in den Eckpunkten der Triangulierung festgelegt. Aufgrund der homogenen Dirichlet-Randbedingung ist v_h in Randpunkten immer Null, sodass es genügt je eine Basisfunktion für einen inneren Eckpunkt zu definieren. Daher können wir $\mathcal{J} := \{n \in \mathcal{N} : n \notin \partial\Omega\}$ benutzen. Weil die Eckpunkte der Triangulierung auch *Knoten* des Gitters heißen, wird die im Folgenden definierte Basis auch *Knotenbasis* genannt.

Definition 3.22 (Knotenbasis)**.** *Es sei $\mathcal{T}$ eine zulässige Triangulierung von Ω, V_h der in* (3.14) *definierte Finite Elemente Raum und $\mathcal{J}$ die Menge der inneren Knoten von $\mathcal{T}$. Für $i \in \mathcal{J}$ ist $\varphi_i \in V_h$ durch*

$$\varphi_i(j) := \delta_{ij} \qquad \text{für alle } j \in \mathcal{J}$$

definiert. $(\varphi_i)_{i\in\mathcal{J}}$ heißt Knotenbasis *von V_h.*

Die Knotenbasis heißt auch *Lagrange-Basis*. Die Basisfunktionen besitzen einen lokalen Träger, wie in Abbildung 3.1 dargestellt ist. Daraus folgt, dass die entstehende Steifigkeitsmatrix schwachbesetzt ist.

Lemma 3.23. *Es sei $\mathcal{T}$ eine zulässige Triangulierung von Ω, V_h der in* (3.14) *definierte Finite Elemente Raum und $\mathcal{J}$ die Menge der inneren Knoten von $\mathcal{T}$. Es existiert eine Konstante $C_P \in \mathbb{R}_{>0}$ so, dass für $v \in V_h$ mit $v = \sum_{i\in\mathcal{J}} v_i\varphi_i$ mit $v_i \in \mathbb{R}$ für alle $i \in \mathcal{J}$*

$$C_P^{-1} \|v\|_{L^2(\Omega)} \leq \left(\max_{T\in\mathcal{T}} \operatorname{diam}(T)^d \sum_{i\in\mathcal{J}} |v_i|^2 \right)^{1/2} \leq C_P \|v\|_{L^2(\Omega)} \tag{3.15}$$

gilt.

Beweis. Siehe [Hac96, Satz 8.8.1]. □

Um die Güte einer Finite Elemente Lösung $u_h \in V_h$ abzuschätzen, ist es nach Satz 3.17 ausreichend, den Approximationsfehler abzuschätzen. Für die Wahl von stückweise linearen Finiten Elementen ist dabei die Abschätzung des Interpolationsfehlers der schwachen Lösung von zentraler Bedeutung. Abschätzungen hierfür sind beispielsweise in [Bra07] und [Hac96] enthalten. Wir benötigen noch die beiden folgenden inversen Abschätzungen der stückweisen linearen Finiten Elemente.

Lemma 3.24 (Lokale inverse Abschätzung)**.** *Sei $(\mathcal{T}_i)_{i\in\mathbb{N}_0}$ eine Folge quasiuniformer Triangulierungen eines Gebiebts $\Omega \subset \mathbb{R}^d$. Für $u \in V_h$ und $T \in \mathcal{T}_i$, für ein $i \in \mathbb{N}_0$, gilt die lokale inverse Ungleichung*

$$|u|_{H^1(T)} \leq c_L m_T^{-1} \|u\|_{L^2(T)} \tag{3.16}$$

mit einer von m_T unabhängigen Konstanten $c_L \in \mathbb{R}_{>0}$.

Beweis. Siehe [Ste03, Lemma 9.4]. □

Lemma 3.25 (Inverse Abschätzung)**.** *Es sei $(\mathcal{T}_i)_{i\in\mathbb{N}_0}$ eine Folge quasiuniformer Triangulierungen eines Gebiets $\Omega \subset \mathbb{R}^d$. Für alle $u \in V_h$ und $i \in \mathbb{N}_0$ gilt mit einer von m_T unabhängigen Konstanten $c_I \in \mathbb{R}_{>0}$ die inverse Abschätzung*

$$|u|^2_{H^1(\Omega)} \leq c_I \sum_{T\in\mathcal{T}_i} m_T^{-2} \|u\|^2_{L^2(T)}\,. \tag{3.17}$$

Beweis. Siehe [Ste03, Lemma 9.6] □

3.5 Gitterverfeinerung

Weil der Approximationsfehler von der Auflösung der Triangulierung abhängt, sind Verfahren zur Verfeinerung von Triangulierungen gewünscht, um bessere Approximationen

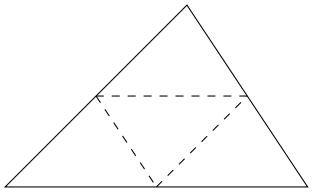


Abbildung 3.2: Rot-Verfeinerung eines Dreiecks T.

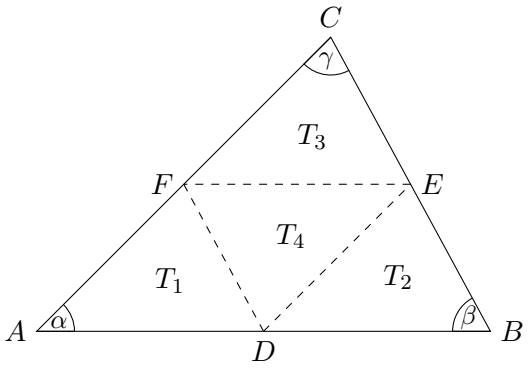


Abbildung 3.3: Das Dreieck T mit den Eckpunkten A, B, C und Seitenlängen a, b, c wurde mittels Rot-Verfeinerung verfeinert und die daraus resultierenden Dreiecke $T_1, \ldots, T_4$ dargestellt.

der Lösung zu erhalten. Wir behandeln hier nur Verfeinerungen von Triangulierungen im zweidimensionalen Raum, die aus Dreiecken bestehen. Für ein Dreieck $T \in \mathcal{T}$, wobei $\mathcal{T}$ eine zulässige Triangulierung von Ω ist, gilt $\operatorname{diam}(T) = \max\{a, b, c\}$, wenn a, b und c die Seitenlängen von T bezeichnen. Das bedeutet, dass der Durchmesser eines Dreiecks die Länge der längsten Seite ist.

Bei der *Rot-Verfeinerung* zerlegen wir ein Dreieck $T \in \mathcal{T}$ in vier Dreiecke, indem wir die drei Kantenmittelpunkte miteinander verbinden, wie in Abbildung 3.2 dargestellt. Die Besonderheit dieser Verfeinerung ist im folgenden Satz aufgeführt.

Satz 3.26. *Es sei $\mathcal{T}$ eine Triangulierung eines Gebiets $\Omega \subset \mathbb{R}^2$ und $T \in \mathcal{T}$. Die vier Dreiecke $T_1, \ldots, T_4$, die aus der Rot-Verfeinerung von T entstehen, sind paarweise kongruent und jeweils ähnlich zu T.*

Beweis. Das Dreieck T besitze wie in Abbildung 3.3 die Ecken A, B, C, die Seitenlängen $a = \|C - B\|_2$, $b = \|C - A\|_2$ und $c = \|B - A\|_2$ und die Winkel α, β, γ. Die vier Dreiecke $T_1, \ldots, T_4$, die aus der Rot-Verfeinerung von T entstehen, seien ebenfalls wie in Abbildung 3.3 gegeben, wobei das Dreieck T_1 die Ecken A, D, F, das Dreieck T_2 die Ecken D, B, E, das Dreieck T_3 die Ecken F, E, C und das Dreieck T_4 die Ecken D, E, F

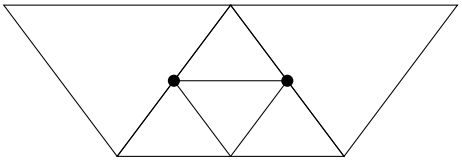


Abbildung 3.4: Bei der Rot-Verfeinerung eines einzelnen Dreiecks entstehende hängende Knoten, die als Knoten dargestellt sind.

besitze.

Aufgrund der Rot-Verfeinerung von T sind die Punkte D, E und F jeweils die Kantenmittelpunkte. Da das Dreieck T_1 wie das Dreieck T den Winkel α besitzt und für die Verhältnisse der Längen der beiden anliegenden Seiten

$$\frac{\frac{b}{2}}{\frac{c}{2}} = \frac{b}{c}$$

gilt, sind diese beiden Dreiecke nach Satz 2.15 ähnlich. Somit besitzt die Kante DF des Dreiecks T_1 die Länge $\frac{a}{2}$. Analog folgt, dass die Dreiecke T_2 und T_3 ähnlich zu T sind, weil jeweils der Winkel β beziehungsweise γ und die Verhältnisse der beiden anliegenden Seiten übereinstimmen. Daraus ergibt sich für die Kante DE die Länge $\frac{b}{2}$ und für die Kante EF die Länge $\frac{c}{2}$. Da alle Verhältnisse der Seitenlängen der Dreicke T_4 und T gleich sind, sind diese Dreiecke nach Satz 2.15 ebenfalls ähnlich.

Da die Dreiecke $T_1, \ldots, T_4$ jeweils die Seitenlängen $\frac{a}{2}, \frac{b}{2}$ und $\frac{c}{2}$ aufweisen, sind diese vier Dreiecke nach Satz 2.14 jeweils paarweise kongruent. □

Häufig sind wir an einer *lokalen Verfeinerung* der Triangulierung $\mathcal{T}$ des Gebiets Ω interessiert, falls sich die Lösung in Teilgebieten schneller als in anderen Teilgebieten verändert oder die Lösung in bestimmten Teilgebieten genauer dargestellt werden soll. Dazu eignet sich die Rot-Verfeinerung eines einzelnen Dreiecks nicht, weil die entstehenden Kantenmittelpunkte keinen Eckpunkten der benachbarten Dreiecke entsprechen und somit keine zulässige Triangulierung entsteht, wie in Abbildung 3.4 dargestellt ist. Diese Knoten heißen *hängende Knoten*. Sollen bei der reinen Rot-Verfeinerung hängende Kno-

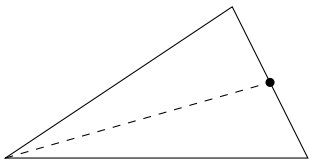
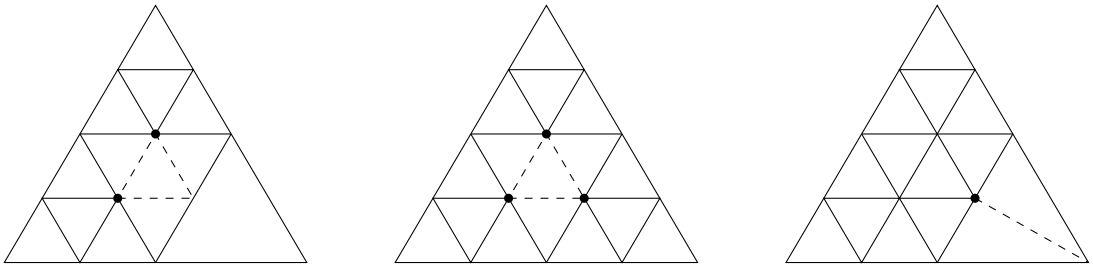


Abbildung 3.5: Grün-Verfeinerung eines Dreiecks T.



(a) Rot-Verfeinerung wegen zwei hängender Knoten

(b) Rot-Verfeinerung wegen drei hängender Knoten

(c) Grün-Verfeinerung wegen eines hängenden Knotens

Abbildung 3.6: Resultierende Rot- beziehungsweise Grün-Verfeinerung aufgrund hängender Knoten: Dabei sind jeweils die hängenden Knoten markiert und die resultierende Verfeinerung ist gestrichelt dargestellt.

ten vermieden werden, führt das dazu, dass die gesamte Triangulierung verfeinert werden muss. Die *Grün-Verfeinerung*, die in Abbildung 3.5 dargestellt ist, bietet eine Möglichkeit, nur lokal zu verfeinern und trotzdem eine zulässige Triangulierung zu erhalten. Dabei zerteilen wir ein Dreieck, in dem ein Kantenmittelpunkt eines verfeinerten Nachbardreiecks liegt, durch Verbinden des Kantenmittelpunkts mit dem gegenüberliegenden Eckpunkt. Da die Grün-Verfeinerung zur Erzeugung einer zulässigen Triangulierung aus einer Triangulierung mit hängenden Knoten verwendet wird, bezeichnen wir sie auch als Grüner-Abschluss. Falls ein Dreieck zwei Kantenmittelpunkte von verfeinerten Nachbardreiecken enthält, verfeinern wir dieses Dreieck mittels Rot-Verfeinerung. Um eine zulässige Triangulierung zu erhalten, sind eventuell weitere Rot- oder Grün-Verfeinerungen notwendig.

Ein Verfahren zur Gitterverfeinerung, das auf der Rot- und Grün-Verfeinerung basiert, ist in [BSW83] dargestellt. Dabei entsteht eine Verfeinerung einer Triangulierung eines Gebiets, indem zunächst ausgewählte Elemente mittels Rot-Verfeinerung verfeinert werden. Dabei können die zu verfeinernden Elemente beispielsweise durch einen adaptiven Fehlerschätzer bestimmt werden. Adaptive Fehlerschätzer werden zum Beispiel in [Bra07] behandelt. Da bei der Verfeinerung dieser Elemente hängende Knoten entstehen können, sodass keine zulässige Triangulierung entsteht, verfeinern wir im Folgenden alle Dreiecke, die zwei oder drei hängende Knoten besitzen, mit der Rot-Verfeinerung, wie in Abbildung 3.6(a) und 3.6(b) dargestellt. Insbesondere können hierbei weitere Dreiecke mit hängenden Knoten, wie in Abbildung 3.6(a) gezeigt, entstehen. Folglich sind jetzt nur noch Dreiecke ohne oder mit einem hängenden Knoten vorhanden. Um eine zulässige Triangulierung zu erhalten, verfeinern wir abschließend alle Dreiecke, die einen hängenden Knoten enthalten, mittels der Grün-Verfeinerung, wie in Abbildung

3.6(c) dargestellt. Damit die Innenwinkel der Dreiecke, die durch die Grün-Verfeinerung entstehen, nicht beliebig klein werden, dürfen Dreiecke, die mittels Grün-Verfeinerung verfeinert wurden, nicht weiter verfeinert werden. Deshalb entfernen wir vor jeder Verfeinerung einer Triangulierung mit dieser Verfeinerungsstrategie die Grünen-Abschlüsse aus der Triangulierung und beginnen dann den Verfeinerungsprozess.

Abschätzungen des Durchmessers, Inkreisradius und der Maschenweite von verfeinerten Dreiecken im Bezug auf das herkömmliche Dreieck sind in der folgenden Proposition aufgeführt.

Proposition 3.27. *Sei $\mathcal{T}$ eine Triangulierung eines Gebiets $\Omega \subset \mathbb{R}^2$ und $T \in \mathcal{T}$. Für ein Dreieck S, das durch Rot- oder Grün-Verfeinerung von T entstanden ist, gilt für die Inkreisradien*

$$\rho_S \geq \frac{1}{2} \cdot \rho_T, \tag{3.18}$$

für die lokalen Maschenweiten

$$m_S \geq \frac{1}{2} \cdot m_T \tag{3.19}$$

und für die Durchmesser

$$\operatorname{diam}(S) \geq \frac{1}{2} \cdot \operatorname{diam}(T). \tag{3.20}$$

Beweis. Zunächst nehmen wir an, dass das Dreieck S durch Rot-Verfeinerung, wie sie in Abbildung 3.3 dargestellt ist, aus T entstanden ist. Da die vier Dreiecke, die bei der Rot-Verfeinerung entstehen, nach Satz 3.26 kongruent sind, zeigen wir die Behauptung ohne Beschränkung der Allgemeinheit nur für das Dreieck mit den Eckpunkten A, D, F. Für den Flächeninhalt von S gilt $|S| = \frac{1}{4} \cdot |T|$, weil die vier durch Rot-Verfeinerung entstandenen Dreiecke kongruent sind. Da das Dreieck S die Seitenlängen $\frac{a}{2}, \frac{b}{2}$ und $\frac{c}{2}$ besitzt, gilt nach (2.3)

$$\rho_S = \frac{2 \cdot |S|}{\frac{a+b+c}{2}} = \frac{\frac{2}{4} \cdot |T|}{\frac{a+b+c}{2}} = \frac{1}{2} \cdot \frac{2 \cdot |T|}{a+b+c} = \frac{1}{2} \cdot \rho_T$$

und damit gilt insbesondere $\rho_S \geq \frac{1}{2} \cdot \rho_T$. Für die lokalen Maschenweiten von S und T gilt in diesem Fall

$$m_S = |S|^{1/2} = \left(\frac{1}{4} \cdot |T|\right)^{1/2} = \frac{1}{2} \cdot |T|^{1/2} = \frac{1}{2} \cdot m_T$$

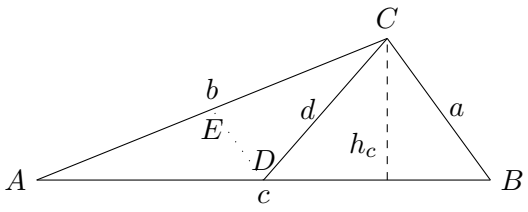


Abbildung 3.7: Das Dreieck mit den Eckpunkten A, B, C und den Seitenlängen a, b, c wurde mittels Grün-Verfeinerung verfeinert und die beiden resultierenden Dreiecke wurden dargestellt. Zusätzlich ist die Höhe h_c des Dreiecks dargestellt.

und somit gilt insbesondere $m_S \geq \frac{1}{2} \cdot m_T$. Für den Durchmesser gelte ohne Beschränkung der Allgemeinheit $\operatorname{diam}(T) = a$, sodass die längste Seite des Dreiecks S die Länge $\frac{a}{2}$ besitzt und somit

$$\operatorname{diam}(S) = \frac{a}{2} = \frac{1}{2} \cdot \operatorname{diam}(T)$$

gilt. Damit gilt insbesondere $\operatorname{diam}(S) \geq \frac{1}{2} \cdot \operatorname{diam}(T)$.

Nun nehmen wir an, dass das Dreieck S durch Grün-Verfeinerung, wie in Abbildung 3.7 dargestellt, aus dem Dreieck T entstanden ist und die Ecken A, D, C besitzt. Weil für den Flächeninhalt $|S| = \frac{1}{2} \cdot h_c \cdot \frac{c}{2} = \frac{1}{2} \cdot |T|$ und die Abschätzung $d \leq \frac{c}{2} + a$ gilt, folgt daraus mit (2.3)

$$\rho_S = \frac{2 \cdot |S|}{b + \frac{c}{2} + d} = \frac{1}{2} \cdot \frac{2 \cdot |T|}{b + \frac{c}{2} + d} \geq \frac{1}{2} \cdot \frac{2 \cdot |T|}{b + \frac{c}{2} + \frac{c}{2} + a} = \frac{1}{2} \cdot \frac{2 \cdot |T|}{a + b + c} = \frac{1}{2} \cdot \rho_T.$$

Für die lokalen Maschenweiten gilt

$$m_S = |S|^{1/2} = \left(\frac{1}{2} \cdot |T|\right)^{1/2} = \frac{1}{\sqrt{2}} \cdot |T|^{1/2} = \frac{1}{\sqrt{2}} \cdot m_T \geq \frac{1}{2} \cdot m_T.$$

Da $\operatorname{diam}\left(\tilde{S}\right) = \frac{1}{2} \cdot \operatorname{diam}(T)$ und $\tilde{S} \subseteq S$ für das Dreieck $\tilde{S}$ mit den Eckpunkten A, D, E aus der Abbildung 3.7 gilt, folgt für den Durchmesser von S

$$\operatorname{diam}(S) \geq \operatorname{diam}\left(\tilde{S}\right) = \frac{1}{2} \cdot \operatorname{diam}(T).$$

Falls S das Dreieck mit den Ecken D, B, C ist, folgen die drei Abschätzungen analog, indem wir für die Abschätzung des Inkreisradius $d \leq \frac{c}{2} + b$ und $|S| = \frac{1}{2} \cdot |T|$ verwenden. □

Da sich bei der Rot-Verfeinerung das Verhältnis von Umkreisradius zu Inkreisradius

der verfeinerten Dreiecke nicht ändert, ist eine Folge von Triangulierungen, die nur aus Rot-Verfeinerungen entstanden ist, quasiuniform. Bei der Grün-Verfeinerung kann sich hingegen das Verhältnis vom Umkreisradius zu Inkreisradius der verfeinerten Dreiecke vergrößern, bleibt aber beschränkt, da wir nur eine Grün-Verfeinerung eines Dreiecks zulassen. Damit können wir mit dem Verfahren zur Gitterverfeinerung aus [BSW83] eine quasiuniforme Folge von Triangulierungen erzeugen.

Alternative Verfeinerungsstrategien im zweidimensionalen Raum beruhen zum Beispiel auf der Bisektion, wobei ein Knoten eines Dreiecks mit dem Mittelpunkt der gegenüberliegenden Seite verbunden wird. Zur Auswahl des Knotens gibt es verschiedene Strategien. [Riv84] verwendet den Knoten, der gegenüber der längsten Seite liegt. Die Strategie der „newest vertex bisection“, die in [Mit91] benutzt wird, verwendet hingegen den neuesten Knoten. Eine Verfeinernungsstrategie für dreidimensionale Triangulierungen ist in [Bey95] dargestellt.

3.6 Modellproblem

Als Modellproblem betrachten wir in dieser Arbeit die *Poisson-Gleichung*

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega, \\ u &= g && \text{auf } \partial\Omega \end{aligned} \tag{3.21}$$

auf einem beschränkten, polygonalen Gebiet $\Omega \subset \mathbb{R}^d$ für $d \in \{1, 2, 3\}$, wobei $f : \Omega \to \mathbb{R}$ und $g : \partial\Omega \to \mathbb{R}$ vorgegebene Funktionen sind und die Lösung $u \in C^2(\Omega) \cap C\left(\bar{\Omega}\right)$ gesucht ist. Der *Laplace-Operator* ist für eine Funktion $u \in C^2(\Omega)$ durch

$$\Delta u := \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}$$

definiert. Für den Spezialfall $f = 0$ heißt die Poisson-Gleichung auch *Laplace-Gleichung*.

Die Poisson-Gleichung tritt in vielen Zusammenhängen in den Naturwissenschaften auf. Beispielsweise beschreibt die Poisson-Gleichung, wie etwa in [DW11, Kapitel 2.1.1] dargestellt, in der Physik im Bereich der Elektrostatik das elektrische Potential, das von einer Ladungsverteilung erzeugt wird. Die Funktion f beschreibt in diesem Zusammenhang die Ladungsdichte im Gebiet Ω und homogene Dirichlet-Randwerte bedeuten, dass am Rand keine Potentialunterschiede auftreten können. Ein weiteres Beispiel ist die Beschreibung der Auslenkung einer Membran im Gebiet Ω, auf die eine Kraft f wirkt. Die Randbedingung beschreibt dabei, dass die Membran am Rande eingespannt ist. Die

stationäre Temperaturverteilung in einem Körper Ω ist eine weitere Interpretation der Poisson-Gleichung, wobei f hierbei eine Energiequelle und die Randwerte die vorgegebene Randtemperatur darstellen. Diese beiden letzten Beispiele stammen aus [Dob10, Kapitel 7.1].

Da für alle $x \in \Omega$ und $\xi \in \mathbb{R}^d$

$$\sum_{i,j=1}^{d} \alpha_{ij}(x)\xi_i\xi_j = \sum_{i=1}^{d} \xi_i^2 = \|\xi\|_2^2 = 1 \cdot \|\xi\|_2^2$$

wegen $\alpha_{ij} = \delta_{ij}$ gilt, ist der Laplace-Operator gleichmäßig elliptisch, sodass die Poisson-Gleichung (3.21) eine elliptische partielle Differentialgleichung ist. Damit können wir die Theorie aus Kapitel 3.2 anwenden, um eine Variationsformulierung der Poisson-Gleichung zu erhalten. Wir erhalten für die Poisson-Gleichung die Bilinearform

$$a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}, (u, v) \mapsto \int_\Omega \sum_{i=1}^{d} D_i u(x) D_i v(x)\, dx = \int_\Omega \langle \nabla u(x), \nabla v(x) \rangle_2\, dx.$$

Da diese Bilinearform $a(\cdot, \cdot)$ symmetrisch und elliptisch ist, besitzt die Variationsformulierung der Poisson-Gleichung nach dem Satz 3.16 eine eindeutige Lösung.

Wir betrachten ebenfalls die erweiterte Poisson-Gleichung. Diese ist durch

$$\begin{aligned} -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_j}\left(\alpha_{ij}\frac{\partial}{\partial x_i}u\right) &= f && \text{in } \Omega, \\ u &= g && \text{auf } \partial\Omega \end{aligned}$$

mit differenzierbaren Funktionen $\alpha_{ij} : \Omega \to \mathbb{R}$ für alle $i, j \in \{1, \ldots, d\}$, $f : \Omega \to \mathbb{R}$ und $g : \partial\Omega \to \mathbb{R}$ gegeben. Hierbei betrachten wir für alle $i, j \in \{1, \ldots, d\}$ nur Funktionen $\alpha_{ij} := \delta_{ij} \cdot \sigma$, wobei $\sigma : \Omega \to \mathbb{R}$ eine stückweise konstante Funktion ist. Weil dieser Differentialoperator auch elliptisch ist, erhalten wir auch hier eine elliptische Bilinearform

$$\begin{aligned} a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}, (u, v) &\mapsto \int_\Omega \sum_{i,j=1}^{d} \alpha_{ij}(x) D_i u(x) D_j v(x)\, dx \\ &= \int_\Omega \langle \nabla u(x), \sigma(x) \nabla v(x) \rangle_2\, dx. \end{aligned}$$

Nach Satz 3.16 existiert mithin eine eindeutige Lösung der zugehörigen Variationsformulierung.

4 Unterraumkorrekturverfahren

Die Beschreibung der allgemeinen Unterraumkorrekturverfahren in diesem Kapitel beruht auf [Yse93] und [Xu92]. Einige ergänzende Resultate und Notationen sind aus [DW11] entnommen, die sich an [Yse93] und [Xu92] orientieren. Die Theorie der Unterraumkorrekturverfahren wurde Ende der 80er Jahre entwickelt und viele klassische Verfahren, wie das Jacobi-Verfahren, das Gauß-Seidel-Verfahren, Gebietszerlegungsverfahren oder Mehrgitterverfahren, lassen sich als Unterraumkorrekturverfahren darstellen.

Sei $\mathcal{S}$ ein endlichdimensionaler Vektorraum und $(\cdot,\cdot)$ das Dualitätsprodukt. Weiter sei $A : \mathcal{S} \to \mathcal{S}'$ ein positiv definiter, elliptischer linearer Operator. Dieser Operator erzeugt das Energieskalarprodukt $(\cdot,\cdot)_A$ und die induzierte Energienorm $\|u\|_A = (u,u)_A^{1/2}$ für $u \in \mathcal{S}$, wobei das Energieskalarprodukt durch

$$(u,v)_A = (Au,v)$$

für $u, v \in \mathcal{S}$ definiert ist. In diesem Kapitel stellen wir eine Klasse von Verfahren vor, um die lineare Gleichung

$$Au = f \tag{4.1}$$

mit $f \in \mathcal{S}'$ und $u \in \mathcal{S}$ zu lösen.

Sei $m \in \mathbb{N}$. Seien $\mathcal{W}_0, \ldots, \mathcal{W}_m \subseteq \mathcal{S}$ Unterräume derart, dass für jedes $u \in \mathcal{S}$ Elemente $w_k \in \mathcal{W}_k$ für alle $k \in \{0,\ldots,m\}$ mit

$$u = w_0 + \ldots + w_m$$

existieren. Diese Unterräume bilden eine Zerlegung von $\mathcal{S}$ mit $\sum_{k=0}^{m} \mathcal{W}_k = \mathcal{S}$. Wir fordern für diese Zerlegung des Raums $\mathcal{S}$ dabei weder, dass die Darstellung eindeutig ist, noch, dass $\mathcal{W}_k \subseteq \mathcal{W}_\ell$ für $k, \ell \in \{0,\ldots,m\}$ mit $k \leq \ell$ gilt, also die Unterräume von $\mathcal{S}$ geschachelt sind. Im Folgenden definieren wir für $k \in \{0,\ldots,m\}$ jeweils eine orthogonale Projektion auf den Unterraum $\mathcal{W}_k$ bezüglich des Energieskalarprodukts $(\cdot,\cdot)_A$ und eine Einbettung des Dualraums S' in den Dualraum $\mathcal{W}_k'$.

Definition 4.1. *Für $k \in \{0, \dots, m\}$ definieren wir die Einbettung $Q_k : \mathcal{S}' \to \mathcal{W}_k'$ und die orthogonale Projektion $P_k : \mathcal{S} \to \mathcal{W}_k$ für $u \in \mathcal{S}'$ und $v \in \mathcal{S}$ durch*

$$(Q_k u, w_k) = (u, w_k), \qquad \text{für } w_k \in \mathcal{W}_k, \tag{4.2}$$

$$(P_k v, w_k)_A = (v, w_k)_A, \qquad \text{für } w_k \in \mathcal{W}_k. \tag{4.3}$$

Wir benötigen die Restriktion des Operators A auf die Unterräume $\mathcal{W}_k$ für $k \in \{0, \dots, m\}$.

Definition 4.2. *Sei $k \in \{0, \dots, m\}$. Der Operator $A_k : \mathcal{W}_k \to \mathcal{W}_k'$ definiert die Restriktion des Operators A auf den Unterraum $\mathcal{W}_k$. Der Operator A_k ist für $u, v \in \mathcal{W}_k$ definiert durch*

$$(A_k u, v) = (u, v)_A. \tag{4.4}$$

Die orthogonale Projektion, die Einbettung und die Restriktion des Operators A sind für alle $k \in \{0, \dots, m\}$ durch die Gleichung

$$A_k P_k = Q_k A \tag{4.5}$$

miteinander verbunden. Diese Gleichheit ergibt sich für $k \in \{0, \dots, m\}$ und für alle $u \in \mathcal{S}$ und $w_k \in \mathcal{W}_k$ wegen

$$(A_k P_k u, w_k) = (P_k u, w_k)_A = (u, w_k)_A = (Au, w_k) = (Q_k A u, w_k).$$

Die Idee der Unterraumkorrektur auf dem Unterraum $\mathcal{W}_k$ für $k \in \{0, \dots, m\}$ besteht darin, den Teil des Fehlers, der orthogonal bezüglich des Energieskalarprodukts auf den Unterraum $\mathcal{W}_k$ ist, zwischen der Lösung $u^* := A^{-1} f \in \mathcal{S}$ und einer gegebenen Näherungslösung $\tilde{u} \in \mathcal{S}$ zu der Näherungslösung zu addieren.

Definition 4.3 (Unterraumkorrektur)**.** *Es sei $k \in \{0, \dots, m\}$. Die* Unterraumkorrektur *auf dem Unterraum $\mathcal{W}_k$ ist durch*

$$\bar{u} = \tilde{u} + P_k(u^* - \tilde{u}) \tag{4.6}$$

für die Lösung $u^* := A^{-1} f \in \mathcal{S}$ und eine Näherungslösung $\tilde{u} \in \mathcal{S}$ gegeben, wobei $\bar{u} \in \mathcal{S}$ eine neue Näherungslösung darstellt.

Sei $\tilde{u} \in \mathcal{S}$ eine Näherungslösung der Lösung $u^* := A^{-1} f \in \mathcal{S}$ und $k \in \{0, \dots, m\}$. Die Lösung ergibt sich aus der Näherungslösung durch $u^* = \tilde{u} + e$, wobei $e = u^* - \tilde{u}$ der Fehler

zwischen der Lösung u^* und der Näherungslösung $\tilde{u}$ ist. Bei der Unterraumkorrektur auf dem Unterraum $\mathcal{W}_k$ ergibt sich die neue Näherungslösung $\bar{u} \in \mathcal{S}$ durch $\bar{u} = \tilde{u} + e_k$ mit $e_k = P_k(u^* - \tilde{u})$. Deswegen ist der Fehler $u^* - \bar{u}$ hoffentlich kleiner als der Fehler $u^* - \tilde{u}$, weil wir den bezüglich des Energieskalarprodukts auf dem Unterraum $\mathcal{W}_k$ orthogonalen Teil des Fehlers $u^* - \tilde{u}$ zu der Näherungslösung $\tilde{u}$ addieren.

Um die Unterraumkorrektur auf dem Unterraum $\mathcal{W}_k$ berechnen zu können, lässt sich der Term $P_k(u^* - \tilde{u})$ mithilfe von (4.5) als

$$\begin{aligned} P_k\left(u^* - \tilde{u}\right) &= A_k^{-1} A_k P_k\left(u^* - \tilde{u}\right) = A_k^{-1} Q_k A\left(u^* - \tilde{u}\right) = A_k^{-1} Q_k\left(Au^* - A\tilde{u}\right) \\ &= A_k^{-1} Q_k\left(f - A\tilde{u}\right) \end{aligned}$$

schreiben. Somit können wir die Unterraumkorrektur auf dem Unterraum $\mathcal{W}_k$ durch

$$\bar{u} = \tilde{u} + P_k\left(u^* - \tilde{u}\right) = \tilde{u} + A_k^{-1} Q_k\left(f - A\tilde{u}\right) \tag{4.7}$$

berechnen. Dabei wird das lineare Gleichungssystem $A_k e_k = r_k$ mit dem Residuum $r_k = Q_k\left(f - A\tilde{u}\right)$, das in den Dualraum $\mathcal{W}_k'$ eingebettet wurde, auf dem Unterraum $\mathcal{W}_k$ gelöst und der Fehler e_k zu der Näherungslösung $\tilde{u}$ addiert. Das lineare Gleichungssystem auf dem Unterraum $\mathcal{W}_k$ ist kleiner als das ursprüngliche lineare Gleichungssystem (4.1), sodass wir die Lösung mit geringerem Aufwand berechnen können.

Das Lösen des linearen Gleichungssystems auf dem Unterraum $\mathcal{W}_k$ kann für einen großen Unterraum $\mathcal{W}_k$ weiterhin aufwändig sein. Daher führen wir positiv definite Operatoren $B_k : \mathcal{W}_k \to \mathcal{W}_k'$ für alle $k \in \{0, \ldots, m\}$ ein, die jeweils eine Approximation der Operatoren A_k sind. Der Operator B_k auf dem Unterraum $\mathcal{W}_k$ sollte dabei so gewählt werden, dass der Korrekturterm $d_k = B_k^{-1} Q_k\left(f - A\tilde{u}\right)$ einfach als Lösung des linearen Gleichungssystems $B_k d_k = Q_k\left(f - A\tilde{u}\right)$ berechnet werden kann. Mithilfe der Operatoren B_k für $k \in \{0, \ldots, m\}$ können wir die approximative Unterraumkorrektur definieren.

