Randkonzentrierte und adaptive hp-FEM

Von der Fakultät für Mathematik der Technischen Universität Chemnitz genehmigte

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium

(Dr. rer. nat.)

Vorgelegt von

Dipl.-Math.-techn. Tino Eibner

geboren am 25. April 1978 in Karl-Marx-Stadt

Eingereicht am	:	30.01.2006
Betreuer	:	Prof. Dr.(USA) J.M. Melenk
Gutachter	:	Prof. Dr.(USA) J.M. Melenk (Technische Universität Wien) Prof. Dr. A. Meyer (Technische Universität Chemnitz) PD Dr. B.N. Khoromskij (MPI Leipzig)
Verteidigt am	:	19.06.2006

Inhaltsverzeichnis

Liste der benutzen Symbole

Die folgende Liste enthält eine kurze Aufstellung der für die Arbeit wichtigsten Notationen.

Allgemeines

Symbol	kurze Erläuterung	Seite
$\mathbb{N}, \mathbb{R}, \mathbb{C}$	natürliche, reelle, komplexe Zahlen	—
$\langle u, v\rangle$, $u \cdot v$	$\mathbb{R}^d$-Skalarprodukt für $u, v \in \mathbb{R}^d$	—
C	generische Konstante	—
$\#M$	Kardinalität der Menge M	—
A_{ij}, $A[i][j]$	Eintrag an der ij-Position der Matrix A	—

Gebiete, Räume, Normen

Symbol	kurze Erläuterung	Seite
Ω	Gebiet des $\mathbb{R}^d$	9
$\Gamma = \partial\Omega$	Rand von Ω	9
$\overline{\Omega}$	Abschluss von Ω	9
$\Omega_0 \subset\subset \Omega$	Ω_0 ist kompakt enthalten in Ω	9
$r, r(x)$	Abstand von x zu $\partial\Omega$	62
$\chi_\Omega(x)$	charakteristische Funktion	66
	$\chi_\Omega(x) = \begin{cases} 1: & x \in \Omega \\ 0: & \text{sonst} \end{cases}$	
$C^k(\Omega)$, $C^\infty(\Omega)$, $C^k(\overline{\Omega})$, $C_0(\Omega)$, $C_0^\infty(\Omega)$	Räume stetiger Funktionen	10
$L^p(\Omega)$	L^p-Räume	10
$L^p(\Omega, \mathbb{R}^d)$, $L^p(\Omega, \mathbb{R}^{d\times d})$	Vektoren bzw. Matrizen, deren Einträge L^p-Funktionen sind	18
$\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$	Multiindex	9
$D^\alpha u$	α-te schwache Ableitung	11
$W^{k,p}(\Omega)$,$W_0^{k,p}(\Omega)$	Sobolev-Räume	11
$H^k(\Omega)$, $H_0^k(\Omega)$, $H_D^k(\Omega)$	Sobolev-Räume	11, 18
$\|u\|_{W^{k,p}(\Omega)}$, $\|u\|_{W^{k,p}(\Omega)}$,	Norm, Seminorm im Raum $W^{k,p}(\Omega)$	11
$\|u\|_{H^k(\Omega)}$, $\|u\|_{H^k(\Omega)}$,	Norm, Seminorm im Raum $H^k(\Omega)$	11
$(u, v)_{H^k(\Omega)}$	Skalarprodukt im Raum $H^k(\Omega)$	11

Vernetzungen

Symbol	kurze Erläuterung	Seite
$\mathcal{T}^2, \mathcal{T}^3$	Referenzdreieck, Referenztetraeder	20
$\mathcal{Q}^2, \mathcal{Q}^3$	Einheitsquadrat, Einheitswürfel	20
K, F_K	Element, Elementtransformation	20
h_K	Durchmesser des Elements K	22
$p(K)$	Polynomgradverteilung auf K	22
p_K	innerer Polynomgrad von K	22
$\mathcal{N}$	$\mathcal{N} = \{(K, F_K)\}$ Vernetzung	21
$\mathcal{T}(\mathcal{N})$	Menge aller Elemente von $\mathcal{N}$	21
$e(\mathcal{N})$	Menge aller Kanten von $\mathcal{N}$	21
$f(\mathcal{N})$	Menge aller Seitenflächen von $\mathcal{N}$	21
$v(\mathcal{N})$	Menge aller Knoten von $\mathcal{N}$	21
$p(\mathcal{N})$	Polynomgradverteilung zu $\mathcal{N}$	22

Dualräume, FE-Räume, Spur

Symbol	kurze Erläuterung	Seite
$\mathcal{V}^*$	Dualraum zum Raum $\mathcal{V}$	—
$H^{-k}(\Omega)$	Dualraum zu $H_0^k(\Omega)$	11
$\langle u, v \rangle_{\mathcal{V}^* \times \mathcal{V}}$	Dualitätsprodukt $u(v)$	—
$\gamma_0 u$, $\gamma_1 u$	Spuroperatoren	12
$S^p(\Omega, \mathcal{N})$, $S_0^p(\Omega, \mathcal{N})$, $S_D^p(\Omega, \mathcal{N})$	FE-Räume	23
$Y_D^p(\Omega, \mathcal{N})$	Einschränkung von $S^p(\Omega, \mathcal{N})$ auf $\Gamma_D \subset \partial\Omega$	23
$\Pi_{p(K)}(\mathcal{T}^d)$,	Polynomraum auf $\mathcal{T}^d$ bezüglich $p(K)$	23

Polynome, Interpolation

Symbol	kurze Erläuterung	Seite
$\mathcal{P}_p(\Omega)$	Raum der Polynome mit Gesamtgrad $\leq p$ auf Ω	23
$\mathcal{Q}_p(\Omega)$	Raum der Polynome mit maximalem Polynomgrad $\leq p$ für jede Variable	—
$P_n^{(\alpha,\beta)}$	Jacobi-Polynome	12
L_n	Legendre-Polynome	12
l_i	Lagrange-Interpolationspolynom	28
i_p	Gauß-Lobatto-Interpolationsoperator	68

Kapitel 1

Einleitung

1.1 Zur Finiten-Element-Methode

Zahlreiche Probleme aus den Naturwissenschaften, der Technik und der mathematischen Physik werden als elliptische Differentialgleichung formuliert. Angefangen im Automobil- und Flugzeugbau bis hin zur Geophysik und Medizintechnik überall treffen wir auf partielle Differentialgleichungen. In der Regel ist die Herleitung einer exakten Lösung nur für sehr einfache Beispiele möglich. Für praxisrelevante Probleme hingegen wird der Einsatz von Näherungsverfahren unumgänglich. Neben der Randelemente-Methode (kurz BEM) (siehe z.B. [Hac95, SS04, STW90, Ste03]), der Finiten-Volumen-Methode (FVM) (siehe z.B. [CMM91, EGH00]) und der Finiten-Differenzen-Methode (FDM) (siehe z.B. [GR92, Hac96, Sam84]) ist die Finite-Element-Methode (FEM) (siehe z.B. [Bra97, Cia76, GR92, JL01, QV97]) eines der wichtigsten Verfahren zur approximativen Berechnung von Lösungen partieller Differentialgleichungen.

Die Finite-Element-Methode wurde ursprünglich in den fünfziger Jahren von Ingenieuren des Flugzeugbaus entwickelt und umfasst heute Anwendungen in Festigkeits- und Stabilitätsuntersuchungen, Strömungs- und Magnetfeldberechnungen sowie Crashsimulationen, um nur einige zu nennen. Die Leistungsfähigkeit der FEM liegt vor allem darin begründet, selbst für komplizierte Geometrien auf einfache Art und Weise Näherungslösungen zu berechnen und deren Approximationsgüte abschätzen zu können.

Es existieren drei Versionen der Finiten-Element-Methode: die h-Version, die p-Version und die hp-Version. Die älteste und wohl noch immer am weitesten verbreitete Methode ist die h-Version. Hierbei wird das zu Grunde liegende Gebiet mit einem Netz aus geometrischen Basisobjekten, wie zum Beispiel Dreiecken und Vierecken im 2-Dimensionalen bzw. Tetraedern, Pyramiden, Prismen und Quadern im 3-Dimensionalen, überzogen, um auf diesen, den so genannten finiten Elementen, die Lösung durch transformierte Polynome niedrigen Grades ($p = 1, 2, 3$) zu approximieren. Die Konvergenz der Näherungslösungen gegen die exakte Lösung wird in der h-FEM mittels immer feiner werdender Netz erreicht.

Einen gänzlich anderen Weg beschreiten die p-Version der Finiten-Element-Methode [BD81, BS94, BG00, BSK81] und die nahe verwandte Spektralmethode [BM92, BM97, Ors80]. [BSK81] zeigt, dass auf einem fest vorgegebenen Gitter die Konvergenz der Näherungslösungen gegen die exakte Lösung auch durch Erhöhen des Polynomgrades erreicht werden kann. Vielmehr kann sogar gezeigt werden, dass die asymptotische Konvergenzrate der p-Version nicht schlechter als die der h-Version ist. Oft ist die Konvergenzrate der p-Version, gemessen im Verhältnis von Fehler gegenüber Freiheitsgraden, sogar deutlich besser als die der h-Version. Speziell für

auf einer Umgebung des Gebiets Ω analytische Lösungen erhalten wir mittels p-FEM exponentielle Konvergenz, wohingegen eine h-FEM stets nur zu algebraischen Konvergenzraten führt.
Die dritte und jüngste Version der Finiten-Element-Methode ist schließlich die hp-FEM [BG86a, BG86b, BG86c, BG86d, BG88, KS99, Mel02, Sch98]. Die hp-FEM stellt eine Verschmelzung von klassischer h-FEM und p-FEM dar und besitzt somit das Potential, die Vorzüge beider Methoden in sich zu vereinen. Wesentlich für die hp-FEM ist eine lokale Gitterverfeinerung in Kombination mit einer variablen Approximationsordnung. Natürlich sind für den erfolgreichen Einsatz der hp-FEM die lokale Gittergröße und Approximationsordnung aneinander gekoppelt und hängen vor allem von der Regularität der Lösung ab. In Regionen mit lokal glatter Lösung sollte das Gitter verhältnismäßig grob und der verwendete Polynomgrad hoch sein, wohingegen bei einer lokal weniger glatten Lösung das Gitter fein und der Polynomgrad niedrig zu wählen ist.
Für eine große Klasse von Aufgaben kann gezeigt werden, dass auf geeigneten Netzen mit geeigneter Polynomgradverteilung die hp-FEM exponentielle Konvergenz erreicht. Unter anderem zeigen die Arbeiten von Babuška und Guo [BG86c, BG88], dass für Randwertaufgaben mit stückweise analytischen Koeffizienten und rechter Seite sowie stückweise analytischen Randbedingungen und einem Rand bestehend aus endlich vielen analytischen Kurvenbögen die hp-FEM bei Verwendung von geometrischen Gittern in Verbindung mit einer linearen Polynomgradverteilung exponentiell konvergiert. Im Vergleich zur reinen h- und p-Methode kann die hp-FEM daher im Allgemeinen mit deutlich weniger Freiheitsgraden eine wesentlich bessere Approximation der exakten Lösung erzielen.

1.2 Gliederung der Arbeit

Die Arbeit umfasst fünf Kapitel, welche zum Ziel haben, verschiedene sowohl theoretische als auch implementatorische Aspekte der hp-FEM genauer zu untersuchen.
Das sich direkt an die Einleitung anschließende Kapitel 2 trägt grundlegende und zum Verständnis zwingend notwendige Definitionen und Beziehungen kurz und knapp zusammen.
In Kapitel 3 betrachten wir die hp-FEM im Allgemeinen und bereiten damit die Grundlage für die nachfolgenden Kapitel. Neben der Überführung einer elliptischen Randwertaufgabe in ein hp-FE-Problem sowie Existenz- und Eindeutigkeitsaussagen der Lösung, gehen wir in Kapitel 3 auch verstärkt auf implementatorische Aspekte ein. Insbesondere beschäftigen wir uns hierbei mit dem sehr wichtigen Punkt des effizienten Generierens der Steifigkeitsmatrix bzw. der Alternative einer „on the fly“ Matrix-Vektor-Multiplikation. Im Gegensatz zur h-FEM kann, insbesondere für höhere Polynomgrade, das Aufstellen der Steifigkeitsmatrix in der p- und hp-FEM sehr rechenaufwändig werden. Wir werden hierfür sowohl bereits bekannte Algorithmen (siehe [KS99, Ors80]) vorstellen als auch den in [MGS01] entwickelten Spektral-Galerkin-Algorithmus auf Netze bestehend aus Dreiecks- oder Tetraederelementen verallgemeinern. Anhand numerischer Testrechnungen werden wir die Vor- und Nachteile der verschiedenen Algorithmen verdeutlichen und aufzeigen, welche Methode unter welchen Umständen die geeignetste ist.
In Kapitel 4 beschäftigen wir uns mit der erstmalig in [KM03] vorgestellten randkonzentrierten Finiten-Element-Methode. Diese spezielle Version der hp-FEM ist, durch ihre a priori Vorgabe eines zum gesamten Rand hin stark verfeinerten Netzes, vor allem für Randwertprobleme mit komplizierter Randgeometrie, Randbedingungen geringer Regularität oder allgemein für Probleme, die aus verschiedenen anderen Gründen hochauflösende Gitter am Rand benötigen,

geeignet. Im ersten Abschnitt von Kapitel 4 werden wir kurz auf die aus [KM03] stammenden grundlegenden Ideen und Fakten eingehen. Anschließend werden wir eine für die randkonzentrierte FEM verbesserte Konvergenz im Gebietsinneren beweisen und im letzten Teil des Kapitels zwei auf der Additiv-Schwarz-Methode basierende Multilevel-Vorkonditionierer vorstellen. Unsere theoretischen Ergebnisse werden wir durch zahlreiche numerische Beispiele verifizieren.
In Kapitel 5 wenden wir uns der adaptiven *hp*-FEM zu. Wir beweisen, dass eine auf dem Referenztetraeder definierte L^2-Funktion genau dann analytisch auf einer Umgebung des Referenztetraeders ist, wenn ihre Zerlegungskoeffizienten bezüglich einer geeignet gewählten L^2-Orthogonalbasis exponentiell abklingen. Zusammen mit der analogen 2D Aussage für Funktionen auf dem Referenzdreieck, liefert dies die theoretische Grundlage einer adaptiven *hp*-Strategie für Dreiecks- bzw. Tetraedernetze. Wir werden diese Strategie kurz vorstellen und sie anschließend zwei weiteren adaptiven *hp*-Verfahren zu Vergleichszwecken gegenüberstellen. Die Wirksamkeit der Strategie werden wir anhand numerischer Testrechnungen demonstrieren.

Danksagung

Die vorliegende Arbeit beinhaltet im Wesentlichen die in den letzten drei Jahren im SFB393 Projekt A13 - „Randkonzentrierte Finite-Elemente-Methoden“ erzielten Forschungsergebnisse und wäre nicht ohne die tatkräftige Unterstützung von Markus Melenk möglich gewesen. An dieser Stelle möchte ich mich daher bei Markus Melenk insbesondere für die sehr gute Betreuung sowie für zahlreiche Tipps, Diskussionen und Verbesserungsvorschläge bedanken. Des Weiteren möchte ich mich auch bei all meinen Chemnitzer Kollegen für das angenehme Arbeitsklima und die moralische Unterstützung bedanken. Ein besonderer Dank geht zudem an Arnd Meyer für Hinweise und Diskussionen sowie an Sven Beuchler. Zu guter Letzt sei zudem noch der DFG für die finanzielle Unterstützung des Projekts gedankt.

Kapitel 2

Grundlegende Definitionen

Dieses Kapitel enthält die wichtigsten Definitionen und Zusammenhänge, auf die wir in den späteren Kapiteln des Öfteren zurückgreifen werden. Alle hier aufgeführten Aussagen sind weithin bekannte Sachverhalte, die ohne Probleme der Literatur [Ada75, AF03, Bra97, GR92, KS99, QV97, Sch98] entnommen werden können.