Definition 4.4 (Approximative Unterraumkorrektur)**.** *Es sei $k \in \{0, \ldots, m\}$. Die* approximative Unterraumkorrektur *auf dem Unterraum $\mathcal{W}_k$ ist durch*

$$\bar{u} = \tilde{u} + B_k^{-1} Q_k(f - A\tilde{u}) \tag{4.8}$$

für die Lösung $u^* := A^{-1} f \in \mathcal{S}$ und eine Näherungslösung $\tilde{u} \in \mathcal{S}$ gegeben. $\bar{u} \in \mathcal{S}$ stellt eine neue Näherungslösung dar.

Die Berechnung der Unterraumkorrektur in einem Unterraum $\mathcal{W}_k$ für $k \in \{0, \ldots, m\}$ benötigt keine Kenntnis des Operators A, sondern nur die Kenntnis des Energieskalarpro-

dukts. Das Lösen des linearen Gleichungssystems für die approximative Unterraumkorrektur auf dem Unterraum $\mathcal{W}_k$ kann beispielsweise durch ein in Kapitel 2.2 eingeführtes iteratives Lösungsverfahren erfolgen.

Indem wir die Unterraumkorrekturen für alle Unterräume $\mathcal{W}_k$ für $k \in \{0, \dots, m\}$ kombinieren, erhalten wir ein Unterraumkorrekturverfahren. In den beiden folgenden Kapiteln 4.1 und 4.2 stellen wir zwei Varianten der Unterraumkorrekturverfahren vor.

4.1 Multiplikative Unterraumkorrekturverfahren

Das multiplikative Unterraumkorrekturverfahren verbindet die Unterraumkorrekturen in den Unterräumen der Reihe nach. Das heißt, zunächst berechnen wir die Unterraumkorrektur für eine gegebene Näherungslösung $w_{-1} \in \mathcal{S}$ auf dem Unterraum $\mathcal{W}_0$ und erhalten somit eine neue Näherungslösung $w_0 \in \mathcal{S}$. Mit dieser neuen Näherungslösung w_0 führen wir die Unterraumkorrektur auf dem Unterraum $\mathcal{W}_1$ durch, sodass wir eine zweite Näherungslösung $w_1 \in \mathcal{S}$ erhalten. Nach diesem Schema werden die Unterraumkorrekturen auf den Unterräumen $\mathcal{W}_2$ bis $\mathcal{W}_m$ der Reihe nach durchgeführt. Abschließend erhalten wir eine Näherungslösung $w_m \in \mathcal{S}$ des multiplikativen Unterraumkorrekturverfahrens, welche die Unterraumkorrekturen in allen Unterräumen enthält. Die Näherungslösungen $w_0, \dots, w_{m-1}$ sind bei diesem Verfahren nur Teillösungen. Die Näherungslösung w_m ist die Näherungslösung nach einem vollständigen Schritt des multiplikativen Unterraumkorrekturverfahrens.

Definition 4.5 (Multiplikatives Unterraumkorrekturverfahren)**.** *Für alle $k \in \{0, \dots, m\}$ und $n \in \mathbb{N}_0$ sind die Folgenglieder der Folge $\left(w_k^{(n)}\right)_{k \in \{0,\dots,m\}}$ der Unterraumkorrekturen auf den Unterräumen $\mathcal{W}_0$ bis $\mathcal{W}_m$ mit $w_{-1}^{(n)} \in \mathcal{S}$ durch*

$$w_k^{(n)} = w_{k-1}^{(n)} + B_k^{-1} Q_k \left(f - A w_{k-1}^{(n)}\right)$$

gegeben. Es sei $u^{(0)} \in \mathcal{S}$. Die Folge $\left(u^{(n)}\right)_{n \in \mathbb{N}_0}$ mit $w_{-1}^{(n)} = u^{(n)}$ und $u^{(n+1)} = w_m^{(n)}$ für $n \in \mathbb{N}_0$ bezeichnet die Iterierten des multiplikativen Unterraumkorrekturverfahrens*.*

Das Verfahren wird auch sequentielles Unterraumkorrekturverfahren genannt. Eine geometrische Interpretation des Verfahrens, die aus [DW11, Abbildung 7.2] entnommen wurde, ist in Abbildung 4.1 dargestellt. In dieser Abbildung ist zu erkennen, dass der Fehler $u^* - u^{(n)}$ für eine Näherungslösung $u^{(n)} \in \mathcal{S}$ mit $n \in \mathbb{N}_0$ bei jeder Unterraumkorrektur orthogonal bezüglich des Energieskalarprodukts auf den jeweiligen Unterraum projiziert wird. Dabei sind die Orthogonalprojektionen bezüglich des euklidischen Skalarprodukts

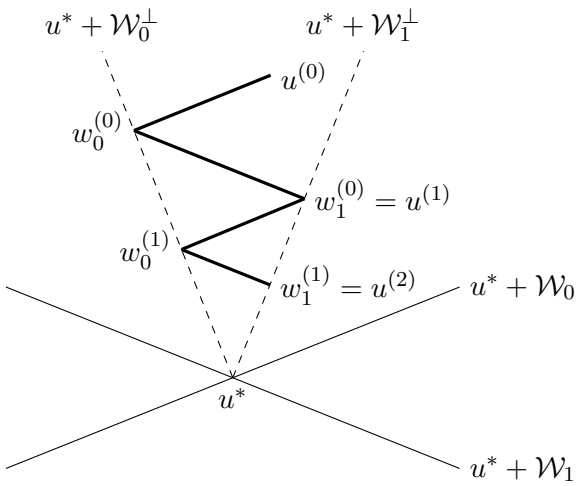


Abbildung 4.1: Geometrische Interpretation des multiplikativen Unterraumkorrekturverfahrens für zwei Unterräume $\mathcal{W}_0$ und $\mathcal{W}_1$. Für eine gegebene Näherungslösung $u^{(0)} \in \mathcal{S}$ werden zwei Iterierte des multiplikativen Unterraumkorrekturverfahrens berechnet. Die Orthogonalprojektionen auf $\mathcal{W}_0$ und $\mathcal{W}_1$ bezüglich des Energieskalarprodukts sind in dieser Abbildung als orthogonal bezüglich des euklidischen Skalarprodukts dargestellt.

dargestellt. Die Unterraumkorrekturen, welche die orthogonalen Projektionen darstellen, werden bei diesem Verfahren nacheinander jeweils mit der Lösung der vorherigen Unterraumkorrektur durchgeführt, was in der Abbildung 4.1 zu erkennen ist.

Der Algorithmus des multiplikativen Verfahrens ist in Algorithmus 2 dargestellt. Der Korrekturterm $d_k^{(n)} = B_k^{-1} Q_k \left(f - A w_{k-1}^{(n)} \right)$ für $k \in \{0, \ldots, m\}$ und $n \in \mathbb{N}_0$ wird aus der Näherungslösung der vorherigen Unterraumkorrektur $w_{k-1}^{(n)} \in \mathcal{S}$ als Lösung des linearen Gleichungssystems $B_k d_k^{(n)} = Q_k \left(f - A w_{k-1}^{(n)} \right)$ berechnet. Dieses Gleichungssystem kann durch Multiplikation mit Funktionen $v \in \mathcal{W}_k$ durch

$$(B_k d_k^{(n)}, v) = (Q_k \left(f - A w_{k-1}^{(n)} \right), v) = (f, v) - (w_{k-1}^{(n)}, v)_A$$

ohne Kenntnis des Operators A mit der Bilinearform $(\cdot, \cdot)_A$ gelöst werden. Für $k \in \{0, \ldots, m\}$ definieren wir den Operator

$$T_k : \mathcal{S} \to \mathcal{W}_k, u \mapsto B_k^{-1} A_k P_k u, \tag{4.9}$$

das heißt, es gilt nach (4.5)

$$T_k = B_k^{-1} A_k P_k = B_k^{-1} Q_k A. \tag{4.10}$$

Algorithmus 2 : Multiplikatives Unterraumkorrekturverfahren

Eingabe : $u^{(0)} \in \mathcal{S}$, $e \in \mathbb{R}_{>0}$
Ausgabe : Näherungslösung $u^{(n)}$
$n \leftarrow 0$
while $\left\| f - Au^{(n)} \right\|_A \geq e$ **do**
  $w_{-1}^{(n)} \leftarrow u^{(n)}$
  for $k = 0$ **to** m **do**
    bestimme $d_k^{(n)} \in \mathcal{W}_k$ als Lösung von $B_k d_k^{(n)} = Q_k \left(f - A w_{k-1}^{(n)} \right)$
    $w_k^{(n)} \leftarrow w_{k-1}^{(n)} + d_k^{(n)}$
  $u^{(n+1)} \leftarrow w_m^{(n)}$
  $n \leftarrow n + 1$

Lemma 4.6. *Für alle $k \in \{0, \ldots, m\}$ sind die Operatoren T_k selbstadjungiert bezüglich des Energieskalarprodukts $(\cdot, \cdot)_A$.*

Beweis. Es sei $k \in \{0, \ldots, m\}$ und $u, v \in \mathcal{S}$. Da B_k und somit auch B_k^{-1} selbstadjungiert bezüglich $(\cdot, \cdot)$ ist, gilt

$$\begin{aligned}(T_k u, v)_A &= (B_k^{-1} Q_k A u, v)_A = (B_k^{-1} Q_k A u, Av) \overset{(4.2)}{=} (B_k^{-1} Q_k A u, Q_k A v) \\ &= (Q_k A u, B_k^{-1} Q_k A v) \overset{(4.2)}{=} (Au, B_k^{-1} Q_k A v) = (u, B_k^{-1} Q_k A v)_A \\ &= (u, T_k v)_A.\end{aligned}$$

□

Anhand dieser Operatoren T_k für $k \in \{0, \ldots, m\}$ können wir die Teilschritte der Iterierten des multiplikativen Unterraumkorrekturverfahrens wie folgt darstellen:

Proposition 4.7. *Für $n \in \mathbb{N}_0$, $k \in \{0, \ldots, m\}$ und $u^{(n)} \in \mathcal{S}$ gilt für die Darstellung der Iterierten nach der Unterraumkorrektur auf dem Unterraum $\mathcal{W}_k$*

$$w_k^{(n)} = u^{(n)} + \left(I - (I - T_k) \cdot \ldots \cdot (I - T_0) \right) A^{-1} \left(f - Au^{(n)} \right). \tag{4.11}$$

Beweis. Sei $n \in \mathbb{N}_0$ und $u^{(n)} \in \mathcal{S}$ gegeben. Beweis per Induktion über k.
Induktionsanfang: Sei $k = 0$. Dann gilt

$$\begin{aligned}w_k^{(n)} = w_0^{(n)} &= w_{-1}^{(n)} + B_0^{-1} Q_0 \left(f - A w_{-1}^{(n)} \right) \\ &= u^{(n)} + B_0^{-1} Q_0 \left(f - Au^{(n)} \right) \\ &= u^{(n)} + B_0^{-1} Q_0 A A^{-1} \left(f - Au^{(n)} \right)\end{aligned}$$

$$\begin{aligned}
&\overset{(4.10)}{=} u^{(n)} + T_0 A^{-1}\left(f - Au^{(n)}\right) \\
&= u^{(n)} + \left(I - (I - T_0)\right) A^{-1}\left(f - Au^{(n)}\right).
\end{aligned}$$

Induktionsvoraussetzung: Sei $k \in \{0, \ldots, m-1\}$ und es gelte

$$w_k^{(n)} = u^{(n)} + \left(I - (I - T_k) \cdot \ldots \cdot (I - T_0)\right) A^{-1}\left(f - Au^{(n)}\right).$$

Induktionsschritt: Es gilt

$$\begin{aligned}
w_{k+1}^{(n)} &= w_k^{(n)} + B_{k+1}^{-1} Q_{k+1}\left(f - Aw_k^{(n)}\right) \\
&\overset{\text{I.V.}}{=} u^{(n)} + \left(I - (I - T_k) \cdot \ldots \cdot (I - T_0)\right) A^{-1}\left(f - Au^{(n)}\right) \\
&\qquad + B_{k+1}^{-1} Q_{k+1}\left(f - A\left(u^{(n)} + \left(I - (I - T_k) \cdot \ldots \cdot (I - T_0)\right) A^{-1}\left(f - Au^{(n)}\right)\right)\right) \\
&= u^{(n)} + \left(I - (I - T_k) \cdot \ldots \cdot (I - T_0)\right) A^{-1}\left(f - Au^{(n)}\right) \\
&\qquad + B_{k+1}^{-1} Q_{k+1}\left(f - Au^{(n)} - A\left(I - (I - T_k) \cdot \ldots \cdot (I - T_0)\right) A^{-1}\left(f - Au^{(n)}\right)\right) \\
&= u^{(n)} + \left(I - (I - T_k) \cdot \ldots \cdot (I - T_0)\right) A^{-1}\left(f - Au^{(n)}\right) \\
&\qquad + B_{k+1}^{-1} Q_{k+1} A A^{-1}\left(f - Au^{(n)}\right) \\
&\qquad - B_{k+1}^{-1} Q_{k+1} A\left(I - (I - T_k) \cdot \ldots \cdot (I - T_0)\right) A^{-1}\left(f - Au^{(n)}\right) \\
&\overset{(4.10)}{=} u^{(n)} + \left(I - (I - T_k) \cdot \ldots \cdot (I - T_0) + T_{k+1}\right. \\
&\qquad \left. - T_{k+1}\left(I - (I - T_k) \cdot \ldots \cdot (I - T_0)\right)\right) A^{-1}\left(f - Au^{(n)}\right) \\
&= u^{(n)} + \left(I - (I - T_k) \cdot \ldots \cdot (I - T_0) + T_{k+1} - T_{k+1}\right. \\
&\qquad \left. + T_{k+1}\left(I - T_k\right) \cdot \ldots \cdot (I - T_0)\right) A^{-1}\left(f - Au^{(n)}\right) \\
&= u^{(n)} + \left(I - (I - T_{k+1})\left(I - T_k\right) \cdot \ldots \cdot (I - T_0)\right) A^{-1}\left(f - Au^{(n)}\right).
\end{aligned}$$

□

Mithilfe dieser Proposition 4.7 lässt sich für eine gegebene Iterierte $u^{(n)} \in \mathcal{S}$ für $n \in \mathbb{N}_0$ des multiplikativen Unterraumkorrekturverfahrens die nächste Iterierte folgendermaßen darstellen:

$$\begin{aligned}
u^{(n+1)} = w_m^{(n)} &= u^{(n)} + \left(I - (I - T_m) \cdot \ldots \cdot (I - T_0)\right) A^{-1}\left(f - Au^{(n)}\right) \\
&= u^{(n)} - \left(I - (I - T_m) \cdot \ldots \cdot (I - T_0)\right) A^{-1}\left(Au^{(n)} - f\right).
\end{aligned}$$

Somit ist das multiplikative Unterraumkorrekturverfahren nach Lemma 2.23 ein konsistentes lineares Iterationsverfahren mit $N_{\text{mul}} := \left(I - (I - T_m) \cdot \ldots \cdot (I - T_0)\right) A^{-1}$.

Das Gauß-Seidel-Verfahren kann, wie im folgenden Beispiel erläutert wird, als multiplikatives Unterraumkorrekturverfahren mit eindimensionalen Unterräumen aufgefasst werden.

Beispiel (Gauß-Seidel-Verfahren)**.** *Sei $m \in \mathbb{N}$ und $\mathcal{S} = \mathbb{R}^{m+1}$ mit dem euklidischen Skalarprodukt $\langle \cdot, \cdot \rangle$. Für die Unterräume $\mathcal{W}_k := \operatorname{span}(e_{k+1})$ für alle $k \in \{0, \dots, m\}$, wobei $e_\ell \in \mathbb{R}^{m+1}$ für $\ell \in \{1, \dots, m+1\}$ der ℓ-te Einheitsvektor sei, gilt*

$$\mathbb{R}^{m+1} = \sum_{k=0}^{m} \mathcal{W}_k.$$

Mit den approximativen Operatoren $B_k : \mathcal{W}_k \to \mathcal{W}_k'$ für $k \in \{0, \dots, m\}$, die definiert sind durch $B_k = \langle Ae_{k+1}, e_{k+1} \rangle$ und somit die Diagonalelemente der Matrix darstellen, ist das zugehörige multiplikative Unterraumkorrekturverfahren in diesem Fall das Gauß-Seidel-Verfahren.

Um das multiplikative Unterraumkorrekturverfahren als Vorkonditionierer für das Verfahren der konjugierten Gradienten zu verwenden, benötigen wir einen positiv definiten Operator. Da N_{mul} im Allgemeinen nicht selbstadjungiert und damit nicht positiv definit ist, bedienen wir uns einer symmetrischen Variante des multiplikativen Unterraumkorrekturverfahrens. Dazu wenden wir die Unterraumkorrekturen in den Unterräumen nochmals in umgekehrter Reihenfolge an, nachdem die Unterraumkorrekturen wie bei dem multiplikativen Unterraumkorrekturverfahren der Reihe nach durchgeführt worden sind.

Definition 4.8 (Symmetrisches multiplikatives Unterraumkorrekturverfahren)**.** *Für $n \in \mathbb{N}_0$ sind die Folgenglieder der Folge $\left(w_k^{(n)}\right)_{k \in \{0,\dots,2m+1\}}$ der Unterraumkorrekturen auf den Unterräumen $\mathcal{W}_0$ bis $\mathcal{W}_m$ mit $w_{-1}^{(n)} \in \mathcal{S}$ durch*

$$w_k^{(n)} = w_{k-1}^{(n)} + B_k^{-1} Q_k \left(f - A w_{k-1}^{(n)}\right)$$

für $k \in \{0, \dots, m\}$ und den anschließenden Unterraumkorrekturen in umgekehrter Reihenfolge auf den Unterräumen $\mathcal{W}_m$ bis $\mathcal{W}_0$ durch

$$w_k^{(n)} = w_{k-1}^{(n)} + B_{2m+1-k}^{-1} Q_{2m+1-k} \left(f - A w_{k-1}^{(n)}\right)$$

für $k \in \{m+1, \dots, 2m+1\}$ gegeben. Es sei $u^{(0)} \in \mathcal{S}$. Die Folge $\left(u^{(n)}\right)_{n \in \mathbb{N}_0}$ mit $w_{-1}^{(n)} = u^{(n)}$ und $u^{(n+1)} = w_{2m+1}^{(n)}$ für $n \in \mathbb{N}_0$ bezeichnet die Iterierten des symmetrischen multiplikativen Unterraumkorrekturverfahrens*.*

Wie in (4.11) für die Darstellung der Iterierten des multiplikativen Unterraumkorrekturverfahrens gezeigt, lässt sich für $n \in \mathbb{N}_0$ und einer gegebenen Iterierten $u^{(n)} \in \mathcal{S}$ die nächste Iterierte des symmetrischen multiplikativen Unterraumkorrekturverfahrens mit $N_{\text{mul, sym}} := (I - (I - T_0) \cdot \ldots \cdot (I - T_m)\,(I - T_m) \cdot \ldots \cdot (I - T_0))\,A^{-1}$ durch

$$\begin{aligned} u^{(n+1)} &= u^{(n)} + (I - (I - T_0) \cdot \ldots \cdot (I - T_m)\,(I - T_m) \cdot \ldots \cdot (I - T_0))\,A^{-1}\left(f - Au^{(n)}\right) \\ &= u^{(n)} - (I - (I - T_0) \cdot \ldots \cdot (I - T_m)\,(I - T_m) \cdot \ldots \cdot (I - T_0))\,A^{-1}\left(Au^{(n)} - f\right) \\ &= u^{(n)} - N_{\text{mul, sym}}\left(Au^{(n)} - f\right) \end{aligned}$$

darstellen. Somit ist das symmetrische multiplikative Unterraumkorrekturverfahren als eigenständiges Verfahren nach Lemma 2.23 ein konsistentes lineares Iterationsverfahren.

4.2 Additive Unterraumkorrekturverfahren

Bei dem additiven Unterraumkorrekturverfahren werden alle Unterraumkorrekturen auf den Unterräumen mit derselben Näherungslösung durchgeführt. Das heißt, für eine gegebene Näherungslösung $u^{(n)} \in \mathcal{S}$ für $n \in \mathbb{N}_0$ werden jeweils die orthogonalen Projektionen des Fehlers zwischen der Lösung $u^* := A^{-1}f \in \mathcal{S}$ und der Näherungslösung $u^{(n)}$ auf die Unterräume bezüglich des Energieskalarprodukts berechnet und diese Fehleranteile zu der Näherungslösung $u^{(n)}$ addiert. Auf diese Weise liefert das additive Unterraumkorrekturverfahren eine neue Näherungslösung $u^{(n+1)} \in \mathcal{S}$.

Definition 4.9 (Additives Unterraumkorrekturverfahren)**.** *Für $n \in \mathbb{N}_0$ und $u^{(n)} \in \mathcal{S}$ ist die Folge $\left(w_k^{(n)}\right)_{k\in\{0,\ldots,m\}}$ der Unterraumkorrekturen auf den Unterräumen $\mathcal{W}_0$ bis $\mathcal{W}_m$ definiert durch*

$$w_k^{(n)} = B_k^{-1} Q_k \left(f - Au^{(n)}\right)$$

für alle $k \in \{0, \ldots, m\}$. Es sei $u^{(0)} \in \mathcal{S}$. Die Folge $\left(u^{(n)}\right)_{n\in\mathbb{N}_0}$ mit

$$u^{(n+1)} = u^{(n)} + \sum_{k=0}^{m} w_k^{(n)}$$

für $n \in \mathbb{N}_0$ bezeichnet die Iterierten des additiven Unterraumkorrekturverfahrens*.*

Das Verfahren wird auch paralleles Unterraumkorrekturverfahren genannt, weil die Unterraumkorrekturen auf den verschiedenen Unterräumen aufgrund ihrer Unabhängigkeit voneinander parallel ausgeführt werden können. Eine geometrische Interpretation

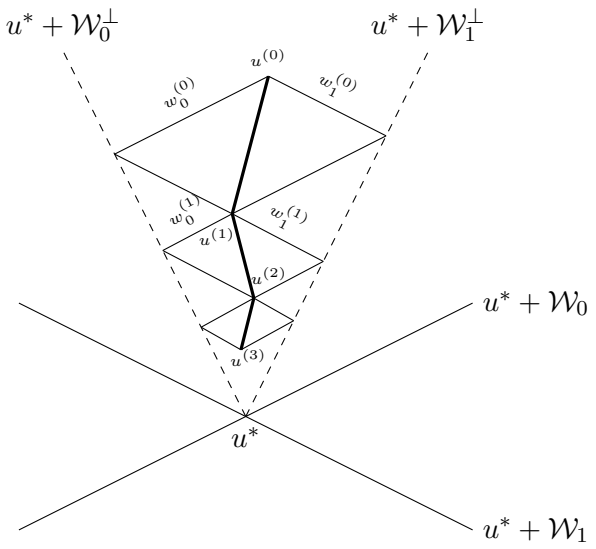


Abbildung 4.2: Geometrische Interpretation des additiven Unterraumkorrekturverfahrens für zwei Unterräume $\mathcal{W}_0$ und $\mathcal{W}_1$. Für eine gegebene Näherungslösung $u^{(0)} \in \mathcal{S}$ werden drei Iterierte des additiven Unterraumkorrekturverfahrens berechnet. Die Orthogonalprojektionen auf $\mathcal{W}_0$ und $\mathcal{W}_1$ bezüglich des Energieskalarprodukts sind in dieser Abbildung als orthogonal bezüglich des euklidischen Skalarprodukts dargestellt.

des Verfahrens, die aus [DW11, Abbildung 7.4] entnommen wurde, ist in Abbildung 4.2 dargestellt. In dieser Abbildung ist zu erkennen, dass der Fehler $u^* - u^{(n)}$ für eine Näherungslösung $u^{(n)} \in \mathcal{S}$ für $n \in \mathbb{N}_0$ jeweils bezüglich des Energieskalarprodukts auf die einzelnen Unterräume projiziert wird und danach diese Fehleranteile zu der Näherungslösung $u^{(n)}$ addiert werden, um die neue Näherungslösung nach einem vollständigen Schritt des additiven Unterraumkorrekturverfahrens zu erhalten. Die orthogonalen Projektionen sind in der Abbildung 4.2 bezüglich des euklidischen Skalarprodukts dargestellt. In dieser Abbildung ist ebenfalls zu sehen, dass die einzelnen Unterraumkorrekturen in einem Schritt des Verfahrens unabhängig voneinander sind und somit parallel durchgeführt werden können.

Der Algorithmus des additiven Unterraumkorrekturverfahrens ist in Algorithmus 3 dargestellt. Für eine gegebene Näherungslösung werden dabei die Unterraumkorrekturen auf jedem Unterraum als Lösung eines linearen Gleichungssystems berechnet. Abschließend wird durch Addition der für jeden Unterraum berechneten Korrekturterme die neue Näherungslösung als nächste Iterierte des additiven Unterraumkorrekturverfahrens berechnet.

Algorithmus 3 : Additives Unterraumkorrekturverfahren

Eingabe : $u^{(0)} \in \mathcal{S}$, $e \in \mathbb{R}_{>0}$
Ausgabe : Näherungslösung $u^{(n)}$
$n \leftarrow 0$
while $\left\| f - Au^{(n)} \right\|_A \geq e$ **do**
  for $k = 0$ **to** m **do**
    bestimme $w_k^{(n)} \in \mathcal{W}_k$ als Lösung von $B_k w_k^{(n)} = Q_k \left(f - Au^{(n)} \right)$
  $u^{(n+1)} \leftarrow u^{(n)} + \sum_{k=0}^{m} w_k^{(n)}$
  $n \leftarrow n + 1$

Für eine gegebene Näherungslösung $u^{(n)} \in \mathcal{S}$ für $n \in \mathbb{N}_0$ gilt mit $C := \sum_{k=0}^{m} B_k^{-1} Q_k$

$$\begin{aligned} u^{(n+1)} &= u^{(n)} + \sum_{k=0}^{m} B_k^{-1} Q_k \left(f - Au^{(n)} \right) = u^{(n)} + C \left(f - Au^{(n)} \right) \\ &= u^{(n)} - C \left(Au^{(n)} - f \right). \end{aligned}$$

Also ist das additive Unterraumkorrekturverfahren nach Lemma 2.23 ein konsistentes lineares Iterationsverfahren mit $N_{\text{add}} := C$.

Das Jacobi-Verfahren kann mit eindimensionalen Unterräumen, die mit denen des als multiplikatives Unterraumkorrekturverfahren interpretierten Gauß-Seidel-Verfahrens übereinstimmen, als additives Unterraumkorrekturverfahren betrachtet werden, wie im folgenden Beispiel erläutert wird.

Beispiel (Jacobi-Verfahren)**.** *Sei $m \in \mathbb{N}$ und $\mathcal{S} = \mathbb{R}^{m+1}$ mit dem euklidischen Skalarprodukt $\langle \cdot, \cdot \rangle$. Für die Unterräume $\mathcal{W}_k = \operatorname{span}(e_{k+1})$ für alle $k \in \{0, \ldots, m\}$, wobei $e_\ell \in \mathbb{R}^{m+1}$ für $\ell \in \{1, \ldots, m+1\}$ der ℓ-te Einheitsvektor sei, gilt*

$$\mathbb{R}^{m+1} = \sum_{k=0}^{m} \mathcal{W}_k.$$

Mit den approximativen Operatoren $B_k : \mathcal{W}_k \to \mathcal{W}_k'$ für $k \in \{0, \ldots, m\}$, die definiert sind durch $B_k = \langle Ae_{k+1}, e_{k+1} \rangle$, ist das additive Unterraumkorrekturverfahren das Jacobi-Verfahren.

5 Konvergenztheorie

In diesem Kapitel stellen wir die Konvergenzbeweise für das additive und multiplikative Unterraumkorrekturverfahren, die in Kapitel 4 eingeführt wurden, vor. Dazu sind die Bezeichnungen wie in Kapitel 4. Diese beiden Konvergenzbeweise orientieren sich an [Yse93], wobei die Ideen aus [Xu92] stammen. Die Notation orientiert sich an der Notation der Unterraumkorrekturverfahren in [DW11]. Die beiden Konvergenzbeweise beruhen jeweils auf einer Fehlerdarstellung zwischen einer Näherungslösung des Verfahrens und der exakten Lösung. Diese Fehlerdarstellungen wurden aus [DW11] entnommen. Unter zusätzlichen Annahmen, die Einschränkungen an die approximativen Lösungsverfahren auf den Unterräumen darstellen, können wir für beide Unterraumkorrekturverfahren zeigen, dass sich die neu berechnete Näherungslösung nach einem Schritt des Unterraumkorrekturverfahrens der Lösung annähert. Das heißt, dass der Fehler in einer geeigneten Norm kleiner wird und das Verfahren somit konvergiert.

5.1 Konvergenz multiplikativer Unterraumkorrekturverfahren

Der Fehler eines vollständigen Zyklus des multiplikativen Unterraumkorrekturverfahrens kann durch die in (4.9) eingeführten Operatoren T_k für $k \in \{0, \dots, m\}$ folgendermaßen dargestellt werden.

Lemma 5.1. *Für $n \in \mathbb{N}_0$ und einer gegebenen Näherungslösung $u^{(n)} \in \mathcal{S}$ lässt sich der Fehler eines vollständigen Zyklus des multiplikativen Unterraumkorrekturverfahrens durch*

$$u^* - u^{(n+1)} = (I - T_m) \cdot \ldots \cdot (I - T_0) \left(u^* - u^{(n)} \right) \tag{5.1}$$

mit der exakten Lösung $u^* \in \mathcal{S}$ darstellen, wobei $u^{(n+1)}$ die Näherungslösung nach einem vollständigen Zyklus des multiplikativen Unterraumkorrekturverfahrens ist.

Beweis. Sei $n \in \mathbb{N}_0$ und $u^{(n)} \in \mathcal{S}$ gegeben. Dann gilt für den Fehler zwischen der Lösung $u^* := A^{-1} f$ und der neuen Näherungslösung $u^{(n+1)}$ des multiplikativen Unterraumkor-

rekturverfahrens

$$\begin{aligned}
u^* - u^{(n+1)} &= u^* - w_m^{(n)} \\
&\overset{(4.11)}{=} u^* - \left(u^{(n)} + (I - (I - T_m) \cdot \ldots \cdot (I - T_0))\, A^{-1} \left(f - Au^{(n)}\right)\right) \\
&= A^{-1} f - u^{(n)} - A^{-1} \left(f - Au^{(n)}\right) + (I - T_m) \ldots (I - T_0)\, A^{-1} \left(f - Au^{(n)}\right) \\
&= (I - T_m) \cdot \ldots \cdot (I - T_0) \left(A^{-1} f - u^{(n)}\right) \\
&= (I - T_m) \cdot \ldots \cdot (I - T_0) \left(u^* - u^{(n)}\right).
\end{aligned}$$

□

5.1.1 Annahmen

Die abstrakte Konvergenztheorie für multiplikative Unterraumkorrekturverfahren basiert auf einer Zerlegung des Raums $\mathcal{S}$ in eine Summe von Unterräumen $\mathcal{V}_k \subseteq \mathcal{W}_k$ für alle $k \in \{0, \ldots, m\}$ mit

$$\mathcal{S} = \mathcal{V}_0 + \ldots + \mathcal{V}_m. \tag{5.2}$$

Diese Unterräume $\mathcal{V}_k$ sind dabei nur ein Hilfsmittel für die theoretische Analyse, die wir für die Implementierung des Verfahrens nicht benötigen. Für die Konvergenztheorie benötigen wir die folgenden drei Annahmen.

Annahme 1 (Stabilität der Zerlegung)**.** *Es existiert eine Konstante $K_1 \in \mathbb{R}_{>0}$, sodass für alle $v \in \mathcal{S}$ mit $v = \sum_{k=0}^{m} v_k$ und $v_k \in \mathcal{V}_k$ für alle $k \in \{0, \ldots, m\}$*

$$\sum_{k=0}^{m} (B_k v_k, v_k) \leq K_1 \|v\|_A^2 \tag{5.3}$$

gilt.

Die Annahme 1 beschränkt, wie für exakte Unterraumlösungsverfahren zu sehen ist, die Summe der Energienormen auf den Unterräumen bezüglich der approximativen Lösungsverfahren durch die Energienorm bezüglich A, sodass die Summe nicht zu groß wird und somit die Unterraumkorrekturen einen signifikanten Schritt in Richtung Lösung machen. Die Konvergenzgeschwindigkeit des Verfahrens hängt von der Stabilität der Zerlegung ab. Dies ist in Abbildung 5.1, die aus [DW11, Abbildung 7.3] entnommen wurde, dargestellt. Dabei ist die Wahl der Unterräume entscheidend, was durch den Winkel zwischen den Unterräumen in der Abbildung 5.1 dargestellt ist.

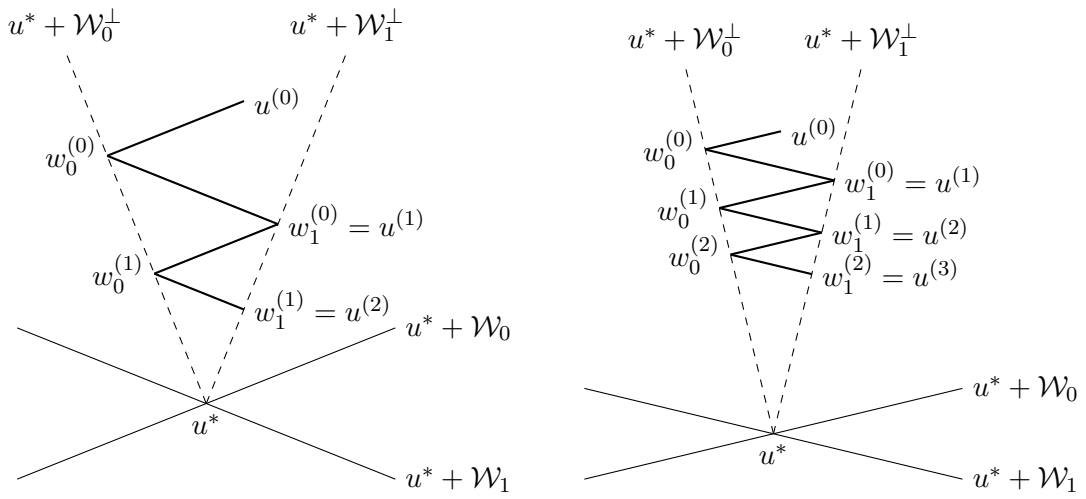


Abbildung 5.1: Geometrische Interpretation der Konvergenzgeschwindigkeit des multiplikativen Unterraumkorrekturverfahrens für zwei Unterräume $\mathcal{W}_0$ und $\mathcal{W}_1$. Die Orthogonalprojektionen auf $\mathcal{W}_0$ und $\mathcal{W}_1$ bezüglich des Energieskalarprodukts sind in dieser Abbildung als orthogonal bezüglich des euklidischen Skalarprodukts dargestellt.