2.1 Gebiete

Mit $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$, sind stets beschränkte Gebiete gemeint, wobei ein Gebiet eine offene und zusammenhängende Punktmenge darstellt. Für $\Omega \subset \mathbb{R}^d$ bezeichnen wir mit $\overline{\Omega}$ den Abschluss und mit $\partial\Omega$ den Rand von Ω. Die Schreibweise $\Omega_0 \subset\subset \Omega$ bedeutet, dass Ω_0 kompakt in Ω enthalten ist, d.h. $\overline{\Omega}_0 \subset \Omega$. In dieser Arbeit gehen wir stets davon aus, dass Ω ein Lipschitz-Gebiet ist:

Definition 2.1.1. Ein Gebiet $\Omega \subset \mathbb{R}^d$ heißt Lipschitz-Gebiet, wenn es ein $k \in \mathbb{N}$ und offene Mengen $U_1, \ldots, U_k \subset \mathbb{R}^d$ gibt, so dass:

1. $\partial\Omega \subset \bigcup_{i=1}^k U_i$.
2. Für alle $1 \leq i \leq k$ ist $\partial\Omega \cap U_i$ darstellbar als Graph einer Lipschitz-stetigen Funktion.

Lemma 2.1.2. *Sei Ω ein Lipschitz-Gebiet, so existiert fast überall auf $\partial\Omega$ das äußere Einheitsnormalenfeld zu Ω.*

2.2 Funktionenräume

Im Folgenden bezeichnet $\Omega \subset \mathbb{R}^d$ ein Gebiet mit $d \in \{1,2,3\}$ und $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ einen Multiindex. Wir schreiben

$$|\alpha| = \sum_{i=1}^{d} \alpha_i, \qquad D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}.$$

2.2.1 Räume stetiger Funktionen

Mit $C^k(\Omega)$, $k = 0, 1, 2, \ldots$, bezeichnen wir den Vektorraum aller auf Ω stetigen Funktionen u, für die sämtliche partielle Ableitungen $D^\alpha u$ mit $\alpha \in \mathbb{N}_0^d$, $|\alpha| \leq k$ existieren und stetig sind.

Wir schreiben auch $C(\Omega) := C^0(\Omega)$ und definieren $C^\infty(\Omega) := \cap_{m=0}^\infty C^m(\Omega)$. Für eine Funktion $u : \Omega \to \mathbb{R}/\mathbb{C}$ heißt

$$\text{supp}(u) = \overline{\{x \in \Omega \mid u(x) \neq 0\}}$$

der Träger von u und wir bezeichnen mit $C_0(\Omega)$ und $C_0^\infty(\Omega)$ die Unterräume aller Funktionen aus $C(\Omega)$ bzw. $C^\infty(\Omega)$ mit kompaktem Träger in Ω ([AF03, Abschnitt 1.26]).
Die Funktionen $u \in C^k(\Omega)$ sind nicht notwendigerweise beschränkt. Wir definieren daher (siehe [AF03, Abschnitt 1.28]):

Definition 2.2.1. Für ein Gebiet $\Omega \subset \mathbb{R}^d$ sei $C(\overline{\Omega}) \subset C(\Omega)$ der Raum der in Ω gleichmäßig stetigen und beschränkten Funktionen sowie

$$C^k(\overline{\Omega}) := \{u \in C^k(\Omega) \mid D^\alpha u \in C(\overline{\Omega}) \text{ für alle Multiindizes } \alpha \text{ mit } |\alpha| \leq k\}$$

der Raum der Funktionen mit gleichmäßig stetigen und beschränkten Ableitungen bis einschließlich k-ter Ordnung.

2.2.2 L^p-Räume

Neben den Räumen der stetigen und k-fach stetig differenzierbaren Funktionen stellen die so genannten L^p-Räume und die darauf aufbauenden Sobolev-Räume die für diese Arbeit wichtigsten Klassen von Funktionen dar. Beginnen wir mit der Definition der L^p-Räume.

Definition 2.2.2. Für ein Gebiet $\Omega \subset \mathbb{R}^d$ und $1 \leq p < \infty$ besteht der Raum $L^p(\Omega)$ aus allen auf Ω messbaren Funktionen u, für die $|u|^p$ auf Ω Lebesgue-integrierbar ist. $L^\infty(\Omega)$ ist der Raum aller auf Ω messbaren Funktionen u mit der Eigenschaft

$$\text{ess sup}_\Omega |u| := \inf_N \sup_{x \in \Omega \setminus N} |u(x)| < \infty,$$

wobei $N \subset \Omega$ alle Mengen vom Maß Null durchläuft. Die Funktionen $u \in L^\infty(\Omega)$ heißen wesentlich beschränkt. $L^p_{loc}(\Omega)$, $1 \leq p \leq \infty$, sei der Raum aller Funktionen u, für die $u \in L^p(\Omega_0)$ für alle $\Omega_0 \subset\subset \Omega$ gilt.

Streng genommen sind die $L^p(\Omega)$-Räume Räume von Äquivalenzklassen messbarer Funktionen, die sich nur auf einer Menge vom Maß Null unterscheiden. D.h. wir identifizieren zwei Funktionen u und v, falls sich diese nur auf einer Nullmenge unterscheiden. Ausgestattet mit den Normen

$$\|u\|_{L^p(\Omega)} = \left(\int_\Omega |u|^p \mathrm{d}\Omega \right)^{1/p} \quad \text{für } 1 \leq p < \infty \text{ und } \|u\|_{L^\infty(\Omega)} = \text{ess sup}_\Omega |u|$$

werden die Räume $L^p(\Omega)$, $1 \leq p \leq \infty$, zu Banach-Räumen. Mit dem inneren Produkt

$$(u, v)_{L^2(\Omega)} := \int_\Omega u \, v \mathrm{d}\Omega$$

wird der Raum $L^2(\Omega)$ sogar zu einem Hilbert-Raum. Für beschränkte Lipschitz-Gebiete $\Omega \subset \mathbb{R}^d$ gilt die Einbettung $L^p(\Omega) \subset L^1(\Omega)$ für alle $1 \leq p \leq \infty$.

2.2.3 Sobolev-Räume

Bevor wir die Sobolev-Räume definieren können, müssen wir noch den Begriff der verallgemeinerten bzw. schwachen Ableitung einführen (siehe [AF03, Abschnitt 1.62]).

Definition 2.2.3. Sei $\Omega \subset \mathbb{R}^d$ ein Gebiet, $u, v \in L^1_{loc}(\Omega)$ und $\alpha \in \mathbb{N}_0^d$ ein Multiindex. Die Funktion v heißt die α-te schwache Ableitung von u, kurz $v = D^\alpha u$, wenn für alle $\phi \in C_0^\infty(\Omega)$ gilt

$$\int_\Omega u D^\alpha \phi \mathrm{d}\Omega = (-1)^{|\alpha|} \int_\Omega v\phi \mathrm{d}\Omega.$$

Sofern sie existiert, ist die α-te schwache Ableitung eindeutig im Sinne von $L^1_{loc}(\Omega)$-Funktionen und für $u \in C^{|\alpha|}(\Omega)$ stimmen die α-te schwache Ableitung und die α-te klassische Ableitung überein. Auf Grund dieser Übereinstimmung werden wir in Zukunft nicht zwischen klassischer und verallgemeinerter Ableitung unterscheiden.
Kommen wir nun zur Definition der Sobolev-Räume.

Definition 2.2.4 (Sobolev-Räume)**.** Sei $\Omega \subset \mathbb{R}^d$ ein Gebiet. Für $k \in \mathbb{N}_0$ und $1 \le p \le \infty$ sei der Sobolev-Raum $W^{k,p}(\Omega)$ und seine Norm $\|\cdot\|_{W^{k,p}(\Omega)}$ gegeben durch:

$$W^{k,p}(\Omega) = \{u \in L^p(\Omega) \mid D^\alpha u \in L^p(\Omega) \ \forall\, |\alpha| \le k\},$$

$$\|u\|_{W^{k,p}(\Omega)} = \begin{cases} \left(\sum_{|\alpha|\le k} \|D^\alpha u\|^p_{L^p(\Omega)}\right)^{1/p} & : \ 1 \le p < \infty \\ \max_{|\alpha|\le k} \|D^\alpha u\|_{L^\infty(\Omega)} & : \ p = \infty \end{cases}.$$

Neben den Normen $\|u\|_{W^{k,p}(\Omega)}$ definieren wir auch noch Seminormen

$$|u|_{W^{k,p}(\Omega)} = \begin{cases} \left(\sum_{|\alpha|= k} \|D^\alpha u\|^p_{L^p(\Omega)}\right)^{1/p} & : \ 1 \le p < \infty \\ \max_{|\alpha|= k} \|D^\alpha u\|_{L^\infty(\Omega)} & : \ p = \infty \end{cases}.$$

Die eben eingeführten Räume $W^{k,p}(\Omega)$ sind allesamt Banach-Räume. Speziell für $p = 2$ wird mit dem inneren Produkt

$$(u, v)_{H^k(\Omega)} := \sum_{|\alpha|\le k} (D^\alpha u, D^\alpha v)_{L^2(\Omega)}$$

der Raum $W^{k,2}(\Omega)$ sogar zu einem Hilbert-Raum und wir schreiben $H^k(\Omega)$ an Stelle von $W^{k,2}(\Omega)$. In Verbindung mit Dirichlet-Randbedingungen werden wir auch die Sobolev-Räume vom Typ $W_0^{k,p}(\Omega)$ benötigen:

Definition 2.2.5. Für $k \in \mathbb{N}$ und $1 \le p \le \infty$ sei $W_0^{k,p}(\Omega)$ die Vervollständigung von $C_0^\infty(\Omega)$ bezüglich der $\|\cdot\|_{W^{k,p}(\Omega)}$ Norm. Für den Fall $p = 2$ schreiben wir wieder $H_0^k(\Omega)$ an Stelle von $W_0^{k,p}(\Omega)$.

Die Räume $W_0^{k,p}(\Omega)$ sind abgeschlossene Teilräume von $W^{k,p}(\Omega)$ und ihrerseits ebenfalls Banach- bzw. Hilbert-Räume. Mittels Interpolation ist es möglich Sobolev-Räume auch für nicht-ganzzahlige Ordnungen einzuführen. Die von uns benutzte Variante ist die so genannte K-Methode der Interpolation. Für eine genaue Definition verweisen wir an dieser Stelle jedoch auf die Bücher [BL76, Tri95, AF03, Sch98]. Sobolev-Räume H^{-s} mit $s \in \mathbb{R}_+$ verstehen wir als Dualräume zu $H_0^s(\Omega)$. D.h. $H^{-s}(\Omega) = (H_0^s(\Omega))^*$.

2.2.4 Sobolev-Räume auf dem Gebietsrand

Wollen wir die „Randwerte“ einer Funktion $u \in W^{k,p}(\Omega)$ auf $\partial\Omega$ betrachten, so stehen wir vor dem Problem, dass u im Allgemeinen nicht stetig zu sein braucht und nur bis auf Mengen vom Maß Null eindeutig definiert ist. Folgender Spursatz erlaubt es uns jedoch, die Spur einer $H^1(\Omega)$-Funktion als $L^2(\partial\Omega)$-Funktion (genauer $H^{1/2}(\partial\Omega)$) zu betrachten:

Theorem 2.2.6. *Sei $\Omega \subset \mathbb{R}^d$, $d \geq 2$, ein beschränktes Lipschitz-Gebiet. Dann gibt es eine beschränkte und eindeutig bestimmte lineare Abbildung*

$$\gamma_0 \ : \ H^1(\Omega) \to H^{1/2}(\partial\Omega) \subset L^2(\partial\Omega),$$

mit

$$\begin{aligned} \gamma_0 u &= u|_{\partial\Omega} \qquad \forall\, u \in H^1(\Omega) \cap C^0(\overline{\Omega}), \\ \|\gamma_0 u\|_{H^{1/2}(\partial\Omega)} &\leq C_\Omega \|u\|_{H^1(\Omega)} \qquad \forall\, u \in H^1(\Omega), \end{aligned}$$

wobei

$$H^{1/2}(\partial\Omega) := \{w \in L^2(\partial\Omega) \mid \exists\, v \in H^1(\Omega) \ : \ \omega = \gamma_0 v, \}$$

$$\|w\|_{H^{1/2}(\partial\Omega)} := \inf\{\|v\|_{H^1(\Omega)} \mid v \in H^1(\Omega),\ \gamma_0 v = w\}.$$

Für $u \in H^1_\Delta(\Omega) := \{u \in H^1(\Omega) \,|\, \Delta u \in L^2(\Omega)\}$ bezeichne $\gamma_1 : H^1_\Delta(\Omega) \to H^{-1/2}(\partial\Omega)$ die Normalenableitung $\gamma_1 u = \partial_n u|_{\partial\Omega}$. Δu ist hierbei im distributiven Sinn zu verstehen (siehe [Sch98, Anhang A]).

2.3 Jacobi-Polynome

Eine wichtige Klasse von Polynomen sind die Jacobi-Polynome [GR80, KS99, Sch98, Sze75]. Jacobi-Polynome sind Polynome $P_i^{(\alpha,\beta)}$ vom Grad $i = 0, 1, \ldots$, die zu vorgegebenen Parametern $\alpha, \beta > -1$ bezüglich einer Gewichtsfunktionen $\omega^{\alpha,\beta}(x) = (1-x)^\alpha (1+x)^\beta$ orthogonal im Sinne des L_2-Skalarproduktes auf $(-1, 1)$ sind. Für $P_n^{(\alpha,\beta)}(x)$ existieren mehrere äquivalente Definitionen und Darstellungsformeln:

- **Rodrigues Formel**

$$P_n^{(\alpha,\beta)}(x) = \frac{(-1)^n}{2^n n!}(1-x)^{-\alpha}(1+x)^{-\beta}\frac{d^n}{dx^n}[(1-x)^{\alpha+n}(1+x)^{\beta+n}]$$

- **3-gliedrige Rekursionsformel für orthogonale Polynome**

$$\begin{aligned} P_0^{(\alpha,\beta)}(x) &= 1, \\ P_1^{(\alpha,\beta)}(x) &= \frac{1}{2}[\alpha - \beta + (\alpha+\beta+2)x], \\ P_{n+1}^{(\alpha,\beta)}(x) &= \frac{b_n + c_n x}{a_n} P_n^{(\alpha,\beta)}(x) - \frac{d_n}{a_n} P_{n-1}^{(\alpha,\beta)}(x), \end{aligned}$$

mit

$$\begin{aligned} a_n &= 2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta), \\ b_n &= (2n+\alpha+\beta+1)(\alpha^2-\beta^2), \\ c_n &= (2n+\alpha+\beta)(2n+\alpha+\beta+1)(2n+\alpha+\beta+2), \\ d_n &= 2(n+\alpha)(n+\beta)(2n+\alpha+\beta+2). \end{aligned}$$

Als Spezialfälle sind in den Jacobi-Polynomen insbesondere enthalten:

- Tschebychev-Polynome $T_n = P_n^{(-1/2,-1/2)}$,
- Legendre-Polynome $L_n = P_n^{(0,0)}$.

Wie bereits angeführt, sind Jacobi-Polynome orthogonal auf $(-1,1)$ bezüglich der Gewichtsfunktion $\omega^{\alpha,\beta}(x) = (1-x)^\alpha(1+x)^\beta$. Genauer gilt:

$$\int_{-1}^{1} \omega^{\alpha,\beta}(x) P_n^{(\alpha,\beta)}(x) P_m^{(\alpha,\beta)}(x)\mathrm{d}x = \begin{cases} 0 & n \neq m \\ \frac{2^{\alpha+\beta+1}}{2n+\alpha+\beta+1}\frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{n!\Gamma(n+\alpha+\beta+1)} & n = m \end{cases}.$$

Definieren wir die $L^2_{\alpha,\beta}(-1,1)$ Norm durch

$$\|u\|^2_{L^2_{\alpha,\beta}(-1,1)} = \int_{-1}^{1} \omega^{\alpha,\beta}(x)|u|^2\mathrm{d}x$$

und bezeichnen mit $L^2_{\alpha,\beta}(-1,1)$ den Hilbert-Raum aller auf $(-1,1)$ messbaren Funktionen, für die die $L^2_{\alpha,\beta}(-1,1)$-Norm existiert und endlich ist. Dann gilt: Jedes $u \in L^2_{\alpha,\beta}(-1,1)$ kann entwickelt werden als

$$u(x) = \sum_{n=0}^{\infty} a_n P_n^{(\alpha,\beta)}(x)$$

mit

$$a_n = \frac{2n+\alpha+\beta+1}{2^{\alpha+\beta+1}} \frac{n!\Gamma(n+\alpha+\beta+1)}{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)} \int_{-1}^{1} \omega^{\alpha,\beta}(x) P_n^{\alpha,\beta}(x) u(x)\mathrm{d}x$$

und

$$\lim_{N\to\infty} \|u - \sum_{n=0}^{N} a_n P_n^{\alpha,\beta}(x)\|_{L^2_{\alpha,\beta}(-1,1)} = 0.$$

2.4 Die Gauß-Lobatto-Jacobi-Quadratur

Die GLJ-Quadraturregeln (siehe z.B. [KS99, Sze75]) umfassen Quadraturen vom Gauß-Typ bezüglich der Gewichtsfunktion $\omega^{(\alpha,\beta)}$ (siehe Abschnitt zu Jacobi-Polynomen), bei der die beiden Intervallendpunkte zur Stützstellenmenge gehören. Speziell für diese Arbeit benötigen wir Integrationsregeln für $\alpha \in \mathbb{N}_0$ und $\beta = 0$. D.h. wir approximieren

$$\int_{-1}^{1} (1-x)^\alpha f(x)\mathrm{d}x \approx \sum_{i=0}^{n} \omega_i f(x_i) =: \mathrm{GLJ}_{(\alpha,n)}(f).$$

Die Stützstellen x_i sind hierbei die Nullstellen von $(1-x^2)P_{n-1}^{(1+\alpha,1)}$ und die zugehörigen Gewichte ω_i berechnen sich zu [KS99, Anhang B]:

$$\omega_i = \begin{cases} \frac{2^{\alpha+1}}{n(n+\alpha+1)\left(P_n^{(\alpha,0)}(x_i)\right)^2} & : \quad i = 0, \ldots, n-1 \\ \frac{(1+\alpha)2^{\alpha+1}}{n(n+\alpha+1)\left(P_n^{(\alpha,0)}(x_i)\right)^2} & : \quad i = n \end{cases} .$$

Es gilt [KS99, Anhang B]:

- Die Nullstellen $\{x_i \mid i = 0, \ldots, n\}$ von $(1-x^2)P_{n-1}^{(1+\alpha,1)}$ sind reell und paarweise verschieden mit $-1 = x_0 < x_1 < \ldots < x_{n-1} < x_n = 1$.
- Die Gauß-Lobatto-Jacobi-Quadratur ist exakt, falls f ein Polynom vom Grade kleiner oder gleich $2n-1$ ist.
- Für genügend oft differenzierbare Funktionen f gilt die Fehlerabschätzung:

$$\begin{aligned} R_n(f) &= \left| \int_{-1}^{1} (1-x)^\alpha f(x) \mathrm{d}x - \sum_{i=0}^{n} \omega_i f(x_i) \right| \\ &\leq \frac{2^{2n+\alpha+1}(n-1)!n!(n+\alpha)!(n+\alpha+1)!}{(2n-1)!(2n+\alpha+1)[(2n+\alpha)!]^2} \sup_{x\in[-1,1]} \left| f^{(2n)}(x) \right| \end{aligned}$$

  und aus den Abschätzungen

$$\frac{(n-1)!n!}{(2n-1)!} \leq \left(\frac{1}{2}\right)^{n-1}, \quad \frac{(n+\alpha+1)!(n+\alpha)!}{(2n+\alpha+1)[(2n+\alpha)!]^2} \leq \left(\frac{1}{n+\alpha+1}\right)^{2n}$$

  folgt

$$R_n(f) \leq 2^{\alpha+2} \left(\frac{\sqrt{2}}{(n+\alpha+1)}\right)^{2n} \sup_{x\in[-1,1]} \left| f^{(2n)}(x) \right| .$$

Lemma 2.4.1. *Für beliebiges $\alpha > -1$ ist die Gauß-Lobatto-Jacobi-Quadratur* $\mathrm{GLJ}_{(\alpha,n)}$ *mit $n \geq 1$ auf dem Polynomraum $\mathcal{P}_n(-1,1)$ äquivalent zur gewichteten $L^2_{\alpha,0}(-1,1)$-Norm:*

$$\|P\|^2_{L^2_{\alpha,0}(-1,1)} \leq \mathrm{GLJ}_{(\alpha,n)}(P^2) \leq \left(2 + \frac{\alpha+1}{n}\right) \|P\|^2_{L^2_{\alpha,0}(-1,1)} \qquad \forall P \in \mathcal{P}_n(-1,1).$$

Beweis. Sei $P \in \mathcal{P}_n$ ein Polynom vom Grad n und $\alpha > -1$. Wir zerlegen P bezüglich einer Basis aus Jacobi-Polynomen

$$P = \sum_{i=0}^{n} a_i P_i^{(\alpha,0)}$$

und aus den Orthogonalitätseigenschaften der $P_i^{(\alpha,0)}$ folgt:

$$\|P\|^2_{L^2_{\alpha,0}(-1,1)} = \sum_{i=0}^{n} a_i^2 \left\| P_i^{(\alpha,0)} \right\|^2_{L^2_{\alpha,0}(-1,1)} = \sum_{i=0}^{n} \frac{2^{\alpha+1}}{2i+\alpha+1} a_i^2. \tag{2.1}$$

Des Weiteren, auf Grund der Exaktheit der Quadraturformel, gilt mit $\tilde{P} := \sum_{i=0}^{n-1} a_i P_i^{(\alpha,0)}$:

$$\begin{aligned}\mathrm{GLJ}_{(\alpha,n)}(P^2) &= \mathrm{GLJ}_{(\alpha,n)}\left(\tilde{P}^2 + 2a_n P_n^{(\alpha,0)}\tilde{P} \quad + \left(a_n P_n^{(\alpha,0)}\right)^2\right) \\ &= \|\tilde{P}\|^2_{L^2_{\alpha,0}(-1,1)} + 2a_n \int_{-1}^{1} \omega^{\alpha,0} P_n^{(\alpha,0)}\tilde{P}\mathrm{d}x + \mathrm{GLJ}_{(\alpha,n)}\left(a_n P_n^{(\alpha,0)}\right)^2.\end{aligned}$$

Nach Definition von $\mathrm{GLJ}_{(\alpha,n)}$ gilt:

$$\begin{aligned}\mathrm{GLJ}_{(\alpha,n)}\left(a_n P_n^{(\alpha,0)}\right)^2 &= a_n^2 \sum_{i=0}^{n} \omega_i \left(P_n^{(\alpha,0)}(x_i)\right)^2 \\ &= \frac{a_n^2(1+\alpha)2^{\alpha+1}}{n(n+\alpha+1)} + a_n^2 \sum_{i=0}^{n-1} \frac{2^{\alpha+1}}{n(n+\alpha+1)} = a_n^2 \frac{2^{\alpha+1}}{n}\end{aligned}$$

und wir erhalten

$$\mathrm{GLJ}_{(\alpha,n)}(P^2) = \|\tilde{P}\|^2_{L^2_{\alpha,0}(-1,1)} + 2a_n \int_{-1}^{1} \omega^{\alpha,0} P_n^{(\alpha,0)}\tilde{P}\mathrm{d}x + a_n^2 \frac{2^{\alpha+1}}{n}.$$

Aus

$$\frac{2^{\alpha+1}}{n} = \frac{2^{\alpha+1}(2n+\alpha+1)}{n2^{\alpha+1}} \|P_n^{(\alpha,0)}\|^2_{L^2_{\alpha,0}(-1,1)} = \left(2 + \frac{\alpha+1}{n}\right) \|P_n^{(\alpha,0)}\|^2_{L^2_{\alpha,0}(-1,1)}$$

folgt schließlich

$$\mathrm{GLJ}_{(\alpha,n)}(P^2) = \|\tilde{P} + a_n P_n^{(\alpha,0)}\|^2_{L^2_{\alpha,0}(-1,1)} + a_n^2 \left(1 + \frac{\alpha+1}{n}\right) \|P_n^{(\alpha,0)}\|^2_{L^2_{\alpha,0}(-1,1)}$$

und da $P = \tilde{P} + a_n P_n^{(\alpha,0)}$, ist die untere Abschätzung $\|P\|^2_{L^2_{\alpha,0}(-1,1)} \leq \mathrm{GLJ}_{(\alpha,n)}(P^2)$ nun offensichtlich. Die obere Abschätzung ergibt sich in Verbindung mit (2.1):

$$\begin{aligned}\mathrm{GLJ}_{(\alpha,n)}(P^2) &= \|P\|^2_{L^2_{\alpha,0}(-1,1)} + a_n^2 \left(1 + \frac{\alpha+1}{n}\right) \|P_n^{(\alpha,0)}\|^2_{L^2_{\alpha,0}(-1,1)} \\ &\leq \|P\|^2_{L^2_{\alpha,0}(-1,1)} + \left(1 + \frac{\alpha+1}{n}\right) \|P\|^2_{L^2_{\alpha,0}(-1,1)}.\end{aligned}$$

□

Kapitel 3

hp-FEM

In diesem Kapitel wollen wir die Grundzüge der *hp*-FEM im 2- und 3-Dimensionalen vorstellen und damit sowohl eine Basis für das spätere Kapitel zur randkonzentrierten Finiten-Element-Methode als auch für das Kapitel über adaptive *hp*-Strategien bereitstellen. Neben den Grundlagen wie klassische und schwache Formulierung von Randwertaufgaben, Existenz und Eindeutigkeit einer Lösung sowie dem Überführen in ein diskretes Problem - werden wir in diesem Kapitel speziell auch einige algorithmische und implementatorische Aspekte in den Vordergrund rücken. Insbesondere setzen wir uns hierbei mit dem effizienten Generieren der Steifigkeitsmatrix - beziehungsweise der Möglichkeit einer „on the fly“ Matrix-Vektor-Multiplikation auseinander.

Das allgemein übliche Vorgehen beim Aufstellen der Steifigkeitsmatrix ist, die Steifigkeitsmatrix aus lokalen, für jedes Element einer Vernetzung einzeln zu berechnenden Elementsteifigkeitsmatrizen zu assemblieren. Betrachten wir hierbei ein Element mit zugeordnetem Polynomgrad p, so besitzt im d-dimensionalen Fall der allereinfachste Algorithmus zum Aufstellen der lokalen Elementsteifigkeitsmatrix mittels numerischer Quadratur (siehe Algorithmus 3.6.12) eine Laufzeit von $O(p^{3d})$ bei $O(p^{2d})$ Matrixelementen und $O(p^d)$ Quadraturpunkten. Gegenüber der h-FEM erscheint das Aufstellen der Steifigkeitsmatrix in der p- und *hp*-FEM bei wachsendem Polynomgrad daher sehr rechenintensiv. Ein entscheidender Punkt ist jedoch, dass unter gewissen Zusatzvoraussetzungen an die auf den Elementen definierten Formfunktionen die Komplexität für das Aufstellen der Elementsteifigkeitsmatrizen signifikant reduziert werden kann.

Typischerweise werden sowohl alle zum Aufstellen der Elementsteifigkeitsmatrix notwendigen Quadraturen als auch die Definition der Elementformfunktionen mittels einer Transformation auf ein Referenzelement bewerkstelligt. Als Erster zeigte ([Ors80]), dass, falls sowohl das Referenzelemente als auch die darauf definierten Formfunktionen eine Tensorproduktstruktur besitzen, d.h. insbesondere bei Verwendung von Rechtecks- und Quaderelementen mit geeigneten Formfunktionen, die Laufzeit für das Aufstellen der Elementsteifigkeitsmatrix mittels Summenfaktorisierung auf $O(p^{2d+1})$ reduziert werden kann. Später gelang Karniadakis und Sherwin [SK95, SK96, KS99] mittels Duffy-Transformation die Verallgemeinerung dieser Idee auch auf Dreiecks- und Tetraederelemente.

Die durch Summenfaktorisierung eingesparte Rechenzeit gegenüber dem Standardalgorithmus ist, speziell bei hohem Polynomgrad, enorm (siehe Abschnitt 3.11). Nichtsdestotrotz werden jedoch noch immer $O(p^{2d+1})$ Rechenoperationen für lediglich $O(p^{2d})$ Matrixeinträge benötigt und es stellt sich die Frage, ob nicht auch eine Laufzeit von $O(p^{2d})$ erreicht werden

kann. In [MGS01] wurde diese Frage für Rechtecks- und Quaderelemente beantwortet. Der dort vorgestellte Spektral-Galerkin-Algorithmus generiert mittels speziell angepassten Formfunktionen die Elementsteifigkeitsmatrix mit einer Rechenzeit von $O(p^{2d})$ und führt somit zu einer weiteren Rechenzeiteinsparungen gegenüber der Summenfaktorisierung. Noch offen ist die Verallgemeinerung dieses Spektral-Galerkin-Algorithmus auf Dreiecks- und Tetraederelemente sowie die Frage nach der hierbei tatsächlich erzielbaren Rechenzeitersparnis gegenüber der Summenfaktorisierung.
Im Verlaufe des nun folgenden Kapitels werden wir den Spektral-Galerkin-Algorithmus auf Dreiecks- und Tetraederelemente verallgemeinern, auf Vor- und Nachteile der verschiedenen Algorithmen eingehen sowie, aufbauend auf den Ideen der Summenfaktorisierung und des Spektral-Galerkin-Algorithmus, effiziente Algorithmen für eine „on the fly“ Matrix-Vektor-Multiplikation konstruieren. Anhand von Testrechnungen werden wir zudem die konkreten Rechenzeiten der einzelnen Algorithmen genauer betrachten und miteinander vergleichen. [1]

3.1 Grundlagen

Sei $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, ein beschränktes Gebiet mit $\Gamma = \partial\Omega = \overline{\Gamma_D \cup \Gamma_N}$, wobei $\Gamma_D \cap \Gamma_N = \emptyset$ und die Mengen Γ_D, Γ_N entweder leer sind oder sich aus endlich vielen offenen disjunkten Teilmengen von positivem Maß zusammensetzen. Zu vorgegebenen Koeffizienten

$$\hat{A}(x) = \hat{A}^T(x) \in \mathbb{R}^{d\times d} \text{ mit } \inf_{x\in\Omega} \left\langle \xi, \hat{A}(x)\xi \right\rangle \geq \alpha_{\hat{A}}|\xi|^2 \quad \forall \xi \in \mathbb{R}^d \quad \text{und } \alpha_{\hat{A}} > 0,$$
$$b(x) \in \mathbb{R}^d, \quad a_0(x) \in \mathbb{R}$$

sowie gegebener rechter Seite $f(x)$ und Randbedingungen $g_D(x)$, $g_N(x)$ betrachten wir die folgende elliptische Randwertaufgabe:

Problem 3.1.1 (Klassische Formulierung)**.** *Finde eine Funktion $u \in C^2(\Omega) \cap C^1(\Omega \cup \overline{\Gamma}_N) \cap C(\overline{\Omega})$, welche den Gleichungen*

$$\begin{aligned} -\nabla \cdot (\hat{A}\nabla u) + b \cdot \nabla u + a_0 u &= f && \text{in } \Omega, \\ u &= g_D && \text{auf } \Gamma_D \\ \left\langle \hat{A}\nabla u, n \right\rangle &= g_N && \text{auf } \Gamma_N \end{aligned}$$

genügt.

Den Fall $\Gamma_N = \emptyset$ bezeichnen wir als reines Dirichlet-Problem, den Fall $\Gamma_D = \emptyset$ als reines Neumann-Problem und $\Gamma_D, \Gamma_N \neq \emptyset$ als Randwertproblem mit gemischten Randbedingungen. Eine dritte Art von Randbedingungen, auf die wir in dieser Arbeit nicht weiter eingehen, sind die so genannten Robin-Randbedingungen, d.h. $\langle \hat{A}\nabla u, n\rangle + \sigma u = g_R$ auf Γ_R.
Der große Nachteil des klassischen Zugangs ist, dass für eine Vielzahl von Problemen aus der Physik und den Ingenieurwissenschaften die Glattheitsforderungen an eine etwaige Lösung zu restriktiv sind und somit keine solche Lösung existiert. Dennoch besitzen diese Probleme aber durchaus Lösungen, jedoch von geringerer Glattheit. Die in dieser Hinsicht daher wesentlich besser geeignete Formulierung von Problem 3.1.1 ist die so genannte schwache Formulierung:

[1] Kapitel 3 enthält die wesentlichen Ergebnisse aus [EM05a].