Annahme 2 (Verschärfte Cauchy-Schwarz-Ungleichung)**.** *Es existiert eine symmetrische Matrix* $\gamma = (\gamma_{k\ell})_{k,\ell\in\{0,\dots,m\}} \in \mathbb{R}^{(m+1)\times(m+1)}$ *mit*

$$(w_k, v_\ell)_A \leq \gamma_{k\ell}(B_k w_k, w_k)^{1/2}(B_\ell v_\ell, v_\ell)^{1/2} \tag{5.4}$$

für alle $k, \ell \in \{0, \dots, m\}$ *mit* $k \leq \ell$, $w_k \in \mathcal{W}_k$ *und* $v_\ell \in \mathcal{V}_\ell$, *sodass*

$$\sum_{k,\ell=0}^{m} \gamma_{k\ell} x_k y_\ell \leq K_2 \left(\sum_{k=0}^{m} x_k^2\right)^{1/2} \left(\sum_{\ell=0}^{m} y_\ell^2\right)^{1/2} \tag{5.5}$$

für alle $x_k, y_\ell \in \mathbb{R}$ *für* $k, \ell \in \{0, \dots, m\}$ *mit einer Konstanten* $K_2 \in \mathbb{R}_{\geq 0}$ *gilt.*

Die Annahme 2 beschränkt die Energienormen der approximativen Lösungsverfahren, sodass diese nicht zu klein sind. Diese Annahme hängt zudem von der Reihenfolge der Zerlegung ab. Aus der folgenden Annahme 3 und der Cauchy-Schwarz-Ungleichung folgt $0 \leq \gamma_{k\ell} \leq 2$ für alle $k, \ell \in \{0, \dots, m\}$.

Annahme 3 (Lokale Konvergenz)**.** *Es existiert ein* $\lambda \in \mathbb{R}$ *mit* $0 < \lambda < 2$, *sodass*

$$(A_k w_k, w_k) \leq \lambda (B_k w_k, w_k) \tag{5.6}$$

für alle $k \in \{0, \dots, m\}$ und für alle $w_k \in \mathcal{W}_k$ gilt.

Die Annahme 3 ist äquivalent dazu, dass die approximativen Lösungsverfahren B_k für alle $k \in \{0, \dots, m\}$ konvergent sind. Sei $k \in \{0, \dots, m\}$. Da B_k und A_k positiv definit und somit selbstadjungiert sind, gilt für alle $v, w \in \mathcal{W}_k$

$$\begin{aligned}(B_k(I - B_k^{-1}A_k)v, w) &= ((B_k - B_kB_k^{-1}A_k)v, w) = (B_kv, w) - (A_kv, w)\\ &= (v, B_kw) - (v, A_kw) = (v, (B_k - A_k)w)\\ &= (v, B_k(I - B_k^{-1}A_k)w) = (B_kv, (I - B_k^{-1}A_k)w).\end{aligned}$$

Somit ist die Iterationsmatrix $I - B_k^{-1}A_k$ des approximativen Lösungsverfahrens selbstadjungiert bezüglich des von B_k erzeugten Energieskalarprodukts. Deshalb sind die Eigenwerte der Iterationsmatrix reell. Aus der positiv Definitheit von B_k und A_k folgt für alle $v \in \mathcal{W}_k \backslash \{0\}$ mit der Annahme 3

$$0 < \frac{(A_kv, v)}{(B_kv, v)} \leq \lambda < 2.$$

Mithilfe des Rayleigh-Quotienten folgt daher für den Spektralradius der Iterationsmatrix

$$\varrho\left(I - B_k^{-1}A_k\right) = \max_{v \in \mathcal{W}_k \backslash \{0\}} \left| \frac{(B_k(I - B_k^{-1}A_k)v, v)}{(B_kv, v)} \right| = \max_{v \in \mathcal{W}_k \backslash \{0\}} \left| 1 - \frac{(A_kv, v)}{(B_kv, v)} \right| < 1.$$

Daraus ergibt sich mit Satz 2.24, dass das approximative Lösungsverfahren B_k konvergent ist. Mit denselben Argumenten erfüllt ein konvergentes Lösungsverfahren B_k die Annahme 3.

5.1.2 Konvergenzbeweis

Um den Konvergenzsatz zu beweisen, benötigen wir die folgenden vier Lemmata, deren Aussagen aus [Yse93] stammen.

Lemma 5.2. *Für $v \in \mathcal{S}$ mit $v = \sum_{k=0}^{m} v_k$ mit $v_k \in \mathcal{V}_k$ und beliebigen $u_k \in \mathcal{S}$ für alle $k \in \{0, \dots, m\}$ gilt unter Annahme 1*

$$\sum_{k=0}^{m} (v_k, u_k)_A \leq \sqrt{K_1}\, \|v\|_A \left(\sum_{k=0}^{m} (T_ku_k, u_k)_A \right)^{1/2}. \tag{5.7}$$

Beweis. Es sei $v \in \mathcal{S}$ mit $v = \sum_{k=0}^{m}$ und $v_k \in \mathcal{V}_k$ für alle $k \in \{0, \dots, m\}$. Weiter seien für $k \in \{0, \dots, m\}$ beliebige $u_k \in \mathcal{S}$ gegeben und es gelte die Annahme 1. Da B_k für

alle $k \in \{0, \dots, m\}$ selbstadjungiert ist und daher B_k^{-1} ebenfalls selbstadjungiert ist, ist auch $B_k^{-1/2}$ selbstadjungiert. Somit erhalten wir die Abschätzung

$$\begin{aligned}
\sum_{k=0}^{m}(v_k, u_k)_A &= \sum_{k=0}^{m}(v_k, Au_k) \overset{(4.2)}{=} \sum_{k=0}^{m}(v_k, Q_kAu_k) \overset{(4.5)}{=} \sum_{k=0}^{m}(v_k, A_kP_ku_k) \\
&= \sum_{k=0}^{m}(B_k^{-1/2}B_k^{1/2}v_k, A_kP_ku_k) = \sum_{k=0}^{m}(B_k^{1/2}v_k, B_k^{-1/2}A_kP_ku_k) \\
&\overset{\substack{\text{Cauchy-}\\\text{Schwarz}}}{\leq} \left(\sum_{k=0}^{m}(B_k^{1/2}v_k, B_k^{1/2}v_k)\right)^{1/2} \\
&\qquad \left(\sum_{k=0}^{m}(B_k^{-1/2}A_kP_ku_k, B_k^{-1/2}A_kP_ku_k)\right)^{1/2} \\
&= \left(\sum_{k=0}^{m}(B_kv_k, v_k)\right)^{1/2} \left(\sum_{k=0}^{m}(B_k^{-1}A_kP_ku_k, A_kP_ku_k)\right)^{1/2} \\
&\overset{(4.10)}{=} \left(\sum_{k=0}^{m}(B_kv_k, v_k)\right)^{1/2} \left(\sum_{k=0}^{m}(T_ku_k, A_kP_ku_k)\right)^{1/2} \\
&\overset{(4.4)}{=} \left(\sum_{k=0}^{m}(B_kv_k, v_k)\right)^{1/2} \left(\sum_{k=0}^{m}(T_ku_k, P_ku_k)_A\right)^{1/2} \\
&\overset{(4.3)}{=} \left(\sum_{k=0}^{m}(B_kv_k, v_k)\right)^{1/2} \left(\sum_{k=0}^{m}(T_ku_k, u_k)_A\right)^{1/2} \\
&\overset{(5.3)}{\leq} \left(K_1 \|v\|_A^2\right)^{1/2} \left(\sum_{k=0}^{m}(T_ku_k, u_k)_A\right)^{1/2} \\
&= \sqrt{K_1}\,\|v\|_A \left(\sum_{k=0}^{m}(T_ku_k, u_k)_A\right)^{1/2}.
\end{aligned}$$

□

Um den Fehler des multiplikativen Unterraumkorrekturverfahrens nach jedem Teilschritt des Verfahrens zu beschreiben, führen wir den Fehleroperator E_k für alle $k \in \{0, \dots, m\}$ ein, der durch $E_{-1} := I$ und für $k \in \{0, \dots, m\}$ rekursiv durch $E_k := (I - T_k)E_{k-1}$ definiert ist. Der Gesamtfehler eines vollständigen Zyklus des multiplikativen Unterraumkorrekturverfahrens zwischen der exakten Lösung $u^* \in \mathcal{S}$ und der aus einer Näherungslösung $u^{(n)} \in \mathcal{S}$ durch das multiplikative Unterraumkorrekturverfahren berechneten Näherungslösung $u^{(n+1)} \in \mathcal{S}$ für $n \in \mathbb{N}_0$ lässt sich damit nach Lemma 5.1 durch den Fehleroperator E_m ausdrücken. Diese Fehleroperatoren können wir folgendermaßen mithilfe der Operatoren T_k, $k \in \{0, \dots, m\}$, ausdrücken:

Lemma 5.3. *Für $\ell \in \{0, \dots, m\}$ gilt*

$$I - E_{\ell-1} = \sum_{k=0}^{\ell-1} T_k E_{k-1}. \tag{5.8}$$

Beweis. Sei $\ell \in \{0, \dots, m\}$. Für alle $k \in \{0, \dots, \ell-1\}$ gilt

$$E_{k-1} - E_k = E_{k-1} - (I - T_k) E_{k-1} = E_{k-1} - E_{k-1} + T_k E_{k-1} = T_k E_{k-1}.$$

Daraus folgt

$$I - E_{\ell-1} = E_{-1} - E_{\ell-1} = \sum_{k=0}^{\ell-1} (E_{k-1} - E_k) = \sum_{k=0}^{\ell-1} T_k E_{k-1}. \qquad \square$$

Lemma 5.4. *Es sei $k \in \{0, \dots, m\}$. Unter Annahme 3 gilt für alle $u \in \mathcal{S}$*

$$\|T_k u\|_A^2 \leq \lambda (T_k u, u)_A. \tag{5.9}$$

Beweis. Es sei $u \in \mathcal{S}$ und es gelte Annahme 3. Dann gilt

$$\begin{aligned} \|T_k u\|_A^2 = (T_k u, T_k u)_A &\overset{(4.4)}{=} (T_k u, A_k T_k u) \overset{(5.6)}{\leq} \lambda (T_k u, B_k T_k u) \\ &\overset{(4.10)}{=} \lambda (T_k u, B_k B_k^{-1} A_k P_k u) = \lambda (T_k u, A_k P_k u) \overset{(4.4)}{=} \lambda (T_k u, P_k u)_A \\ &\overset{(4.3)}{=} \lambda (T_k u, u)_A. \end{aligned}$$

$\square$

Lemma 5.5. *Für $\ell \in \{0, \dots, m\}$ und $v \in \mathcal{S}$ gilt*

$$\|E_\ell v\|_A^2 = \|v\|_A^2 - \sum_{k=0}^{\ell} \left(2(T_k E_{k-1} v, E_{k-1} v)_A - (T_k E_{k-1} v, T_k E_{k-1} v)_A \right). \tag{5.10}$$

Beweis. Beweis per Induktion über ℓ.

Induktionsanfang: Es sei $\ell = 0$. Dann gilt

$$\begin{aligned} \|E_\ell v\|_A^2 = \|E_0 v\|_A^2 &= \|(I - T_0) v\|_A^2 = (Iv - T_0 v, Iv - T_0 v)_A \\ &= (Iv, Iv)_A - (Iv, T_0 v)_A - (T_0 v, Iv)_A + (T_0 v, T_0 v)_A \\ &= (v, v)_A - 2(T_0 v, Iv)_A + (T_0 v, T_0 v)_A \\ &= (v, v)_A - 2(T_0 Iv, Iv)_A + (T_0 Iv, T_0 Iv)_A \\ &= (v, v)_A - 2(T_0 E_{-1} v, E_{-1} v)_A + (T_0 E_{-1} v, T_0 E_{-1} v)_A \end{aligned}$$

$$= \|v\|_A^2 - \sum_{k=0}^{0} (2(T_k E_{k-1} v, E_{k-1} v)_A - (T_k E_{k-1} v, T_k E_{k-1} v)_A)$$
$$= \|v\|_A^2 - \sum_{k=0}^{\ell} (2(T_k E_{k-1} v, E_{k-1} v)_A - (T_k E_{k-1} v, T_k E_{k-1} v)_A) .$$

Induktionsvoraussetzung: Es sei $\ell \in \{0, \dots, m-1\}$ und es gelte

$$\|E_\ell v\|_A^2 = \|v\|_A^2 - \sum_{k=0}^{\ell} (2(T_k E_{k-1} v, E_{k-1} v)_A - (T_k E_{k-1} v, T_k E_{k-1} v)_A) .$$

Induktionsschritt: Es gilt

$$\begin{aligned}
\|E_{\ell+1} v\|_A^2 &= \|(I - T_{\ell+1}) E_l v\|_A^2 = ((I - T_{\ell+1}) E_\ell v, (I - T_{\ell+1}) E_\ell v)_A \\
&= (E_\ell v - T_{\ell+1} E_\ell v, E_\ell v - T_{\ell+1} E_\ell v)_A \\
&= (E_\ell v, E_\ell v)_A - (E_\ell v, T_{\ell+1} E_\ell v)_A - (T_{\ell+1} E_\ell v, E_\ell v)_A + (T_{\ell+1} E_\ell v, T_{\ell+1} E_\ell v)_A \\
&= \|E_\ell v\|_A^2 - 2(T_{\ell+1} E_\ell v, E_\ell v)_A + (T_{\ell+1} E_\ell v, T_{\ell+1} E_\ell v)_A \\
&\overset{\text{I.V.}}{=} \|v\|_A^2 - \sum_{k=0}^{\ell} (2(T_k E_{k-1} v, E_{k-1} v)_A - (T_k E_{k-1} v, T_k E_{k-1} v)_A) \\
&\quad - 2(T_{\ell+1} E_\ell v, E_\ell v)_A + (T_{\ell+1} E_\ell v, T_{\ell+1} E_\ell v)_A \\
&= \|v\|_A^2 - \sum_{k=0}^{\ell+1} (2(T_k E_{k-1} v, E_{k-1} v)_A - (T_k E_{k-1} v, T_k E_{k-1} v)_A) .
\end{aligned}$$

□

Mit den in diesem Kapitel bewiesenen Lemmata können wir nun den Konvergenzsatz für multiplikative Unterraumkorrekturverfahren unter den Annahmen aus Kapitel 5.1.1 beweisen.

Satz 5.6 (Konvergenz)**.** *Sei $n \in \mathbb{N}_0$, $u^{(n)} \in \mathcal{S}$ eine Näherungslösung und $u^* := A^{-1} f \in \mathcal{S}$ bezeichne die exakte Lösung. Unter den Annahmen 1, 2 und 3 gilt für einen vollständigen Zyklus des multiplikativen Unterraumkorrekturverfahrens die Fehlerabschätzung*

$$\left\| u^* - u^{(n+1)} \right\|_A^2 \leq \left(1 - \frac{2 - \lambda}{K_1 (1 + K_2)^2} \right) \left\| u^* - u^{(n)} \right\|_A^2 \tag{5.11}$$

und für die n-te Iterierte des multiplikativen Unterraumkorrekturverfahrens mit einer Näherungslösung $u^{(0)} \in \mathcal{S}$ gilt die Fehlerabschätzung

$$\left\| u^* - u^{(n)} \right\|_A^2 \leq \left(1 - \frac{2 - \lambda}{K_1 (1 + K_2)^2} \right)^n \left\| u^* - u^{(0)} \right\|_A^2 . \tag{5.12}$$

Beweis. Es gelten die Annahmen 1, 2 und 3. Es sei $v \in \mathcal{S}\backslash\{0\}$. Da $\mathcal{S} = \mathcal{V}_0 + \ldots + \mathcal{V}_m$ gilt, existieren $v_k \in \mathcal{V}_k$ für alle $k \in \{0, \ldots, m\}$ mit $v = \sum_{k=0}^{m} v_k$. Dann gilt

$$\begin{aligned}
\|v\|_A^2 = (v,v)_A = (v, \sum_{k=0}^{m} v_k)_A &= \sum_{k=0}^{m} (v, v_k)_A = (v, v_0)_A + \sum_{k=1}^{m} (v, v_k)_A \\
&= (E_{-1}v, v_0)_A + \sum_{k=1}^{m} (v, v_k)_A \\
&= (E_{-1}v, v_0)_A + \sum_{k=1}^{m} \left((E_{k-1}v, v_k)_A - (E_{k-1}v, v_k)_A + (v, v_k)_A\right) \\
&= (E_{-1}v, v_0)_A + \sum_{k=1}^{m} \left((E_{k-1}v, v_k)_A + ((I - E_{k-1})v, v_k)_A\right) \\
&= (E_{-1}v, v_0)_A + \sum_{k=1}^{m} (E_{k-1}v, v_k)_A + \sum_{k=1}^{m} ((I - E_{k-1})v, v_k)_A \\
&= \sum_{k=0}^{m} (E_{k-1}v, v_k)_A + \sum_{k=1}^{m} ((I - E_{k-1})v, v_k)_A. \qquad (5.13)
\end{aligned}$$

Wir schätzen die beiden Summen im Folgenden getrennt voneinander ab. Die erste Summe schätzen wir mit dem Lemma 5.2 ab. Es gilt

$$\begin{aligned}
\sum_{k=0}^{m} (E_{k-1}v, v_k)_A &= \sum_{k=0}^{m} (v_k, E_{k-1}v)_A \\
&\overset{(5.7)}{\leq} \sqrt{K_1}\, \|v\|_A \left(\sum_{k=0}^{m} (T_k E_{k-1}v, E_{k-1}v)_A \right)^{1/2}. \qquad (5.14)
\end{aligned}$$

Die Abschätzung der zweiten Summe beruht auf der verschärften Cauchy-Schwarz-Ungleichung aus Annahme 2. Zudem verwenden wir in dieser Abschätzung das Lemma 5.3. Es gilt

$$\begin{aligned}
\sum_{\ell=1}^{m} ((I - E_{\ell-1})v, v_\ell)_A &\overset{(5.8)}{=} \sum_{\ell=1}^{m} \left(\left(\sum_{k=0}^{\ell-1} T_k E_{k-1} \right) v, v_\ell \right)_A = \sum_{\ell=1}^{m} \left(\sum_{k=0}^{\ell-1} T_k E_{k-1}v, v_\ell \right)_A \\
&= \sum_{\ell=1}^{m} \sum_{k=0}^{\ell-1} (T_k E_{k-1}v, v_\ell)_A \\
&\overset{(5.4)}{\leq} \sum_{\ell=1}^{m} \sum_{k=0}^{\ell-1} \gamma_{k\ell} (B_k T_k E_{k-1}v, T_k E_{k-1}v)^{1/2} (B_\ell v_\ell, v_\ell)^{1/2} \\
&\leq \sum_{\ell=0}^{m} \sum_{k=0}^{m} \gamma_{k\ell} (B_k T_k E_{k-1}v, T_k E_{k-1}v)^{1/2} (B_\ell v_\ell, v_\ell)^{1/2}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(5.5)}{\leq} K_2 \left(\sum_{k=0}^{m} (B_k T_k E_{k-1} v, T_k E_{k-1} v) \right)^{1/2} \left(\sum_{\ell=0}^{m} (B_\ell v_\ell, v_\ell) \right)^{1/2} \\
&\overset{(4.10)}{=} K_2 \left(\sum_{k=0}^{m} (B_k B_k^{-1} A_k P_k E_{k-1} v, T_k E_{k-1} v) \right)^{1/2} \\
&\qquad \left(\sum_{\ell=0}^{m} (B_\ell v_\ell, v_\ell) \right)^{1/2} \\
&= K_2 \left(\sum_{k=0}^{m} (A_k P_k E_{k-1} v, T_k E_{k-1} v) \right)^{1/2} \left(\sum_{\ell=0}^{m} (B_\ell v_\ell, v_\ell) \right)^{1/2} \\
&\overset{(4.4)}{=} K_2 \left(\sum_{k=0}^{m} (P_k E_{k-1} v, T_k E_{k-1} v)_A \right)^{1/2} \left(\sum_{\ell=0}^{m} (B_\ell v_\ell, v_\ell) \right)^{1/2} \\
&\overset{(4.3)}{=} K_2 \left(\sum_{k=0}^{m} (E_{k-1} v, T_k E_{k-1} v)_A \right)^{1/2} \left(\sum_{\ell=0}^{m} (B_\ell v_\ell, v_\ell) \right)^{1/2} \\
&= K_2 \left(\sum_{\ell=0}^{m} (B_\ell v_\ell, v_\ell) \right)^{1/2} \left(\sum_{k=0}^{m} (T_k E_{k-1} v, E_{k-1} v)_A \right)^{1/2} \\
&\overset{(5.3)}{\leq} K_2 \left(K_1 \|v\|_A^2 \right)^{1/2} \left(\sum_{k=0}^{m} (T_k E_{k-1} v, E_{k-1} v)_A \right)^{1/2} \\
&= \sqrt{K_1} K_2 \|v\|_A \left(\sum_{k=0}^{m} (T_k E_{k-1} v, E_{k-1} v)_A \right)^{1/2}. \qquad (5.15)
\end{aligned}
$$

Mithilfe dieser Abschätzungen für die beiden Summen ergibt sich

$$
\begin{aligned}
\|v\|_A^2 &\overset{(5.13)}{=} \sum_{k=0}^{m} (E_{k-1} v, v_k)_A + \sum_{k=1}^{m} ((I - E_{k-1}) v, v_k)_A \\
&\overset{(5.14)}{\leq} \sqrt{K_1} \|v\|_A \left(\sum_{k=0}^{m} (T_k E_{k-1} v, E_{k-1} v)_A \right)^{1/2} + \sum_{k=1}^{m} ((I - E_{k-1}) v, v_k)_A \\
&\overset{(5.15)}{\leq} \sqrt{K_1} \|v\|_A \left(\sum_{k=0}^{m} (T_k E_{k-1} v, E_{k-1} v)_A \right)^{1/2} \\
&\qquad + \sqrt{K_1} K_2 \|v\|_A \left(\sum_{k=0}^{m} (T_k E_{k-1} v, E_{k-1} v)_A \right)^{1/2} \\
&= \sqrt{K_1}(1 + K_2) \|v\|_A \left(\sum_{k=0}^{m} (T_k E_{k-1} v, E_{k-1} v)_A \right)^{1/2}.
\end{aligned}
$$

Mittels Division durch $\|v\|_A$ und anschließendem Quadrieren erhalten wir

$$\|v\|_A^2 \leq K_1(1+K_2)^2 \left(\sum_{k=0}^{m} (T_k E_{k-1} v, E_{k-1} v)_A \right). \tag{5.16}$$

Der Fehler in einem Zyklus des multiplikativen Unterraumkorrekturverfahrens lässt sich mit der vorbereitenden Abschätzung (5.16) und der Annahme der lokalen Konvergenz abschätzen. Als weitere Hilfsmittel für diese Abschätzung verwenden wir die Lemmata 5.4 und 5.5. Es gilt

$$\begin{aligned}
\|E_m v\|_A^2 &\overset{(5.10)}{=} \|v\|_A^2 - \sum_{k=0}^{m} \left(2(T_k E_{k-1} v, E_{k-1} v)_A - (T_k E_{k-1} v, T_k E_{k-1} v)_A\right) \\
&= \|v\|_A^2 - \sum_{k=0}^{m} \left(2(T_k E_{k-1} v, E_{k-1} v)_A - \|T_k E_{k-1} v\|_A^2\right) \\
&\overset{(5.9)}{\leq} \|v\|_A^2 - \sum_{k=0}^{m} \left(2(T_k E_{k-1} v, E_{k-1} v)_A - \lambda (T_k E_{k-1} v, E_{k-1} v)_A\right) \\
&= \|v\|_A^2 - (2-\lambda) \sum_{k=0}^{m} (T_k E_{k-1} v, E_{k-1} v)_A \qquad (5.17) \\
&\overset{(5.16)}{\leq} \|v\|_A^2 - (2-\lambda) \frac{\|v\|_A^2}{K_1 (1+K_2)^2} \\
&= \left(1 - \frac{2-\lambda}{K_1 (1+K_2)^2}\right) \|v\|_A^2 .
\end{aligned}$$

Mit $v = u^* - u^{(n)} \in \mathcal{S}$ erhalten wir

$$\left\|u^* - u^{(n+1)}\right\|_A^2 \overset{(5.1)}{=} \left\|E_m \left(u^* - u^{(n)}\right)\right\|_A^2 \leq \left(1 - \frac{2-\lambda}{K_1 (1+K_2)^2}\right) \left\|u^* - u^{(n)}\right\|_A^2 .$$

Durch Induktion erhalten wir aus dieser Fehlerabschätzung (5.11) für einen vollständigen Zyklus die Abschätzung (5.12) für n Zyklen des multiplikativen Unterraumkorrekturverfahrens. □

Der Faktor in der Konvergenzabschätzung in Satz 5.6 hängt nur von der Konstanten K_1 aus der Stabilitätsannahme (5.3), der Konstanten K_2 aus der verschärften Cauchy-Schwarz-Ungleichung (5.5) und der Konstanten λ aus der lokalen Konvergenzannahme (5.6) ab.

Da $0 < \lambda < 2$ nach Annahme 3 gilt, folgt daraus $2 - \lambda > 0$. Für die beiden Konstanten K_1 und K_2 aus den Annahmen 1 und 2 gilt $K_1 > 0$ und $K_2 \geq 0$, sodass damit

$K_1 (1 + K_2)^2 > 0$ gilt. Daraus folgt $\frac{2-\lambda}{K_1(1+K_2)^2} > 0$ und somit gilt $1 - \frac{2-\lambda}{K_1(1+K_2)^2} < 1$. Da für alle $n \in \mathbb{N}_0$ und $u^{(n)} \in \mathcal{S}$ nach (5.11)

$$0 \leq \left\| u^* - u^{(n+1)} \right\|_A^2 \leq \left(1 - \frac{2-\lambda}{K_1 (1 + K_2)^2} \right) \left\| u^* - u^{(n)} \right\|_A^2$$

gilt, folgt auch $0 \leq 1 - \frac{2-\lambda}{K_1(1+K_2)^2}$. Insgesamt erfüllt die Konvergenzrate im Satz 5.6 die Abschätzung

$$0 \leq 1 - \frac{2-\lambda}{K_1 (1 + K_2)^2} < 1.$$

Für die Konstanten λ, K_1 und K_2 aus den Annahmen 1, 2 und 3 ist es wünschenswert, dass diese unabhängig von der Anzahl der Unterräume m sind, um eine von m unabhängige Konvergenzrate zu erhalten.

Um die Konvergenz des symmetrischen multiplikativen Unterraumkorrekturverfahrens zu zeigen, verwenden wir die Konvergenz des multiplikativen Unterraumkorrekturverfahrens.

Satz 5.7 (Konvergenz)**.** *Es sei $n \in \mathbb{N}_0$, $u^{(n)} \in \mathcal{S}$ eine Näherungslösung und $u^* := A^{-1} f \in \mathcal{S}$ bezeichne die exakte Lösung. Unter den Annahmen 1, 2 und 3 gilt für einen vollständigen Zyklus des symmetrischen multiplikativen Unterraumkorrekturverfahrens die Fehlerabschätzung*

$$\left\| u^* - u^{(n+1)} \right\|_A \leq \left(1 - \frac{2-\lambda}{K_1(1 + K_2)^2} \right) \left\| u^* - u^{(n)} \right\|_A \tag{5.18}$$

und für die n-te Iterierte des symmetrischen multiplikativen Unterraumkorrekturverfahrens mit einer Näherungslösung $u^{(0)} \in \mathcal{S}$ gilt die Fehlerabschätzung

$$\left\| u^* - u^{(n)} \right\|_A \leq \left(1 - \frac{2-\lambda}{K_1(1 + K_2)^2} \right)^n \left\| u^* - u^{(0)} \right\|_A . \tag{5.19}$$

Beweis. Es gelten die Annahmen 1, 2 und 3. Der Fehler zwischen der Lösung u^* und der neuen Näherungslösung $u^{(n+1)}$ des symmetrischen multiplikativen Unterraumkorrekturverfahrens lässt sich durch

$$\begin{aligned}
&u^* - u^{(n+1)} = u^* - w_{2m+1}^{(n)} \\
&= u^* - \left(u^{(n)} + (I - (I - T_0) \cdot \ldots \cdot (I - T_m)\,(I - T_m) \cdot \ldots \cdot (I - T_0))\, A^{-1} \left(f - A u^{(n)} \right) \right) \\
&= A^{-1} f - u^{(n)} - A^{-1} \left(f - A u^{(n)} \right)
\end{aligned}$$

$$
\begin{aligned}
&+ (I - T_0) \cdot \ldots \cdot (I - T_m) (I - T_m) \cdot \ldots \cdot (I - T_0) A^{-1} \left(f - A u^{(n)} \right) \\
&= (I - T_0) \cdot \ldots \cdot (I - T_m) (I - T_m) \cdot \ldots \cdot (I - T_0) \left(A^{-1} f - u^{(n)} \right) \\
&= (I - T_0) \cdot \ldots \cdot (I - T_m) (I - T_m) \cdot \ldots \cdot (I - T_0) \left(u^* - u^{(n)} \right) \\
&= E_m^* E_m \left(u^* - u^{(n)} \right)
\end{aligned}
$$

darstellen. Da wegen (5.17) die Abschätzung

$$
\| E_m^* E_m \|_A = \| E_m \|_A^2 = \sup_{v \in \mathcal{S} \setminus \{0\}} \frac{\| E_m v \|_A^2}{\| v \|_A^2} \leq 1 - \frac{2 - \lambda}{K_1 (1 + K_2)^2}
$$

gilt, erhalten wir insbesondere für $u^* - u^{(n+1)} \in \mathcal{S}$ die Abschätzung

$$
\left\| u^* - u^{(n+1)} \right\|_A = \left\| E_m^* E_m \left(u^* - u^{(n)} \right) \right\|_A \leq \left(1 - \frac{2 - \lambda}{K_1 (1 + K_2)^2} \right) \left\| u^* - u^{(n)} \right\|_A .
$$

Durch Induktion erhalten wir aus dieser Fehlerabschätzung (5.18) für einen vollständigen Zyklus die Abschätzung (5.19) für n Zyklen des symmetrischen multiplikativen Unterraumkorrekturverfahrens. □

Um das symmetrische multiplikative Unterraumkorrekturverfahren als Vorkonditionierer für das Verfahren der konjugierten Gradienten zu verwenden, bleibt es die positiv Definitheit des Operators $N_{\text{mul, sym}}$ zu zeigen.

Lemma 5.8. *$N_{mul,\ sym}$ ist positiv definit.*

Beweis. Mit dem Fehleroperator E_m erhalten wir die Darstellung

$$
N_{\text{mul, sym}} = \left(I - (I - T_0) \ldots (I - T_m) (I - T_m) \ldots (I - T_0) \right) A^{-1} = \left(I - E_m^* E_m \right) A^{-1}.
$$

Weil nach (5.17) für alle $v \in \mathcal{S} \setminus \{0\}$ die Abschätzung

$$
(E_m v, E_m v)_A = \| E_m v \|_A^2 \leq \left(1 - \frac{2 - \lambda}{K_1 (1 + K_2)^2} \right) \| v \|_A^2 < \| v \|_A^2 = (v, v)_A
$$

gilt, gilt für jedes $v \in \mathcal{S} \setminus \{0\}$

$$
\begin{aligned}
0 &< (v, v)_A - (E_m v, E_m v)_A = (v, v)_A - (v, E_m^* E_m v)_A = \left(v, \left(I - E_m^* E_m \right) v \right)_A \\
&= \left(Av, \left(I - E_m^* E_m \right) A^{-1} Av \right) = (Av, N_{\text{mul, sym}} Av).
\end{aligned}
$$

Aus der Regularität von A folgt somit, dass $N_{\text{mul, sym}}$ positiv definit ist. □

Der folgende Satz enthält eine Abschätzung der Konditionszahl des vorkonditionierten Operators $N_{\text{mul, sym}}A$, die für die Konvergenzgeschwindigkeit des vorkonditionierten Verfahrens der konjugierten Gradienten entscheidend ist.

Satz 5.9. *Für die Konditionszahl des vorkonditionierten Operators $N_{mul,\ sym}A$ gilt*

$$\kappa\left(N_{mul,\ sym}A\right) \leq \frac{\left(1+2\lambda^2 K_2^2\right) K_1}{2-\lambda}. \tag{5.20}$$

Beweis. Siehe [SBG96, Kapitel 5, Lemma 4]. □

5.2 Konvergenz additiver Unterraumkorrekturverfahren

Die Konvergenztheorie für additive Unterraumkorrekturverfahren basiert auf einer Zerlegung des Raums $\mathcal{S}$ in eine Summe von Unterräumen $\mathcal{V}_k \subseteq \mathcal{W}_k$ für alle $k \in \{0, \ldots, m\}$ mit

$$\mathcal{S} = \mathcal{V}_0 + \ldots + \mathcal{V}_m. \tag{5.21}$$

Der Fehler des additiven Unterraumkorrekturverfahrens, das in Kapitel 4.2 eingeführt wurde, kann durch die in (4.9) eingeführten Operatoren T_k für $k \in \{0, \ldots, m\}$ folgendermaßen dargestellt werden.

Lemma 5.10. *Sei $n \in \mathbb{N}_0$ und eine Näherungslösung $u^{(n)} \in \mathcal{S}$ gegeben. Es bezeichne $u^* := A^{-1}f \in \mathcal{S}$ die exakte Lösung. Für den Fehler des additiven Unterraumkorrekturverfahrens gilt*

$$u^* - u^{(n+1)} = (I - CA)\left(u^* - u^{(n)}\right),$$

wobei $C := \sum_{k=0}^{m} B_k^{-1} Q_k : \mathcal{S}' \to \mathcal{S}$ selbstadjungiert und positiv definit ist.

Beweis. Es gilt

$$\begin{aligned} u^* - u^{(n+1)} &= u^* - \left(u^{(n)} + \sum_{k=0}^{m} B_k^{-1} Q_k \left(f - Au^{(n)}\right)\right) \\ &= u^* - u^{(n)} - \sum_{k=0}^{m} B_k^{-1} Q_k \left(Au^* - Au^{(n)}\right) \\ &= u^* - u^{(n)} - \sum_{k=0}^{m} B_k^{-1} Q_k A \left(u^* - u^{(n)}\right) \end{aligned}$$

$$= \left(I - \sum_{k=0}^{m} B_k^{-1} Q_k A \right) \left(u^* - u^{(n)} \right) = (I - CA) \left(u^* - u^{(n)} \right).$$

Weil die B_k und damit auch die B_k^{-1} für alle $k \in \{0, \ldots, m\}$ selbstadjungiert sind, gilt für alle $u, v \in \mathcal{S}'$

$$\begin{aligned}(Cu, v) = (\sum_{k=0}^{m} B_k^{-1} Q_k u, v) = \sum_{k=0}^{m} (B_k^{-1} Q_k u, v) &\overset{(4.2)}{=} \sum_{k=0}^{m} (B_k^{-1} Q_k u, Q_k v) \\ = \sum_{k=0}^{m} (Q_k u, B_k^{-1} Q_k v) &\overset{(4.2)}{=} \sum_{k=0}^{m} (u, B_k^{-1} Q_k v) = (u, \sum_{k=0}^{m} B_k^{-1} Q_k v) = (u, Cv),\end{aligned}$$

sodass C selbstadjungiert ist. Weiter gilt für $u \in \mathcal{S}' \backslash \{0\}$

$$(Cu, u) = (\sum_{k=0}^{m} B_k^{-1} Q_k u, u) = \sum_{k=0}^{m} (B_k^{-1} Q_k u, u) \overset{(4.2)}{=} \sum_{k=0}^{m} (B_k^{-1} Q_k u, Q_k u) \geq 0,$$

da die B_k und somit B_k^{-1} für alle $k \in \{0, \ldots, m\}$ positiv definit sind. Damit ist C positiv semidefinit. Sei $u \in \mathcal{S}'$ mit

$$0 = (Cu, u) = (\sum_{k=0}^{m} B_k^{-1} Q_k u, u) = \sum_{k=0}^{m} (B_k^{-1} Q_k u, u) \overset{(4.2)}{=} \sum_{k=0}^{m} (B_k^{-1} Q_k u, Q_k u).$$

Da B_k^{-1} positiv definit ist, gilt dann $Q_k u = 0$ für alle $k \in \{0, \ldots, m\}$ und somit

$$0 = (Q_k u, v_k) \overset{(4.2)}{=} (u, v_k)$$

für alle $v_k \in \mathcal{V}_k$. Daraus folgt für alle $w \in \mathcal{S}$ mit $w = \sum_{k=0}^{m} v_k$ und $v_k \in \mathcal{V}_k$ für alle $k \in \{0, \ldots, m\}$

$$0 = \sum_{k=0}^{m} (u, v_k) = (u, \sum_{k=0}^{m} v_k) = (u, w)$$

und damit $u = 0$. Somit ist C positiv definit. □

5.2.1 Annahmen

Für die Konvergenztheorie benötigen wir zwei weitere Annahmen, wobei die erste Annahme mit der Annahme 1 aus der Konvergenztheorie für multiplikative Unterraumkorrekturverfahren übereinstimmt.