Problem 3.1.2 (Schwache Formulierung)**.** *Finde ein $u \in H^1(\Omega)$, so dass*

$$u|_{\Gamma_D} = g_D \quad \textit{und} \quad a(u,v) = \langle \mathsf{f}, v\rangle \qquad \forall\, v \in H^1_D(\Omega) := \{u \in H^1(\Omega) \,|\, u|_{\Gamma_D} = 0\},$$

wobei

$$a(u,v) := \int_\Omega \left\langle \nabla u, \hat{A}\nabla v \right\rangle + \langle b, \nabla u\rangle\, v + a_0 uv \, d\Omega, \tag{3.1}$$

$$\langle \mathsf{f}, v\rangle := \int_\Omega fv\, d\Omega + \int_{\Gamma_N} g_N v\, d\Gamma. \tag{3.2}$$

Für den Rest der Arbeit vereinbaren wir folgende Regularitätsvoraussetzungen, welche mindestens erfüllt sein sollen:

Annahme 3.1.3 (Regularitätsvoraussetzungen)**.**

- *Rechte Seite $f \in L^2(\Omega)$.*
- *Koeffizienten $a_0 \in L^\infty(\Omega)$, $b \in L^\infty(\Omega, \mathbb{R}^d)$, $\hat{A} \in L^\infty(\Omega, \mathbb{R}^{d\times d})$, d.h. die Komponenten von b und $\hat{A}$ sind Funktionen aus $L^\infty(\Omega)$. Außerdem soll* $\operatorname{div} b \in L^\infty(\Omega)$ *gelten, wobei* div *im distributiven Sinn zu verstehen ist.*
- *Randbedingungen $g_D \in H^{1/2}(\Gamma_D)$, $g_N \in \left(H^{1/2}_{00}(\Gamma_N)\right)^*$ (siehe [Sch98, Abschnitt 1.4.3] für eine Definition und Bemerkungen zu $H^{1/2}_{00}(\Gamma_N)$).*

Im Gegensatz zur klassischen Formulierung reichen die obigen Regularitätsvoraussetzungen bereits aus, um für die schwache Formulierung folgende Existenz- und Eindeutigkeitsaussagen zu beweisen:

Theorem 3.1.4 (Existenz und Eindeutigkeit schwacher Lösungen)**.** *Seien die Annahmen 3.1.3 erfüllt, so gilt:*

1. *Das Problem 3.1.2 mit $\Gamma_N = \emptyset$ und homogenen Dirichlet-Randbedingungen besitzt eine eindeutig bestimmte schwache Lösung, falls*

$$-\frac{1}{2}\operatorname{div} b + a_0 \geq 0 \quad \textit{fast überall (f.ü.) auf } \Omega. \tag{3.3}$$

2. *Das Problem 3.1.2 mit $\Gamma_D = \emptyset$ besitzt eine eindeutig bestimmte schwache Lösung, falls*

$$-\frac{1}{2}\operatorname{div} b + a_0 \geq 0 \quad \textit{f.ü. auf } \Omega, \quad a_0 \geq \alpha_0 > 0 \quad \textit{und} \quad \langle b, n\rangle \geq 0 \quad \textit{f.ü. auf } \partial\Omega. \tag{3.4}$$

3. *Das Problem 3.1.2 mit $\Gamma_D, \Gamma_N \neq \emptyset$ und homogenen Dirichlet-Randbedingungen besitzt eine eindeutig bestimmte schwache Lösung, falls*

$$-\frac{1}{2}\operatorname{div} b + a_0 \geq 0 \quad \textit{f.ü. auf } \Omega \quad \textit{und} \quad \langle b, n\rangle \geq 0 \quad \textit{f.ü. auf } \Gamma_N. \tag{3.5}$$

Beweis. [Ver98, Satz 3.3] □

Bemerkung 3.1.5. Die in Theorem 3.1.4 getätigte Beschränkung auf homogene Dirichlet-Randbedingungen ist nicht wesentlich. Den Fall inhomogener Dirichlet-Randbedingungen kann man mittels Reduktion von inhomogenen auf homogene Dirichlet-Randbedingungen (siehe [Bra97]) abdecken.

Damit wissen wir nun, dass unter geeigneten Voraussetzungen an die Koeffizienten stets eine eindeutig bestimmte schwache Lösung zu Problem 3.1.2 existiert. Bleibt noch der Zusammenhang zwischen schwacher und klassischer Lösung zu klären.
Anhand der Überführung der klassischen Formulierung in die schwache Formulierung (siehe [GR92]) erkennt man, dass jede Lösung des klassischen Problems auch Lösung der schwachen Formulierung ist. Umgekehrt gilt, dass jede schwache Lösung auch eine klassische Lösung ist, sofern sie die notwendige Glattheit besitzt (siehe z.Bsp. [GR92, Satz 3.3]).

3.2 Das diskretisierte Problem

Das Problem 3.1.2 ist im Allgemeinen nicht analytisch lösbar. Um mittels geeigneter numerischer Verfahren Näherungslösungen zu berechnen, ersetzen wir die unendlich-dimensionalen Räume $H^1(\Omega)$, $H^1_D(\Omega)$ durch endlich-dimensionale Teilräume $\mathcal{V} \subset H^1(\Omega)$, $\mathcal{V}_D = \mathcal{V} \cap H^1_D(\Omega)$ und erhalten eine diskretisierte Version von Problem 3.1.2.

Problem 3.2.1 (Diskrete Formulierung)**.** *Sei $\tilde{g}_D \in Y_D := \{u|_{\Gamma_D} \mid u \in \mathcal{V}\}$ die $L^2(\Gamma_D)$-Projektion der Dirichlet-Bedingungen g_D, gegeben durch*

$$\int_{\Gamma_D} \tilde{g} v \, d\Gamma = \int_{\Gamma_D} g_D v \, d\Gamma \qquad \forall\, v \in Y_D.$$

Finde ein $u \in \mathcal{V}$, so dass mit $a(\cdot,\cdot)$ und $\langle \mathsf{f}, \cdot \rangle$ gegeben durch (3.1) bzw. (3.2) gilt

$$\begin{aligned} u|_{\Gamma_D} &= \tilde{g}_D \\ a(u,v) &= \langle \mathsf{f}, v \rangle \qquad \forall\, v \in \mathcal{V}_D. \end{aligned}$$

Als endlich-dimensionale Teilräume von Hilbert-Räumen sind $\mathcal{V}$ und $\mathcal{V}_D$ ihrerseits ebenfalls Hilbert-Räume und die Existenz- und Eindeutigkeitsaussagen übertragen sich auf das diskrete Problem. Über die Güte der Näherungslösung lässt sich folgende a priori Aussage machen:

Lemma 3.2.2. *Sei $a(u,u) \sim \|u\|^2_{H^1(\Omega)}$ für alle $u \in H^1_D(\Omega)$. Für $\mathcal{V} \subset H^1(\Omega)$ sei $u_\mathcal{V} \in \mathcal{V}$ die eindeutig bestimmte Lösung von Problem 3.2.1 und $u \in H^1(\Omega)$ die eindeutig bestimmte Lösung von Problem 3.1.2. Dann existiert $C > 0$, unabhängig von $\mathcal{V}$, so dass:*

$$\|u - u_V\|_{H^1(\Omega)} \;\leq\; C \left(\inf_{v \in \mathcal{V}, v|_{\Gamma_D} = \tilde{g}_D} \|u - v\|_{H^1(\Omega)} + \|g_D - \tilde{g}_D\|_{H^{1/2}(\Gamma_D)} \right).$$

Beweis. Für das reine Neumann-Problem entfällt der $\|g_D - \tilde{g}_D\|$ Term und obige Aussage entspricht dem Lemma von Céa. Für den Fall $\Gamma_D \neq \emptyset$ betrachten wir die Hilfsgröße $\tilde{u} \in H^1(\Omega)$, gegeben durch

$$\tilde{u}|_{\Gamma_D} = \tilde{g}_D \quad \text{und} \quad a(\tilde{u}, v) = \langle \mathsf{f}, v \rangle \quad \forall\, v \in H^1_D(\Omega).$$

Nach Dreiecksungleichung und dem Lemma von Céa existiert ein $C > 0$, unabhängig von $\mathcal{V}$, so dass

$$\begin{aligned} \|u - u_V\|_{H^1(\Omega)} &\leq \|u - \tilde{u}\|_{H^1(\Omega)} + \|\tilde{u} - u_V\|_{H^1(\Omega)} \\ &\leq \|u - \tilde{u}\|_{H^1(\Omega)} + C\|\tilde{u} - v\|_{H^1(\Omega)} \end{aligned} \tag{3.6}$$

für beliebiges $v \in \mathcal{V}$ mit $v|_{\Gamma_D} = \tilde{g}_D$ gilt. Setzen wir $w := u - \tilde{u}$, so ist $w \in H^1(\Omega)$ die eindeutig bestimmte Lösung von

$$w|_{\Gamma_D} = g_D - \tilde{g}_D \quad \text{und} \quad a(w, v) = 0 \quad \forall\, v \in H^1_D(\Omega).$$

Da $g_D - \tilde{g}_D \in H^{1/2}(\Gamma_D)$, existiert ein $C > 0$, welches nur von Ω und Γ_D abhängt, sowie $\overline{w} \in H^1(\Omega)$ mit $\overline{w}|_{\Gamma_D} = g_D - \tilde{g}_D$ und $\|\overline{w}\|_{H^1(\Omega)} \leq C\|g_D - \tilde{g}_D\|_{H^{1/2}(\Gamma_D)}$. Setzen wir nun $w = w_D + \overline{w}$, so erhalten wir $w_D \in H^1_D(\Omega)$ aus:

$$a(w_D, v) = -a(\overline{w}, v) \quad \forall\, v \in H^1_D(\Omega).$$

Nach Dreiecksungleichung und dem Lax-Milgram-Lemma [GR92, Lem.3.6] gilt somit

$$\begin{aligned} \|w\|_{H^1(\Omega)} &\leq \|\overline{w}\|_{H^1(\Omega)} + \|w_D\|_{H^1(\Omega)} \leq \|\overline{w}\|_{H^1(\Omega)} + C \sup_{v \in H^1_D(\Omega)} \frac{|a(\overline{w}, v)|}{\|v\|_{H^1(\Omega)}} \\ &\leq (1 + C)\|\overline{w}\|_{H^1(\Omega)} \leq C\|g_D - \tilde{g}_D\|_{H^{1/2}(\Gamma_D)}, \end{aligned} \tag{3.7}$$

wobei die Konstante C nur von $a(\cdot, \cdot)$, Ω und Γ_D abhängt. Aus (3.6) und (3.7) folgt die Behauptung. □

Wir sehen also, dass die Güte unserer Näherungslösung im Wesentlichen davon abhängt, wie gut sich die exakte Lösung im Raum $\mathcal{V}$ approximieren lässt. Es stellt sich daher die Frage nach der Konstruktion geeigneter endlich-dimensionaler Teilräume $\mathcal{V}$.

3.3 *hp*-FE-Räume

Das Wesen der Finiten-Element-Methode besteht darin, das Grundgebiet Ω in geometrisch einfache Teilgebiete K_i, $i = 1, \ldots, N$, zu zerlegen und den Finiten-Element-Raum $\mathcal{V} \subset H^1(\Omega)$ als Raum von Funktionen $u \in H^1(\Omega)$ zu definieren, die eingeschränkt auf K_i einer einfach zu handhabenden Klasse von Funktionen, in der Regel transformierten Polynomen, angehören. In dieser Arbeit wollen wir Zerlegungen von Ω in Dreiecks- bzw. Tetraederelemente betrachten. Diese Zerlegungen bieten große Flexibilität in der Vernetzung, haben jedoch speziell bei Quadraturen und somit beim Aufstellen der Steifigkeitsmatrix den Nachteil, dass Dreiecks- und Tetraederelemente keine natürliche Tensorproduktstruktur aufweisen.
Wir beginnen mit der Definition der Referenzelemente sowie der Definition einer in unserem Sinn zulässigen Vernetzung von Ω.

Definition 3.3.1 (Referenzelemente)**.** Das Referenzdreieck $\mathcal{T}^2$ und das Referenztetraeder $\mathcal{T}^3$ seien gegeben durch

$$\begin{aligned} \mathcal{T}^2 &= \{(x, y) \mid -1 < x, y \ \wedge\ x + y < 0\}, \\ \mathcal{T}^3 &= \{(x, y, z) \mid -1 < x, y, z \ \wedge\ x + y + z < -1\}. \end{aligned}$$

Definition 3.3.2 (Dreiecks/Tetraeder-Vernetzung von Ω)**.** Sei $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, ein Gebiet. Wir bezeichnen $\mathcal{N} = \{(K_i, F_i) \mid i = 1, \ldots, N\}$ als eine Dreiecks- oder Tetraedervernetzung von Ω, falls folgende Voraussetzungen erfüllt sind:

1. Alle $K_i \subset \Omega$ sind offene, nicht-leere und zusammenhängende Punktmengen mit $\overline{\Omega} = \cup_{i=1}^{N} \overline{K_i}$.

2. Die Abbildungen $F_{K_i} : \overline{\mathcal{T}}^d \to \overline{K}$ sind stetig differenzierbar, es existieren stetig differenzierbare Inverse F_i^{-1} und es gilt

   $$K_i = F_i(\mathcal{T}^d) \qquad \forall\, i = 1, \ldots, N.$$

   Als Eckpunkte v_l, Kanten e_l und Seitenflächen f_l (für d=3) von K_i bezeichnen wir die Bilder der entsprechenden Objekte des Referenzelements $\mathcal{T}^d$.

3. Die Mengen K_i sind paarweise durchschnittsfremd und für $i \neq j$ trifft auf die Schnittmenge $S_{ij} = \overline{K}_i \cap \overline{K}_j$ genau eine der folgenden Aussagen zu:

   (a) $S_{ij} = \emptyset$,

   (b) S_{ij} ist gemeinsamer Eckpunkt von K_i und K_j,

   (c) S_{ij} ist gemeinsame Kante von K_i und K_j,

   (d) S_{ij} ist gemeinsame Seitenfläche von K_i und K_j (für $d = 3$).

4. Sei $S_{ij} = \overline{K}_i \cap \overline{K}_j$ gemeinsame Kante oder gemeinsame Seitenfläche (für $d = 3$) von K_i und K_j. Bezeichne $\{P_1, \ldots, P_n\}$, $n \in \{2, 3\}$, die Eckpunkte von S_{ij} und P einen beliebigen Punkt auf S_{ij}, so gilt

   $$\Lambda_i := (\lambda_{i,1}, \ldots, \lambda_{i,n}) = (\lambda_{j,1}, \ldots, \lambda_{j,n}) =: \Lambda_j,$$

   wobei Λ_k, $k \in \{i, j\}$, die baryzentrischen Koordinaten des Punktes $F_k^{-1}(P)$ auf der Strecke $\overline{F_k^{-1}(P_1)F_k^{-1}(P_2)}$ bzw. im Dreieck $\Delta(F_k^{-1}(P_1), F_k^{-1}(P_2), F_k^{-1}(P_3))$ bezeichne. D.h.

   $$F_k^{-1}(P) = \sum_{l=1}^{n} \lambda_{k,l} F_k^{-1}(P_l), \qquad \sum_{l=1}^{n} \lambda_{k,l} = 1, \qquad \lambda_{k,l} \geq 0.$$

Für eine Vernetzung $\mathcal{N}$ von Ω führen wir noch folgende Bezeichnungen ein:

Definition 3.3.3. Sei $\mathcal{N} = \{(K_i, F_i) \mid i = 1, \ldots, N\}$ eine Vernetzung von $\Omega \subset \mathbb{R}^d$ gemäß Definition 3.3.2. Wir bezeichnen

- die Menge aller Elemente von $\mathcal{N}$ mit

  $$\mathcal{T}(\mathcal{N}) = \{K \subset \Omega \mid \exists i \in 1, \ldots, N \text{ mit } (K, F_i) \in \mathcal{N}\},$$

- die Menge aller Seitenflächen (faces) von $\mathcal{N}$ mit

  $$f(\mathcal{N}) = \{S \subset \overline{\Omega} \mid \exists K \in \mathcal{T}(\mathcal{N}), \text{ welches } S \text{ als Seitenfläche hat}\},$$

- die Menge aller Kanten (edges) von $\mathcal{N}$ mit

$$e(\mathcal{N}) = \{E \subset \overline{\Omega} \mid \exists K \in \mathcal{T}(\mathcal{N}),\ \text{welches } E \text{ als Kante hat}\},$$

- die Menge aller Ecken (vertices) von $\mathcal{N}$ mit

$$v(\mathcal{N}) = \{v \in \overline{\Omega} \mid \exists K \in \mathcal{T}(\mathcal{N}),\ \text{welches } v \text{ als Ecke hat}\}.$$

Bemerkung 3.3.4. Für $K \in \mathcal{T}(\mathcal{N})$ bezeichnen wir mit $h_K := diam(K)$ den Durchmesser von K und die zu K gehörige Elementtransformation auch als $F_K : \mathcal{T}^d \to K$.

Bemerkung 3.3.5. Definition 3.3.2 lässt recht allgemeine Elementtransformationen zu, womit das zu vernetzende Gebiet Ω nicht notwendigerweise polynomial berandet zu sein braucht. Eine Einschränkung auf affine Transformation werden wir erst in späteren Kapiteln vornehmen und an betreffender Stelle darauf hinweisen. Mit Punkt 3 der obigen Definition schließen wir hängende Knoten aus. Punkt 4 stellt sicher, dass auf gemeinsamen Kanten und Flächen stets gleiche metrische Verhältnisse herrschen, und ist damit wesentlich für das spätere Assemblieren von lokal auf $K \in \mathcal{T}(\mathcal{N})$ definierten Funktionen zu globalen und auf Ω stetigen Funktionen.

Zusätzlich zur Zerlegung des Gebietes Ω in Teilgebiete K_i ordnen wir im Kontext der *hp*-FEM noch jedem Element $K \in \mathcal{T}(N)$, jeder Kante $e \in e(N)$ und für $d = 3$ jeder Seitenfläche $f \in f(N)$ einen Polynomgrad zu. Diese Polynomgradverteilung wollen wir mit $p(\mathcal{N})$ bezeichnen und sie soll stets zulässig im Sinne der folgenden Definition sein:

Definition 3.3.6. Es sei $\mathcal{N}$ eine Vernetzung von $\Omega \subset \mathbb{R}^d$ gemäß Definition 3.3.2. Die Polynomgradverteilung

$$p(\mathcal{N}) = \begin{cases} \{(p_e)_{e\in e(\mathcal{N})}, (p_K)_{K\in\mathcal{T}(\mathcal{N})}\} & : \quad d = 2 \\ \{(p_e)_{e\in e(\mathcal{N})}, (p_f)_{f\in f(\mathcal{N})}, (p_K)_{K\in\mathcal{T}(\mathcal{N})}\} & : \quad d = 3 \end{cases} \tag{3.8}$$

ist eine zulässige Polynomgradverteilung auf $\mathcal{N}$, falls

- jedem Element $K \in \mathcal{T}(N)$ ein Polynomgrad $p_K \in \mathbb{N}$,
- jeder Seitenfläche $f \in f(\mathcal{N})$ der Polynomgrad

$$p_f = \min\{p_K \mid \text{f ist Seitenfläche des Elements } K\},$$

- jeder Kante $e \in e(\mathcal{N})$ der Polynomgrad

$$p_e = \min\{p_K \mid \text{e ist Kante des Elements } K\}$$

zugeordnet ist. Mit $\mathbf{p} := (p_K)_{K\in\mathcal{T}}$ bezeichnen wir den Vektor der Polynomgradverteilung von $\mathcal{T}(\mathcal{N})$ und mit

$$p(K) := \begin{cases} (p_{e_1}, p_{e_2}, p_{e_3}, p_K) & : \quad d = 2 \\ (p_{e_1}, \ldots, p_{e_6}, p_{f_1}, \ldots, p_{f_4}, p_K) & : \quad d = 3 \end{cases}$$

den Vektor der Polynomgradverteilung des Elements $K \in \mathcal{T}(\mathcal{N})$.

Ausgehend von dem Grundgebiet Ω, einer Vernetzung $\mathcal{N}$ und einer Polynomgradverteilung $p(\mathcal{N})$ definieren wir Finite-Element-Räume

$$S_D^{\mathbf{p}}(\Omega,\mathcal{N}) \subset S^{\mathbf{p}}(\Omega,\mathcal{N}) \subset H^1(\Omega)$$

wie folgt:

Definition 3.3.7 (*hp*-FE-Räume)**.** Sei Ω ein Gebiet, $\mathcal{N}$ eine Vernetzung von Ω gemäß Definition 3.3.2 und $p(\mathcal{N})$ eine Polynomgradverteilung gemäß Definition 3.3.6. Wir definieren

$$\begin{aligned} S^{\mathbf{p}}(\Omega,\mathcal{N}) &:= \{u \in H^1(\Omega) \mid u|_K \circ F_K \in \Pi_{\mathbf{p}(K)}(\mathcal{T}^d) \quad \forall K \in \mathcal{T}(\mathcal{N})\}, \\ S_D^{\mathbf{p}}(\Omega,\mathcal{N}) &:= S^{\mathbf{p}}(\Omega,\mathcal{N}) \cap H_D^1(\Omega), \end{aligned}$$

$$\Pi_{p(K)}(\mathcal{T}^d) := \left\{ u \in P_{p_K}(\mathcal{T}^d) \,\middle|\, \begin{array}{l} u|_{e_l} \in P_{p_{e_l}}(e_l) \text{ für alle Kanten } e_l \text{ von } \mathcal{T}^d \\ u|_{f_l} \in P_{p_{f_l}}(f_l) \text{ für alle Seiten } f_l \text{ von } \mathcal{T}^3 \end{array} \right\}.$$

Bemerkung 3.3.8. Die Definition von $\Pi_{p(K)}(\mathcal{T}^d)$ geschieht mittels Formfunktionen auf $\mathcal{T}^d$ (siehe Abschnitt 3.6.1). Im weiteren Verlauf des Kapitels werden wir auch noch erweiterte Räume $\tilde{\Pi}_{p(K)}(\mathcal{T}^d)$ einführen, welche dann an die Stelle von $\Pi_{p(K)}(\mathcal{T}^d)$ gesetzt werden können.

Mit obigen Definitionen lautet die FE-Diskretisierung von Problem 3.1.2:

Problem 3.3.9 (hp-FEM Formulierung)**.** *Sei $\tilde{g}_D \in Y_D^{\mathbf{p}}(\Omega,\mathcal{N}) := \{u|_{\Gamma_D} \mid u \in S^{\mathbf{p}}(\Omega,\mathcal{N})\}$ die $L^2(\Gamma_D)$-Projektion der Dirichlet-Bedingungen g_D, gegeben durch*

$$\int_{\Gamma_D} \tilde{g} v \, d\Gamma = \int_{\Gamma_D} g_D v \, d\Gamma \qquad \forall\, v \in Y^{\mathbf{p}}(\Omega,\mathcal{N}).$$

Finde ein $u \in S^{\mathbf{p}}(\Omega,\mathcal{N})$, so dass mit $a(\cdot,\cdot)$ und $\langle \mathsf{f}, \cdot \rangle$ gegeben durch (3.1) bzw. (3.2) gilt

$$\begin{aligned} u|_{\Gamma_D} &= \tilde{g}_D \\ a(u,v) &= \langle \mathsf{f}, v \rangle \qquad \forall\, v \in S_D^{\mathbf{p}}(\Omega,\mathcal{N}). \end{aligned}$$

3.4 Assemblieren von Steifigkeitsmatrix und Lastvektor

Um Problem 3.3.9 numerisch zu lösen, müssen wir die Räume $S^{\mathbf{p}}(\Omega,\mathcal{N})$ und $S_D^{\mathbf{p}}(\Omega,\mathcal{N})$ noch mit Basen Θ bzw. Θ_D ausstatten. Zweckmäßigerweise soll dabei insbesondere gelten:

$$\Theta_D = \{\theta_i \mid i = 1,\ldots,N_D\} \subset \Theta = \{\theta_i \mid i = 1,\ldots,N\}.$$

Die hierbei übliche Vorgehensweise ist die folgende:

1. Wir definieren für jede mögliche Polynomgradverteilung $p(K)$ auf dem Referenzelement $\mathcal{T}^d$ eine Basis $\Psi^{(p(K))} = \{\psi_i^{(p(K))} \mid i = 1,\ldots,N_{p(K)}\}$ für den Raum $\Pi_{p(K)}(\mathcal{T}^d)$.

2. Für jedes $K \in \mathcal{T}(\mathcal{N})$ definieren wir mittels der Transformation F_K eine lokale Basis $\Theta^{(K)} = \{\theta_i^{(K)} \mid i = 1,\ldots,N_{p(K)}\}$ und setzen die Basisfunktionen außerhalb von K identisch Null fort. D.h.

$$\theta_i^{(K)}(x) = \begin{cases} [\psi_i^{(p(K))} \circ F_K^{-1}](x) & : \quad x \in \overline{K} \\ 0 & : \quad \text{sonst} \end{cases}.$$

3. Wir assemblieren die lokalen Basen $\Theta^{(K)}$, $K \in \mathcal{T}(\mathcal{N})$ zu einer Basis für $S^{\mathbf{p}}(\Omega, \mathcal{N})$:

$$\Theta = \mathcal{A}_{K \in \mathcal{T}} \Theta^{(K)}.$$

4. Wir bestimmen Θ_D als Teilmenge von Θ.

Geeignete Basisfunktionen $\Psi^{(p(K))}$ sowie eine detaillierte Beschreibung der Arbeitsweise des Assemblierungsoperators sind in [KS99] zu finden.
Haben wir erst einmal eine Basis für $S^{\mathbf{p}}(\Omega, \mathcal{T})\}$ bzw. $S^{\mathbf{p}}_D(\Omega, \mathcal{N})$ definiert, so kann die Lösung von Problem 3.3.9 mittels eines linearen Gleichungssystems bestimmt werden. Für das reine Neumann-Problem ergibt sich hierfür

$$\underbrace{[a(\theta_j, \theta_i)]^N_{i,j=1}}_{A} [u_i]^N_{i=1} = \underbrace{[\langle \mathsf{f}, \theta_i \rangle]^N_{i=1}}_{l}. \tag{3.9}$$

Die Matrix A heißt Steifigkeitsmatrix, l Lastvektor und $[u_i]^N_{i=1}$ enthält die Zerlegungskoeffizienten der Lösung bezüglich der Basis $\{\theta_i \mid i = 1, \ldots, N\}$.

Bemerkung 3.4.1. Für Dirichlet-Problem und Probleme mit gemischten Randbedingungen erhalten wir ein zu (3.9) analoges Gleichungssystem. Die zugehörige Steifigkeitsmatrix kann hierbei aus der Steifigkeitsmatrix des reinen Neumann-Problems durch Streichen aller Zeilen und Spalten, welche zu den Basisfunktionen $\theta \in \Theta \backslash \Theta_D$ gehören, bestimmt werden. Auf der rechten Seite des Gleichungssystems müssen vor dem Streichen der zu $\theta \in \Theta \backslash \Theta_D$ gehörigen Zeilen noch eventuelle nicht-homogene Dirichlet-Randbedingungen berücksichtigt werden (siehe [GR92]).
Führen wir die Vektoren

$$[\Theta] = (\theta_1, \ldots, \theta_N), \quad \left[\Theta^{(K)}\right] = \left(\theta^{(K)}_1, \ldots, \theta^{(K)}_{N_{p(K)}}\right) \qquad \forall\, K \in \mathcal{T}(\mathcal{N})$$

ein, so können wir das Wirken des Operators $\mathcal{A}$ auch durch Matrizen T_K, $K \in \mathcal{T}(\mathcal{N})$, ausdrücken:

$$[\Theta]^T = \sum_{K \in \mathcal{T}(\mathcal{N})} T_K \left[\Theta^{(K)}\right]^T.$$

Die sehr schwach besetzten Matrizen $T_K \in \mathbb{R}^{N \times N_{p(K)}}$, $K \in \mathcal{T}(\mathcal{N})$, realisieren hierbei die so genannte „local to global“-Abbildung. Jede Spalte j von T_K enthält genau einen von Null verschiedenen Eintrag e_{ij} (in der Regel ± 1), welcher den Anteil der lokalen Basisfunktion $\theta^{(K)}_j$ zur globalen Basisfunktion θ_i addiert (für eine detaillierte Beschreibung verweisen wir abermals auf [KS99]). Als Folge des Aufbauens der globalen Basis Θ aus den lokalen Basen $\Theta^{(K)}$ können auch die globale Steifigkeitsmatrix A und der Lastvektor l aus den lokalen Steifigkeitsmatrizen A_K und Lastvektoren l_K, $K \in \mathcal{T}(\mathcal{N})$ zusammengesetzt werden. Aus den Definitionen von A und l ergibt sich

$$A = \mathcal{A}_{K \in \mathcal{T}(\mathcal{N})} A_K = \sum_{K \in \mathcal{T}(\mathcal{N})} T_K A_K T_K^T, \qquad l = \mathcal{A}_{K \in \mathcal{T}(\mathcal{N})} l_K = \sum_{K \in \mathcal{T}(\mathcal{N})} T_K l_K$$

mit

$$A_K = \left[a(\theta^{(K)}_j, \theta^{(K)}_i)\right]^{N_{p(K)}}_{i,j=1}, \qquad l_K = \left[\left\langle \mathsf{f}, \theta^{(K)}_i \right\rangle\right]^{N_{p(K)}}_{i=1}.$$

Ab dieser Stelle können die einzelnen Elementsteifigkeitsmatrizen A_K und Lastvektoren l_K völlig unabhängig voneinander aufgestellt werden.

3.5 Einträge der lokalen Steifigkeitsmatrizen

Im Gegensatz zur h-FEM mit niedrigem Polynomgrad kann das Berechnen der lokalen Steifigkeitsmatrizen und Lastvektoren in der p- und hp-FEM speziell für hohe Polynomgrade sehr rechenintensiv werden. In den nun folgenden Abschnitten wollen wir uns daher mit verschiedenen Algorithmen beschäftigen, die diese Aufgabe mit möglichst optimalem Aufwand bewältigen. Wir beginnen mit dem einfachst denkbaren Algorithmus, welcher die lokale Steifigkeitsmatrix A_K mit einem Arbeitsaufwand von $O(p_K^{3d})$ assembliert, stellen die Methode der Summenfaktorisierung vor und verallgemeinern schließlich die in [MGS01] für Elemente mit Tensorproduktgestalt vorgestellte Spektral-Galerkin-Methode auf Dreiecks- bzw. Tetraederelemente.
Um die Darstellungen zu vereinfachen und damit insbesondere auch die wesentlichen Ideen klarer herauszustellen, beschränken wir uns von nun an auf die Berechnung der lokalen Steifigkeitsmatrizen A_K und hierbei speziell auf den für skalare Probleme typischen Fall

$$a(u,v) := \int_\Omega (\hat{A}(x)\nabla u)\cdot\nabla v \mathrm{d}\Omega.$$

Wir weisen an dieser Stelle jedoch explizit darauf hin, dass alle Algorithmen ohne große Mühe auch auf das Auswerten der Funktionale $\langle \mathsf{f}, \theta_i^{(K)}\rangle$ sowie auf vektorwertige Probleme und Terme niedrigerer Ordnung, d.h. $\int_\Omega b(x)\nabla uv + c(x)uv\mathrm{d}\Omega$, übertragen werden können.
Um die Einträge der lokalen Steifigkeitsmatrix zu berechnen, müssen wir die Bilinearform $a(\theta_j^{(K)}, \theta_i^{(K)})$ numerisch auswerten. D.h. wir müssen die Integrale

$$(A_K)_{ij} = a(\theta_j^{(K)}, \theta_i^{(K)}) = \int_K \left\langle \nabla\theta_j^{(K)}, \hat{A}\nabla\theta_i^{(K)} \right\rangle \mathrm{d}\Omega, \tag{3.10}$$

mit $i,j = 1,\ldots,N_{p(K)}$ und $\Theta^{(K)} = \{\theta_i^{(K)} \mid i = 1,\ldots,N_{p(K)}\}$ eine lokale Basis auf $K \in \mathcal{T}(K)$ bestimmen. Da wir variable Koeffizienten $\hat{A}(x)$ betrachten, können wir nicht davon ausgehen, die Integrale analytisch bestimmen zu können und müssen somit auf geeignete numerische Quadraturverfahren zurückgreifen, wobei die zu verwendende Quadraturordnung neben den Koeffizienten $\hat{A}(x)$ auch wesentlich vom Polynomgrad p_K des betreffenden Elementes $K \in \mathcal{T}(\mathcal{N})$ abhängt (siehe hierzu auch Abschnitt 3.10). Für Elemente mit hohem Polynomgrad benötigen wir zwangsläufig Quadraturverfahren hoher Ordnung und für Elemente mit niedrigem Polynomgrad sollte aus Effizienzgründen ein Quadraturverfahren niedrigerer Ordnung benutzt werden. Allgemein werden wir von einer Quadraturordnung $O(p_K)^d$ ausgehen, d.h. wir benutzen $O(p_K)$ Stützstellen für jede Raumrichtung.
Die Konstruktion geeigneter Quadraturverfahren beliebiger Ordnung wird dadurch erschwert, als dass Dreiecke und Tetraeder keine natürliche Tensorproduktstruktur aufweisen. Der Ausweg besteht daher in einer Transformation der Elemente $K \in \mathcal{T}(\mathcal{N})$ auf das Referenzelement $\mathcal{T}^d$ und anschließender Transformation des Referenzelements auf den i-dimensionalen Einheitswürfel $\mathcal{Q}^i$. Hier lassen sich dann leicht numerische Quadraturformeln als Tensorprodukte von 1D-Quadraturformeln zusammensetzen.
Beginnen wir mit der Definition des d-dimensionalen Einheitswürfels $\mathcal{Q}^d$ und der Transformation von $\mathcal{T}^d$ auf $\mathcal{Q}^d$.