Annahme 1 (Stabilität der Zerlegung)**.** *Es existiert eine Konstante $K_1 \in \mathbb{R}_{>0}$, sodass*

für alle $v \in \mathcal{S}$ *mit* $v = \sum_{k=0}^{m} v_k$ *und* $v_k \in \mathcal{V}_k$ *für alle* $k \in \{0, \dots, m\}$

$$\sum_{k=0}^{m} (B_k v_k, v_k) \leq K_1 \|v\|_A^2 \tag{5.22}$$

gilt.

Annahme 2' (Verschärfte Cauchy-Schwarz-Ungleichung)**.** *Es existiert eine symmetrische Matrix* $\gamma = (\gamma_{k\ell})_{k,\ell \in \{0,\dots,m\}} \in \mathbb{R}^{(m+1)\times(m+1)}$*, sodass*

$$(w_k, v_\ell)_A \leq \gamma_{k\ell} (B_k w_k, w_k)^{1/2} (B_\ell v_\ell, v_\ell)^{1/2} \tag{5.23}$$

für alle $k, \ell \in \{0, \dots, m\}$*,* $w_k \in \mathcal{W}_k$ *und* $v_\ell \in \mathcal{W}_\ell$ *gilt. Weiter existiert eine Konstante* $K_2 \in \mathbb{R}_{\geq 0}$ *mit*

$$\|w\|_A^2 \leq K_2 \sum_{k=0}^{m} (B_k w_k, w_k) \tag{5.24}$$

für alle $w \in \mathcal{S}$ *mit der Zerlegung* $w = \sum_{k=0}^{m} w_k$ *und* $w_k \in \mathcal{W}_k$ *für alle* $k \in \{0, \dots, m\}$*.*

Die Annahme 1 beschränkt, wie bei dem multiplikativen Unterraumkorrekturverfahren, die Summe der Energienormen der approximativen Lösungsverfahren, sodass diese nicht zu groß ist. Die Annahme 2' ist eine stärkere Annahme als Annahme 2 des multiplikativen Verfahrens und beschränkt die Energienormen der approximativen Lösungsverfahren nach unten, damit diese nicht zu klein sind. Diese Annahme ist im Gegensatz zu Annahme 2 unabhängig von der Reihenfolge der Zerlegung.

5.2.2 Konvergenzbeweis

Für die Konvergenz des Verfahrens kann es notwendig sein, die Schrittweite des additiven Unterraumkorrekturverfahrens anzupassen, indem die neue Näherungslösung für $n \in \mathbb{N}_0$ aus einer gegebenen Näherungslösung $u^{(n)} \in \mathcal{S}$ durch

$$u^{(n+1)} = u^{(n)} - \theta C \left(A u^{(n)} - f \right)$$

mit $C := \sum_{k=0}^{m} B_k^{-1} Q_k$ und einem Dämpfungsparameter $\theta \in \mathbb{R}_{>0}$ berechnet wird. Mit den Annahmen aus Kapitel 5.2.1 können wir den Konvergenzsatz für additive Unterraumkorrekturverfahren beweisen.

Satz 5.11 (Konvergenz)**.** *Es seien die Annahmen 1 und 2' erfüllt und* $K_2 > 0$*. Für* $\theta \in \left(0, \frac{2}{K_1 K_2^2}\right)$ *ist das gedämpfte additive Unterraumkorrekturverfahren konvergent.*

Beweis. Sei $u \in \mathcal{S}$. Dann gilt mit Lemma 5.10

$$(CAu, u)_A = (CAu, Au) = (Au, CAu) = (u, CAu)_A.$$

Sei $v \in \mathcal{S}\backslash\{0\}$. Dann existieren $v_k \in \mathcal{V}_k$ für alle $k \in \{0, \dots, m\}$ mit $v = \sum_{k=0}^{m} v_k$. Dann gilt mit Annahme 1 und Lemma 5.2

$$\begin{aligned}
\|v\|_A^2 = (v,v)_A = (\sum_{k=0}^{m} v_k, v)_A = \sum_{k=0}^{m} (v_k, v)_A &\overset{(4.3)}{=} \sum_{k=0}^{m} (v_k, P_k v)_A \\
&\overset{(5.7)}{\leq} \sqrt{K_1}\, \|v\|_A \left(\sum_{k=0}^{m} (T_k P_k v, P_k v)_A \right)^{1/2} \\
&\overset{(4.10)}{=} \sqrt{K_1}\, \|v\|_A \left(\sum_{k=0}^{m} (B_k^{-1} A_k P_k P_k v, P_k v)_A \right)^{1/2} \\
&= \sqrt{K_1}\, \|v\|_A \left(\sum_{k=0}^{m} (B_k^{-1} A_k P_k v, P_k v)_A \right)^{1/2} \overset{(4.10)}{=} \sqrt{K_1}\, \|v\|_A \left(\sum_{k=0}^{m} (T_k v, P_k v)_A \right)^{1/2} \\
&\overset{(4.3)}{=} \sqrt{K_1}\, \|v\|_A \left(\sum_{k=0}^{m} (T_k v, v)_A \right)^{1/2} = \sqrt{K_1}\, \|v\|_A \, (\sum_{k=0}^{m} T_k v, v)_A^{1/2} \\
&\overset{(4.10)}{=} \sqrt{K_1}\, \|v\|_A \, (\sum_{k=0}^{m} B_k^{-1} Q_k A v, v)_A^{1/2} = \sqrt{K_1}\, \|v\|_A \, (CAv, v)_A^{1/2}.
\end{aligned}$$

Daraus folgt nach Division durch $\|v\|_A$ und quadrieren der Gleichung

$$(v,v)_A = \|v\|_A^2 \leq K_1 (CAv, v)_A.$$

Für alle $v \in \mathcal{S}\backslash\{0\}$ gilt mit Annahme 2'

$$\begin{aligned}
\|CAv\|_A^2 = \left\| \sum_{k=0}^{m} T_k v \right\|_A^2 &\overset{(5.24)}{\leq} K_2 \sum_{k=0}^{m} (B_k T_k v, T_k v) \overset{(4.10)}{=} K_2 \sum_{k=0}^{m} (B_k B_k^{-1} Q_k A v, T_k v) \\
&= K_2 \sum_{k=0}^{m} (Q_k A v, T_k v) \overset{(4.2)}{=} K_2 \sum_{k=0}^{m} (Av, T_k v) = K_2 \sum_{k=0}^{m} (v, T_k v)_A \\
&= K_2 (v, \sum_{k=0}^{m} T_k v)_A = K_2 (v, CAv)_A = K_2 (CAv, v)_A \overset{\text{Cauchy-Schwarz}}{\leq} K_2 \, \|CAv\|_A \, \|v\|_A .
\end{aligned}$$

Damit erhalten wir $\|CAv\|_A \leq K_2 \|v\|_A$. Für $v \in \mathcal{S}\backslash\{0\}$ und $\theta \in \left(0, \frac{2}{K_1 K_2^2}\right)$ gilt

$$\|(I - \theta CA)\, v\|_A^2 = ((I - \theta CA)\, v, (I - \theta CA)\, v)_A$$

$$
\begin{aligned}
&= (v,v)_A - \theta(CAv,v)_A - \theta(v,CAv)_A + \theta^2(CAv,CAv)_A \\
&= \|v\|_A^2 - 2\theta(CAv,v)_A + \theta^2 \|CAv\|_A^2 \\
&\leq \|v\|_A^2 - 2\theta(CAv,v)_A + \theta^2 K_2^2 \|v\|_A^2 \\
&\leq \|v\|_A^2 - 2\theta\frac{1}{K_1}\|v\|_A^2 + \theta^2 K_2^2 \|v\|_A^2 \\
&= \left(1 - 2\theta\frac{1}{K_1} + \theta^2 K_2^2\right)\|v\|_A^2 \\
&< \left(1 - 2\cdot\frac{2}{K_1 K_2^2}\frac{1}{K_1} + \frac{4}{K_1^2 K_2^4}K_2^2\right)\|v\|_A^2 \\
&= \left(1 - \frac{4}{K_1^2 K_2^2} + \frac{4}{K_1^2 K_2^2}\right)\|v\|_A^2 \\
&= \|v\|_A^2 .
\end{aligned}
$$

Daraus folgt

$$
\|I - \theta CA\|_A = \sup_{v\in\mathcal{S}\setminus\{0\}} \frac{\|(I-\theta CA)\,v\|_A}{\|v\|_A} < \sup_{v\in\mathcal{S}\setminus\{0\}} \frac{\|v\|_A}{\|v\|_A} = 1.
$$

Nach Satz 2.10 gilt $\varrho\,(I-\theta CA) \leq \|I-\theta CA\|_A < 1$ und daher folgt die Konvergenz des additiven Unterraumkorrekturverfahrens aus Satz 2.24. □

Der folgende Satz enthält eine Abschätzung der Konditionszahl des vorkonditionierten Operators CA, die für die Konvergenzgeschwindigkeit des vorkonditionierten Verfahrens der konjugierten Gradienten entscheidend ist.

Satz 5.12. *Für die Konditionszahl des vorkonditionierten Operators CA gilt*

$$
\kappa\,(CA) \leq \lambda\,(1+K_2)\,K_1. \tag{5.25}
$$

Beweis. Siehe [SBG96, Kapitel 5, Lemma 3]. □

6 Verfahren der hierarchischen Basen

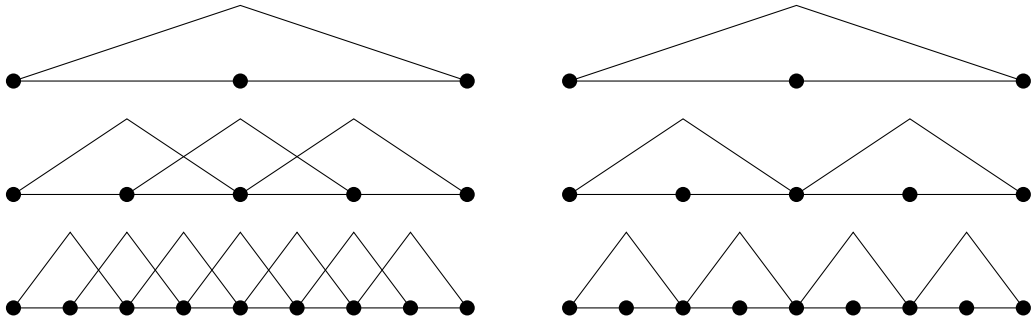
In diesem Kapitel führen wir als spezielles Unterraumkorrekturverfahren das Verfahren der hierarchischen Basen ein. Dazu beschreiben wir das Verfahren zunächst allgemein und definieren die Unterräume $\mathcal{W}_k$ für $k \in \{0, \ldots, m\}$ mit $m \in \mathbb{N}$, die eine Zerlegung des zugrundeliegenden, endlichdimensionalen Vektorraums $\mathcal{S}$ bilden und das Unterraumkorrekturverfahren auszeichnen. Die Konvergenz dieses Verfahrens beweisen wir für das in Kapitel 3.6 eingeführte Modellproblem, indem wir die Annahmen der Konvergenztheorie für Unterraumkorrekturverfahren aus Kapitel 5 beweisen.

6.1 Beschreibung des Verfahrens der hierarchischen Basen

Die Idee des Verfahrens der hierarchischen Basen geht, wie in [Yse92] erläutert, auf die hierarchische Darstellung von stetigen Funktionen, die bereits in [Fab08] entwickelt wurde, zurück. Das Verfahren benötigt ein hierarchisches System von Triangulierungen, das beispielsweise aus der Diskretisierung einer partiellen Differentialgleichung mit der Finiten Elemente Methode entsteht. Anstatt nur das lineare Gleichungssystem auf der feinsten Gitterebene zu lösen, werden die entsprechenden linearen Gleichungssysteme auf allen Gitterebenen simultan gelöst.

Das Verfahren der hierarchischen Basen gibt es als multiplikative und als additive Variante, die jeweils auf der Verwendung der hierarchischen Basis anstelle der Knotenbasis beruhen. Die additive Variante wurde in [Yse86b] im zweidimensionalen Raum analysiert, während das Verfahren im eindimensionalen Raum bereits 1982 in [ZKB82] analysiert wurde. In [BDY88] wurde die multiplikative Variante des Verfahrens entwickelt und analysiert. Diese beiden Varianten wurden in [Xu92] und [Yse93] als Unterraumkorrekturverfahren aufgefasst und die Konvergenz dieser Verfahren mittels der Theorie der Unterraumkorrekturverfahren gezeigt.

Im Folgenden sei $\Omega \subset \mathbb{R}^2$ ein beschränktes, polygonales Gebiet und $m \in \mathbb{N}_0$. Zudem sei $\mathcal{T}_0$ eine grobe zulässige Anfangstriangulierung des Gebiets Ω und $(\mathcal{T}_k)_{k=0,\ldots,m}$ eine geschachtelte Familie von Triangulierungen, die durch Verfeinerungen aus $\mathcal{T}_0$ entsteht. Um die Triangulierungen zu erzeugen, sind verschiedene Verfeinerungsstrategien anwendbar,



(a) Knotenbasis (b) Hierarchische Basis

Abbildung 6.1: Darstellung der eindimensionalen Knotenbasis und hierarchischen Basis für stückweise lineare Finite Elemente auf einem hierarchischen Gitter, wobei für $k \in \{1, 2\}$ für die hierarchische Basis jeweils nur die Basisfunktionen $\psi_i^{(k)}$ für $i \in \mathcal{N}_k^{\mathrm{HB}}$ abgebildet sind.

wie zum Beispiel die in Kapitel 3.5 beschriebenen Strategien zur Gitterverfeinerung. Für alle $k \in \{0, \ldots, m\}$ bezeichnet $\mathcal{S}_k$ den zur Triangulierung $\mathcal{T}_k$ gehörenden Finite Elemente Raum der stückweise linearen Finiten Elemente, das heißt

$$\mathcal{S}_k := \left\{ u \in C\left(\bar{\Omega}\right) : u|_{\partial\Omega} = 0,\, u|_T \in \mathcal{P}_1^2 \text{ für alle } T \in \mathcal{T}_k \right\},$$

und $\mathcal{N}_k$ bezeichnet die Menge der inneren Knoten der Triangulierung $\mathcal{T}_k$. Insbesondere gilt $\mathcal{S}_0 \subset \mathcal{S}_1 \subset \ldots \subset \mathcal{S}_m$.

Eine Basis des Finite Elemente Raums $\mathcal{S}_k$ für $k \in \{0, \ldots, m\}$ ist die Knotenbasis $\left(\varphi_i^{(k)}\right)_{i \in \mathcal{N}_k}$. In Abbildung 6.1(a) ist diese Basis für eine Folge von drei geschachtelten Triangulierungen im eindimensionalen Raum dargestellt. Um die Gitterhierarchie der geschachtelten Triangulierungen auszunutzen, beruht das Verfahren der hierarchischen Basen auf der Verwendung der hierarchischen Basis anstatt der Knotenbasis für die Finiten Elemente Räume $\mathcal{S}_k$ für alle $k \in \{0, \ldots, m\}$. Während die Knotenbasis des Finite Elemente Raums $\mathcal{S}_k$ für $k \in \{0, \ldots, m\}$ von den Knotenbasen der Finiten Elemente Räume $\mathcal{S}_\ell$ für $\ell \in \{0, \ldots, k-1\}$ unabhängig ist, ergänzt die hierarchische Basis dieses Finite Elemente Raums $\mathcal{S}_k$ die hierarchischen Basen der Finiten Elemente Räume $\mathcal{S}_\ell$ für $\ell \in \{0, \ldots, k-1\}$.

Definition 6.1 (Hierarchische Basis)**.** *Sei $k \in \{0, \ldots, m\}$ und der Finite Elemente Raum $\mathcal{S}_k$ gegeben. Setze $\mathcal{N}_0^{HB} := \mathcal{N}_0$ und $\mathcal{N}_\ell^{HB} := \mathcal{N}_\ell \backslash \mathcal{N}_{\ell-1}$ für alle $\ell \in \{1, \ldots, m\}$. Für $k = 0$ besteht die* hierarchische Basis *des Finite Elemente Raums $\mathcal{S}_0$ aus den Basisfunktionen $\left\{\psi_i^{(0)} : i \in \mathcal{N}_0^{HB}\right\}$ mit $\psi_i^{(0)} := \varphi_i^{(0)}$ für $i \in \mathcal{N}_0^{HB}$. Für $k > 0$ besteht die* hierarchi-

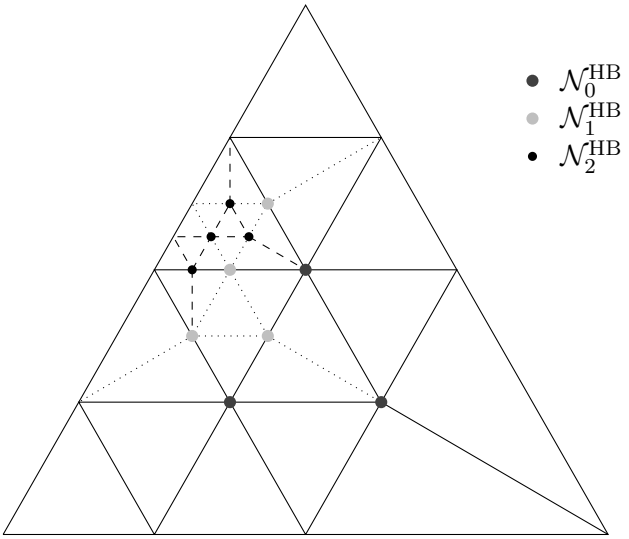


Abbildung 6.2: Disjunkte Zerlegung der Knotenmenge einer zweidimensionalen, nicht gleichmäßig verfeinerten Triangulierung: Dargestellt ist eine Folge von drei geschachtelten Triangulierungen, wobei die Anfangstriangulierung $\mathcal{T}_0$ mit durchgezogenen Linien, die zusätzlichen Kanten der Triangulierung $\mathcal{T}_1$ gepunktet und der Triangulierung $\mathcal{T}_2$ gestrichelt dargestellt sind. Für jede Triangulierung $\mathcal{T}_k$ für $k \in \{0,1,2\}$ ist die Teilmenge $\mathcal{N}_k^{\mathrm{HB}}$ der inneren Knoten dieser Triangulierung markiert.

sche Basis *des Finite Elemente Raums* $\mathcal{S}_k$ *aus den Basisfunktionen* $\left\{\psi_i^{(k)} \,:\, i \in \mathcal{N}_{k-1}\right\} \cup \left\{\psi_i^{(k)} \,:\, i \in \mathcal{N}_k^{HB}\right\}$ *mit* $\psi_i^{(k)} := \psi_i^{(k-1)}$ *für alle* $i \in \mathcal{N}_{k-1}$ *und mit* $\psi_i^{(k)} := \varphi_i^{(k)}$ *für alle* $i \in \mathcal{N}_k^{HB}$.

Für $k = 0$ entspricht die hierarchische Basis somit der Knotenbasis des Finite Elemente Raums $\mathcal{S}_0$. Die hierarchische Basis des Finite Elemente Raums $\mathcal{S}_k$ für $k \in \{1,\ldots,m\}$ besteht aus den hierarchischen Basisfunktionen des Finite Elemente Raums $\mathcal{S}_{k-1}$ und den Basisfunktionen der Knotenbasis, die zu den Knoten aus der Menge $\mathcal{N}_k^{\mathrm{HB}}$ gehören. In der Abbildung 6.1(b) ist die hierarchische Basis für die eindimensionalen Finiten Elemente Räume, die zu einer Folge von drei geschachtelten Triangulierungen gehören, dargestellt. Für $k \in \{1,2\}$ sind zu der Triangulierung $\mathcal{T}_k$ dabei nur die hierarchischen Basisfunktionen des Finite Elemente Raums $\mathcal{S}_k$ abgebildet, die zu den hierarchischen Basisfunktionen des Finite Elemente Raums $\mathcal{S}_{k-1}$ hinzukommen, also die Basisfunktionen der Knotenbasis zu den Knoten aus der Menge $\mathcal{N}_k^{\mathrm{HB}}$.

Aufgrund der geschachtelten Triangulierungen können wir jeden inneren Knoten $x \in \mathcal{N}_m$ der feinsten Triangulierung $\mathcal{T}_m$ mit genau einem inneren Knoten $x \in \mathcal{N}_k^{\mathrm{HB}}$ der Triangulierung $\mathcal{T}_k$ für ein $k \in \{0,\ldots,m\}$ assoziieren. Damit können wir die Menge

der inneren Knoten durch $\mathcal{N}_m = \bigcup_{k=0}^{m} \mathcal{N}_k^{\mathrm{HB}}$ disjunkt zerlegen. In Abbildung 6.2, die aus [DW11, Abbildung 7.7] entnommen ist, ist die Zuordnung der inneren Knoten für eine geschachtelte Folge von drei Triangulierungen im zweidimensionalen Raum gegeben, wobei als Verfeinerungsstrategie die Rot- und Grün-Verfeinerung verwendet wurde.

Für $k \in \{0, \ldots, m\}$ ist der Interpolationsoperator $\mathcal{I}_k : \mathcal{S}_m \to \mathcal{S}_k$ für $u \in \mathcal{S}_m$ durch $(\mathcal{I}_k u)(x) = u(x)$ für alle $x \in \mathcal{N}_k$ definiert. Die Unterräume $\mathcal{W}_k$ für $k \in \{0, \ldots, m\}$ des Unterraumkorrekturverfahrens definieren wir durch $\mathcal{W}_0 := \{\mathcal{I}_0 u : u \in \mathcal{S}_m\}$ und

$$\mathcal{W}_k := \{\mathcal{I}_k u - \mathcal{I}_{k-1} u : u \in \mathcal{S}_m\} \tag{6.1}$$

für alle $k \in \{1, \ldots, m\}$. Da für $k \in \{1, \ldots, m\}$ eine Funktion $u \in \mathcal{W}_k$ an den Knoten $x \in \mathcal{N}_{k-1}$ verschwindet, ist diese Funktion durch ihre Funktionswerte an den Knoten $x \in \mathcal{N}_k^{\mathrm{HB}}$ eindeutig bestimmt, sodass wir zur Darstellung des Raums $\mathcal{W}_k$ die hierarchischen Basisfunktionen verwenden können, um die Darstellung $\mathcal{W}_k = \operatorname{span}\left(\left\{\psi_i^{(m)} \;:\; i \in \mathcal{N}_k^{\mathrm{HB}}\right\}\right)$ zu erhalten. Weil für $k \in \{0, \ldots, m\}$ die Räume $\mathcal{W}_k$ jeweils von der Teilmenge der Basisfunktionen der hierarchischen Basis von $\mathcal{S}_m$, die zu den Knoten $x \in \mathcal{N}_k^{\mathrm{HB}}$ gehören, aufgespannt werden, partitionieren diese Räume den Finite Elemente Raum $\mathcal{S}_m$, sodass

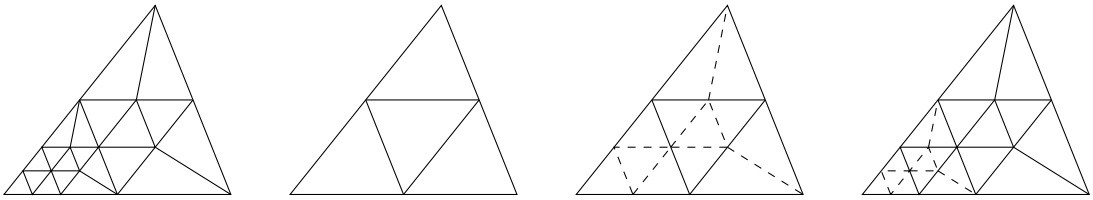
$$\mathcal{S}_m = \mathcal{W}_0 \oplus \mathcal{W}_1 \oplus \ldots \oplus \mathcal{W}_m$$

gilt. Aus dieser Darstellung folgt insbesondere $\mathcal{S}_m = \sum_{k=0}^{m} \mathcal{W}_k$. Für $k \in \{0, \ldots, m\}$ definieren wir $\mathcal{V}_k := \mathcal{W}_k$, sodass folglich auch $\mathcal{S}_m = \sum_{k=0}^{m} \mathcal{V}_k$ gilt. Mit dieser Zerlegung des Finite Elemente Raums $\mathcal{S}_m$ können wir jede Funktion $u \in \mathcal{S}_m$ durch

$$u = \mathcal{I}_m u = \mathcal{I}_0 u + \sum_{k=1}^{m} (\mathcal{I}_k u - \mathcal{I}_{k-1} u) = \sum_{k=0}^{m} w_k$$

mit $w_0 := \mathcal{I}_0 u \in \mathcal{W}_0$ und $w_k := \mathcal{I}_k u - \mathcal{I}_{k-1} u \in \mathcal{W}_k$ für $k \in \{1, \ldots, m\}$ darstellen. Weil das Unterraumkorrekturverfahren jeweils den Fehler auf jedem Unterraum $\mathcal{W}_k$ mit $k \in \{0, \ldots, m\}$ behandelt, werden in jedem Unterraum unterschiedliche Fehleranteile reduziert.

Zwei entscheidende Eigenschaften bei der Verwendung der hierarchischen Basis anstelle der Knotenbasis sind, dass die Steifigkeitsmatrix, die aus der Diskretisierung einer partiellen Differentialgleichung mittels der Finiten Elemente Methode entsteht, gut konditioniert ist und dass der Basiswechsel zwischen der Knotenbasis und der hierarchischen Basis algorithmisch effizient realisierbar ist. Wie in [Yse86b] gezeigt wurde, wächst die Kondition der Steifigkeitsmatrix im Gegensatz zum exponentiellen Wachs-



(a) Triangulierung $\tilde{\mathcal{T}}$ (b) Triangulierung $\mathcal{T}_0$ (c) Triangulierung $\mathcal{T}_1$ (d) Triangulierung $\mathcal{T}_2$

Abbildung 6.3: Eine Folge von geschachtelten Triangulierungen, die einer vorgegebenen Triangulierung entsprechen: Die Elemente, die in einer Triangulierung verfeinert wurden, sind gestrichelt dargestellt.

tum in der Knotenbasis in der hierarchischen Basis nur quadratisch mit der Anzahl der Verfeinerungsebenen. Zudem können wir die multiplikative Variante des Verfahrens der hierarchischen Basen als ein Mehrgitterverfahren auffassen mit dem Unterschied, dass für $k \in \{0, \ldots, m\}$ auf der Ebene k nur die Knoten in $\mathcal{N}_k^{\mathrm{HB}}$ anstatt alle Knoten in $\mathcal{N}_k$ geglättet werden. Die Unterraumkorrekturen entsprechen dabei jeweils den Glättungsschritten.

6.2 Beweis der Annahmen für die Konvergenztheorie

In diesem Kapitel zeigen wir, dass das Verfahren der hierarchischen Basen für das erweiterte Poisson-Problem, welches in Kapitel 3.6 eingeführt wurde, auf einem beschränkten, polygonalen Gebiet $\Omega \subset \mathbb{R}^2$ konvergiert. Dazu sei $\sigma : \Omega \to \mathbb{R}$ eine stückweise konstante Funktion. Wir wenden außerdem als Randbedingungen die homogenen Dirichlet-Randbedingungen an.

$\mathcal{T}_0$ sei eine zulässige Anfangstriangulierung des Gebiets Ω, die Funktion σ sei auf allen Dreiecken $T \in \mathcal{T}_0$ konstant und es sei $m \in \mathbb{N}$. Aus dieser initialen Triangulierung erzeugen wir eine Triangulierung $\tilde{\mathcal{T}}$ mithilfe des in [BSW83] vorgestellten Verfahrens für Gitterverfeinerungen, das in Kapitel 3.5 beschrieben wurde und in [Yse86a] bereits in Zusammenhang mit hierarchischen Basen verwendet wurde. Beim Erstellen der Triangulierung $\tilde{\mathcal{T}}$ erzeugen wir eine Folge $(\mathcal{T}_k)_{k=0,\ldots,m}$ von geschachtelten Triangulierungen mit $\tilde{\mathcal{T}} = \mathcal{T}_m$. Die Triangulierung $\mathcal{T}_{k+1}$ entsteht für $k \in \{0, \ldots, m-1\}$ aus der Triangulierung $\mathcal{T}_k$, indem ein Dreieck $T \in \mathcal{T}_k$ übernommen oder durch Rot- beziehungsweise Grün-Verfeinerung verfeinert wird. Dabei erlauben wir eine Verfeinerung eines Dreiecks $T \in \mathcal{T}_k$ für $k > 0$ nur dann, wenn T durch Rot-Verfeinerung aus einem Dreieck aus $\mathcal{T}_{k-1}$ entstanden ist. Somit wird ein durch Grün-Verfeinerung entstandenes Dreieck nicht weiter verfeinert. In Abbildung 6.3 ist für eine Triangulierung $\tilde{\mathcal{T}}$ die Folge $(\mathcal{T}_k)_{k=0,1,2}$

der entsprechenden geschachtelten Triangulierungen dargestellt. Die so erzeugte Folge $(\mathcal{T}_k)_{k=0,\ldots,m}$ von Triangulierungen ist quasiuniform. Mit dieser Verfeinerungsstrategie lassen sich insbesondere adaptive, nicht uniforme Triangulierungen erzeugen.

6.2.1 Stabilitätsannahme

Die Stabilitätsannahme benötigen wir sowohl für die multiplikative als auch für die additive Variante des Verfahrens der hierarchischen Basen. Diese Annahme beruht auf einer Abschätzung des Interpolationsoperators, die allerdings von der Dimension des Problems abhängt. Deswegen zeigen wir die folgenden Aussagen nur für Raumdimension zwei. Für diese Abschätzung verwenden wir eine Abschätzung aus [Yse85], die aus dem folgenden Lemma 6.2, welches ebenfalls aus [Yse85] stammt, folgt. Für den Finite Elemente Raum $\mathcal{S}_m$ schreiben wir im Folgenden einfach $\mathcal{S}$.

Lemma 6.2. *Sei $O \subseteq \mathbb{R}^2$ eine offene und beschränkte Teilmenge mit dem Durchmesser $R := \operatorname{diam}(O)$. Dann gilt für alle Punkte $x \in O$, alle $\nu \in (0,\ R]$ und für alle Funktionen $v \in H_0^1(O)$*

$$\frac{1}{\pi\nu^2} \int_{B(x,\nu)\cap O} |v(x)|\, dx \leq \frac{1}{\sqrt{2\pi}} \left(\log_2 \frac{R}{\nu} + \frac{1}{4} \right)^{1/2} |v|_{H^1(O)}\,.$$

Beweis. Siehe [Yse85, Satz 2.1]. □

Lemma 6.3. *Es sei $k \in \{0,\ldots,m\}$, $T \in \mathcal{T}_k$ und $v \in \mathcal{S}$. Dann gilt die Abschätzung*

$$\|v\|_{L^\infty(T)} \leq C_1 \left(\log_2 \frac{\operatorname{diam}(T)}{h} + \frac{1}{4} \right)^{1/2} \|v\|_{H^1(T)} \tag{6.2}$$

mit $h := \min_{S\in\mathcal{T}_m,\, S\subseteq T} \operatorname{diam}(S)$ und einer Konstanten $C_1 \in \mathbb{R}_{\geq 0}$.

Beweis. Siehe [Yse85, Satz 3.2]. □

Aus dem Lemma 6.3 folgt eine Abschätzung des Interpolationsoperators auf einem Element einer Triangulierung $\mathcal{T}_k$ für $k \in \{0,\ldots,m\}$, deren Beweis sich an [Yse86b, Lemma 2.2] orientiert.

Lemma 6.4. *Es sei $k \in \{0,\ldots,m\}$, $T \in \mathcal{T}_k$ und $v \in \mathcal{S}$. Dann gilt*

$$|\mathcal{I}_k v|_{H^1(T)} \leq C_T \left(\log_2 \frac{\operatorname{diam}(T)}{h} + \frac{1}{4} \right)^{1/2} |v|_{H^1(T)} \tag{6.3}$$

mit $h := \min_{S\in\mathcal{T}_m,\, S\subseteq T} \operatorname{diam}(S)$ und einer Konstanten $C_T \in \mathbb{R}_{\geq 0}$.