Definition 3.5.1 (Einheitswürfel)**.** Der d-dimensionale Einheitswürfel $\mathcal{Q}^d$ sei gegeben durch:

$$\mathcal{Q}^d = (-1,1)^d.$$

Lemma 3.5.2 (Duffy-Transformation $\mathcal{T}^d \leftrightarrow \mathcal{Q}^d$)**.** *Seien die Abbildungen* $D_2 : \mathbb{R}^2 \mapsto \mathbb{R}^2$ *und* $D_3 : \mathbb{R}^3 \mapsto \mathbb{R}^3$ *gegeben durch:*

$$D_2 \ : \ (\eta_1, \eta_2) \quad \mapsto \left(\frac{1}{2}(1+\eta_1)(1-\eta_2) - 1, \eta_2 \right),$$
$$D_3 \ : \ (\eta_1, \eta_2, \eta_3) \mapsto \left(\frac{1}{4}(1+\eta_1)(1-\eta_2)(1-\eta_3) - 1, \frac{1}{2}(1+\eta_2)(1-\eta_3) - 1, \eta_3 \right).$$

Dann gilt

$$|\det D_2'| = \left(\frac{1-\eta_2}{2} \right), \qquad |\det D_3'| = \left(\frac{1-\eta_2}{2} \right) \left(\frac{1-\eta_3}{2} \right)^2$$

und

$$\mathcal{T}^d = D_d(\mathcal{Q}^d) \quad \textit{für } d = 2, 3. \tag{3.11}$$

Beweis. Einfaches Nachrechnen bzw. [KS99]. □

Mit den aus der Analysis bekannten Gesetzen für die Variablensubstitution in Integralen erhalten wir für die Berechnung von (3.10) somit:

Lemma 3.5.3. *Sei* $\mathcal{N}$ *eine Vernetzung von* $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, *und* $K \in \mathcal{T}(\mathcal{N})$ *mit* $K = F_K(\mathcal{T}^d)$. *Sei die Abbildung* $D_d : \mathbb{R}^d \mapsto \mathbb{R}^d$ *gegeben durch Lemma 3.5.2. Sei* $\theta_i^{(K)} = \psi_i \circ F_K^{-1}$ *und* $A_K(\theta_j^{(K)}, \theta_i^{(K)})$ *gegeben durch (3.10). Dann gilt:*

$$(A_K)_{ij} \ = \ \int_{\mathcal{T}^d} \left\langle \nabla\psi_i, (F_K')^{-1}(\hat{A} \circ F_K)(F_K')^{-T} \nabla\psi_j \right\rangle |\det F_K'| \, d\vec{\xi} \tag{3.12}$$
$$= \ \int_{\mathcal{Q}^d} \left\langle \nabla(\psi_i \circ D_d), \tilde{A} \nabla(\psi_j \circ D_d) \right\rangle |\det D_d'| \, d\vec{\eta}, \tag{3.13}$$

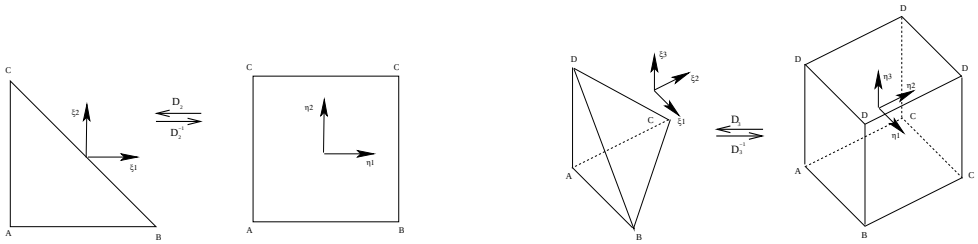
wobei

$$\tilde{A} \ := \ (D_d')^{-1}(F_K')^{-1}(\hat{A} \circ F_K \circ D_d)(F_K')^{-T}(D_d')^{-T} |\det F_K'|$$

und

$$F_K' := \begin{bmatrix} \frac{\partial x_1}{\partial \xi_1} & \cdots & \frac{\partial x_1}{\partial \xi_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_d}{\partial \xi_1} & \cdots & \frac{\partial x_d}{\partial \xi_d} \end{bmatrix}, \qquad D_d' := \begin{bmatrix} \frac{\partial \xi_1}{\partial \eta_1} & \cdots & \frac{\partial \xi_1}{\partial \eta_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial \xi_d}{\partial \eta_1} & \cdots & \frac{\partial \xi_d}{\partial \eta_d} \end{bmatrix}.$$

Abbildung 3.1: Transformationen D_2 und D_3



Bemerkung 3.5.4. In den folgenden Abschnitten werden für $K \in \mathcal{T}(\mathcal{N})$ die Transformationen

$$K \underset{F_K^{-1}}{\overset{F_K}{\leftrightarrows}} \mathcal{T}^d \underset{D_d^{-1}}{\overset{D_d}{\leftrightarrows}} \mathcal{Q}^d$$

eine große Rolle spielen. Um es etwas einfacher zu machen den Überblick zu behalten, treffen wir folgende Vereinbarungen:

- Die Variablen η beziehen sich stets auf den Einheitswürfel $\mathcal{Q}^d$, die Variablen ξ beziehen sich auf $\mathcal{T}^d$ und die Variablen x auf K.
- Formfunktionen auf $\mathcal{Q}^d$ werden mit $\phi(\eta)$ bezeichnet.
- Transformieren wir Formfunktionen von $\mathcal{Q}^d$ auf $\mathcal{T}^d$, so bezeichnen wir diese mit $\psi(\xi)$, d.h. $\psi(\xi) = [\psi \circ D_d](\eta) = \phi(\eta)$.
- Transformieren wir die Formfunktionen auf K, so bezeichnen wir diese mit $\theta(x)$, d.h. $\theta(x) = [\theta \circ F_K](\xi) = \psi(\xi)$.

3.6 Algorithmen zum Aufstellen der Elementsteifigkeitsmatrizen

Lemma 3.5.3 versetzt uns in die Lage, die Eintrage der Elementsteifigkeitsmatrizen numerisch mittels leicht zu konstruierender Quadraturregeln auszuwerten. Wir ersetzen hierbei das Integral über $\mathcal{Q}^d$ durch eine Tensorproduktkonstruktion eindimensionaler Quadraturregeln bezüglich des Intervalls $(-1, 1)$. Eine geeignete Wahl sind dabei die Gauß-Lobatto-Jacobi-Quadraturen (siehe Abschnitt 2.4), welche es insbesondere gestatten, die $|\det D_d'|$-Terme (siehe Lemma 3.5.2) in die Quadraturregel zu integrieren. D.h. mit

$$\begin{aligned} \mathrm{QR} &= \mathrm{QR}^1 \times \ldots \times \mathrm{QR}^d \\ &= \left\{ (\eta_{l_1}^{(1)}, \omega_{l_1}^{(1)}) \mid l_1 = 0, \ldots, q_1 \right\} \times \ldots \times \left\{ (\eta_{l_d}^{(d)}, \omega_{l_d}^{(d)}) \mid l_d = 0, \ldots, q_d \right\} \end{aligned}$$

erhalten wir

$$(A_K)_{ij} \approx \sum_{l_1=0}^{q_1} \ldots \sum_{l_d=0}^{q_d} \omega_{l_1}^{(1)} \ldots \omega_{l_d}^{(d)} \left\langle \nabla(\psi_j \circ D_d), \tilde{A} \nabla(\psi_i \circ D_d) \right\rangle |_{(\eta_{l_1}^{(1)}, \ldots, \eta_{l_d}^{(d)})}. \tag{3.14}$$

Nachdem wir nun wissen, wie sich die Einträge der Elementsteifigkeitsmatrix A_K mittels numerischer Quadratur beliebiger Ordnung berechnen lassen, stellt sich natürlich die Frage, wie wir dies möglichst effektiv umsetzen können. Die elementarste und einfachste Vorgehensweise stellt Algorithmus 3.6.12 dar, bei dem wir die Formel (3.10) schlicht und einfach auf jeden Matrixeintrag separat anwenden. Da wir für ein Element $K \in \mathcal{T}(\mathcal{N})$ mit Polynomgrad p_K ein Gitter aus $O(p_K^d)$ Quadraturpunkten annehmen, erhalten wir für die Berechnung aller $O(p_K^d \times p_K^d)$ Einträge der lokalen Steifigkeitsmatrix A_K einen Arbeitsaufwand von $O(p_K^{3d})$. Der Vorteil von Algorithmus 3.6.12 liegt in seiner Einfachheit. Er ist leicht zu implementieren und stellt keine speziellen Forderungen an die lokalen Formfunktionen. Andererseits ist ein Aufwand von $O(p_K^{3d})$ für die lediglich $O(p_K^{2d})$ Einträge der Steifigkeitsmatrix jedoch alles andere als optimal und es stellt sich die Frage, ob und wie sich diese Komplexität reduzieren

lässt. Eine Möglichkeit ist die so genannte Summenfaktorisierung (siehe [KS99, Ors80]). Diese Methode reduziert die Komplexität für das Aufstellen von A_K auf $O(p_K^{2d+1})$ und setzt Tensorproduktstruktur der Formfunktionen $\phi_i := (\psi_i \circ D_d)$ auf dem Einheitswürfel $\mathcal{Q}^d$ voraus. Eine weitere Möglichkeit, die die Komplexität sogar auf optimale $O(p^{2d})$ reduziert, ist eine Verallgemeinerung der in [MGS01] vorgestellten Spektral-Galerkin-Methode auf den für uns interessanten Fall von Dreiecks- und Tetraederelementen. Wie wir sehen werden, geschieht diese Reduktion der Komplexität von $O(p^{2d+1})$ auf $O(p^{2d})$ jedoch mittels einer Erweiterung der Räume $\Pi_{p(K)}(\mathcal{T}^d)$, d.h. sie wird mit einer vergrößerten Anzahl innerer Formfunktionen erkauft. Insbesondere im Zusammenhang mit der statischen Kondensation (siehe Abschnitt 3.7) wird es daher von Interesse sein zu erforschen, welche der beiden Methoden, Summenfaktorisierung oder Spektral-Galerkin, unter welchen Umständen die geeignetere ist.

3.6.1 Formfunktionen für $\mathcal{T}^2$ und $\mathcal{T}^3$

Wir beginnen mit der Definition der aus [KS99] stammenden Formfunktionen $\Psi^{(KS)}$ für die Referenzelemente $\mathcal{T}^2$ und $\mathcal{T}^3$ sowie der Definition der von uns modifizierten Formfunktionen $\Psi^{(Lag)}$. Die Definition der Formfunktionen erfolgt hierbei auf dem jeweiligen Referenzwürfel $\mathcal{Q}^d$ und mittels Duffy-Transformation werden die Formfunktionen anschließend auf $\mathcal{T}^d$ übertragen. Diese Konstruktion der Formfunktionen ist dahingehend von Vorteil, als dass all unsere Quadraturen auf den Referenzwürfel $\mathcal{Q}^d$ transformiert werden (siehe Lemma 3.5.3).

Abkürzung 3.6.1.

$$f_1(x) := \left(\frac{1-x}{2}\right)\left(\frac{1+x}{2}\right), \quad f_2(x) := \left(\frac{1-x}{2}\right), \quad f_3(x) := \left(\frac{1+x}{2}\right).$$

Definition 3.6.2 (Formfunktionen für $\mathcal{T}^2$)**.** Es sei $\mathcal{T}^2$ das Referenzdreieck und $p(K) = (p_{AB}, p_{AC}, p_{BC}, p_K)$ eine Polynomgradverteilung, wobei p_{AB} den der Kante AB zugeordneten Polynomgrad bezeichnet und Analoges für die weiteren Einträge von $p(K)$ gilt. Für $i = 1, 2$ bezeichne $N_i = \{\eta_k^{(i)} \mid k = 1, \ldots, p_K - i\}$ eine Stützstellenmenge mit

$$-1 < \eta_1^{(i)} < \ldots < \eta_{p_K - i}^{(i)} < 1$$

und es bezeichne $l_k^{(N_i)}$ das k-te Lagrange-Interpolationspolynom bezüglich der Stützstellenmenge N_i. Mit den Abkürzungen aus 3.6.1 definieren wir

$$\Psi^{(KS)} = \bigcup_{B=0}^{5} \Psi_B^{(KS)} \quad \text{und} \quad \Psi^{(Lag)} = \bigcup_{B=0}^{5} \Psi_B^{(Lag)},$$

wobei

$$\begin{aligned} \Psi_B^{(KS)} &:= \Phi_B^{(KS)} \circ D_2^{-1} := \left\{\phi \circ D_2^{-1} \mid \phi \in \Phi_B^{(KS)}\right\}, \\ \Psi_B^{(Lag)} &:= \Phi_B^{(Lag)} \circ D_2^{-1} := \left\{\phi \circ D_2^{-1} \mid \phi \in \Phi_B^{(Lag)}\right\} \end{aligned}$$

und die Mengen $\Phi_B^{(KS)}$, $\Phi_B^{(Lag)}$ gegeben sind durch:

$$
\begin{aligned}
\Phi_0^{(KS)} = \Phi_0^{(Lag)} = \Phi_0 &:= \{f_3(\eta_2)\}\,,\\
\Phi_1^{(KS)} = \Phi_1^{(Lag)} = \Phi_1 &:= \{f_2(\eta_1)f_2(\eta_2), f_3(\eta_1)f_2(\eta_2)\}\,,\\
\Phi_2^{(KS)} &:= \left\{ f_1(\eta_1) f_2^{p+1}(\eta_2) P_{p-1}^{(1,1)}(\eta_1) \mid p = 1,\dots,p_{AB}-1 \right\},\\
\Phi_2^{(Lag)} &:= \left\{ f_1(\eta_1) f_2^{2}(\eta_2) P_{p-1}^{(1,1)}(\eta_1) \mid p = 1,\dots,p_{AB}-1 \right\},\\
\Phi_3^{(KS)} = \Phi_3^{(Lag)} = \Phi_3 &:= \left\{ f_2(\eta_1) f_1(\eta_2) P_{q-1}^{(1,1)}(\eta_2) \mid q = 1,\dots,p_{AC}-1 \right\},\\
\Phi_4^{(KS)} = \Phi_4^{(Lag)} = \Phi_4 &:= \left\{ f_3(\eta_1) f_1(\eta_2) P_{q-1}^{(1,1)}(\eta_2) \mid q = 1,\dots,p_{BC}-1 \right\},\\
\Phi_5^{(KS)} &:= \left\{ f_1(\eta_1) f_1(\eta_2) f_2^{p}(\eta_2) P_{p-1}^{(1,1)}(\eta_1) P_{q-1}^{(2p+1,1)}(\eta_2) \middle| \begin{array}{l} 1 \le p \le p_K - 2 \\ 1 \le q \le p_K - p - 1 \end{array} \right\},\\
\Phi_5^{(Lag)} &:= \left\{ f_1(\eta_1) f_1(\eta_2) f_2(\eta_2) C_p l_p^{(N_1)}(\eta_1) C_q l_q^{(N_2)}(\eta_2) \middle| \begin{array}{l} 1 \le p \le p_K - 1 \\ 1 \le q \le p_K - 2 \end{array} \right\}.
\end{aligned}
$$

C_p, C_q bezeichnen frei wählbare Skalierungsfaktoren.