Beweis. Setze $\bar{v} := \frac{1}{|T|}\int_T v(x)\,dx$. Dann gilt $\bar{v} \in \mathbb{R}$ und

$$\begin{aligned}
|\mathcal{I}_k(v-\bar{v})|_{H^1(T)} &= \left(\sum_{|\alpha|=1} \|D^\alpha(\mathcal{I}_k(v-\bar{v}))\|^2_{L^2(T)}\right)^{1/2} \\
&= \left(\sum_{|\alpha|=1} \int_T |D^\alpha \mathcal{I}_k(v-\bar{v})(x)|^2\,dx\right)^{1/2} \\
&= \left(\sum_{|\alpha|=1} \int_T |D^\alpha \mathcal{I}_k v(x) - D^\alpha \mathcal{I}_k \bar{v}(x)|^2\,dx\right)^{1/2} \\
&= \left(\sum_{|\alpha|=1} \int_T |D^\alpha \mathcal{I}_k v(x)|^2\,dx\right)^{1/2} \\
&= |\mathcal{I}_k v|_{H^1(T)}.
\end{aligned}$$

Ebenso folgt, dass auch $|v-\bar{v}|_{H^1(T)} = |v|_{H^1(T)}$ gilt. Außerdem gilt die Abschätzung

$$|\mathcal{I}_k v|_{H^1(T)} \leq 2\,\|\mathcal{I}_k v\|_{L^\infty(T)}.$$

Mithilfe dieser beiden Aussagen, dem Lemma 6.3 und der Poincaré-Ungleichung aus Satz 3.12, die gilt, da T konvex ist und $T \subset B(x_0, \operatorname{diam}(T))$ für alle $x_0 \in T$ sowie

$$\int_T v-\bar{v}\,dx = \int_T v\,dx - \bar{v}\int_T dx = \int_T v\,dx - \frac{1}{|T|}\int_T v\,dx\,|T| = 0$$

gilt, erhalten wir mit $C_T := 2\cdot C_1 \cdot \left(16\cdot\left(\max_{k\in\{0,\dots,m\}, S\in\mathcal{T}_k} \operatorname{diam}(S)\right)^2 + 1\right)^{1/2}$ auf dem Dreieck T die Abschätzung

$$\begin{aligned}
|\mathcal{I}_k v|^2_{H^1(T)} &= |\mathcal{I}_k(v-\bar{v})|^2_{H^1(T)} \leq 4\cdot\|\mathcal{I}_k(v-\bar{v})\|^2_{L^\infty(T)} \leq 4\cdot\|v-\bar{v}\|^2_{L^\infty(T)} \\
&\overset{(6.2)}{\leq} 4\cdot C_1^2\left(\log_2\frac{\operatorname{diam}(T)}{h} + \frac{1}{4}\right)\|v-\bar{v}\|^2_{H^1(T)} \\
&= 4\cdot C_1^2\left(\log_2\frac{\operatorname{diam}(T)}{h} + \frac{1}{4}\right)\left(\|v-\bar{v}\|^2_{L^2(T)} + |v-\bar{v}|^2_{H^1(T)}\right) \\
&\overset{(3.3)}{\leq} 4\cdot C_1^2\left(\log_2\frac{\operatorname{diam}(T)}{h} + \frac{1}{4}\right)\left((2\cdot 2\cdot\operatorname{diam}(T))^2\,|v|^2_{H^1(T)} + |v|^2_{H^1(T)}\right) \\
&= 4\cdot C_1^2\cdot\left(4^2\cdot\operatorname{diam}(T)^2+1\right)\left(\log_2\frac{\operatorname{diam}(T)}{h} + \frac{1}{4}\right)|v|^2_{H^1(T)} \\
&\leq C_T^2\left(\log_2\frac{\operatorname{diam}(T)}{h} + \frac{1}{4}\right)|v|^2_{H^1(T)}.
\end{aligned}$$

□

Das folgende Lemma stellt die Basis der Stabilitätsabschätzung dar. Dabei hängt diese Abschätzung von der Raumdimension ab, sodass die Anzahl der Verfeinerungsebenen für Dimension drei, wie in [Yse86b] erläutert, exponentiell eingeht. Wir betrachten deswegen das Resultat nur für Dimension zwei. Der Beweis orientiert sich an [Yse86b, Lemma 2.3].

Lemma 6.5. *Es existiert eine Konstante $C_I \in \mathbb{R}_{\geq 0}$, sodass für alle $v \in \mathcal{S}$*

$$|\mathcal{I}_k v|^2_{H^1(\Omega)} \leq C_I(m-k+1)\,|v|^2_{H^1(\Omega)}\,, \tag{6.4}$$

$$\|\mathcal{I}_k v\|^2_{L^2(\Omega)} \leq C_I(m-k+1)\,\|v\|^2_{H^1(\Omega)} \tag{6.5}$$

für alle $k \in \{0,\ldots,m\}$ gelten.

Beweis. Es sei $k \in \{0,\ldots,m\}$ und $v \in \mathcal{S}$. Da $\mathcal{I}_k v$ auf $T \in \mathcal{T}_k$ stetig und auf allen Dreiecken $S \in \mathcal{T}_m$ mit $S \subseteq T$ linear ist, gilt mit $h := \min_{S\in\mathcal{T}_m,\, S\subseteq T} \operatorname{diam}(S)$ nach Lemma 6.4 und wegen $h \overset{(3.20)}{\geq} 2^{k-m}\operatorname{diam}(T)$

$$\begin{aligned}
|\mathcal{I}_k v|^2_{H^1(T)} &\overset{(6.3)}{\leq} C_T^2\left(\log_2 \frac{\operatorname{diam}(T)}{h} + \frac{1}{4}\right)|v|^2_{H^1(T)} \\
&\leq C_T^2\left(\log_2 \frac{\operatorname{diam}(T)}{2^{k-m}\operatorname{diam}(T)} + \frac{1}{4}\right)|v|^2_{H^1(T)} \\
&= C_T^2\left(\log_2 2^{m-k} + \frac{1}{4}\right)|v|^2_{H^1(T)} \\
&\leq C_T^2\,(m-k+1)\,|v|^2_{H^1(T)}\,.
\end{aligned}$$

Aus dieser Abschätzung folgt

$$|\mathcal{I}_k v|^2_{H^1(\Omega)} = \sum_{T\in\mathcal{T}_k} |\mathcal{I}_k v|^2_{H^1(T)} \leq \sum_{T\in\mathcal{T}_k} C_T^2\,(m-k+1)\,|v|^2_{H^1(T)} = C_T^2\,(m-k+1)\,|v|^2_{H^1(\Omega)}$$

und somit die Abschätzung (6.4). Es gilt für $T \in \mathcal{T}_k$

$$\begin{aligned}
\|\mathcal{I}_k v\|_{L^2(T)} &= \left(\int_T |\mathcal{I}_k v(x)|^2\; dx\right)^{1/2} \leq \left(\int_T \max_{y\in T} |\mathcal{I}_k v(y)|^2\; dx\right)^{1/2} \\
&= \max_{y\in T} |\mathcal{I}_k v(y)| \cdot \left(\int_T dx\right)^{1/2} \leq \|\mathcal{I}_k v\|_{L^\infty(T)} \cdot \left(\operatorname{diam}(T)^2\right)^{1/2} \\
&= \operatorname{diam}(T) \cdot \|\mathcal{I}_k v\|_{L^\infty(T)}\,.
\end{aligned}$$

Daraus folgt mit Lemma 6.3 für $T \in \mathcal{T}_k$ und mit $h \overset{(3.20)}{\geq} 2^{k-m}\operatorname{diam}(T)$

$$\|\mathcal{I}_k v\|^2_{L^2(T)} \leq \operatorname{diam}(T)^2 \cdot \|\mathcal{I}_k v\|^2_{L^\infty(T)} \leq \operatorname{diam}(T)^2 \cdot \|v\|^2_{L^\infty(T)}$$

$$\begin{aligned}
&\overset{(6.2)}{\leq} \operatorname{diam}(T)^2 \cdot C_1^2 \left(\log_2 \frac{\operatorname{diam}(T)}{h} + \frac{1}{4} \right) \|v\|_{H^1(T)}^2 \\
&\leq \operatorname{diam}(T)^2 \cdot C_1^2 \left(\log_2 \frac{\operatorname{diam}(T)}{2^{k-m} \operatorname{diam}(T)} + \frac{1}{4} \right) \|v\|_{H^1(T)}^2 \\
&= \operatorname{diam}(T)^2 \cdot C_1^2 \left(\log_2 2^{m-k} + \frac{1}{4} \right) \|v\|_{H^1(T)}^2 \\
&\leq \operatorname{diam}(T)^2 \cdot C_1^2 \, (m-k+1) \, \|v\|_{H^1(T)}^2 .
\end{aligned}$$

Daher gilt mit $H := \max_{k \in \{0,\dots,m\}, S \in \mathcal{T}_k} \operatorname{diam}(S)$

$$\begin{aligned}
\|\mathcal{I}_k v\|_{L^2(\Omega)}^2 &= \sum_{T \in \mathcal{T}_k} \|\mathcal{I}_k v\|_{L^2(T)}^2 \leq \sum_{T \in \mathcal{T}_k} \operatorname{diam}(T)^2 \cdot C_1^2 \, (m-k+1) \, \|v\|_{H^1(T)}^2 \\
&\leq \sum_{T \in \mathcal{T}_k} H^2 \cdot C_1^2 \, (m-k+1) \, \|v\|_{H^1(T)}^2 \\
&= H^2 \cdot C_1^2 \, (m-k+1) \, \|v\|_{H^1(\Omega)}^2 .
\end{aligned}$$

Insbesondere gelten die Abschätzungen (6.4) und (6.5) mit $C_I := \max\{C_T^2, H^2 \cdot C_1^2\}$. □

Als nächstes beweisen wir ein Lemma, das wie die Poincaré-Ungleichung die Norm $\|\cdot\|_{L^2(\Omega)}$ durch die Halbnorm $|\cdot|_{H^1(\Omega)}$ beschränkt. Dabei beschränken wir uns auf Funktionen, die auf Ω stetig und auf jedem Dreieck einer Triangulierung $\mathcal{T}_k$ für $k \in \{1, \dots, m\}$ linear sind.

Lemma 6.6. *Sei $k \in \{1, \dots, m\}$. Dann existiert eine Konstante $c_P \in \mathbb{R}_{>0}$, sodass*

$$\|v_k\|_{L^2(\Omega)}^2 \leq c_P \, 4^{-k} \, |v_k|_{H^1(\Omega)}^2 \tag{6.6}$$

für alle $v_k \in \mathcal{V}_k$ gilt.

Beweis. Definiere $H := \max\{\operatorname{diam}(S) : S \in \mathcal{T}_0\}$. Sei $T \in \mathcal{T}_{k-1}$ und $v_k \in \mathcal{V}_k$. Da $v_k(\xi) = 0$ für alle $\xi \in \mathcal{N}_{k-1}$ gilt, folgt insbesondere $v_k(\xi) = 0$ für die Ecken $\xi \in T \cap \mathcal{N}_{k-1}$ von T. Wir unterscheiden im Folgenden drei Fälle in Abhängigkeit von der Verfeinerung des Dreiecks T im Verfeinerungsschritt von der Triangulierung $\mathcal{T}_{k-1}$ zur Triangulierung $\mathcal{T}_k$. Setze $c_P := 4 \cdot H^2 \in \mathbb{R}_{>0}$.
Fall 1: $T \in \mathcal{T}_{k-1} \cap \mathcal{T}_k$
In diesem Fall wurde T nicht weiter verfeinert. Da $v_k(\xi) = 0$ für die Ecken $\xi \in \mathcal{N}_{k-1} \cap T$ von T gilt, gilt $v_k(x) = 0$ für alle $x \in T$ und daher

$$\|v_k\|_{L^2(T)}^2 = 0 \leq c_P \, 4^{-k} \, |v_k|_{H^1(T)}^2 .$$

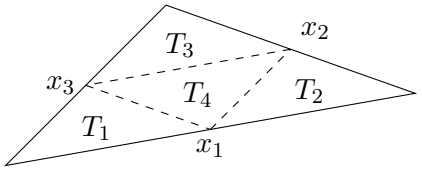


Abbildung 6.4: Dargestellt ist ein Dreieck $T \in \mathcal{T}_{k-1}$ mit den aus der Rot-Verfeinerung resultierenden Dreiecken $T_1, \ldots, T_4 \in \mathcal{T}_k$ und den Kantenmittelpunkten x_1, x_2, x_3.

Fall 2: $\exists\, T_1, T_2 \in \mathcal{T}_k : T_1 \cup T_2 = T$
In diesem Fall wurde T mittels Grün-Verfeinerung verfeinert. Da $v_k(x) = 0$ für alle $x \in \mathcal{N}_{k-1} \cap T$ gilt, gilt für die schwachen Ableitungen, weil v_k linear auf den Dreiecken T_1 und T_2 ist,

$$\left|D^{(1,0)} v_k(x)\right| \geq \frac{\max_{y \in \mathcal{N}_k^{\mathrm{HB}} \cap T} |v_k(y)|}{\operatorname{diam}(T)}$$

oder

$$\left|D^{(0,1)} v_k(x)\right| \geq \frac{\max_{y \in \mathcal{N}_k^{\mathrm{HB}} \cap T} |v_k(y)|}{\operatorname{diam}(T)}$$

für alle $x \in T$. Weil T nur durch Rot-Verfeinerungen aus einem Dreieck $S \in \mathcal{T}_0$ entstanden sein kann, gilt $\operatorname{diam}(T) \overset{(3.20)}{\leq} 2^{-(k-1)} \operatorname{diam}(S) \leq 2^{-(k-1)} H$. Daraus folgt

$$\begin{aligned}
\|v_k\|^2_{L^2(T)} &= \int_T |v_k(x)|^2 \, dx \leq |T| \cdot \max_{x \in \mathcal{N}_k^{\mathrm{HB}} \cap T} |v_k(x)|^2 \\
&= \operatorname{diam}(T)^2 \cdot |T| \cdot \left(\frac{\max_{x \in \mathcal{N}_k^{\mathrm{HB}} \cap T} |v_k(x)|}{\operatorname{diam}(T)} \right)^2 \\
&\leq \operatorname{diam}(T)^2 \int_T \left|D^{(1,0)} v_k(x)\right|^2 + \left|D^{(0,1)} v_k(x)\right|^2 \, dx \\
&= \operatorname{diam}(T)^2 \left(\left\|D^{(1,0)} v_k\right\|^2_{L^2(T)} + \left\|D^{(0,1)} v_k\right\|^2_{L^2(T)} \right) \\
&= \operatorname{diam}(T)^2 \, |v_k|^2_{H^1(T)} \\
&\leq \left(2^{-(k-1)} H\right)^2 |v_k|^2_{H^1(T)} = H^2 \cdot 4^{-(k-1)} \, |v_k|^2_{H^1(T)} \\
&= 4 \cdot H^2 \cdot 4^{-k} \, |v_k|^2_{H^1(T)} = c_P \cdot 4^{-k} \, |v_k|^2_{H^1(T)} \, .
\end{aligned}$$

Fall 3: $\exists\, T_1, \ldots, T_4 \in \mathcal{T}_k : T_1 \cup \ldots \cup T_4 = T$
In diesem Fall wurde T per Rot-Verfeinerung verfeinert. Ohne Beschränkung der Allge-

meinheit seien die Dreiecke $T_1, \ldots, T_4$ wie in Abbildung 6.4 mit den Kantenmittelpunkten $x_1, x_2, x_3 \in T$ des Dreiecks T gegeben. Ohne Einschränkung gelte $|v_k(x_1)| \geq |v_k(x_2)| \geq |v_k(x_3)|$. Für $i \in \{1,2,3\}$ gilt für die schwache Ableitung auf dem Dreieck T_i

$$\left|D^{(1,0)}v_k(x)\right| \geq \frac{\max_{x\in\mathcal{N}_k^{\mathrm{HB}}\cap T_i}|v_k(x)|}{\operatorname{diam}(T_i)}$$

oder

$$\left|D^{(0,1)}v_k(x)\right| \geq \frac{\max_{x\in\mathcal{N}_k^{\mathrm{HB}}\cap T_i}|v_k(x)|}{\operatorname{diam}(T_i)}$$

für alle $x \in T_i$, da v_k auf T_i linear ist. Weil T_i für $i \in \{1,2,3\}$ aus einem Dreieck $S \in \mathcal{T}_0$ durch mehrfache Rot-Verfeinerungen entsteht und deshalb $\operatorname{diam}(T_i) \overset{(3.20)}{\leq} 2^{-k}\operatorname{diam}(S) \leq 2^{-k}H$ gilt, folgt

$$\begin{aligned}
\int_{T_i}|v_k(x)|^2\,dx &\leq |T_i|\cdot\max_{x\in\mathcal{N}_k^{\mathrm{HB}}\cap T_i}|v_k(x)|^2\\
&= \operatorname{diam}(T_i)^2\cdot|T_i|\cdot\left(\frac{\max_{x\in\mathcal{N}_k^{\mathrm{HB}}\cap T_i}|v_k(x)|}{\operatorname{diam}(T_i)}\right)^2\\
&\leq \operatorname{diam}(T_i)^2\int_{T_i}\left|D^{(1,0)}v_k(x)\right|^2+\left|D^{(0,1)}v_k(x)\right|^2\,dx\\
&= \operatorname{diam}(T_i)^2\,|v_k|^2_{H^1(T_i)} \leq \left(2^{-k}\cdot H\right)^2|v_k|^2_{H^1(T_i)}\\
&= H^2\cdot 4^{-k}\,|v_k|^2_{H^1(T_i)}\,.
\end{aligned}$$

Da $\max_{x\in\mathcal{N}_k^{\mathrm{HB}}\cap T_1} = |v_k(x_1)| = \max_{x\in\mathcal{N}_k^{\mathrm{HB}}\cap T_4}$ gilt und T_1 und T_4 nach Lemma 3.26 kongruent sind, folgt zudem

$$\begin{aligned}
\int_{T_1}|v_k(x)|^2\,dx+\int_{T_4}|v_k(x)|^2\,dx &\leq |T_1|\cdot\max_{x\in\mathcal{N}_k^{\mathrm{HB}}\cap T_1}|v_k(x)|^2\\
&\quad+|T_4|\cdot\max_{x\in\mathcal{N}_k^{\mathrm{HB}}\cap T_4}|v_k(x)|^2\\
&= (|T_1|+|T_4|)\cdot\max_{x\in\mathcal{N}_k^{\mathrm{HB}}\cap T_1}|v_k(x)|^2\\
&= 2\cdot|T_1|\cdot\max_{x\in\mathcal{N}_k^{\mathrm{HB}}\cap T_1}|v_k(x)|^2\\
&\leq 2\cdot H^2\cdot 4^{-k}\,|v_k|^2_{H^1(T_1)}\,.
\end{aligned}$$

Daraus folgt insgesamt die Abschätzung

$$\begin{aligned}
\|v_k\|^2_{L^2(T)} &= \int_T |v_k(x)|^2 \, dx \\
&= \int_{T_1} |v_k(x)|^2 \, dx + \int_{T_4} |v_k(x)|^2 \, dx + \int_{T_2} |v_k(x)|^2 \, dx + \int_{T_3} |v_k(x)|^2 \, dx \\
&\leq 2 \cdot H^2 \cdot 4^{-k} \cdot |v_k|^2_{H^1(T_1)} + H^2 \cdot 4^{-k} \cdot |v_k|^2_{H^1(T_2)} + H^2 \cdot 4^{-k} \cdot |v_k|^2_{H^1(T_3)} \\
&\leq 4 \cdot H^2 \cdot 4^{-k} \cdot \left(|v_k|^2_{H^1(T_1)} + |v_k|^2_{H^1(T_2)} + |v_k|^2_{H^1(T_3)} \right) \\
&\leq 4 \cdot H^2 \cdot 4^{-k} \cdot \left(|v_k|^2_{H^1(T_1)} + |v_k|^2_{H^1(T_2)} + |v_k|^2_{H^1(T_3)} + |v_k|^2_{H^1(T_4)} \right) \\
&= 4 \cdot H^2 \cdot 4^{-k} \, |v_k|^2_{H^1(T)} \, .
\end{aligned}$$

Aus diesen drei Fällen ergibt sich die Abschätzung

$$\begin{aligned}
\|v_k\|^2_{L^2(\Omega)} &= \sum_{T \in \mathcal{T}_{k-1}} \|v_k\|^2_{L^2(T)} \leq \sum_{T \in \mathcal{T}_{k-1}} 4 \cdot H^2 \cdot 4^{-k} \, |v_k|^2_{H^1(T)} \\
&= 4 \cdot H^2 \cdot 4^{-k} \sum_{T \in \mathcal{T}_{k-1}} |v_k|^2_{H^1(T)} = 4 \cdot H^2 \cdot 4^{-k} \, |v_k|^2_{H^1(\Omega)} \\
&= c_P \cdot 4^{-k} \, |v_k|^2_{H^1(\Omega)} \, .
\end{aligned}$$

□

In [Yse86b] wurde bereits gezeigt, dass die Zerlegung von $\mathcal{S}$ in die Unterräume $\mathcal{V}_k$, $k \in \{0, \ldots, m\}$, stabil ist, wobei unser Beweis sich an [DW11, Lemma 7.24] orientiert.

Lemma 6.7. *Es existiert eine von m unabhängige Konstante $C_S \in \mathbb{R}_{\geq 0}$, sodass*

$$\|v_0\|^2_{H^1(\Omega)} + \sum_{k=1}^{m} 4^k \, \|v_k\|^2_{L^2(\Omega)} \leq C_S (m+1)^2 \, \|v\|^2_{H^1(\Omega)} \tag{6.7}$$

für alle $v \in \mathcal{S}$ mit der Zerlegung $v = \sum_{k=0}^{m} v_k$ und $v_k \in \mathcal{V}_k$ für alle $k \in \{0, \ldots, m\}$ gilt.

Beweis. Ohne Beschränkung der Allgemeinheit nehmen wir für die Konstante aus Lemma 6.6 $c_P \geq 1$ an. Es sei $v \in \mathcal{S}$. Dann existieren $v_k \in \mathcal{V}_k$ für alle $k \in \{0, \ldots, m\}$ mit $v = \sum_{k=0}^{m} v_k$. Dann gilt

$$\begin{aligned}
|v_0|^2_{H^1(\Omega)} + \sum_{k=1}^{m} 4^k \, \|v_k\|^2_{L^2(\Omega)} &\overset{(6.6)}{\leq} |v_0|^2_{H^1(\Omega)} + \sum_{k=1}^{m} c_P \, |v_k|^2_{H^1(\Omega)} \\
&= |\mathcal{I}_0 v|^2_{H^1(\Omega)} + c_P \sum_{k=1}^{m} |\mathcal{I}_k v - \mathcal{I}_{k-1} v|^2_{H^1(\Omega)} \\
&\overset{(2.1)}{=} |\mathcal{I}_0 v|^2_{H^1(\Omega)} + c_P \sum_{k=1}^{m} \left(2 \cdot \left(|\mathcal{I}_k v|^2_{H^1(\Omega)} + |\mathcal{I}_{k-1} v|^2_{H^1(\Omega)} \right) \right.
\end{aligned}$$

$$
\begin{aligned}
&\quad - |\mathcal{I}_k v + \mathcal{I}_{k-1} v|^2_{H^1(\Omega)}\Big) \\
&\leq 2c_P \, |\mathcal{I}_0 v|^2_{H^1(\Omega)} + 2c_P \sum_{k=1}^{m} \Big(|\mathcal{I}_k v|^2_{H^1(\Omega)} + |\mathcal{I}_{k-1} v|^2_{H^1(\Omega)} \Big) \\
&\leq 4c_P \sum_{k=0}^{m} |\mathcal{I}_k v|^2_{H^1(\Omega)} \\
&\overset{(6.4)}{\leq} 4c_P \sum_{k=0}^{m} C_I(m-k+1) \, |v|^2_{H^1(\Omega)} \\
&= 4c_P \, C_I \, |v|^2_{H^1(\Omega)} \sum_{k=0}^{m} (m-k+1) \\
&= 4c_P \, C_I \, |v|^2_{H^1(\Omega)} \sum_{k=1}^{m+1} k \\
&= 4c_P \, C_I \, |v|^2_{H^1(\Omega)} \, \frac{(m+1)(m+1+1)}{2} \\
&= 2c_P \, C_I \, |v|^2_{H^1(\Omega)} \Big((m+1)^2 + (m+1) \Big) \\
&\leq 2c_P \, C_I \, |v|^2_{H^1(\Omega)} \, 2(m+1)^2 \\
&= 4c_P \, C_I (m+1)^2 \, |v|^2_{H^1(\Omega)} \, .
\end{aligned}
$$

Daraus folgt mit $C_1 := 4c_P \, C_I$ und $C_S := C_I + C_1$

$$
\begin{aligned}
\|v_0\|^2_{H^1(\Omega)} + \sum_{k=1}^{m} 4^k \, \|v_k\|^2_{L^2(\Omega)} &= \|v_0\|^2_{L^2(\Omega)} + |v_0|^2_{H^1(\Omega)} + \sum_{k=1}^{m} 4^k \, \|v_k\|^2_{L^2(\Omega)} \\
&\leq \|v_0\|^2_{L^2(\Omega)} + C_1 (m+1)^2 \, |v|^2_{H^1(\Omega)} \\
&\leq \|v_0\|^2_{L^2(\Omega)} + C_1 (m+1)^2 \, \|v\|^2_{H^1(\Omega)} \\
&= \|\mathcal{I}_0 v\|^2_{L^2(\Omega)} + C_1 (m+1)^2 \, \|v\|^2_{H^1(\Omega)} \\
&\overset{(6.5)}{\leq} C_I (m+1)^2 \, \|v\|^2_{H^1(\Omega)} + C_1 (m+1)^2 \, \|v\|^2_{H^1(\Omega)} \\
&= C_S (m+1)^2 \, \|v\|^2_{H^1(\Omega)} \, .
\end{aligned}
$$

□

Mithilfe des vorbereitenden Lemmas 6.7 können wir die Stabilität der Zerlegung zeigen, wobei diese Abschätzung nur suboptimal ist, da die Stabilitätskonstante von der Anzahl der Verfeinerungsebenen der Triangulierung abhängt. Der Beweis der Stabilitätsaussage orientiert sich an [DW11, Satz 7.25].

Satz 6.8. *Es sei A ein $H^1(\Omega)$-elliptischer Operator auf $\Omega \subset \mathbb{R}^2$. Es gelte $B_0 = A_0$ und*

$$
c_{B_1} 2^k \, \|w_k\|_{L^2(\Omega)} \leq \|w_k\|_{B_k} \leq c_{B_2} 2^k \, \|w_k\|_{L^2(\Omega)} \tag{6.8}
$$

für alle $k \in \{1, \dots, m\}$ und für alle $w_k \in \mathcal{W}_k$ mit Konstanten $c_{B_1}, c_{B_2} \in \mathbb{R}_{\geq 0}$. Dann existiert eine Konstante $K_1 \in \mathbb{R}_{>0}$, sodass für jedes $v \in \mathcal{S}$ mit der Zerlegung $v = \sum_{k=0}^{m} v_k$ mit $v_k \in \mathcal{V}_k$ für alle $k \in \{0, \dots, m\}$

$$\sum_{k=0}^{m} (B_k v_k, v_k) \leq K_1 \|v\|_A^2 \tag{6.9}$$

gilt.

Beweis. Sei $v \in \mathcal{S}$. Dann existieren $v_k \in \mathcal{V}_k$ für alle $k \in \{0, \dots, m\}$ mit $v = \sum_{k=0}^{m} v_k$. Dann gilt mit $\tilde{c} := \max\{\alpha_1, c_{B_2}^2\}$ und $K_1 := \frac{\tilde{c} \cdot C_S}{\alpha_2} \cdot (m+1)^2$

$$\begin{aligned}
\sum_{k=0}^{m} (B_k v_k, v_k) &= (B_0 v_0, v_0) + \sum_{k=1}^{m} (B_k v_k, v_k) \overset{(6.8)}{\leq} (B_0 v_0, v_0) + \sum_{k=1}^{m} c_{B_2}^2 4^k \|v_k\|_{L^2(\Omega)}^2 \\
&= (A_0 v_0, v_0) + \sum_{k=1}^{m} c_{B_2}^2 4^k \|v_k\|_{L^2(\Omega)}^2 \overset{(4.4)}{=} (A v_0, v_0) + \sum_{k=1}^{m} c_{B_2}^2 4^k \|v_k\|_{L^2(\Omega)}^2 \\
&\overset{(3.8)}{\leq} \alpha_1 \|v_0\|_{H^1(\Omega)}^2 + c_{B_2}^2 \sum_{k=1}^{m} 4^k \|v_k\|_{L^2(\Omega)}^2 \\
&\leq \tilde{c} \left(\|v_0\|_{H^1(\Omega)}^2 + \sum_{k=1}^{m} 4^k \|v_k\|_{L^2(\Omega)}^2 \right) \overset{(6.7)}{\leq} \tilde{c}\, C_S (m+1)^2 \|v\|_{H^1(\Omega)}^2 \\
&\overset{(3.9)}{\leq} \frac{\tilde{c}\, C_S}{\alpha_2} (m+1)^2 \langle Av, v \rangle = \frac{\tilde{c}\, C_S}{\alpha_2} (m+1)^2 \|v\|_A^2 \\
&= K_1 \|v\|_A^2 .
\end{aligned}$$

□

6.2.2 Verschärfte Cauchy-Schwarz-Ungleichung

Wir zeigen die verschärfte Cauchy-Schwarz-Ungleichung für die additive Variante des Verfahrens der hierarchischen Basen, woraus die verschärfte Cauchy-Schwarz-Ungleichung für die multiplikative Variante des Verfahrens der hierarchischen Basen folgt. In [Yse86b, Lemma 2.7] wurde die hierfür entscheidende Abschätzung bewiesen. Wir orientieren uns hier aber an [DW11, Satz 7.26].

Satz 6.9. *Es sei A ein $H^1(\Omega)$-elliptischer Operator auf $\Omega \subset \mathbb{R}^2$. Es gelte $B_0 = A_0$ und für alle $k \in \{1, \dots, m\}$*

$$c_{B_1} 2^k \|w_k\|_{L^2(\Omega)} \leq \|w_k\|_{B_k} \leq c_{B_2} 2^k \|w_k\|_{L^2(\Omega)} \tag{6.10}$$

für alle $w_k \in \mathcal{W}_k$ mit Konstanten $c_{B_1}, c_{B_2} \in \mathbb{R}_{>0}$. Dann existieren $\gamma_{k\ell} \in \mathbb{R}_{\geq 0}$ für alle

$k, \ell \in \{0, \dots, m\}$ mit

$$(v_k, w_\ell)_A \leq \gamma_{k\ell} (B_k v_k, v_k)^{1/2} (B_\ell w_\ell, w_\ell)^{1/2} \tag{6.11}$$

für alle $v_k \in \mathcal{W}_k$ und $w_\ell \in \mathcal{W}_\ell$. Weiter existiert eine Konstante $K_2 \in \mathbb{R}_{>0}$, sodass für alle $w \in \mathcal{S}$ mit $w = \sum_{k=0}^m w_k$ und $w_k \in \mathcal{W}_k$ für alle $k \in \{0, \dots, m\}$

$$\|w\|_A^2 \leq K_2 \sum_{k=0}^m (B_k w_k, w_k)$$

gilt.

Beweis. Seien $k, \ell \in \{0, \dots, m\}$ mit $k \leq \ell$ und $v_k \in \mathcal{W}_k$ sowie $w_\ell \in \mathcal{W}_\ell$. Sei $T \in \mathcal{T}_k$. Wir zeigen zunächst

$$(v_k, w_\ell)_A \leq \gamma_{k\ell} \, |v_k|_{H^1(\Omega)} \, \|w_\ell\|_{L^2(\Omega)} \tag{6.12}$$

mit $\gamma_{k\ell} = c \cdot \sqrt{2}^{k-\ell} \cdot 2^\ell$ und einer von m unabhängigen Konstante $c \in \mathbb{R}_{\geq 0}$.