Definition 3.6.3 (Formfunktionen für $\mathcal{T}^3$)**.** Sei $\mathcal{T}^3$ das Referenztetraeder und $p(K) = (p_{AB},\dots,p_{CD},p_{ABC},\dots,p_{BCD},p_K)$ eine Polynomgradverteilung, wobei p_{AB} den der Kante AB zugeordneten Polynomgrad bezeichnet und Analoges für die weiteren Einträge von $p(K)$ gilt. Für $i = 1,2,3$ bezeichne $N_i = \{\eta_k^{(i)} \mid k = 1,\dots,p_K - 3\}$ eine Stützstellenmenge mit

$$-1 < \eta_1^{(i)} < \ldots < \eta_{p_K-3}^{(i)} < 1$$

und es bezeichne $l_k^{(N_i)}$ das k-te Lagrange-Interpolationspolynom bezüglich der Stützstellenmenge N_i. Mit den Abkürzungen aus 3.6.1 definieren wir

$$\Psi^{(KS)} = \bigcup_{B=0}^{13} \Psi_B^{(KS)} \quad \text{und} \quad \Psi^{(Lag)} = \bigcup_{B=0}^{13} \Psi_B^{(Lag)},$$

wobei

$$
\begin{aligned}
\Psi_B^{(KS)} &:= \Phi_B^{(KS)} \circ D_3^{-1} := \left\{ \phi \circ D_3^{-1} \mid \phi \in \Phi_B^{(KS)} \right\},\\
\Psi_B^{(Lag)} &:= \Phi_B^{(Lag)} \circ D_3^{-1} := \left\{ \phi \circ D_3^{-1} \mid \phi \in \Phi_B^{(Lag)} \right\}
\end{aligned}
$$

und die Mengen $\Phi_B^{(KS)}$, $\Phi_B^{(Lag)}$ gegeben sind durch:

$$
\begin{aligned}
\Phi_0^{(KS)} = \Phi_0^{(Lag)} = \Phi_0 &:= \{f_3(\eta_3)\}\,,\\
\Phi_1^{(KS)} = \Phi_1^{(Lag)} = \Phi_1 &:= \{f_3(\eta_2)f_2(\eta_3)\}\,,\\
\Phi_2^{(KS)} = \Phi_2^{(Lag)} = \Phi_2 &:= \{f_2(\eta_1)f_2(\eta_2)f_2(\eta_3), f_3(\eta_1)f_2(\eta_2)f_2(\eta_3)\}\,,\\
\Phi_3^{(KS)} = \Phi_3^{(Lag)} = \Phi_3 &:= \left\{ f_1(\eta_1) P_{p-1}^{(1,1)}(\eta_1) f_2^{p+1}(\eta_2) f_2^{p+1}(\eta_3) \mid 1 \le p \le p_{AB}-1 \right\},\\
\Phi_4^{(KS)} = \Phi_4^{(Lag)} = \Phi_4 &:= \left\{ f_2(\eta_1) f_1(\eta_2) P_{q-1}^{(1,1)}(\eta_2) f_2^{q+1}(\eta_3) \mid 1 \le q \le p_{AC}-1 \right\},
\end{aligned}
$$

$$
\begin{aligned}
\Phi_5^{(KS)} = \Phi_5^{(Lag)} = \Phi_5 \;&:= \Big\{ f_3(\eta_1) f_1(\eta_2) P_{q-1}^{(1,1)}(\eta_2) f_2^{q+1}(\eta_3) \mid 1 \le q \le p_{BC} - 1 \Big\}, \\
\Phi_6^{(KS)} = \Phi_6^{(Lag)} = \Phi_6 \;&:= \Big\{ f_2(\eta_1) f_2(\eta_2) f_1(\eta_3) P_{r-1}^{(1,1)}(\eta_3) \mid 1 \le r \le p_{AD} - 1 \Big\}, \\
\Phi_7^{(KS)} = \Phi_7^{(Lag)} = \Phi_7 \;&:= \Big\{ f_3(\eta_1) f_2(\eta_2) f_1(\eta_3) P_{r-1}^{(1,1)}(\eta_3) \mid 1 \le r \le p_{BD} - 1 \Big\}, \\
\Phi_8^{(KS)} = \Phi_8^{(Lag)} = \Phi_8 \;&:= \Big\{ f_3(\eta_2) f_1(\eta_3) P_{r-1}^{(1,1)}(\eta_3) \mid 1 \le r \le p_{CD} - 1 \Big\}, \\
\Phi_9^{(KS)} = \Phi_9^{(Lag)} = \Phi_9 \;&:= \Big\{ f_1(\eta_1) P_{p-1}^{(1,1)}(\eta_1) f_2^p(\eta_2) f_1(\eta_2) P_{q-1}^{(2p+1,1)}(\eta_2) f_2^{p+q+1}(\eta_3) \Big| \\
&\qquad 1 \le p \le p_{ABC} - 2,\ 1 \le q \le p_{ABC} - p - 1 \Big\}, \\
\Phi_{10}^{(KS)} = \Phi_{10}^{(Lag)} = \Phi_{10} &:= \Big\{ f_1(\eta_1) P_{p-1}^{(1,1)}(\eta_1) f_2^{p+1}(\eta_2) f_1(\eta_3) f_2^p(\eta_3) P_{r-1}^{(2p+1,1)}(\eta_3) \Big| \\
&\qquad 1 \le p \le p_{ABD} - 2,\ 1 \le r \le p_{ABD} - p - 1 \Big\}, \\
\Phi_{11}^{(KS)} = \Phi_{11}^{(Lag)} = \Phi_{11} &:= \Big\{ f_2(\eta_1) f_1(\eta_2) P_{q-1}^{(1,1)}(\eta_2) f_1(\eta_3) f_2^q(\eta_3) P_{r-1}^{(2q+1,1)}(\eta_3) \Big| \\
&\qquad 1 \le q \le p_{ACD} - 2,\ 1 \le r \le p_{ACD} - q - 1 \Big\}, \\
\Phi_{12}^{(KS)} = \Phi_{12}^{(Lag)} = \Phi_{12} &:= \Big\{ f_3(\eta_1) f_1(\eta_2) P_{q-1}^{(1,1)}(\eta_2) f_1(\eta_3) f_2^q(\eta_3) P_{r-1}^{(2q+1,1)}(\eta_3) \Big| \\
&\qquad 1 \le q \le p_{BCD} - 2,\ 1 \le r \le p_{BCD} - q - 1 \Big\}, \\
\Phi_{13}^{(KS)} \qquad &:= \Big\{ f_1(\eta_1) P_{p-1}^{(1,1)}(\eta_1) f_1(\eta_2) f_2^p(\eta_2) P_{q-1}^{(2p+1,1)}(\eta_2) f_1(\eta_3) f_2^{p+q}(\eta_3) P_{r-1}^{(2p+2q+1,1)}(\eta_3) \Big| \\
&\qquad 1 \le p \le p_K - 3,\ 1 \le q \le p_K - p - 2,\ 1 \le r \le p_K - p - q - 1 \Big\}, \\
\Phi_{13}^{(Lag)} \qquad &:= \Big\{ f_1(\eta_1) C_p l_p^{(N_1)}(\eta_1) f_1(\eta_2) f_2(\eta_2) C_q l_q^{(N_2)}(\eta_2) f_1(\eta_3) f_2^2(\eta_3) C_r l_r^{(N_3)}(\eta_3) \Big| \\
&\qquad 1 \le p \le p_K - 3,\ 1 \le q \le p_K - 3,\ 1 \le r \le p_K - 3 \Big\}.
\end{aligned}
$$

C_p, C_q, C_r bezeichnen frei wählbare Skalierungsfaktoren.

Bemerkung 3.6.4. Eingeschränkt auf den Rand $\partial\mathcal{T}^d$ sind die Formfunktionen von $\Phi^{(KS)}$ und $\Phi^{(Lag)}$ identisch. Der wesentlichste Unterschied zwischen den beiden Mengen liegt bei den inneren Formfunktionen und in der Anzahl innerer Formfunktionen:

$$
\#\mathrm{INT}(LAG) = \begin{cases} (p_K - 1)(p_K - 2) & : \quad d = 2 \\ (p_K - 3)^3 & : \quad d = 3 \end{cases},
$$

$$
\#\mathrm{INT}(KS) = \begin{cases} \frac{1}{2}(p_K - 1)(p_K - 2) & : \quad d = 2 \\ \frac{1}{6}(p_K - 1)(p_K - 2)(p_K - 3) & : \quad d = 3 \end{cases}.
$$

Bemerkung 3.6.5. Die Formfunktionen aus $\Psi^{(KS)}$ besitzen bereits eine Tensorproduktstruktur, welche die Anwendung der Summenfaktorisierung erlaubt. Wollen wir jedoch den Spektral-Galerkin-Algorithmus aus [MGS01] auf die Elemente $\mathcal{T}^2$ und $\mathcal{T}^3$ verallgemeinern, so müssen wir zusätzlich in der Lage sein, die Formfunktionen und Quadraturregeln aneinander anzupassen. Zu diesem Zweck haben wir den assymptotisch größten Block der inneren Formfunktionen (Φ^5 bzw. Φ^{13}) zu Lagrange-Polynomen bezüglich noch näher zu definierenden Stützstellenmengen N_i modifiziert.

Bemerkung 3.6.6. Im 2D Fall haben wir zusätzlich zu den inneren Formfunktionen auch die Formfunktionen für die Kante AB leicht verändert. Diese Änderung lässt die Funktionen auf dem Rand unverändert, bewirkt jedoch, dass der Laufindex p jetzt ausschließlich für die η_1-Variable relevant ist. Diese Vereinfachung der Struktur ermöglicht uns später eine sehr effektive Umsetzung des Spektral-Galerkin-Algorithmus. Da für den 3D Fall eine solche Vereinfachung der Struktur nicht möglich wäre ohne die Formfunktionen auf $\partial\mathcal{T}^3$ signifikant zu verändern, beschränken wir uns hier auf die Modifikation der inneren Formfunktionen.

Die Unterteilung der Formfunktionen in verschiedene Gruppen geschieht nach einem für die *hp*-FEM üblichen Muster. Im Falle des Dreiecks haben wir Vertexformfunktionen $\psi \in \Psi_0 \cup \Psi_1$, Kantenformfunktionen $\psi \in \Psi_2 \cup \Psi_3 \cup \Psi_4$ und innere Formfunktionen $\psi \in \Psi_5$. Für das Tetraeder haben wir Vertexformfunktionen $\psi \in \Psi_0 \cup \Psi_1 \cup \Psi_2$, Kantenformfunktionen $\psi \in \Psi_3 \cup \ldots \cup \Psi_8$, Flächenformfunktionen $\psi \in \Psi_9 \cup \ldots \cup \Psi_{12}$ und innere Formfunktionen $\psi \in \Psi_{13}$. Vertexformfunktionen sind die üblichen linearen Formfunktionen, d.h sie sind Eins in genau einem Knoten und Null in allen anderen Knoten. Kantenformfunktionen sind Null in allen Knoten und verschwinden auf allen bis auf einer Kante identisch Null. Flächenformfunktionen sind Null in allen Knoten und verschwinden auf allen bis auf einer Seitenfläche identisch Null. Innere Formfunktionen sind Null auf dem Rand $\partial\mathcal{T}^d$.
Folgendes Lemma fasst die für uns wichtigsten Eigenschaften der Formfunktionen zusammen:

Lemma 3.6.7. *Sei K ein Element einer Vernetzung von $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, mit zugehöriger Polynomgradverteilung $p(K)$. Seien $\Psi^{(KS)}$ und $\Psi^{(Lag)}$ gegeben durch Definition 3.6.2 bzw. Definition 3.6.3. Dann gilt:*

1. *$\Psi^{(KS)}$ und $\Psi^{(Lag)}$ sind Mengen linear unabhängiger Funktionen.*

2. $\Pi_{p(K)}(\mathcal{T}^d) = \operatorname{span}\{\psi \mid \psi \in \Psi^{(KS)}\} \subset \operatorname{span}\{\psi \mid \psi \in \Psi^{(Lag)}\} =: \tilde{\Pi}_{p(K)}(\mathcal{T}^d)$.

3. *Alle $\psi \in \Psi^{(KS)}$ sind Polynome.*

4. *Für beliebiges $\psi = \psi(\xi) \in \Psi^{(Lag)}$ und $i,j \in \{1,..,d\}$ sind*

$$\left[\frac{\partial\psi}{\partial\xi_i}\right] \circ D_d \quad \text{und} \quad \left[\frac{\partial^2\psi}{\partial\xi_i\partial\xi_j}\right] \circ D_d$$

Polynome.

Beweis. Da die Eigenschaften bezüglich $\Psi^{(KS)}$ bereits in [KS99] gezeigt sind, können wir uns auf die Aussagen über $\Psi^{(Lag)}$ beschränken. Die lineare Unabhängigkeit der Funktionen $\psi \in \Psi^{(Lag)}$ ist offensichtlich. Die Mengen $\Psi^{(Lag)}$ und $\Psi^{(KS)}$ unterscheiden sich in den inneren Formfunktionen und für $d = 2$ zusätzlich in den Kantenformfunktionen zur Kante AB. Um zu zeigen, dass $\Pi_{p(K)}(\mathcal{T}^d) \subset \operatorname{span}\{\psi \mid \psi \in \Psi^{(Lag)}\}$, reicht es damit zu zeigen, dass $\Phi_2^{(KS)} \cup \Phi_5^{(KS)} \subset \Phi^{(Lag)}$ für $d = 2$ und $\Phi_{13}^{(KS)} \subset \Phi^{Lag}$ für $d = 3$. Betrachten wir ein beliebiges $\phi \in \Phi_{13}^{(KS)}$, so gilt mit $p \le p_K - 3$, $p+q \le p_K - 2$, $p+q+r \le p_K - 1$ und $p,q,r \ge 1$ (siehe