Wir betrachten zuerst den Fall $k \geq 1$ und $T \in \mathcal{T}_k \backslash \mathcal{T}_{k-1}$ oder $k = 0$. Definiere die Funktion

$$\chi : T \to \mathbb{R}, x \mapsto \max\left\{0, 1 - 2^\ell \operatorname{dist}(x, \partial T)\right\}. \tag{6.13}$$

Da diese Funktion nur auf einem Randstreifen der Breite $2^{-\ell}$ des Dreiecks T von Null verschieden ist, gilt $\operatorname{supp}(\chi) = \overline{\{x \in T : \operatorname{dist}(x, \partial T) < 2^{-\ell}\}}$. Der Träger der Funktion χ ist in Abbildung 6.5 dargestellt. Die Fläche von $\operatorname{supp}(\chi)$ lässt sich mithilfe des Umfangs von T abschätzen, sodass mit dem Umkreisradius h_T

$$|\operatorname{supp}(\chi)| \leq 2^{-\ell} \cdot 3 \cdot 2 \cdot h_T \leq 2^{-\ell} \cdot \pi \cdot 2 \cdot h_T \tag{6.14}$$

gilt, da die Länge jeder Kante von T durch $2 \cdot h_T$ abgeschätzt werden kann. Weil $(1 - \chi)(x) = 0$ für alle $x \in \partial T$ und $v_k, w_\ell \in C^2(T)$ aufgrund der Linearität von v_k und w_ℓ auf T gilt, folgt mittels der Greenschen Formel aus Satz 2.12

$$\begin{aligned}
(v_k, (1-\chi) w_\ell)_A|_T &= ((1-\chi) w_\ell, v_k)_A|_T = a((1-\chi) w_\ell, v_k)|_T \\
&= \int_T \langle \nabla((1-\chi) w_\ell)(x), \sigma(x) \, \nabla v_k(x) \rangle_2 \, dx \\
&\overset{(2.2)}{=} - \int_T ((1-\chi) w_\ell)(x) \cdot \sigma(x) \, \Delta v_k(x) \, dx
\end{aligned}$$

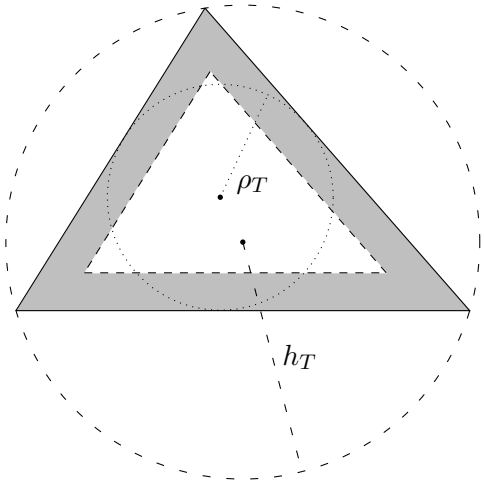


Abbildung 6.5: Darstellung eines Elements $T \in \mathcal{T}_k \backslash \mathcal{T}_{k-1}$ der Triangulierung $\mathcal{T}_k$ und des Trägers $\operatorname{supp}(\chi)$ der Funktion χ, der in grau dargestellt ist und einem Randstreifen der Breite $2^{-\ell}$ des Elements T entspricht. Zusätzlich ist der Umkreis von T mit dem Umkreisradius h_T und der Inkreis mit dem Inkreisradius ρ_T dargestellt.

$$
\begin{aligned}
&\quad + \int_{\partial T} ((1-\chi) w_\ell)(x) \cdot \langle n(x), \sigma(x)\, \nabla v_k(x) \rangle_2 \, dx \\
&= - \int_T ((1-\chi) w_\ell)(x) \cdot \sigma(x)\, \Delta v_k(x) \, dx \\
&= - \int_T ((1-\chi) w_\ell(x) \cdot \operatorname{div}(\sigma \cdot \nabla v_k)(x) \, dx \\
&= 0,
\end{aligned}
$$

da σ auf T konstant und v_k auf T linear ist, sodass ∇v_k auf T konstant ist. Deshalb folgt mit der Cauchy-Schwarz-Ungleichung und der Normäquivalenz der Energienorm $\|\cdot\|_A$ und der Halbnorm $|\cdot|_{H^1(T)}$

$$
\begin{aligned}
(v_k, w_\ell)_A|_T &= (v_k, (\chi + (1-\chi)) w_\ell)_A|_T \\
&= (v_k, \chi w_\ell)_A|_T + (v_k, (1-\chi) w_\ell)_A|_T \\
&= (v_k, \chi w_\ell)_A|_T \\
&= \int_T \langle \nabla v_k(x), \sigma(x) \nabla(\chi w_\ell)(x) \rangle_2 \, dx \\
&= \int_{\operatorname{supp}(\chi)} \langle \nabla v_k, \sigma(x) \nabla(\chi w_\ell)(x) \rangle_2 \, dx \\
&= (v_k, \chi w_\ell)_A|_{\operatorname{supp}(\chi)}
\end{aligned}
$$

$$
\begin{aligned}
&\overset{\text{Cauchy-Schwarz}}{\leq} \|v_k\|_A|_{\operatorname{supp}(\chi)} \, \|\chi w_\ell\|_A|_{\operatorname{supp}(\chi)} \\
&= \|v_k\|_A|_{\operatorname{supp}(\chi)} \, \|\chi w_\ell\|_A|_T \\
&\leq c_A^2 \cdot |v_k|_{H^1(\operatorname{supp}(\chi))} \, |\chi w_\ell|_{H^1(T)} \, .
\end{aligned}
$$

Im Folgenden schätzen wir die beiden Halbnormen $|v_k|_{H^1(\operatorname{supp}(\chi))}$ und $|\chi w_\ell|_{H^1(T)}$ jeweils einzeln ab. Da v_k auf T linear ist, ist somit ∇v_k auf T konstant und es gilt $|T| \geq \pi \cdot \rho_T^2$. Das heißt, dass der Flächeninhalt des Dreiecks T durch den Flächeninhalt des Inkreises mit Radius ρ_T nach unten beschränkt ist. Weiter gilt nach (3.18) $\rho_T \geq 2^{-k} \cdot \rho_{\tilde{T}}$ mit $\tilde{T} \in \mathcal{T}_0$ und $\tilde{T} \cap T = T$. Deshalb folgt mit $c_1 := \sqrt{2 \cdot \kappa} \cdot \max_{S \in \mathcal{T}_0} \rho_S^{-1/2}$

$$
\begin{aligned}
|v_k|_{H^1(\operatorname{supp}(\chi))} &= \left(\sum_{|\alpha|=1} \|D^\alpha v_k\|^2_{L^2(\operatorname{supp}(\chi))} \right)^{1/2} = \left(\sum_{|\alpha|=1} \int_{\operatorname{supp}(\chi)} |D^\alpha v_k(x)|^2 \, dx \right)^{1/2} \\
&= \left(\sum_{|\alpha|=1} \int_{\operatorname{supp}(\chi)} |D^\alpha v_k(x)|^2 \, dx \cdot \frac{\int_T dx}{\int_T dx} \right)^{1/2} \\
&= \left(\sum_{|\alpha|=1} \frac{\int_{\operatorname{supp}(\chi)} dx}{\int_T dx} \cdot \int_T |D^\alpha v_k(x)|^2 \, dx \right)^{1/2} \\
&= \left(\frac{\int_{\operatorname{supp}(\chi)} dx}{\int_T dx} \right)^{1/2} \cdot \left(\sum_{|\alpha|=1} \int_T |D^\alpha v_k(x)|^2 \, dx \right)^{1/2} \\
&= \left(\frac{|\operatorname{supp}(\chi)|}{|T|} \right)^{1/2} \cdot \left(\sum_{|\alpha|=1} \|D^\alpha v_k\|^2_{L^2(T)} \right)^{1/2} \\
&= \left(\frac{|\operatorname{supp}(\chi)|}{|T|} \right)^{1/2} |v_k|_{H^1(T)} \\
&\overset{(6.14)}{\leq} \left(\frac{2^{-\ell} \cdot \pi \cdot 2 \cdot h_T}{\pi \cdot \rho_T^2} \right)^{1/2} |v_k|_{H^1(T)} \\
&= \left(2 \cdot \frac{h_T}{\rho_T} \right)^{1/2} \cdot \left(\frac{1}{\rho_T} \right)^{1/2} \cdot 2^{-\ell/2} \, |v_k|_{H^1(T)} \\
&\leq \sqrt{2 \cdot \kappa} \cdot \left(\frac{1}{2^{-k} \cdot \rho_{\tilde{T}}} \right)^{1/2} \cdot 2^{-\ell/2} \, |v_k|_{H^1(T)} \\
&\leq \sqrt{2 \cdot \kappa} \max_{S \in \mathcal{T}_0} \rho_S^{-1/2} \cdot \sqrt{2}^{k-\ell} \, |v_k|_{H^1(T)} \\
&= c_1 \cdot \sqrt{2}^{k-\ell} \, |v_k|_{H^1(T)} \, .
\end{aligned}
$$

Da $\frac{\partial}{\partial x_i}\mathrm{dist}\,(x,y) \leq 1$ für $x, y \in T$ und $i \in \{1,2\}$ gilt, folgt $\chi(x) \leq 1$ und $\left|\frac{\partial}{\partial x_i}\chi(x)\right| \leq 2^\ell$ für alle $x \in T$ und $i \in \{1,2\}$ und wegen (3.19) gilt $m_T \geq 2^{-k} \cdot m_{\tilde{T}}$. Damit erhalten wir für die zweite Halbnorm $|\chi w_\ell|_{H^1(T)}$ die Abschätzung

$$
\begin{aligned}
|\chi w_\ell|_{H^1(T)} &= \left(\sum_{|\alpha|=1} \|D^\alpha(\chi w_\ell)\|^2_{L^2(T)}\right)^{1/2} \\
&= \left(\sum_{|\alpha|=1} \|(D^\alpha\chi)\cdot w_\ell + \chi\cdot(D^\alpha w_\ell)\|^2_{L^2(T)}\right)^{1/2} \\
&\overset{(2.1)}{=} \left(\sum_{|\alpha|=1} 2\left(\|(D^\alpha\chi)\cdot w_\ell\|^2_{L^2(T)} + \|\chi\cdot(D^\alpha w_\ell)\|^2_{L^2(T)}\right)\right. \\
&\qquad\qquad \left. - \|(D^\alpha\chi)\cdot w_\ell - \chi\cdot(D^\alpha w_\ell)\|^2_{L^2(T)}\right)^{1/2} \\
&\leq \left(\sum_{|\alpha|=1} 2\left(\|(D^\alpha\chi)\cdot w_\ell\|^2_{L^2(T)} + \|\chi\cdot(D^\alpha w_\ell)\|^2_{L^2(T)}\right)\right)^{1/2} \\
&= \sqrt{2}\left(\sum_{|\alpha|=1} \|(D^\alpha\chi)\cdot w_\ell\|^2_{L^2(T)} + \sum_{|\alpha|=1} \|\chi\cdot(D^\alpha w_\ell)\|^2_{L^2(T)}\right)^{1/2} \\
&\leq \sqrt{2}\left(\sum_{|\alpha|=1} \left(2^\ell \|w_\ell\|_{L^2(T)}\right)^2 + \sum_{|\alpha|=1} \|D^\alpha w_\ell\|^2_{L^2(T)}\right)^{1/2} \\
&= \sqrt{2}\left(2\cdot 2^{2\ell}\|w_\ell\|^2_{L^2(T)} + |w_\ell|^2_{H^1(T)}\right)^{1/2} \\
&\overset{(3.16)}{\leq} \sqrt{2}\left(2\cdot 2^{2\ell}\|w_\ell\|^2_{L^2(T)} + c_L^2\cdot m_T^{-2}\|w_\ell\|^2_{L^2(T)}\right)^{1/2} \\
&\leq \sqrt{2}\left(2\cdot 2^{2\ell}\|w_\ell\|^2_{L^2(T)} + c_L^2\cdot 2^{2k}\cdot m_{\tilde{T}}^{-2}\|w_\ell\|^2_{L^2(T)}\right)^{1/2} \\
&\leq \sqrt{2}\left(2\cdot 2^{2\ell}\|w_\ell\|^2_{L^2(T)} + c_L^2\cdot 2^{2\ell}\cdot \max_{S\in\mathcal{T}_0} m_S^{-2}\|w_\ell\|^2_{L^2(T)}\right)^{1/2} \\
&= \sqrt{2}\cdot\left(2 + c_L^2\cdot\max_{S\in\mathcal{T}_0} m_S^{-2}\right)^{1/2}\cdot 2^\ell\|w_\ell\|_{L^2(T)} \\
&= c_2\cdot 2^\ell\|w_\ell\|_{L^2(T)}
\end{aligned}
$$

mit $c_2 := \sqrt{2}\cdot\left(2 + c_L^2\cdot\max_{S\in\mathcal{T}_0} m_S^{-2}\right)^{1/2}$. Mithilfe dieser beiden Abschätzungen erhalten wir mit $c_3 := c_A^2\cdot c_1\cdot c_2$ insgesamt

$$
(v_k, w_\ell)_A|_T \leq c_A^2\,|v_k|_{H^1(\mathrm{supp}(\chi))}\,|\chi w_\ell|_{H^1(T)}
$$

$$
\begin{aligned}
&\leq c_A^2 \cdot c_1 \cdot \sqrt{2}^{k-\ell} \, |v_k|_{H^1(T)} \, |\chi w_\ell|_{H^1(T)} \\
&\leq c_A^2 \cdot c_1 \cdot c_2 \cdot \sqrt{2}^{k-\ell} \cdot 2^\ell \, |v_k|_{H^1(T)} \, \|w_\ell\|_{L^2(T)} \\
&= c_3 \cdot \sqrt{2}^{k-\ell} \cdot 2^\ell \, |v_k|_{H^1(T)} \, \|w_\ell\|_{L^2(T)} \, .
\end{aligned}
$$

Damit haben wir in diesem Fall (6.12) eingeschränkt auf T mit $c := c_3$ bewiesen.

Falls $k \geq 1$ und $T \in \mathcal{T}_k \cap \mathcal{T}_{k-1}$ gilt, gilt $v_k|_T = 0$. Somit folgt

$$
(v_k, w_\ell)_A|_T \leq c_3 \cdot \sqrt{2}^{k-\ell} \cdot 2^\ell \, |v_k|_{H^1(T)} \, \|w_\ell\|_{L^2(T)} \, ,
$$

sodass auch in diesem Fall (6.12) eingeschränkt auf T mit $c := c_3$ gilt.

Mit der Cauchy-Schwarz-Ungleichung und der Abschätzung (6.12) auf jedem Dreieck $T \in \mathcal{T}_k$ erhalten wir

$$
\begin{aligned}
(v_k, w_\ell)_A &= \sum_{T \in \mathcal{T}_k} (v_k, w_\ell)_A|_T \leq \sum_{T \in \mathcal{T}_k} c_3 \cdot \sqrt{2}^{k-\ell} \cdot 2^\ell \, |v_k|_{H^1(T)} \, \|w_\ell\|_{L^2(T)} \\
&= c_3 \cdot \sqrt{2}^{k-\ell} \cdot 2^\ell \sum_{T \in \mathcal{T}_k} |v_k|_{H^1(T)} \, \|w_\ell\|_{L^2(T)} \\
&\overset{\text{Cauchy-Schwarz}}{\leq} c_3 \cdot \sqrt{2}^{k-\ell} \cdot 2^\ell \cdot \left(\sum_{T \in \mathcal{T}_k} |v_k|_{H^1(T)}^2 \right)^{1/2} \cdot \left(\sum_{T \in \mathcal{T}_k} \|w_\ell\|_{L^2(T)}^2 \right)^{1/2} \\
&= c_3 \cdot \sqrt{2}^{k-\ell} \cdot 2^\ell \, |v_k|_{H^1(\Omega)} \, \|w_\ell\|_{L^2(\Omega)} \, ,
\end{aligned}
$$

sodass (6.12) mit $c := c_3$ gilt. Indem wir die inverse Ungleichung mit $m_S := \min_{T \in \mathcal{T}_k} m_T^{-2}$ für $S \in \mathcal{T}_k$, $m_S \geq 2^{-k} \cdot m_{\tilde{S}}$ für $\tilde{S} \in \mathcal{T}_0$ mit $S \cap \tilde{S} = S$ nach (3.19) und die Abschätzung (6.10) der approximativen Lösungsverfahren verwenden, erhalten wir mit $c_4 := c_3 \cdot \sqrt{c_I} \cdot m_{\tilde{S}}^{-1} \cdot c_{B_1}^{-2}$

$$
\begin{aligned}
(v_k, w_\ell)_A &\leq c_3 \cdot \sqrt{2}^{k-\ell} \cdot 2^\ell \, |v_k|_{H^1(\Omega)} \, \|w_\ell\|_{L^2(\Omega)} \\
&\overset{(3.17)}{\leq} c_3 \cdot \sqrt{2}^{k-\ell} \cdot 2^\ell \cdot \sqrt{c_I} \cdot \left(\sum_{T \in \mathcal{T}_k} m_T^{-2} \, \|v_k\|_{L^2(T)}^2 \right)^{1/2} \|w_\ell\|_{L^2(\Omega)} \\
&\leq c_3 \cdot \sqrt{2}^{k-\ell} \cdot 2^\ell \cdot \sqrt{c_I} \cdot \left(\sum_{T \in \mathcal{T}_k} m_S^{-2} \, \|v_k\|_{L^2(T)}^2 \right)^{1/2} \|w_\ell\|_{L^2(\Omega)} \\
&= c_3 \cdot \sqrt{2}^{k-\ell} \cdot 2^\ell \cdot \sqrt{c_I} \cdot m_S^{-1} \cdot \left(\sum_{T \in \mathcal{T}_k} \|v_k\|_{L^2(T)}^2 \right)^{1/2} \|w_\ell\|_{L^2(\Omega)} \\
&\leq c_3 \cdot \sqrt{2}^{k-\ell} \cdot \sqrt{c_I} \cdot m_{\tilde{S}}^{-1} \cdot 2^k \, \|v_k\|_{L^2(\Omega)} \, 2^\ell \, \|w_\ell\|_{L^2(\Omega)}
\end{aligned}
$$

$$\begin{aligned}
&\overset{(6.10)}{\leq} c_3 \cdot \sqrt{c_I} \cdot m_{\tilde{S}}^{-1} \cdot c_{B_1}^{-2} \cdot \sqrt{2}^{k-\ell} \, \|v_k\|_{B_k} \, \|w_\ell\|_{B_\ell} \\
&= c_4 \cdot \sqrt{2}^{k-\ell} \, \|v_k\|_{B_k} \, \|w_\ell\|_{B_\ell} \, .
\end{aligned}$$

Sei $w \in \mathcal{S}$. Dann existieren $w_k \in \mathcal{W}_k$ für alle $k \in \{0, \ldots, m\}$ mit $w = \sum_{k=0}^{m} w_k$. Daraus folgt mit der letzten Abschätzung und $K_2 := \left(\sum_{k,\ell=0}^{m} \gamma_{k\ell}^2\right)^{1/2} \in \mathbb{R}_{>0}$

$$\begin{aligned}
\|w\|_A^2 &= (w,w)_A = (\sum_{k=0}^{m} w_k, \sum_{\ell=0}^{m} w_\ell)_A = \sum_{k=0}^{m}\sum_{\ell=0}^{m} (w_k, w_\ell)_A \\
&\overset{(6.11)}{\leq} \sum_{k=0}^{m}\sum_{\ell=0}^{m} \gamma_{k\ell} (B_k w_k, w_k)^{1/2} (B_\ell w_\ell, w_\ell)^{1/2} \\
&= \sum_{k=0}^{m} (B_k w_k, w_k)^{1/2} \left(\sum_{\ell=0}^{m} \gamma_{k\ell} (B_\ell w_\ell, w_\ell)^{1/2} \right) \\
&\overset{\substack{\text{Cauchy-}\\ \text{Schwarz}}}{\leq} \sum_{k=0}^{m} (B_k w_k, w_k)^{1/2} \left(\sum_{\ell=0}^{m} \gamma_{k\ell}^2 \right)^{1/2} \left(\sum_{\ell=0}^{m} (B_\ell w_\ell, w_\ell) \right)^{1/2} \\
&= \left(\sum_{\ell=0}^{m} (B_\ell w_\ell, w_\ell) \right)^{1/2} \sum_{k=0}^{m} (B_k w_k, w_k)^{1/2} \left(\sum_{\ell=0}^{m} \gamma_{k\ell}^2 \right)^{1/2} \\
&\overset{\substack{\text{Cauchy-}\\ \text{Schwarz}}}{\leq} \left(\sum_{\ell=0}^{m} (B_\ell w_\ell, w_\ell) \right)^{1/2} \cdot \left(\sum_{k=0}^{m} (B_k w_k, w_k) \right)^{1/2} \cdot \left(\sum_{k=0}^{m}\sum_{\ell=0}^{m} \gamma_{k\ell}^2 \right)^{1/2} \\
&= \left(\sum_{k,\ell=0}^{m} \gamma_{k\ell}^2 \right)^{1/2} \cdot \left(\sum_{k=0}^{m} (B_k w_k, w_k) \right)^{1/2} \cdot \left(\sum_{k=0}^{m} (B_k w_k, w_k) \right)^{1/2} \\
&= \left(\sum_{k,\ell=0}^{m} \gamma_{k\ell}^2 \right)^{1/2} \sum_{k=0}^{m} (B_k w_k, w_k) \\
&= K_2 \cdot \sum_{k=0}^{m} (B_k w_k, w_k).
\end{aligned}$$

□

Die verschärfte Cauchy-Schwarz-Ungleichung kann auch für Raumdimension drei bewiesen werden, wir haben uns hier aber auf zwei Raumdimensionen beschränkt, da die Konstante der Stabilitätsannahme für drei Raumdimensionen exponentiell mit der Anzahl der Verfeinerungsebenen wächst. Die verschärfte Cauchy-Schwarz-Ungleichung für die multiplikative Variante des Verfahrens der hierarchischen Basen ist eine Folgerung aus Satz 6.9 und der Cauchy-Schwarz-Ungleichung.

Satz 6.10. *Es sei A ein $H^1(\Omega)$-elliptischer Operator auf $\Omega \subset \mathbb{R}^2$. Es gelte $B_0 = A_0$ und*

für alle $k \in \{1, \dots, m\}$

$$c_{B_1} 2^k \|w_k\|_{L^2(\Omega)} \leq \|w_k\|_{B_k} \leq c_{B_2} 2^k \|w_k\|_{L^2(\Omega)}$$

für alle $w_k \in \mathcal{W}_k$ *mit Konstanten* $c_{B_1}, c_{B_2} \in \mathbb{R}_{>0}$. *Dann existieren* $\gamma_{k\ell} \in \mathbb{R}_{\geq 0}$ *für alle* $k, \ell \in \{0, \dots, m\}$ *mit*

$$(w_k, v_\ell)_A \leq \gamma_{k\ell} (B_k w_k, w_k)^{1/2} (B_\ell v_\ell, v_\ell)^{1/2} \tag{6.15}$$

für alle $w_k \in \mathcal{W}_k$ *und* $v_\ell \in \mathcal{V}_\ell$. *Weiter gibt es eine Konstante* $K_2 \in \mathbb{R}_{>0}$, *sodass für alle* $x, y \in \mathbb{R}^m$

$$\sum_{k,\ell=0}^{m} \gamma_{k\ell} x_k y_\ell \leq K_2 \left(\sum_{k=0}^{m} x_k^2 \right)^{1/2} \left(\sum_{\ell=0}^{m} y_\ell^2 \right)^{1/2}$$

gilt.

Beweis. Die Abschätzung (6.15) folgt aus Satz 6.9, da $\mathcal{V}_\ell \subseteq \mathcal{W}_\ell$ für alle $\ell \in \{0, \dots, m\}$ gilt. Seien $x, y \in \mathbb{R}^m$ gegeben. Es gilt nun mit $K_2 := \left(\sum_{k,\ell=0}^{m} \gamma_{k\ell}^2 \right)^{1/2} \in \mathbb{R}_{>0}$

$$\begin{aligned}
\sum_{k,\ell=0}^{m} \gamma_{k\ell} x_k y_\ell &= \sum_{k=0}^{m} x_k \sum_{\ell=0}^{m} \gamma_{k\ell} y_\ell \overset{\text{Cauchy-Schwarz}}{\leq} \sum_{k=0}^{m} x_k \left(\sum_{\ell=0}^{m} \gamma_{k\ell}^2 \right)^{1/2} \left(\sum_{\ell=0}^{m} y_\ell^2 \right)^{1/2} \\
&= \left(\sum_{\ell=0}^{m} y_\ell^2 \right)^{1/2} \sum_{k=0}^{m} x_k \left(\sum_{\ell=0}^{m} \gamma_{k\ell}^2 \right)^{1/2} \\
&\overset{\text{Cauchy-Schwarz}}{\leq} \left(\sum_{\ell=0}^{m} y_\ell^2 \right)^{1/2} \left(\sum_{k=0}^{m} x_k^2 \right)^{1/2} \left(\sum_{\ell=0}^{m} \left(\sum_{k=0}^{m} \gamma_{k\ell}^2 \right) \right)^{1/2} \\
&= \left(\sum_{k,\ell=0}^{m} \gamma_{k\ell}^2 \right)^{1/2} \left(\sum_{k=0}^{m} x_k^2 \right)^{1/2} \left(\sum_{\ell=0}^{m} y_\ell^2 \right)^{1/2} \\
&= K_2 \left(\sum_{k=0}^{m} x_k^2 \right)^{1/2} \left(\sum_{\ell=0}^{m} y_\ell^2 \right)^{1/2}.
\end{aligned}$$

□

6.2.3 Eigenschaften des iterativen Lösungsverfahrens

Wir verwenden als iteratives Lösungsverfahren B_k ein konvergentes Iterationsverfahren für alle $k \in \{1, \dots, m\}$ und ein exaktes Lösungsverfahren B_0, weil die Annahme 3 für das multiplikative Verfahren der hierarchischen Basen für $A_0 = B_0$ offenbar erfüllt ist und für $k \in \{1, \dots, m\}$ äquivalent zu der Konvergenz des verwendeten Lösungsverfahrens B_k

ist. Sowohl für die Konvergenz des multiplikativen als auch des additiven Verfahrens der hierarchischen Basen benötigen wir die Abschätzung (6.8) des verwendeten iterativen Lösungsverfahrens. Diese Abschätzung geht nicht nur in die Stabilitätsannahme in Satz 6.8 sondern auch in die verschärfte Cauchy-Schwarz-Ungleichung in Satz 6.9 ein.

Satz 6.11. *Das Jacobi-Verfahren $B_k^J : \mathcal{W}_k \to \mathcal{W}_k$ erfüllt*

$$c_{J_1} 2^k \|v_k\|_{L^2(\Omega)} \leq \|v_k\|_{B_k^J} \leq c_{J_2} 2^k \|v_k\|_{L^2(\Omega)} \tag{6.16}$$

mit $c_{J_1}, c_{J_2} \in \mathbb{R}_{>0}$ für alle $v_k \in \mathcal{V}_k$ und $k \in \{1, \ldots, m\}$.

Beweis. Es sei $k \in \{1, \ldots, m\}$ und $v_k \in \mathcal{V}_k$. Für alle $i \in \mathcal{N}_k^{\mathrm{HB}}$ existieren $v_i^{(k)} \in \mathbb{R}$ mit $v_k = \sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} v_i^{(k)} \psi_i^{(k)}$. Das durch B_k^J gegebene Jacobi-Verfahren erfüllt

$$(B_k^J \psi_i^{(k)}, \psi_j^{(k)}) = \delta_{ij} (A\psi_i^{(k)}, \psi_j^{(k)})$$

für alle $i, j \in \mathcal{N}_k^{\mathrm{HB}}$. Somit gilt

$$\begin{aligned}(B_k^J v_k, v_k) &= (B_k^J \sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} v_i^{(k)} \psi_i^{(k)}, \sum_{j \in \mathcal{N}_k^{\mathrm{HB}}} v_j^{(k)} \psi_j^{(k)}) = \sum_{i,j \in \mathcal{N}_k^{\mathrm{HB}}} v_i^{(k)} v_j^{(k)} (B_k^J \psi_i^{(k)}, \psi_j^{(k)}) \\ &= \sum_{i,j \in \mathcal{N}_k^{\mathrm{HB}}} v_i^{(k)} v_j^{(k)} \delta_{ij} (A\psi_i^{(k)}, \psi_j^{(k)}) = \sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} \left(v_i^{(k)}\right)^2 (A\psi_i^{(k)}, \psi_i^{(k)}).\end{aligned}$$

Es sei $i \in \mathcal{N}_k^{\mathrm{HB}}$. Dann gilt mit $\sigma_{\min} := \min\{\sigma(x) \ : \ x \in \Omega\}$ und $c_P := 4 \cdot \max\{\mathrm{diam}\,(S) \ : \ S \in \mathcal{T}_0\}$

$$\begin{aligned}(A\psi_i^{(k)}, \psi_i^{(k)}) &= \int_\Omega \sigma(x) \cdot \left(D^{(1,0)} \psi_i^{(k)}(x)\right)^2 + \sigma(x) \cdot \left(D^{(0,1)} \psi_i^{(k)}(x)\right)^2 \, dx \\ &\geq \sigma_{\min} \int_\Omega \left|D^{(1,0)} \psi_i^{(k)}(x)\right|^2 + \left|D^{(0,1)} \psi_i^{(k)}(x)\right|^2 \, dx \\ &= \sigma_{\min} \cdot \left(\left\|D^{(1,0)} \psi_i^{(k)}\right\|_{L^2(\Omega)}^2 + \left\|D^{(0,1)} \psi_i^{(k)}\right\|_{L^2(\Omega)}^2\right) \\ &= \sigma_{\min} \cdot \left|\psi_i^{(k)}\right|_{H^1(\Omega)}^2 \\ &\overset{(6.6)}{\geq} \sigma_{\min} \cdot c_P^{-1} \cdot 4^k \left\|\psi_i^{(k)}\right\|_{L^2(\Omega)}^2.\end{aligned}$$

Mit der inversen Abschätzung (3.17) erhalten wir mit $\sigma_{\max} := \max\{\sigma(x) \ : \ x \in \Omega\}$ und $s_{\min} := \min\{m_S \ : \ S \in \mathcal{T}_0\}$

$$(A\psi_i^{(k)}, \psi_i^{(k)}) = \int_\Omega \sigma(x) \left(D^{(1,0)} \psi_i^{(k)}(x)\right)^2 + \sigma(x) \left(D^{(0,1)} \psi_i^{(k)}(x)\right)^2 \, dx$$

$$
\begin{aligned}
&\leq \sigma_{\max} \int_\Omega \left|D^{(1,0)}\psi_i^{(k)}(x)\right|^2 + \left|D^{(0,1)}\psi_i^{(k)}(x)\right|^2 \, dx \\
&= \sigma_{\max} \left(\left\|D^{(1,0)}\psi_i^{(k)}\right\|^2_{L^2(\Omega)} + \left\|D^{(0,1)}\psi_i^{(k)}\right\|^2_{L^2(\Omega)} \right) \\
&= \sigma_{\max} \left|\psi_i^{(k)}\right|^2_{H^1(\Omega)} \\
&\overset{(3.17)}{\leq} \sigma_{\max} \cdot c_I \sum_{T\in\mathcal{T}_k} m_T^{-2} \left\|\psi_i^{(k)}\right\|^2_{L^2(T)} \\
&\overset{(3.19)}{\leq} \sigma_{\max} \cdot c_I \sum_{T\in\mathcal{T}_k} \left(2^{-k}\cdot s_{\min}\right)^{-2} \left\|\psi_i^{(k)}\right\|^2_{L^2(T)} \\
&= \sigma_{\max} \cdot c_I \cdot s_{\min}^{-2} \cdot 4^k \sum_{T\in\mathcal{T}_k} \left\|\psi_i^{(k)}\right\|^2_{L^2(T)} \\
&= \sigma_{\max} \cdot c_I \cdot s_{\min}^{-2} \cdot 4^k \left\|\psi_i^{(k)}\right\|^2_{L^2(\Omega)}.
\end{aligned}
$$

Das heißt mit $c_1 := \sigma_{\min}\cdot c_P^{-1}$ und $c_2 := \sigma_{\max}\cdot c_I \cdot s_{\min}^{-2}$ erhalten wir

$$
c_1 \cdot 4^k \left\|\psi_i^{(k)}\right\|^2_{L^2(\Omega)} \leq (A\psi_i^{(k)}, \psi_i^{(k)}) \leq c_2 \cdot 4^k \left\|\psi_i^{(k)}\right\|^2_{L^2(\Omega)}. \tag{6.17}
$$

Mit der Cauchy-Schwarz-Ungleichung ergibt sich

$$
\begin{aligned}
\|v_k\|^2_{L^2(\Omega)} &= \int_\Omega |v_k(x)|^2 \, dx = \int_\Omega \left| \sum_{i\in\mathcal{N}_k^{\mathrm{HB}}} v_i^{(k)}\psi_i^{(k)}(x) \right|^2 dx \\
&\leq \int_\Omega \left(\sum_{i\in\mathcal{N}_k^{\mathrm{HB}}} \left|v_i^{(k)}\psi_i^{(k)}(x)\right| \right)^2 dx = \int_\Omega \left(\sum_{i\in\mathcal{N}_k^{\mathrm{HB}}} \left|v_i^{(k)}\psi_i^{(k)}(x)\cdot 1\right| \right)^2 dx \\
&\overset{\substack{\text{Cauchy-}\\ \text{Schwarz}}}{\leq} \int_\Omega \left(\sum_{i\in\mathcal{N}_k^{\mathrm{HB}}} \left|v_i^{(k)}\psi_i^{(k)}(x)\right|^2 \right) \left(\sum_{i\in\mathcal{N}_k^{\mathrm{HB}}} 1^2 \right) dx \\
&= \left|\mathcal{N}_k^{\mathrm{HB}}\right| \int_\Omega \sum_{i\in\mathcal{N}_k^{\mathrm{HB}}} \left|v_i^{(k)}\right|^2 \cdot \left|\psi_i^{(k)}(x)\right|^2 dx \\
&= \left|\mathcal{N}_k^{\mathrm{HB}}\right| \sum_{i\in\mathcal{N}_k^{\mathrm{HB}}} \left|v_i^{(k)}\right|^2 \int_\Omega \left|\psi_i^{(k)}(x)\right|^2 dx \\
&= \left|\mathcal{N}_k^{\mathrm{HB}}\right| \sum_{i\in\mathcal{N}_k^{\mathrm{HB}}} \left(v_i^{(k)}\right)^2 \left\|\psi_i^{(k)}\right\|^2_{L^2(\Omega)}.
\end{aligned}
$$

Da $\psi_i^{(k)} = \varphi_i^{(k)}$ für alle $i \in \mathcal{N}_k^{\mathrm{HB}}$ gilt, folgt mit $v_i^{(k)} := 0$ für alle $i \in \mathcal{N}_k \backslash \mathcal{N}_k^{\mathrm{HB}}$

$$v_k = \sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} v_i^{(k)} \psi_i^{(k)} = \sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} v_i^{(k)} \varphi_i^{(k)} = \sum_{i \in \mathcal{N}_k} v_i^{(k)} \varphi_i^{(k)}.$$

Daraus folgt

$$\begin{aligned}
\sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} \left(v_i^{(k)}\right)^2 \left\|\psi_i^{(k)}\right\|_{L^2(\Omega)}^2 &= \sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} \left(v_i^{(k)}\right)^2 \sum_{T \in \mathcal{T}_K} \left\|\psi_i^{(k)}\right\|_{L^2(T)}^2 \\
&= \sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} \left(v_i^{(k)}\right)^2 \sum_{T \in \mathcal{T}_K} \int_T \left|\psi_i^{(k)}(x)\right|^2 \, dx \\
&\leq \sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} \left(v_i^{(k)}\right)^2 \sum_{T \in \mathcal{T}_K} \int_T 1^2 \, dx \\
&= \sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} \left(v_i^{(k)}\right)^2 \sum_{T \in \mathcal{T}_K} |T| \\
&\leq \sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} \left(v_i^{(k)}\right)^2 \sum_{T \in \mathcal{T}_k} \frac{1}{2} \operatorname{diam}(T)^2 \\
&= \frac{1}{2} \sum_{T \in \mathcal{T}_k} \operatorname{diam}(T)^2 \sum_{i \in \mathcal{N}_k} \left(v_i^{(k)}\right)^2 \\
&\leq \frac{|\mathcal{T}_k|}{2} \max_{T \in \mathcal{T}_k} \operatorname{diam}(T)^2 \sum_{i \in \mathcal{N}_k} \left(v_i^{(k)}\right)^2 \\
&\overset{(3.15)}{\leq} \frac{|\mathcal{T}_k|}{2} \cdot C_P^2 \left\| \sum_{i \in \mathcal{N}_k} v_i^{(k)} \varphi_i^{(k)} \right\|_{L^2(\Omega)}^2 \\
&= \frac{|\mathcal{T}_k|}{2} \cdot C_P^2 \left\|v_k\right\|_{L^2(\Omega)}^2 .
\end{aligned}$$

Somit gilt mit $c_3 := \left|\mathcal{N}_k^{\mathrm{HB}}\right|^{-1}$ und $c_4 := \frac{|\mathcal{T}_k|}{2} \cdot C_P^2$ die Abschätzung

$$c_3 \left\|v_k\right\|_{L^2(\Omega)}^2 \leq \sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} \left(v_i^{(k)}\right)^2 \left\|\psi_i^{(k)}\right\|_{L^2(\Omega)}^2 \leq c_4 \left\|v_k\right\|_{L^2(\Omega)}^2 . \tag{6.18}$$

Aus (6.17) und (6.18) erhalten wir

$$(B_k^J v_k, v_k) = \sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} \left(v_i^{(k)}\right)^2 (A\psi_i^{(k)}, \psi_i^{(k)}) \overset{(6.17)}{\leq} \sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} \left(v_i^{(k)}\right)^2 \cdot c_2 \cdot 4^k \left\|\psi_i^{(k)}\right\|_{L^2(\Omega)}^2$$

$$= c_2 \cdot 4^k \sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} \left(v_i^{(k)}\right)^2 \left\|\psi_i^{(k)}\right\|_{L^2(\Omega)}^2 \overset{(6.18)}{\leq} c_2 \cdot c_4 \cdot 4^k \left\|v_k\right\|_{L^2(\Omega)}^2$$

und des Weiteren

$$\begin{aligned}(B_k^J v_k, v_k) &= \sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} \left(v_i^{(k)}\right)^2 (A\psi_i^{(k)}, \psi_i^{(k)}) \overset{(6.17)}{\geq} \sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} \left(v_i^{(k)}\right)^2 \cdot c_1 \cdot 4^k \left\|\psi_i^{(k)}\right\|_{L^2(\Omega)}^2 \\ &= c_1 \cdot 4^k \sum_{i \in \mathcal{N}_k^{\mathrm{HB}}} \left(v_i^{(k)}\right)^2 \left\|\psi_i^{(k)}\right\|_{L^2(\Omega)}^2 \overset{(6.18)}{\geq} c_1 \cdot c_3 \cdot 4^k \left\|v_k\right\|_{L^2(\Omega)}^2 .\end{aligned}$$

Insgesamt erhalten wir mit $c_{J_1} := c_1 \cdot c_3$ und $c_{J_2} := c_2 \cdot c_4$ daher

$$c_{J_1} \cdot 4^k \left\|v_k\right\|_{L^2(\Omega)}^2 \leq (B_k^J v_k, v_k) \leq c_{J_2} \cdot 4^k \left\|v_k\right\|_{L^2(\Omega)}^2 . \qquad \square$$

Da wir die Annahmen der Konvergenztheorie für Unterraumkorrekturverfahren aus Kapitel 5 in den Sätzen 6.8, 6.9, 6.10 und 6.11 bewiesen haben, erhalten wir für das Modellproblem mit dem Jacobi-Verfahren als approximatives Lösungsverfahren damit die Konvergenz der multiplikativen und additiven Variante des Verfahrens der hierarchischen Basen. Aus Satz 5.6 ergibt sich die Konvergenz der multiplikativen Variante des Verfahrens der hierarchischen Basen und für die Konvergenzrate erhalten wir

$$1 - \frac{2 - \lambda}{K_1 \left(1 + K_2\right)^2} = 1 - \frac{c}{(m+1)^2}$$

mit einer Konstanten $c \in \mathbb{R}_{>0}$ und der Anzahl der Unterräume $m \in \mathbb{N}$. Die Konstante c ergibt sich aus den Konstanten K_1, K_2 und λ aus den Annahmen 1, 2 und 3 für das multiplikative Verfahren der hierarchischen Basen. Die Konvergenz der additiven Variante des Verfahrens der hierarchischen Basen folgt aus Satz 5.11 für einen geeigneten Dämpfungsparameter $\theta \in \mathbb{R}_{>0}$, der von den Parametern K_1 und K_2 aus den Annahmen 1 und 2' für das additive Verfahren der hierarchischen Basen abhängt.

Aus den beiden Sätzen 5.9 und 5.12 erhalten wir als Abschätzungen für die Konditionszahlen der vorkonditionierten Operatoren

$$\begin{aligned}\kappa(N_{\text{mul, sym}} A) &\leq \frac{\left(1 + 2\lambda^2 K_2^2\right) K_1}{2 - \lambda} = c_1 (m+1)^2, \\ \kappa(CA) &\leq \lambda \left(1 + K_2\right) K_1 = c_2 (m+1)^2\end{aligned}$$

mit $c_1, c_2 \in \mathbb{R}_{\geq 0}$. Daraus folgt aus Satz 2.32, dass die Konvergenzgeschwindigkeit des vorkonditionierten Verfahrens der konjugierten Gradienten, das mit dem Operator $N_{\text{mul, sym}}$

beziehungsweise mit dem Operator C vorkonditioniert wird, jeweils mit steigender Verfeinerungsanzahl m sinkt.

7 Implementierung

Wir stellen in diesem Kapitel die Implementierung des Verfahrens der hierarchischen Basen, das in Kapitel 6.1 eingeführt wurde, sowohl als multiplikatives als auch als additives Unterraumkorrekturverfahren und die Implementierung des vorkonditionierten Verfahrens der konjugierten Gradienten vor. Zudem vergleichen wir die Konvergenzresultate mit numerischen Resultaten anhand des Modellproblems.

7.1 Algorithmen

Als Grundlage der Implementierung dient die Programmbibliothek *SimpleFEM* der Arbeitsgruppe Scientific Computing an der Christian-Albrechts-Universität zu Kiel, die Routinen zur Lösung partieller Differentialgleichungen mit dem Verfahren der Finiten Elemente zur Verfügung stellt.

Im Folgenden sei $m \in \mathbb{N}_0$ und $\Omega \subset \mathbb{R}^2$ ein beschränktes, polygonales Gebiet. In einem Vorbereitungsschritt, der in Algorithmus 4 dargestellt ist, wird das Verfahren der hierarchischen Basen initialisiert. Dabei werden die geschachtelte Familie von Triangulierungen $(\mathcal{T}_k)_{k=0,\dots,m}$, die zu den Triangulierungen $\mathcal{T}_0, \dots, \mathcal{T}_m$ zugehörigen Steifigkeitsmatrizen bezüglich der Knotenbasis und die Prolongationen zwischen den Triangulierungen $\mathcal{T}_i$ und $\mathcal{T}_{i+1}$ für alle $i \in \{0, \dots, m-1\}$ mithilfe von Routinen aus der *SimpleFEM* Programmbi-

Algorithmus 4 : Initialisierung

Eingabe : Anfangstriangulierung $\mathcal{T}_0$, Anzahl der Verfeinerungen m, rechte Seite f
Ausgabe : Datenstruktur hb
for $i = 0$ **to** $m - 1$ **do**
  Berechnung der Steifigkeitsmatrix A_i der Ebene i
  Berechnung der Triangulierung $\mathcal{T}_{i+1}$ durch Verfeinerung der Triangulierung $\mathcal{T}_i$
  Berechnung der Anzahl der Freiheitsgrade der Ebene i
  Berechnung der Prolongation p_i
Berechnung der Steifigkeitsmatrix A_m der Ebene m
Berechnung der Anzahl der Freiheitsgrade der Ebene m
Berechnung der rechten Seite b

bliothek erzeugt. Als Restriktionen benutzen wir jeweils die Adjungierte der Prolongation, um die Galerkin-Eigenschaft zu erfüllen, wie in [Hac85, Note 3.6.6] beschrieben. Um die Unterraumkorrekturen auf den Unterräumen $\mathcal{W}_i$ für $i \in \{0, \ldots, m\}$ durchzuführen, nutzen wir die Eigenschaft der Programmbibliothek *SimpleFEM* aus, dass bei einer Gitterverfeinerung die neu erzeugten Knoten an die bisherigen Knoten angehängt werden. Deshalb gilt $k < \ell$ für alle Knotenindizes $k \in \mathcal{N}_i^{\mathrm{HB}}$ und $\ell \in \mathcal{N}_j^{\mathrm{HB}}$ für $i, j \in \{0, \ldots, m\}$ mit $i < j$. Das heißt die Freiheitsgrade sind aufsteigend mit den Triangulierungen sortiert. Daher speichern wir für jede Ebene die Anzahl der Freiheitsgrade in einem Vektor, um jeden Freiheitsgrad $k \in \mathcal{N}_m$ genau einer Ebene $i \in \{0, \ldots, m\}$ durch $k \in \mathcal{N}_i^{\mathrm{HB}}$ zuordnen zu können. Abschließend wird die rechte Seite mit einer Routine der Programmbibliothek *SimpleFEM* berechnet. Die erzeugten Strukturen werden für die weitere Verwendung in einer Datenstruktur *hb* für das Verfahren der hierarchischen Basen gespeichert.

Für $i \in \{1, \ldots, m\}$ verwenden wir als Lösungsverfahren auf dem Unterraum $\mathcal{W}_i$ das Jacobi-Verfahren, das jeweils nur die Unbekannten aus $\mathcal{N}_i^{\mathrm{HB}}$ berücksichtigt. Als exaktes Lösungsverfahren für den Unterraum $\mathcal{W}_0$ bietet sich die LR-Zerlegung mit Vorwärts- und Rückwärtseinsetzen oder das Verfahren der konjugierten Gradienten, das in Kapitel 2.2.3 eingeführt wurde, an. Wir verwenden im Folgenden das Verfahren der konjugierten Gradienten als exaktes Lösungsverfahren.

Im Folgenden sei A die Steifigkeitsmatrix des betrachteten Randwertproblems bezüglich der hierarchischen Basis und $\hat{A}$ die Steifigkeitsmatrix bezüglich der Knotenbasis. Obwohl die Matrix A gut konditioniert ist, wird diese Eigenschaft dadurch kompensiert, dass aufgrund des großen Trägers der hierarchischen Basisfunktionen die Matrix A viele Nichtnulleinträge im Vergleich zur Matrix $\hat{A}$ enthält. Das heißt die Matrix A ist nicht schwachbesetzt wie die Matrix $\hat{A}$, weshalb in Algorithmen der Einsatz der Matrix $\hat{A}$ erstrebenswert ist. Ein möglicher Ausweg besteht darin, eine Transformationsmatrix S zu verwenden, die die Darstellung einer Finite Elemente Funktion bezüglich der hierarchischen Basis in die Darstellung bezüglich der Knotenbasis transformiert. Folglich gilt der Zusammenhang $A = S^T \hat{A} S$. Wie in [Yse86b] beschrieben, ist die Berechnung von Sx für einen Koeffizientenvektor bezüglich der hierarchischen Basis und $S^T y$ für einen Koeffizientenvektor bezüglich der Knotenbasis mit linearem Aufwand zu realisieren.

Wir nutzen im weiteren Verlauf allerdings aus, dass wir jeweils eine Unterraumkorrektur auf den Unterräumen $\mathcal{W}_k$ für alle $k \in \{0, \ldots, m\}$ durchführen, um die Steifigkeitsmatrix $\hat{A}$ anstatt A zu verwenden. Da $\mathcal{W}_k = \operatorname{span}\left(\left\{\psi_i^{(m)} \;:\; i \in \mathcal{N}_k^{\mathrm{HB}}\right\}\right)$ für $k \in \{0, \ldots, m\}$ gilt, ist $\left(\psi_i^{(m)}\right)_{i \in \mathcal{N}_k^{\mathrm{HB}}}$ für $k \in \{0, \ldots, m\}$ eine Basis von $\mathcal{W}_k$. Für alle $k \in \{0, \ldots, m\}$ gilt nach der Definition der hierarchischen Basis $\psi_i^{(m)} = \psi_i^{(k)} = \varphi_i^{(k)}$ für alle $i \in \mathcal{N}_k^{\mathrm{HB}}$, sodass

Algorithmus 5 : Multiplikative Variante des Verfahrens der hierarchischen Basen

Eingabe : Datenstruktur hb, Näherungslösung x
Ausgabe : neue Näherungslösung x_m
$x_m \leftarrow x$
for $i = m$ **to** 1 **do**
  $x_i \leftarrow \Phi_{\mathrm{Gl},i}\,(x_i, b_i)$
  $d_i \leftarrow b_i - A_i x_i$
  $b_{i-1} \leftarrow r_i d_i$
  $x_{i-1} \leftarrow 0$
Löse $A_0 x_0 = b_0$ exakt
for $i = 1$ **to** m **do**
  $x_i \leftarrow x_i + p_i x_{i-1}$

die Basis $\left(\psi_i^{(m)}\right)_{i \in \mathcal{N}_k^{\mathrm{HB}}}$ mit einer Teilmenge der Knotenbasis $\left(\varphi_i^{(k)}\right)_{i \in \mathcal{N}_k}$ des Finite Elemente Raums $\mathcal{S}_k$ zur Triangulierung $\mathcal{T}_k$ für $k \in \{0, \ldots, m\}$ übereinstimmt. Ausgehend von der Knotenbasis $\left(\varphi_i^{(m)}\right)_{i \in \mathcal{N}_m}$ des Finite Elemente Raums $\mathcal{S}_m$ kann die Unterraumkorrektur auf dem Unterraum $\mathcal{W}_m$ mit der Knotenbasis durchgeführt werden, da die Teilmenge der hierarchischen Basis $\left(\psi_i^{(m)}\right)_{i \in \mathcal{N}_m}$ des Finite Elemente Raums $\mathcal{S}_m$, die den Unterraum $\mathcal{W}_m$ aufspannt, mit einer Teilmenge der Knotenbasis übereinstimmt. Durch Transformation der Knotenbasis mithilfe der Restriktion von dem Finite Elemente Raum $\mathcal{S}_m$ der Ebene m in den Finite Elemente Raum $\mathcal{S}_{m-1}$ der Ebene $m-1$ ist die Unterraumkorrektur auf dem Unterraum $\mathcal{W}_{m-1}$ ebenfalls direkt realisierbar. Die Unterraumkorrektur auf dem Unterraum $\mathcal{W}_k$ ist somit anhand der Knotenbasis auf jeder Ebene k für $k \in \{0, \ldots, m\}$ ohne einen Basiswechsel zu bewerkstelligen.

7.1.1 Algorithmus der multiplikativen Variante des Verfahrens der hierarchischen Basen

Die multiplikative Variante des Verfahrens der hierarchischen Basen entspricht einem Mehrgitterverfahren mit den Parametern $\gamma = 1$, $\nu_1 = 1$ und $\nu_2 = 0$, wie in Algorithmus 5 angegeben ist. Obwohl der Algorithmus mithilfe der Knotenbasis realisiert ist, spiegelt sich der Einfluss der hierarchischen Basis in der Tatsache wider, dass in dem Verfahren der hierarchischen Basen auf jeder Gitterstufe nur ein Teil der Unbekannten von dem Glättungsverfahren behandelt wird. Der Algorithmus wurde in einer iterativen Form im Gegensatz zum Algorithmus 1 des allgemeinen Mehrgitterverfahrens, der in einer rekursiven Variante angegeben wurde, dargestellt. Der einzige Unterschied zwischen dem klassischen Mehrgitterverfahren und der multiplikativen Variante des Verfahrens der

hierarchischen Basen besteht darin, dass für alle $k \in \{0, \ldots, m\}$ auf der Gitterstufe k das Glättungsverfahren nur auf die Knoten in $\mathcal{N}_k^{\mathrm{HB}}$ und nicht auf alle Knoten aus $\mathcal{N}_k$ angewandt wird. Das heißt die Glättung wird für $k \in \{1, \ldots, m\}$ nicht auf allen Gitterpunkten des Gitters der Ebene k durchgeführt, sondern nur auf den Punkten der Ebene k, die nicht bereits zur Ebene $k-1$ gehören. Für das Glättungsverfahren $\Phi_{\mathrm{Gl},k}$ legt der Index $k \in \{1, \ldots, m\}$ fest, dass das Verfahren nur die Knoten aus $\mathcal{N}_k^{\mathrm{HB}}$ behandelt.

Lemma 7.1. *Ein Schritt des in Algorithmus 5 angegebenen Mehrgitterverfahrens entspricht einem Schritt der multiplikativen Variante des Verfahrens der hierarchischen Basen.*

Beweis. Wir zeigen, dass der Glättungsschritt auf der Gitterstufe $k \in \{1, \ldots, m\}$ der Unterraumkorrektur auf dem Raum $\mathcal{W}_k$ entspricht.

Es sei $x_m \in \mathcal{S}$ eine gegebene Näherungslösung. Dann gilt im ersten Schritt des Glättungsverfahrens

$$\begin{aligned} x_m^{(\mathrm{neu})} &= x_m + \Phi_{\mathrm{Gl},m}\left(0, b - A_m x_m\right) = x_m + 0 - N_m\left(A_m \cdot 0 - b + A_m x_m\right) \\ &= x_m - N_m\left(A_m x_m - b\right) = \Phi_{\mathrm{Gl},m}\left(x_m, b\right), \end{aligned}$$

wobei $\Phi_{\mathrm{Gl},m}\left(x_m, b\right) = x_m - N_m\left(A_m x_m - b\right)$ die zweite Normalform des Glättungsverfahrens $\Phi_{\mathrm{Gl},m}$ ist. Das heißt, dass die Unterraumkorrektur mittels Glättungsverfahren auf dem Unterraum $\mathcal{W}_m$ mit dem Startvektor 0 und dem Defekt als rechter Seite dem Glättungsverfahren mit dem Startvektor x_m und der rechten Seite b entspricht.

Für alle $k \in \{0, \ldots, m-1\}$ zeigen wir per vollständiger Induktion, dass der Glättungsschritt auf Gitterstufe k der Unterraumkorrektur auf dem Raum $\mathcal{W}_k$ entspricht.
Induktionsanfang: Es sei $k = 0$. Dann ergibt sich der Defekt für die Grobgitterkorrektur durch $d_{m-k} = d_m = b_m - A_m x_m^{(\mathrm{neu})}$, sodass $b_{m-1} = r_m d_m = r_m\left(b_m - A_m x_m^{(\mathrm{neu})}\right)$ für die rechte Seite des Glättungsverfahrens auf Gitterstufe $m-1$ gilt. Daraus folgt

$$x_{m-1} = \Phi_{\mathrm{Gl},m-1}\left(0, b_{m-1}\right) = \Phi_{\mathrm{Gl},m-1}\left(0, r_m\left(b_m - A_m x_m^{(\mathrm{neu})}\right)\right).$$

Deshalb entspricht der Glättungsschritt auf der Gitterstufe $m-1$ der Unterraumkorrektur auf dem Raum $\mathcal{W}_{m-1}$, da $x_m^{(\mathrm{neu})}$ die Unterraumkorrektur auf dem Raum $\mathcal{W}_m$ bereits enthält.
Induktionsvoraussetzung: Es sei $k \in \{0, \ldots, m-2\}$ und es gelte

$$d_{m-k} = r_{m-k+1} \cdot \ldots \cdot r_m\left(b_m - A_m\left(x_m^{(\mathrm{neu})} + \sum_{i=0}^{k-1} p_m \cdot \ldots \cdot p_{m-i} x_{m-i-1}\right)\right).$$

Induktionsschritt: Auf der Gitterstufe $m-(k+1)$ erhalten wir mithilfe der Galerkin-Eigenschaft für den Defekt

$$\begin{aligned}
d_{m-(k+1)} &= b_{m-(k+1)} - A_{m-(k+1)} x_{m-(k+1)} \\
&= r_{m-k} d_{m-k} - A_{m-(k+1)} x_{m-(k+1)} \\
&= r_{m-k} d_{m-k} - r_{m-k} \cdot \ldots \cdot r_m A_m p_m \cdot \ldots \cdot p_{m-k} x_{m-(k+1)} \\
&\overset{\text{I.V.}}{=} r_{m-k} \cdot r_{m-k+1} \cdot \ldots \cdot r_m \left(b_m - A_m \left(x_m^{(\text{neu})} + \sum_{i=0}^{k-1} p_m \cdot \ldots \cdot p_{m-i} x_{m-i-1} \right) \right) \\
&\quad - r_{m-k} \cdot \ldots \cdot r_m A_m p_m \cdot \ldots \cdot p_{m-k} x_{m-(k+1)} \\
&= r_{m-k} \cdot \ldots \cdot r_m \left(b_m - A_m \left(x_m^{(\text{neu})} + \sum_{i=0}^{k} p_m \cdot \ldots \cdot p_{m-i} x_{m-i-1} \right) \right).
\end{aligned}$$

Im Fall $k \in \{0, \ldots, m-3\}$ gilt somit $x_{m-(k+2)} = \Phi_{\text{Gl},m-(k+2)} \left(0, r_{m-(k+1)} d_{m-(k+1)}\right)$ für das Glättungsverfahren auf der Stufe $m-(k+2)$, das dem Unterraumkorrekturverfahren auf dem Unterraum $\mathcal{W}_{m-(k+2)}$ entspricht. Hierbei wurden insbesondere die Unterraumkorrekturen auf den Räumen $\mathcal{W}_i$ mit $i \in \{m-(k+1), \ldots, m\}$ bereits berechnet und sind im Defekt $d_{m-(k+1)}$ enthalten. Im Fall $k = m-2$ wird die Unterraumkorrektur auf dem Raum $\mathcal{W}_0$ mit der rechten Seite $b_0 = r_1 d_1 = r_{m-(k+1)} d_{m-(k+1)}$ exakt durchgeführt, wobei im Defekt d_1 die Unterraumkorrekturen auf den Räumen $\mathcal{W}_i$ für $i \in \{m-(k+1), \ldots, m\}$ bereits enthalten sind. □

Im Hauptprogramm rufen wir den Algorithmus 5 solange auf, bis die berechnete Näherungslösung ein gegebenes Abbruchkriterium erfüllt, beispielsweise bis die Norm des Residuums kleiner als eine vorgegebene Schranke ist.

Aufgrund der Umsetzung der multiplikativen Variante des Verfahrens der hierarchischen Basen als Mehrgitterverfahren, wobei nur auf einem Teil der Unbekannten auf jeder Gitterstufe geglättet wird, ist der Rechenaufwand dieses Verfahrens kleiner als der Rechenaufwand des allgemeinen Mehrgitterverfahrens und die Voraussetzung an die Gitterverfeinerung entfällt. Allerdings konvergiert dafür das Verfahren der hierarchischen Basen langsamer als das allgemeine Mehrgitterverfahren.

7.1.2 Algorithmus der additiven Variante des Verfahrens der hierarchischen Basen

Die additive Variante des Verfahrens der hierarchischen Basen ist in Algorithmus 6 dargestellt. Dabei wird zunächst der Defekt auf dem feinsten Gitter berechnet und schritt-

Algorithmus 6 : Additive Variante des Verfahrens der hierarchischen Basen

Eingabe : Datenstruktur hb, Näherungslösung x, Dämpfungsparameter θ
Ausgabe : neue Näherungslösung x_m

$x_m \leftarrow x$
$d_m \leftarrow A_m x_m - b_m$
for $i = m$ **to** 1 **do**
  $d_{i-1} \leftarrow r_i d_i$
  $x_i \leftarrow \Phi_{\mathrm{Gl},i,\theta}(x_i, d_i)$
  $x_{i-1} \leftarrow 0$
Löse $A_0 x_0 = b_0$ exakt
for $i = 1$ **to** m **do**
  $x_i \leftarrow x_i + p_i x_{i-1}$

weise auf die gröberen Gitter restringiert. Die berechneten Unterraumkorrekturen gehen hierbei nicht in den Defekt ein, sondern werden erst nach der Berechnung aller Unterraumkorrekturen durch stückweise Prolongation zur gegebenen Näherungslösung addiert, woraus sich die neue Näherungslösung ergibt. Das approximative Lösungsverfahren auf dem Unterraum $\mathcal{W}_k$ für $k \in \{1, \ldots, m\}$, das dem Glättungsverfahren $\Phi_{\mathrm{Gl},k,\theta}$ entspricht, erhält einen Dämpfungsparameter $\theta \in \mathbb{R}_{>0}$ als Parameter, um die Konvergenz des gesamten Verfahrens sicherzustellen. Aufgrund des Startvektors 0 für die Unterraumkorrekturen auf den Unterräumen $\mathcal{W}_i$ für $i \in \{1, \ldots, m-1\}$ entspricht dieses Vorgehen dem additiven Unterraumkorrekturverfahren mit dem Dämpfungsparameter θ.

Dieses Verfahren kann wie die multiplikative Variante als Mehrgitterverfahren aufgefasst werden mit dem Unterschied, dass der Defekt vor dem Glättungsschritt auf das gröbere Gitter transferiert wird. Die Neuberechnung des Defekts kann derweil entfallen, weil für $i \in \{1, \ldots, m-1\}$

$$d_i = b_i - A_i x_i = b_i - A_i \cdot 0 = b_i$$

gilt. Deshalb muss der Defekt nur einmal auf dem Raum $\mathcal{W}_m$ berechnet werden.

7.1.3 Algorithmus des vorkonditionierten cg-Verfahrens

Als Vorkonditionierer für das Verfahren der konjugierten Gradienten können wir sowohl die Matrix C aus der zweiten Normalform des additiven Unterraumkorrekturverfahrens, die nach Lemma 5.10 positiv definit ist, als auch die Matrix $N_{\mathrm{mul,\ sym}}$ der zweiten Normalform des symmetrischen multiplikativen Unterraumkorrekturverfahrens, die nach Lemma 5.8 positiv definit ist, einsetzen. Wie die multiplikative Variante des Verfahrens

Algorithmus 7 : Symmetrische multiplikative Variante des Verfahrens der hierarchischen Basen

Eingabe : Datenstruktur hb, Näherungslösung x
Ausgabe : neue Näherungslösung x_m
$x_m \leftarrow x$
for $i = m$ **to** 1 **do**
  $x_i \leftarrow \Phi_{\mathrm{Gl},i}\,(x_i, b_i)$
  $d_i \leftarrow b_i - A_i x_i$
  $b_{i-1} \leftarrow r_i d_i$
  $x_{i-1} \leftarrow 0$
Löse $A_0 x_0 = b_0$ exakt
for $i = 1$ **to** m **do**
  $x_i \leftarrow x_i + p_i x_{i-1}$
  $x_i \leftarrow \Phi_{\mathrm{Gl},i}\,(x_i, b_i)$

der hierarchischen Basen kann die symmetrische multiplikative Variante des Verfahrens der hierarchischen Basen gleichermaßen als ein Mehrgitterverfahren aufgefasst werden, das in Algorithmus 7 angegeben ist, wobei für alle $k \in \{0, \ldots, m\}$ auf der Gitterstufe k die Glättungsverfahren nur auf die Knoten in $\mathcal{N}_k^{\mathrm{HB}}$ anstatt auf alle Knoten aus $\mathcal{N}_k$ angewandt werden. Die Symmetrie des Verfahrens spiegelt sich darin wider, dass neben den Vorglättungsschritten auch Nachglättungsschritte durchgeführt werden.

Lemma 7.2. *Ein Schritt des in Algorithmus 7 angegebenen Mehrgitterverfahrens entspricht einem Schritt der symmetrischen multiplikativen Variante des Verfahrens der hierarchischen Basen.*

Beweis. Da in Lemma 7.1 gezeigt wurde, dass das in Algorithmus 5 dargestellte Mehrgitterverfahren mit den Parametern $\nu_1 = 1$, $\nu_2 = 0$ und $\gamma = 1$ der multiplikativen Variante des Verfahrens der hierarchischen Basen entspricht, bleibt hier nur noch zu zeigen, dass die Unterraumkorrekturen in der umgekehrten Reihenfolge den Nachglättungsschritten entsprechen. Dazu verwenden wir die Notationen aus Lemma 7.1.

Weil die erste Unterraumkorrektur auf dem Unterraum $\mathcal{W}_0$ die exakte Lösung $x_0 = A_0^{-1} b_0 = A_0^{-1} r_1 d_1$ berechnet, gilt für die rechte Seite der darauf folgenden zweiten Unterraumkorrektur auf dem Unterraum $\mathcal{W}_0$

$$
\begin{aligned}
& r_1 \cdot \ldots \cdot r_m \left(b_m - A_m \left(x_m^{(\mathrm{neu})} \sum_{i=0}^{m-2} p_m \cdot \ldots \cdot p_{m-i} x_{m-i-1} + p_m \cdot \ldots \cdot p_1 x_0 \right) \right) \\
&= r_1 d_1 - r_1 \cdot \ldots \cdot r_m A_m p_m \cdot \ldots \cdot p_1 x_0 = r_1 d_1 - A_0 x_0 = r_1 d_1 - A_0 A_0^{-1} r_1 d_1 \\
&= r_1 d_1 - r_1 d_1 = 0.
\end{aligned}
$$

Da A_0 positiv definit und somit insbesondere regulär ist, ist die Lösung der zweiten Unterraumkorrektur auf dem Unterraum $\mathcal{W}_0$, also die Lösung der Gleichung $A_0\tilde{x}_0 = 0$, $\tilde{x}_0 = 0$, sodass diese Lösung keine Verbesserung der Näherungslösung des Unterraumkorrekturverfahrens darstellt und diese Unterraumkorrektur bei der Implementierung deswegen entfallen kann.

Es sei $k \in \{1, \ldots, m-1\}$ und x_k enthalte jeweils beide Unterraumkorrekturen auf den Unterräumen $\mathcal{W}_0$ bis $\mathcal{W}_{k-1}$ und die erste Unterraumkorrektur auf dem Unterraum $\mathcal{W}_k$. Dann gilt für den Nachglättungsschritt auf der Gitterstufe k

$$\begin{aligned}
x_k^{(\text{neu})} &= \Phi_{\text{Gl},k}\left(x_k, b_k\right) = \Phi_{\text{Gl},k}\left(x_k, r_{k+1}d_{k+1}\right) = x_k - N_k\left(A_k x_k - r_{k+1}d_{k+1}\right) \\
&= x_k - N_k\Bigg(r_{k+1}\cdot\ldots\cdot r_m A_m p_m\cdot\ldots\cdot p_{k+1}x_k \\
&\qquad - r_{k+1}\cdot r_{k+2}\cdot\ldots\cdot r_m\left(b_m - A_m\left(x_m^{(\text{neu})} + \sum_{i=0}^{m-k-2} p_m\cdot\ldots\cdot p_{m-i}x_{m-i-1}\right)\right)\Bigg) \\
&= x_k - N_k\left(r_{k+1}\cdot\ldots\cdot r_m\left(A_m\left(x_m^{(\text{neu})} + \sum_{i=0}^{m-k-1} p_m\cdot\ldots\cdot p_{m-i}x_{m-i-1}\right) - b_m\right)\right).
\end{aligned}$$

Damit entspricht der Nachglättungsschritt auf der Gitterstufe k der zweiten Unterraumkorrektur auf dem Unterraum $\mathcal{W}_k$, wobei die ersten Unterraumkorrekturen auf den Unterräumen $\mathcal{W}_m$ bis $\mathcal{W}_{k+1}$ in der rechten Seite b_k und in x_k jeweils die beiden Unterraumkorrekturen auf den Räumen $\mathcal{W}_0$ bis $\mathcal{W}_{k-1}$ sowie die erste Unterraumkorrektur auf dem Unterraum $\mathcal{W}_k$ enthalten sind.

Für den Nachglättungsschritt auf der Gitterstufe m erhalten wir

$$x_m^{(\text{neu})} = \Phi_{\text{Gl},m}\left(x_m, b_m\right) = x_m - N_m\left(A_m x_m - b_m\right),$$

wobei in x_m jeweils die beiden Unterraumkorrekturen auf den Unterräumen $\mathcal{W}_0$ bis $\mathcal{W}_{m-1}$ und die erste Unterraumkorrektur auf dem Unterraum $\mathcal{W}_m$ enthalten sind, sodass dieser Nachglättungsschritt der zweiten Unterraumkorrektur auf dem Unterraum $\mathcal{W}_m$ entspricht. □

Da wir als Vorkonditionierer jeweils eine Matrix N aus der zweiten Normalform eines linearen Iterationsverfahrens Φ einsetzen, kann die Berechnung von $q^{(m)} = Nr^{(m)}$ für alle $m \in \mathbb{N}_0$ im vorkonditionierten Verfahren der konjugierten Gradienten, das in Definition 2.31 angegeben ist, durch

$$q^{(m)} = Nr^{(m)} = 0 - N\left(A\cdot 0 - r^{(m)}\right) = \Phi\left(0, r^{(m)}\right)$$

Algorithmus 8 : Initialisierung des vorkonditionierten Verfahrens der konjugierten Gradienten

Eingabe : Matrix A, rechte Seite b, Näherungslösung x, Datenstruktur hb
Ausgabe : Residuum r, Suchrichtung p
$r \leftarrow b - Ax$
$q \leftarrow 0$
Berechnung von q durch Aufruf des symmetrischen multiplikativen Unterraumkorrekturverfahrens mit rechter Seite r und Startvektor q
$p \leftarrow q$

Algorithmus 9 : Schritt des vorkonditionierten Verfahrens der konjugierten Gradienten

Eingabe : Matrix A, rechte Seite b, Näherungslösung x, Datenstruktur hb, Residuum r, Suchrichtung p
Ausgabe : neue Näherungslösung x, neues Residuum r, neue Suchrichtung p
$a \leftarrow Ap$
$\gamma \leftarrow \langle p, a \rangle_2$
$\lambda \leftarrow \frac{\langle p,r \rangle_2}{\gamma}$
$x \leftarrow x + \lambda p$
$r \leftarrow r - \lambda a$
$q \leftarrow 0$
Berechnung von q durch Aufruf des symmetrischen multiplikativen Unterraumkorrekturverfahrens mit rechter Seite r und Startvektor q
$p \leftarrow q - \frac{\langle q,a \rangle_2}{\gamma} p$

berechnet werden. Das heißt wir berechnen $q^{(m)}$ durch einen Aufruf des linearen Iterationsverfahrens Φ mit dem Startvektor 0 und der rechten Seite $r^{(m)}$.

Die Algorithmen zur Initialisierung des vorkonditionierten Verfahrens der konjugierten Gradienten und der Durchführung eines Schrittes dieses Verfahrens sind in Algorithmus 8 und 9 dargestellt. Als symmetrisches multiplikatives Unterraumkorrekturverfahren verwenden wir dabei sowohl die additive Variante als auch die symmetrische multiplikative Variante des Verfahrens der hierarchischen Basen.

7.2 Numerische Resultate

Wir führen im Folgenden numerische Experimente anhand zweier Testbeispiele durch, um die Ergebnisse der vorgestellten Algorithmen sowohl für die multiplikative und additive Variante des Verfahrens der hierarchischen Basen als auch des vorkonditionierten Verfahrens der konjugierten Gradienten mit den theoretischen Aussagen zu vergleichen.

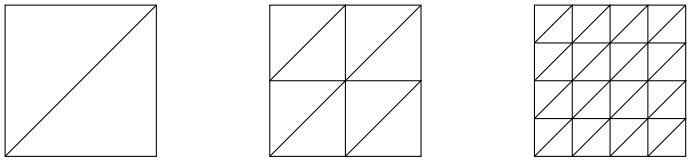


Abbildung 7.1: Familie von Triangulierungen des Einheitsquadrats, die durch Rot-Verfeinerungen aus der Anfangstriangulierung entstanden sind.

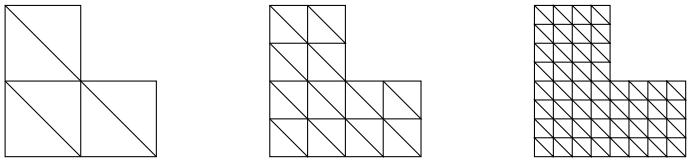


Abbildung 7.2: Familie von Triangulierungen des L-förmigen Gebiets, die durch Rot-Verfeinerungen aus der Anfangstriangulierung entstanden sind.

7.2.1 Testbeispiele

Als erstes Testbeispiel verwenden wir die Poisson-Gleichung, die in Kapitel 3.6 eingeführt wurde, auf dem Einheitsquadrat $\Omega := (-1,1) \times (-1,1) \subset \mathbb{R}^2$ mit der rechten Seite $f : \Omega \to \mathbb{R}, x \mapsto 0$. Gesucht ist somit die eindeutige Lösung $u \in C^2(\Omega) \cap C\left(\bar{\Omega}\right)$ mit

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

Die exakte Lösung dieses Beispiels ist $u : \bar{\Omega} \to \mathbb{R}, x \mapsto 0$. Wir beschränken uns in diesem Beispiel auf eine gleichmäßige Verfeinerung des Gebiets durch Rot-Verfeinerungen, sodass insbesondere eine quasiuniforme Familie von Triangulierungen entsteht, deren ersten drei Elemente in Abbildung 7.1 dargestellt sind. Aus dieser Familie von Triangulierungen wählen wir eine Starttriangulierung $\mathcal{T}_0$ aus. Da der Finite Elemente Raum $\mathcal{S}_0$, der zu dieser Triangulierung $\mathcal{T}_0$ gehört, dem Unterraum $\mathcal{W}_0$ entspricht, wählen wir die dritte Triangulierung aus Abbildung 7.1 als Starttriangulierung, sodass $\dim \mathcal{W}_0 = 9$ gilt. Neben der Starttriangulierung wählen wir noch den Parameter $m \in \mathbb{N}_0$, der die Anzahl der Verfeinerungen der Triangulierung $\mathcal{T}_0$ angibt. Mithilfe des Verfahrens der hierarchischen Basen soll nun das aus der Finite Elemente Methode entstehende lineare Gleichungssystem auf der Triangulierung $\mathcal{T}_m$ gelöst werden.

Als zweites Testbeispiel verwenden wir die Poisson-Gleichung auf dem L-förmigen Gebiet $\Omega := ((-1,1) \times (-1,1)) \setminus ([0,1] \times [0,1]) \subset \mathbb{R}^2$ mit der rechten Seite $f : \Omega \to \mathbb{R}, x \mapsto 0$. Wie im ersten Beispiel ist die exakte Lösung $u : \bar{\Omega} \to \mathbb{R}, x \mapsto 0$. Die ersten

		Norm kleiner 10^{-6}		Norm kleiner 10^{-12}	
m	θ	multiplikativ	additiv	multiplikativ	additiv
1	0,5	37	104	71	200
2	0,25	55	263	109	513
3	0,25	81	416	159	818
4	0,25	111	594	216	1172
5	0,25	143	806	284	1585
6	0,25	183	1041	361	2055
7	0,25	229	1301	449	2566
8	0,25	275	1597	542	3133
9	–	329	–	645	–
10	–	397	–	782	–
11	–	447	–	886	–
12	–	516	–	1024	–

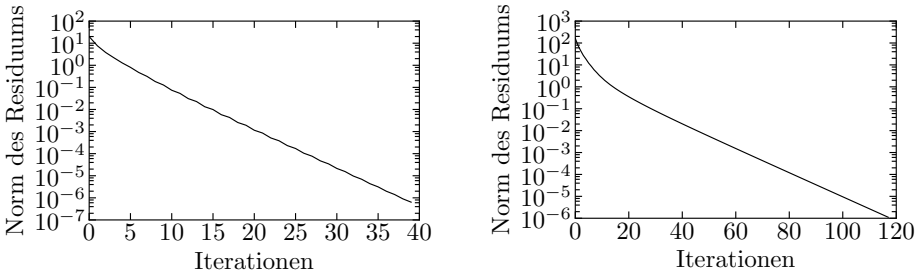
Tabelle 7.1: Anzahl der Iterationsschritte zur Reduktion der Norm des Residuums auf 10^{-6} beziehungsweise 10^{-12} der multiplikativen und additiven Variante des Verfahrens der hierarchischen Basen für das Einheitsquadrat.

drei Elemente einer Familie von gleichmäßigen Triangulierungen sind in Abbildung 7.2 dargestellt, wobei als Verfeinerungsstrategie die Rot-Verfeinerung verwendet wurde. Als Starttriangulierung $\mathcal{T}_0$ wählen wir die zweite Triangulierung aus Abbildung 7.2. Dazu sei $m \in \mathbb{N}_0$ die Anzahl der Verfeinerungen der Triangulierung $\mathcal{T}_0$. Ebenso wie bei dem ersten Testbeispiel lösen wir das entstehende lineare Gleichungssystem auf der Triangulierung $\mathcal{T}_m$ mithilfe des Verfahrens der hierarchischen Basen.

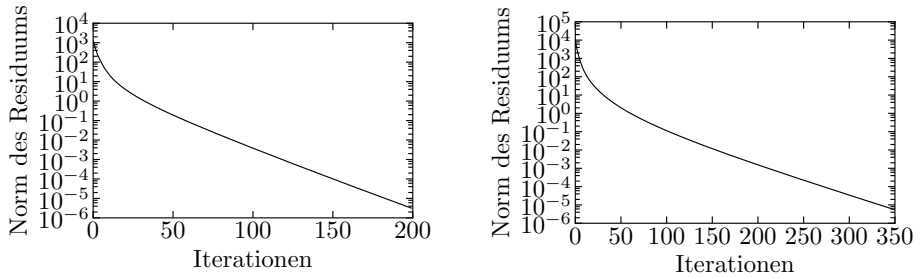
Für die beiden Testbeispiele führen wir für unterschiedliche Anzahlen von Verfeinerungen m je Beispiel mehrere Testläufe mit jeweils zufällig generierten Startvektoren durch. Hierbei wird die Anzahl der Iterationsschritte gezählt, die notwendig sind, bis die Norm des Residuums eine vorgegebene Schranke unterschreitet. Dabei ist in den folgenden Tabellen 7.1, 7.2 und 7.3 jeweils der auf eine ganze Zahl abgerundete Mittelwert angegeben. Für die mit „-“ gekennzeichneten Einträge in diesen Tabellen wurden aufgrund des hohen Rechenaufwands keine Testläufe durchgeführt.

7.2.2 Numerische Resultate für das Verfahren der hierarchischen Basen

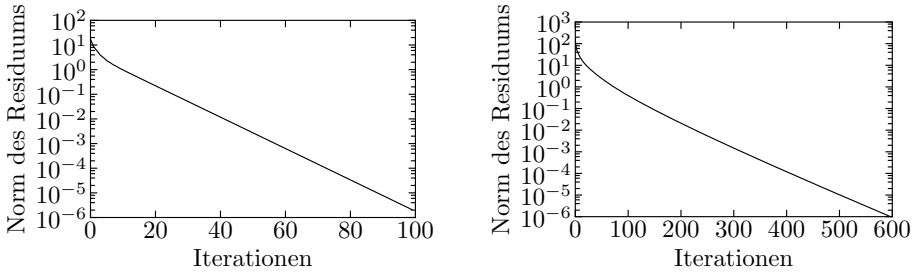
Zunächst beziehen sich die numerischen Resultate ausschließlich auf das erste Testbeispiel auf dem Einheitsquadrat. In der Tabelle 7.1 ist die Anzahl der Iterationsschritte der multiplikativen und additiven Variante des Verfahrens der hierarchischen Basen zur Reduktion der Norm des Residuums für verschiedene Verfeinerungsanzahlen angegeben. Weil die Konvergenz der additiven Variante nach Satz 5.11 von dem Dämpfungspara-



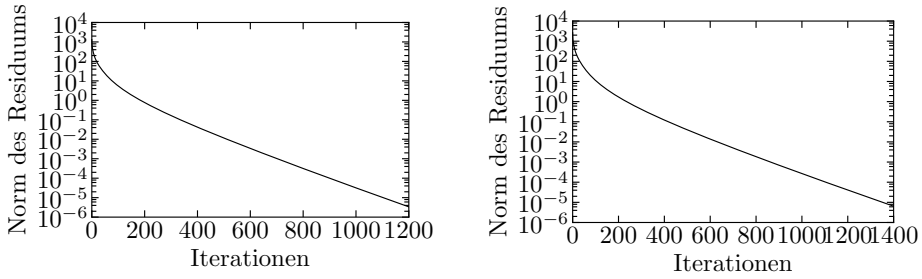
(a) $m = 1$ (b) $m = 4$



(c) $m = 7$ (d) $m = 10$



(e) $m = 1$ (f) $m = 4$



(g) $m = 7$ (h) $m = 8$

Abbildung 7.3: Darstellung der Reduktion der Norm des Residuums für die multiplikative Variante des Verfahrens der hierarchischen Basen in 7.3(a) bis 7.3(d) und für die additive Variante in 7.3(e) bis 7.3(h).

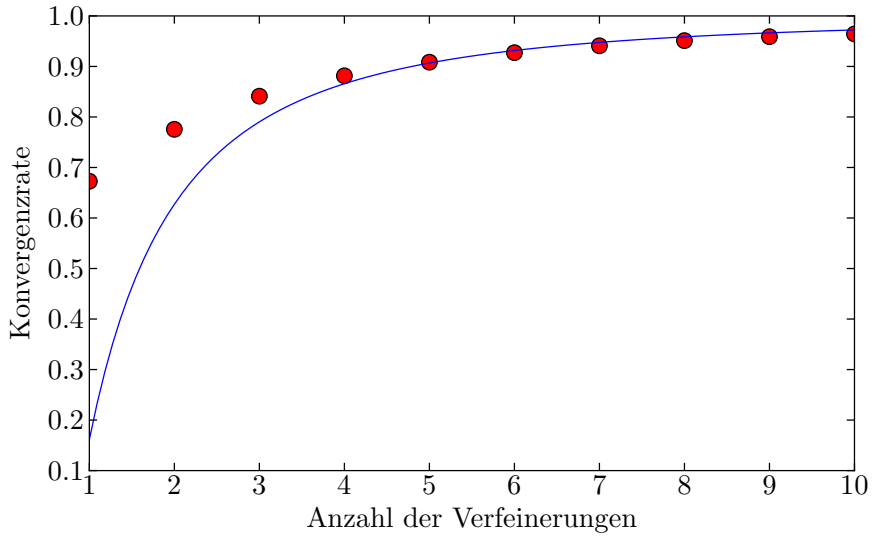


Abbildung 7.4: Konvergenzrate der multiplikativen Variante des Verfahrens der hierarchischen Basen.

meter $\theta \in \mathbb{R}_{>0}$, der von den Konstanten K_1 und K_2 der Annahmen 1 und 2' abhängt, beeinflusst wird, ist zudem in der Tabelle 7.1 der verwendete Dämpfungsparameter aufgeführt. Die Konstante K_1 hängt für das Verfahren der hierarchischen Basen, wie in Satz 6.8 gezeigt, von der Anzahl der Verfeinerungen m ab, sodass für unterschiedliche m der Dämpfungsparameter variiert. In Abbildung 7.3 ist die Reduktion der Norm des Residuums nochmals für jeweils einen Testlauf für unterschiedliche m dargestellt. Wie nach dem Konvergenzsatz 5.6 erwartet, ist in dieser Abbildung aufgrund der logarithmischen Skala zu sehen, dass der Fehler in jedem Iterationsschritt um einen Faktor reduziert wird. Wegen der Konvergenzrate

$$1 - \frac{2-\lambda}{K_1(1+K_2)^2} = 1 - \frac{c}{(m+1)^2}$$

aus Satz 5.6 mit einer Konstanten $c \in \mathbb{R}_{>0}$, der Anzahl der Unterräume $m \in \mathbb{N}$ sowie den Konstanten λ, K_1 und K_2 aus den Annahmen des multiplikativen Verfahrens der hierarchischen Basen ist mit steigender Anzahl der Unterräume ein langsameres Konvergenzverhalten zu erwarten. Dies spiegelt sich in der steigenden Anzahl der Iterationsschritte für eine wachsende Anzahl der Unterräume in Tabelle 7.1 wider. In Abbildung 7.4 sind die aus den Testläufen berechneten Konvergenzraten als Punkte zusammen mit der daraus resultierenden Konvergenzrate des Verfahrens dargestellt, die ebenfalls dem erwarteten Verlauf der Konvergenzrate entspricht.

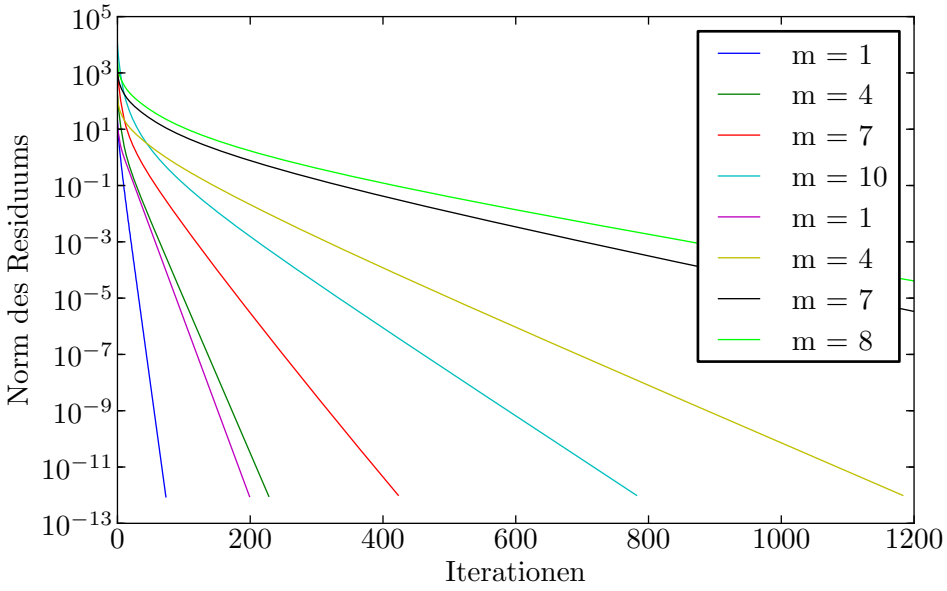


Abbildung 7.5: Vergleich der Reduktion der Norm des Residuums für die multiplikative mit der additiven Variante des Verfahrens der hierarchischen Basen, wobei in blau, grün, rot und cyan Testläufe der multiplikativen Variante dargestellt sind.

Zum Vergleich der multiplikativen Variante mit der additiven Variante des Verfahrens der hierarchischen Basen ist in Abbildung 7.5 die Reduktion der Norm des Residuums für verschiedene Anzahlen der Unterräume für jeweils einen Testlauf für beide Varianten dargestellt. Sowohl diese Abbildung als auch Tabelle 7.1 zeigen, dass die multiplikative Variante schneller als die additive Variante konvergiert, was den Erwartungen entspricht, da bei der multiplikativen Variante gegenüber der additiven Variante nach jeder Unterraumkorrektur mit einer aktualisierten Näherungslösung fortgefahren wird. Die notwendige Wahl des Dämpfungsparameters für die additive Variante des Verfahrens der hierarchischen Basen hat sich gegenüber der multiplikativen Variante als Nachteil herausgestellt, wobei die mögliche Parallelisierung der additiven Variante wiederum einen Vorteil darstellt.

In der Tabelle 7.2 ist für das zweite Testbeispiel auf dem L-förmigen Gebiet die Anzahl der Iterationsschritte zur Reduktion der Norm des Residuums für verschiedene Verfeinerungsanzahlen sowohl der multiplikativen als auch der additiven Variante des Verfahrens der hierarchischen Basen angegeben. Wie für das erste Testbeispiel ist dabei zu erkennen, dass mit steigender Anzahl der Verfeinerungen die Anzahl der Iterationsschritte wächst und dass zudem die multiplikative Variante schneller als die additive Variante konver-

		Norm kleiner 10^{-6}		Norm kleiner 10^{-12}	
m	θ	multiplikativ	additiv	multiplikativ	additiv
1	0,5	33	82	64	200
2	0,25	51	242	100	475
3	0,25	74	396	146	770
4	0,25	102	565	203	1111
5	0,25	136	805	265	1514
6	0,25	173	1047	341	1957
7	0,25	215	1305	422	2465
8	0,25	262	1605	516	3033
9	–	316	–	622	–
10	–	384	–	759	–
11	–	435	–	879	–
12	–	502	–	1001	–

Tabelle 7.2: Anzahl der Iterationsschritte zur Reduktion der Norm des Residuums auf 10^{-6} beziehungsweise 10^{-12} der multiplikativen und additiven Variante des Verfahrens der hierarchischen Basen für das L-förmige Gebiet.

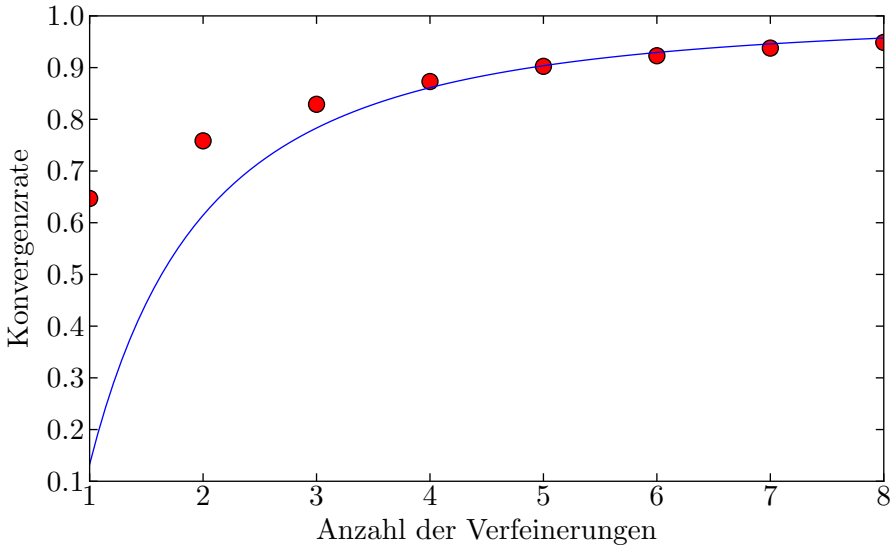


Abbildung 7.6: Konvergenzrate der multiplikativen Variante des Verfahrens der hierarchischen Basen für das L-förmige Gebiet.

giert, was erneut den Erwartungen des Konvergenzverhaltens entspricht. In Abbildung 7.6 sind die aus den Testläufen für das L-förmige Gebiet berechneten Konvergenzraten als Punkte zusammen mit der daraus resultierenden Konvergenzrate des Verfahrens dargestellt, die den erwarteten Verlauf der Konvergenzgeschwindigkeit wiedergibt.

	Norm kleiner 10^{-6}			Norm kleiner 10^{-12}	
m	cg	additiv	multiplikativ	additiv	multiplikativ
1	21	18	9	32	16
2	45	29	14	50	25
3	89	41	18	70	32
4	173	53	22	91	39
5	335	64	26	112	45
6	634	77	30	132	51
7	1192	89	34	154	58
8	2213	102	38	175	64
9	4042	115	42	196	71
10	7745	128	46	217	78
11	–	142	51	240	84
12	–	154	54	261	91

Tabelle 7.3: Anzahl der Iterationsschritte zur Reduktion der Norm des Residuums auf 10^{-6} beziehungsweise 10^{-12} mit dem vorkonditionierten Verfahren der konjugierten Gradienten für das Einheitsquadrat.

7.2.3 Numerische Resultate für das vorkonditionierte Verfahren der konjugierten Gradienten

Das Verfahren der konjugierten Gradienten konditionieren wir sowohl mit der Matrix C der additiven Variante des Verfahrens der hierarchischen Basen als auch mit der Matrix $N_{\text{mul, sym}}$ der symmetrischen multiplikativen Variante des Verfahrens der hierarchischen Basen vor, sodass für die Vorkonditionierung in jedem Schritt die additive Variante beziehungsweise die symmetrische multiplikative Variante des Verfahrens der hierarchischen Basen aufgerufen wird. Als Testbeispiel setzen wir wieder die Poisson-Gleichung auf dem Einheitsquadrat ein. In der Tabelle 7.3 ist für die beiden Vorkonditionierungsvarianten die Anzahl der benötigten Iterationsschritte zur Reduktion der Norm des Residuums für unterschiedliche Verfeinerungsanzahlen m angegeben. Obendrein ist die Anzahl der benötigten Schritte des Verfahrens der konjugierten Gradienten ohne Vorkonditionierung zum Vergleich angegeben. In Abbildung 7.7 ist die Reduktion der Norm des Residuums nochmals jeweils für einen Testlauf für verschiedene Verfeinerungsanzahlen dargestellt. Sowohl aus der Tabelle 7.3 als auch aus der Abbildung 7.7 ist zu erkennen, dass die Anzahl der Iterationsschritte mit zunehmenden Verfeinerungsanzahlen für beide Varianten der Vorkonditionierung steigt und dass für die Vorkonditionierung mit $N_{\text{mul, sym}}$ weniger Iterationsschritte als für die Vorkonditionierung mit C notwendig sind. Diese steigende Anzahl der Iterationsschritte bei der Vorkonditionierung entspricht dem theoretisch erwarteten Konvergenzergebnis aus Satz 2.32, da die Konditionszahlen nach den Sätzen 5.9

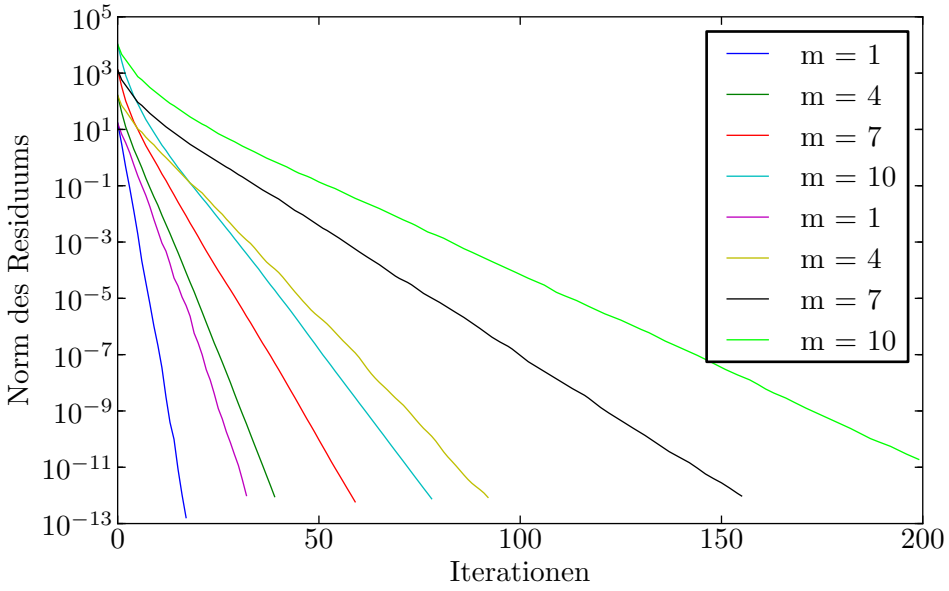


Abbildung 7.7: Vergleich der Reduktion der Norm des Residuums für das mit der additiven beziehungsweise symmetrischen multiplikativen Variante des Verfahrens der hierarchischen Basen vorkonditionierte Verfahren der konjugierten Gradienten, wobei in blau, grün, rot und cyan Testläufe des vorkonditionierten Verfahrens der konjugierten Gradienten dargestellt sind, die mit der symmetrischen multiplikativen Variante des Verfahrens der hierarchischen Basen vorkonditioniert wurden.

und 5.12 größer werden, wenn die Anzahl der Verfeinerungsebenen steigt. Weiter ist im Vergleich der Anzahl der Iterationsschritte des vorkonditionierten mit dem normalen Verfahren der konjugierten Gradienten festzustellen, dass die vorkonditionierten Varianten sehr viel weniger Iterationsschritte benötigen, um das gewünschte Resultat zu erzielen. Das heißt, dass das Spektrum der Steifigkeitsmatrix geeignet transformiert wird. In der Abbildung 7.8 sind die aus den Testläufen berechneten Konvergenzraten als Punkte für das mit der symmetrischen multiplikativen Variante des Verfahrens der hierarchischen Basen vorkonditionierte Verfahren der konjugierten Gradienten und als Dreiecke für das vorkonditionierte Verfahren der konjugierten Gradienten, das mit der additiven Variante des Verfahrens der hierarchischen Basen vorkonditioniert wurde, dargestellt. Diese Konvergenzraten stimmen jeweils mit dem theoretischen Konvergenzergebnis aus Satz 2.32 überein, weil die Konditionszahlen der vorkonditionierten Matrizen $N_{\text{mul, sym}}A$ und CA nach den Sätzen 5.9 und 5.12 von der Anzahl der Verfeinerungen abhängen. Insbesondere geht aus der Abbildung 7.8 hervor, dass das mit der symmetrischen multiplikativen Va-

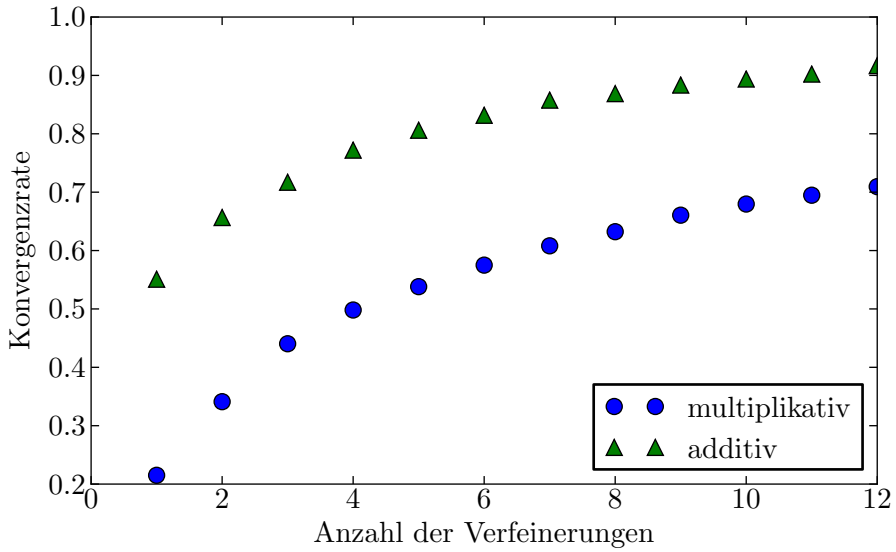


Abbildung 7.8: Konvergenzraten des mit der symmetrischen multiplikativen und der additiven Variante des Verfahrens der hierarchischen Basen vorkonditionierten Verfahrens der konjugierten Gradienten.

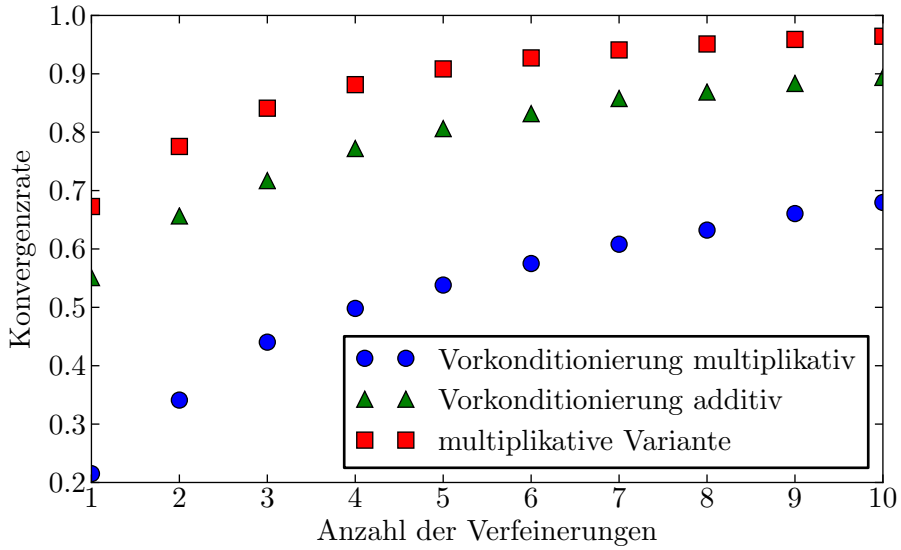


Abbildung 7.9: Vergleich der Konvergenzraten für die multiplikative Variante des Verfahrens der hierarchischen Basen mit den Konvergenzraten der beiden vorkonditionierten Varianten des Verfahrens der konjugierten Gradienten.

riante vorkonditionierte Verfahren der konjugierten Gradienten schneller als das mit der additiven Variante vorkonditionierte Verfahren der konjugierten Gradienten konvergiert.

Der Vergleich der Anzahl der Iterationsschritte zur Reduktion der Norm des Residuums für die beiden Varianten des Verfahrens der hierarchischen Basen, die in Tabelle 7.1 aufgeführt sind, einerseits und für die beiden Varianten des vorkonditionierten Verfahrens der konjugierten Gradienten, die in Tabelle 7.3 dargestellt sind, andererseits zeigt, dass die vorkonditionierten Varianten deutlich effizienter sind. Diese Tatsache spiegelt sich wiederum in Abbildung 7.9 wider, da die Konvergenzraten für die beiden Varianten des vorkonditionierten Verfahrens der konjugierten Gradienten kleiner als die Konvergenzraten der multiplikativen Variante des Verfahrens der hierarchischen Basen sind. Das heißt, dass die Verwendung des Verfahrens der hierarchischen Basen als Vorkonditionierer für das Verfahren der konjugierten Gradienten wirkungsvoller als das eigenständige Verfahren der hierarchischen Basen ist.

8 Fazit

In dieser Arbeit haben wir mit den Unterraumkorrekturverfahren einen flexiblen Rahmen zur Konvergenzanalyse von vielen iterativen Lösungsverfahren vorgestellt, wobei sowohl für additive als auch für multiplikative Unterraumkorrekturen jeweils ein Konvergenzresultat bewiesen wurde, das auf zwei beziehungsweise drei Annahmen beruht. Das Verfahren der hierarchischen Basen haben wir für die Konvergenzanalyse als Unterraumkorrekturverfahren aufgefasst und die Annahmen der allgemeinen Unterraumkorrekturverfahren für das jeweilige Konvergenzresultat für das Poisson-Problem bewiesen. Zudem konnten wir das theoretische Konvergenzverhalten für eine Implementierung des Verfahrens und anschließenden numerischen Experimenten zeigen. Außerdem haben wir das Verfahren der hierarchischen Basen als Vorkonditionierer für das Verfahren der konjugierten Gradienten eingesetzt und für numerische Experimente das theoretische Konvergenzverhalten nachgewiesen.

Die Einschränkung der Konvergenzaussage auf partielle Differentialgleichungen aus dem ein- und zweidimensionalen Raum stellt einen Nachteil des Verfahrens der hierarchischen Basen dar. Im Gegenzug sind die Voraussetzungen an die zugrundeliegenden Triangulierungen in der Praxis relativ einfach zu erfüllen, insbesondere sind sogar adaptiv verfeinerte Gitter zulässig. Die Verwendung des Verfahrens der hierarchischen Basen als Vorkonditionierer für das Verfahren der konjugierten Gradienten besitzt zudem eine bessere Konvergenzgeschwindigkeit als das eigenständige Verfahren der hierarchischen Basen.

Da in dieser Arbeit ausschließlich das Jacobi-Verfahren als approximatives Lösungsverfahren auf den Unterräumen eingesetzt wurde, wäre es interessant, die Verwendung anderer iterativer Lösungsverfahren, wie das Gauß-Seidel-Verfahren, als approximatives Lösungsverfahren auf den Unterräumen zu untersuchen. Aufgrund der Einschränkung der Konvergenz des Verfahrens der hierarchischen Basen auf den ein- oder zweidimensionalen Raum wäre ein Vergleich dieses Verfahrens mit dem Verfahren von Bramble, Pasciak und Xu, das in [BPX90] eingeführt wurde und ebenfalls als Unterraumkorrekturverfahren aufgefasst werden kann, erwähnenswert. Ein Vergleich ist beispielsweise in [Yse90] zu finden, der in dieser Arbeit aber nicht betrachtet wurde. Außerdem wäre die

Anwendung des Verfahrens der hierarchischen Basen auf andere partielle Differentialgleichungen erstrebenswert.

Das Verfahren der hierarchischen Basen und vor allem die Verwendung dieses Verfahrens als Vorkonditionierer für das Verfahren der konjugierten Gradienten ist insgesamt ein effizientes Verfahren zur Lösung linearer Gleichungssysteme, die aus der Diskretisierung einer elliptischen partiellen Differentialgleichung im zweidimensionalen Raum durch die Methode der Finiten Elemente entstehen.

Literatur

[AF11] Ilka Agricola und Thomas Friedrich. *Elementargeometrie. Fachwissen für Studium und Mathematikunterricht*. 3. Aufl. Vieweg+Teubner, 2011.

[Alt12] Hans W. Alt. *Lineare Funktionalanalysis. Eine anwendungsorientierte Einführung*. 6. Aufl. Springer-Lehrbuch Masterclass. Springer Berlin Heidelberg, 2012.

[BDY88] Randolph E. Bank, Todd F. Dupont und Harry Yserentant. „The hierarchical basis multigrid method“. In: *Numerische Mathematik* 52.4 (1988), S. 427–458.

[Bey95] Jürgen Bey. „Tetrahedral grid refinement“. In: *Computing* 55.4 (1995), S. 355–378.

[BPX90] James H. Bramble, Joseph E. Pasciak und Jinchao Xu. „Parallel multilevel preconditioners“. In: *Mathematics of Computation* 55.191 (1990), S. 1–22.

[Bra07] Dietrich Braess. *Finite Elemente. Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. 4. Aufl. Springer Berlin Heidelberg, 2007.

[BSW83] Randolph E. Bank, Andrew H. Sherman und Alan Weiser. „Some refinement algorithms and data structures for regular local mesh refinement“. In: *Scientific Computing. Applications of mathematics and computing to the physical sciences*. Hrsg. von Robert S. Stepleman. Imacs Transactions on Scientific Computation 1. North-Holland, 1983, S. 3–17.

[Dob10] Manfred Dobrowolski. *Angewandte Funktionalanalysis. Funktionalanalysis, Sobolev-Räume und elliptische Differentialgleichungen*. 2. Aufl. Springer-Lehrbuch Masterclass. Springer Berlin Heidelberg, 2010.

[DW11] Peter Deuflhard und Martin Weiser. *Numerische Mathematik 3. Adaptive Lösung partieller Differentialgleichungen*. De Gruyter Lehrbuch Series. De Gruyter, 2011.

[Eva10] Lawrence C. Evans. *Partial differential equations*. 2. Aufl. Bd. 19. Graduate Studies in Mathematics. American Mathematical Society, 2010.

[Fab08] Georg Faber. „Über stetige Funktionen“. In: *Mathematische Annalen* 66.1 (1908), S. 81–94.

[Hac85] Wolfgang Hackbusch. *Multi-grid methods and applications*. Bd. 4. Springer Series in Computational Mathematics. Springer Berlin, 1985.

[Hac93] Wolfgang Hackbusch. *Iterative Lösung großer schwachbesetzter Gleichungssysteme*. 2. Aufl. Leitfäden der angewandten Mathematik und Mechanik; Band 69. Teubner, 1993.

[Hac96] Wolfgang Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen*. 2. Aufl. Teubner-Studienbücher: Mathematik. Teubner, 1996.

[Mit91] William F. Mitchell. „Adaptive refinement for arbitrary finite-element spaces with hierarchical bases“. In: *Journal of Computational and Applied Mathematics* 36.1 (1991), S. 65–78.

[Riv84] María C. Rivara. „Algorithms for refining triangular grids suitable for adaptive and multigrid techniques“. In: *International Journal for Numerical Methods in Engineering* 20.4 (1984), S. 745–756.

[SBG96] Barry F. Smith, Petter E. Bjørstad und William D. Gropp. *Domain decomposition. Parallel multilevel methods for elliptic partial differential equations*. Cambridge University Press, 1996.

[Ste03] Olaf Steinbach. *Numerische Näherungsverfahren für elliptische Randwertprobleme. Finite Elemente und Randelemente*. Advances in Numerical Mathematics. Vieweg+Teubner Verlag, 2003.

[Wer07] Dirk Werner. *Funktionalanalysis*. 6. Aufl. Springer-Lehrbuch. Springer Berlin Heidelberg, 2007.

[Xu92] Jinchao Xu. „Iterative Methods by Space Decomposition and Subspace Correction“. In: *SIAM Review* 34.4 (1992), S. 581–613.

[Yse85] Harry Yserentant. „Über die Maximumnormkonvergenz der Methode der finiten Elemente bei geringsten Regularitätsvoraussetzungen“. In: *Journal of Applied Mathematics and Mechanics* 65.2 (1985), S. 91–100.

[Yse86a] Harry Yserentant. „Hierarchical bases give conjugate gradient type methods a multigrid speed of convergence“. In: *Applied Mathematics and Computation* 19.1–4 (1986), S. 347–358.

[Yse86b] Harry Yserentant. „On the multi-level splitting of finite element spaces“. In: *Numerische Mathematik* 49.4 (1986), S. 379–412.

[Yse90] Harry Yserentant. „Two preconditioners based on the multi-level splitting of finite element spaces“. In: *Numerische Mathematik* 58.1 (1990), S. 163–184.

[Yse92] Harry Yserentant. „Hierarchical Bases“. In: *ICIAM 91. Proceedings of the Second International Conference on Industrial and Applied Mathematics*. Hrsg. von Robert E. O'Malley. Proceedings in Applied Mathematics Series 61. SIAM, 1992, S. 256–276.

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

[ZKB82] Olgierd C. Zienkiewicz, Don W. Kelly und Ivo M. Babuška. „Hierarchical finite element approaches, error estimates and adaptive refinement“. In: *The Mathematics of Finite Elemente and Applications IV. MAFELAP 1981*. Hrsg. von John R. Whiteman. Academic Press, 1982, S. 313–346.

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 12. März 2014