Definition 3.6.3)

$$\begin{aligned}\phi &= \left(\frac{1-\eta_1}{2}\right)\left(\frac{1+\eta_1}{2}\right)P_{p-1}^{(1,1)}(\eta_1)\left(\frac{1-\eta_2}{2}\right)^{p+1}\left(\frac{1+\eta_2}{2}\right)P_{q-1}^{(2p+1,1)}(\eta_2)\times \\ &\quad \left(\frac{1-\eta_3}{2}\right)^{p+q+1}\left(\frac{1+\eta_3}{2}\right)P_{r-1}^{(2p+2q+1,1)}(\eta_3) \\ &= \left(\frac{1-\eta_1}{2}\right)\left(\frac{1+\eta_1}{2}\right)P_{p_K-4}^{(1)}(\eta_1)\left(\frac{1-\eta_2}{2}\right)^{2}\left(\frac{1+\eta_2}{2}\right)P_{p_K-4}^{(2)}(\eta_2)\times \\ &\quad \left(\frac{1-\eta_3}{2}\right)^{3}\left(\frac{1+\eta_3}{2}\right)P_{p_K-4}^{(3)}(\eta_3),\end{aligned}$$

wobei $P_p^{(k)}$ für Polynome vom Grade kleiner oder gleich p stehen. Folglich müssen wir zeigen, dass $f := P_{p_K-4}^{(1)}(\eta_1)P_{p_K-4}^{(2)}(\eta_2)P_{p_K-4}^{(3)}(\eta_3)$ als eine Linearkombination

$$f = \sum_{pqr} c_{pqr} l_p^{(N_1)}(\eta_1) l_q^{(N_2)}(\eta_2) l_r^{(N_3)}(\eta_3)$$

mit geeigneten Koeffizienten c_{pqr} und den Lagrange-Polynomen $l_i^{(N_j)}(\eta_j)$ aus Definition 3.6.3 geschrieben werden kann. Dies ist jedoch leicht möglich, da die $l_i^{(N_j)}(\eta_j)$ mit $i = 1, \ldots, p_K - 3$ jeweils eine Basis des Polynomraums $\mathcal{P}_{p_K-4}[-1,1]$ bilden. Für $d = 2$ können wir völlig analog $\Phi_5^{(KS)} \subset \Phi^{(Lag)}$ beweisen und es bleibt zu zeigen $\Phi_2^{(KS)} \subset \Phi^{(Lag)}$. Für $\phi_p^{(KS)} \in \Phi_2^{(KS)}$ und $\phi_p^{(Lag)} \in \Phi_2^{(Lag)}$ gilt:

$$\begin{aligned}\phi_p^{(KS)} &= \left(\frac{1-\eta_1}{2}\right)\left(\frac{1+\eta_1}{2}\right)P_{p-1}^{(1,1)}(\eta_1)\left(\frac{1-\eta_2}{2}\right)^{p+1}, \\ \phi_p^{(Lag)} &= \left(\frac{1-\eta_1}{2}\right)\left(\frac{1+\eta_1}{2}\right)P_{p-1}^{(1,1)}(\eta_1)\left(\frac{1-\eta_2}{2}\right)^{2}.\end{aligned}$$

Da $\phi_1^{(KS)} = \phi_1^{(Lag)}$, betrachten wir

$$\phi_p^{(KS)} - \phi_p^{(Lag)} = \left(\frac{1-\eta_1}{2}\right)\left(\frac{1+\eta_1}{2}\right)P_{p-1}^{(1,1)}(\eta_1)\left(\frac{1-\eta_2}{2}\right)^{2}\left[\left(\frac{1-\eta_2}{2}\right)^{p-1} - 1\right]$$

für $p \geq 2$. Aus

$$\left(\frac{1-\eta_2}{2}\right)^{p-1} - 1 = \left(\frac{1+\eta_2}{2}\right)P_{p-2}(\eta_2),$$

wobei P_{p-2} wiederum ein Polynom mit maximalem Grade $p-2$ repräsentiert, folgt

$$\phi_p^{(KS)} - \phi_p^{(Lag)} = \left(\frac{1-\eta_1}{2}\right)\left(\frac{1+\eta_1}{2}\right)P_{p-1}^{(1,1)}(\eta_1)\left(\frac{1-\eta_2}{2}\right)^{2}\left(\frac{1+\eta_2}{2}\right)P_{p-2}(\eta_2).$$

Nach Definition 3.3.6 zusammen mit Definition 3.6.2 gilt $p \leq p_K - 1$, und wir können für geeignete Koeffizienten c_{pq} sowie den Lagrange-Polynomen $l_j^{(i)}(\eta_i)$, $j = 1, \ldots, p_K - i$, aus Definition 3.6.2

$$P_{p-1}^{(1,1)}(\eta_1)P_{p-2}(\eta_2) = P_{p_K-2}^{(1)}(\eta_1)P_{p_K-3}^{(2)}(\eta_2) = \sum_{pq} c_{pq} l_p^{(1)}(\eta_1) l_q^{(2)}(\eta_2)$$

schreiben. Damit ist $\Phi^{(KS)} \subset \Phi^{(Lag)}$ auch für $d = 2$ gezeigt und es bleibt Punkt 4. Da alle $\psi \in \Psi^{(KS)}$ Polynome sind (siehe [KS99]), können wir uns auf $\psi \in \Psi_2^{(Lag)} \cup \Psi_5^{(Lag)}$ für $d = 2$ bzw. $\psi \in \Psi_{13}^{(Lag)}$ für $d = 3$ beschränken. Für $D_d : \mathbb{R}^d \to \mathbb{R}^d$ mit $(\eta_1, \ldots, \eta_d) \mapsto (\xi_1(\eta_1, \ldots, \eta_d), \ldots, \xi_d(\eta_1, \ldots, \eta_d))$ gegeben durch Lemma 3.5.2 haben wir

$$\begin{aligned}
\left[\frac{\partial \psi}{\partial \xi_i}\right] \circ D_d &= \sum_{k=1}^{d} \frac{\partial \phi}{\partial \eta_k} \frac{\partial \eta_k}{\partial \xi_i} \circ D_d \\
\left[\frac{\partial^2 \psi}{\partial \xi_i \partial \xi_j}\right] \circ D_d &= \sum_{k,m=1}^{d} \left(\frac{\partial^2 \phi}{\partial \eta_k \partial \eta_m}\right) \left(\frac{\partial \eta_m}{\partial \xi_i} \frac{\partial \eta_k}{\partial \xi_j} \circ D_d\right) + \sum_{k=1}^{d} \left(\frac{\partial \phi}{\partial \eta_k}\right) \left(\frac{\partial^2 \eta_k}{\partial \xi_i \partial \xi_j} \circ D_d\right)
\end{aligned}$$

sowie

$$\begin{aligned}
\begin{bmatrix} \frac{\partial \eta_1}{\partial \xi_1} & \frac{\partial \eta_1}{\partial \xi_2} \\ \frac{\partial \eta_2}{\partial \xi_1} & \frac{\partial \eta_2}{\partial \xi_2} \end{bmatrix} \circ D_2 &= \begin{bmatrix} \frac{2}{1-\eta_2} & \frac{2(1+\eta_1)}{1-\eta_2} \\ 0 & 1 \end{bmatrix} \quad \text{für } d = 2 \\
\begin{bmatrix} \frac{\partial \eta_1}{\partial \xi_1} & \cdots & \frac{\partial \eta_1}{\partial \xi_3} \\ \vdots & \ddots & \vdots \\ \frac{\partial \eta_3}{\partial \xi_1} & \cdots & \frac{\partial \eta_3}{\partial \xi_3} \end{bmatrix} \circ D_3 &= \begin{bmatrix} \frac{4}{(1-\eta_2)(1-\eta_3)} & \frac{2(1+\eta_1)}{(1-\eta_2)(1-\eta_3)} & \frac{2(1+\eta_1)}{(1-\eta_2)(1-\eta_3)} \\ 0 & \frac{2}{(1-\eta_3)} & \frac{(1+\eta_2)}{(1-\eta_3)} \\ 0 & 0 & 1 \end{bmatrix} \quad \text{für } d = 3.
\end{aligned}$$

Weiterhin gilt für $d = 2$ (für $d = 3$ ergeben sich ganz analoge Formeln)

$$\frac{\partial \eta_2^2}{\partial \xi_i \partial \xi_j} = 0 \forall i, j \in 1, 2, \quad \frac{\partial \eta_1^2}{\partial \xi_1 \partial \xi_1} = 0, \quad \frac{\partial \eta_1^2}{\partial \xi_1 \partial \xi_2} = \frac{2}{(1-\xi_2)^2}, \quad \frac{\partial \eta_1^2}{\partial \xi_2 \partial \xi_2} = \frac{4(1+\xi_1)}{(1-\xi_2)^3},$$

d.h.

$$\frac{\partial \eta_1^2}{\partial \xi_1 \partial \xi_2} \circ D_2 = \frac{2}{(1-\eta_2)^2}, \quad \frac{\partial \eta_1^2}{\partial \xi_2 \partial \xi_2} \circ D_2 = \frac{2(1+\eta_1)}{(1-\eta_2)^2}.$$

Wie wir sehen, liefern die ersten und zweiten Ableitungen von η_k nach ξ_i zwar zum Teil Singularitäten der Form $(1-\eta_2)^{-l}$ und $(1-\eta_3)^{-l}$, andererseits sind jedoch alle $\phi \in \Phi_2^{(Lag)} \cup \Phi_5^{(Lag)}$ für $d = 2$ bzw. $\phi \in \Phi_{13}^{(Lag)}$ für $d = 3$ Polynome mit Faktoren $(1-\eta_2)$ und $(1-\eta_3)$ genügend hoher Ordnung, welche die Singularitäten wegkürzen. □

Bemerkung 3.6.8. Die in Lemma 3.6.7 gezeigte Eigenschaft, dass die auf den Referenzwürfel transformierten zweiten Ableitungen frei von Singularitäten sind, ist für die Anwendung des Residuenfehlerschätzers aus Kapitel 5 wichtig.

Korollar 3.6.9. *Sei K ein Element einer Vernetzung von $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, mit zugehöriger Polynomgradverteilung $p(K)$. Seien $\Psi^{(Lag/KS)}$ und $\Phi^{(Lag/KS)}$ gegeben durch Definition 3.6.2 bzw. Definition 3.6.3. Dann lassen sich die Einträge der Elementsteifigkeitsmatrix A_K (siehe (3.13)) berechnen als*

$$(A_K)_{ij} = \int_{\mathcal{Q}^d} \left\langle \tilde{\nabla} \phi_j, \hat{C} \tilde{\nabla} \phi_i \right\rangle |\det D_d'| d\Omega = \sum_{r,r'=1}^{d} \int_{\mathcal{Q}^d} \tilde{\nabla}_{r'} \phi_j \hat{C}_{r',r} \tilde{\nabla}_r \phi_i |\det D_d'| d\Omega, \quad (3.15)$$

mit

$$\tilde{\nabla}\phi_i := \begin{cases} \left[\frac{1}{(1-\eta_2)}\frac{\partial\phi}{\partial\eta_1}, \frac{\partial\phi}{\partial\eta_2}\right]^T & : \quad d=2 \\ \left[\frac{1}{(1-\eta_2)(1-\eta_3)}\frac{\partial\phi}{\partial\eta_1}, \frac{1}{(1-\eta_3)}\frac{\partial\phi}{\partial\eta_2}, \frac{\partial\phi}{\partial\eta_3}\right]^T & : \quad d=3 \end{cases}$$

elementweise polynomiell und

$$\hat{C} = M_d^{-1}(F_K')^{-1}(\hat{A}\circ F_K \circ D_d)(F_K')^{-T}M_d^{-T}|\det F_K'|,$$

$$M_2^{-1} := \begin{bmatrix} 2 & 2(1+\eta_1) \\ 0 & 1 \end{bmatrix}, \qquad M_3^{-1} := \begin{bmatrix} 4 & 2(1+\eta_1) & 2(1+\eta_1) \\ 0 & 2 & (1+\eta_2) \\ 0 & 0 & 1 \end{bmatrix}.$$

Approximationseigenschaften

Lemma 3.6.7 besagt, dass die Karniadakis & Sherwin-Formfunktionen in dem von den Lagrange-Formfunktionen aufgespannten Raum enthalten sind. Benutzen wir die Formfunktionen $\Phi^{(KS)}$ bzw. $\Phi^{(Lag)}$ in einer Finiten-Element-Implementation, so erwarten wir daher, dass

$$|u_{FEM}^{(Lag)} - u_{exakt}|_{H^1(\Omega)} \leq |u_{FEM}^{(KS)} - u_{exakt}|_{H^1(\Omega)}.$$

Folgendes einfache Beispiel soll dies verifizieren:

Beispiel 3.6.10. *Für $\Omega = \mathcal{T}^d$ sei*

$$-\Delta u = 1 \quad \textit{auf } \Omega \quad \textit{und} \quad u = 0 \quad \textit{auf } \partial\Omega.$$

Um die p-Abhängigkeit des H^1-Fehlers zu untersuchen, betrachten wir eine reine p-Methode, beginnend mit $p = 4$, auf Ω. Die Ergebnisse sind in Abbildung 3.2 dargestellt. Wie wir sehen, ist der Fehler bei Verwendung von $\Phi^{(Lag)}$ in der Tat kleiner als bei Verwendung von $\Phi^{(KS)}$.

Bemerkung 3.6.11. Die für die Rechnung benutzten Quadraturen mit $q_i = p_K$ sind nicht exakt. Die optimale Konvergenzordnung bleibt jedoch erhalten (siehe Abschnitt 3.10).

3.6.2 Algorithmen zum Aufstellen der Elementsteifigkeitsmatrizen

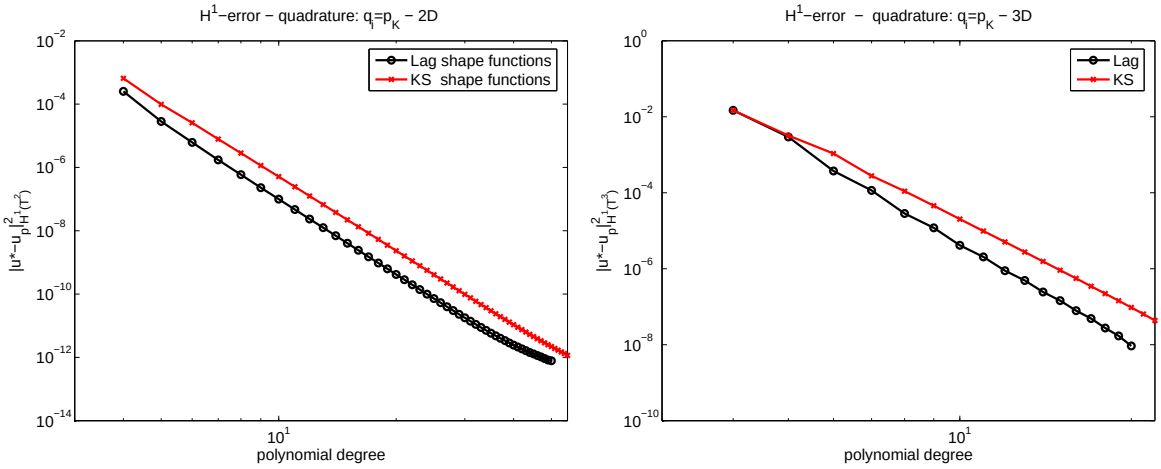
Nachdem wir verschiedene Mengen von Formfunktionen definiert haben, wollen wir uns nun den Algorithmen zum Aufstellen der lokalen Steifigkeitsmatrizen A_K zuwenden. Beginnen wir mit dem allereinfachsten Algorithmus.

Standardalgorithmus

Betrachten wir ein Element $K \in \mathcal{T}(\mathcal{N})$ mit Polynomgradverteilung $p(K)$ und gehen wir von einer zugehörigen Menge $\Psi = \{\psi_i \mid i = 1, \ldots, N\}$ von Formfunktionen auf dem Referenzelement $\mathcal{T}^d$ aus, so lässt sich die Elementsteifigkeitsmatrix A_K nach folgendem Algorithmus berechnen:

Algorithmus 3.6.12 (Standardalgorithmus)**.**

Abbildung 3.2: Approximationseigenschaft Beispiel 3.6.10



1. *Wähle für* $i = 1, \ldots, 3$ *geeignete Quadraturregeln*

$$\mathrm{QR}^{(i)} = S^{(i)} \times W^{(i)} = \{(\eta_0^{(i)}, \omega_0^{(i)}), \ldots, (\eta_{q_i}^{(i)}, \omega_{q_i}^{(i)})\}$$

welche die $\det |D_d'|$*-Terme als Gewichtsfunktion enthalten (siehe Gauß-Jacobi-Lobatto-Quadratur) und setze*

$$\mathrm{QR} = \mathrm{QR}^{(1)} \times \ldots \times \mathrm{QR}^{(d)}\,.$$

2. *Initialisiere* $A_K = 0$.

3. *Für alle* $(\eta^{(1)}, \ldots, \eta^{(d)}) \in S^{(1)} \times \ldots \times S^{(d)}$ *und zugehörigem Gewicht* $(\omega^{(1)}, \ldots, \omega^{(d)})$ *berechne*

$$A_K[i][j] \mathrel{+}= \omega^{(1)} \cdot \ldots \cdot \omega^{(d)} \left(\tilde{\nabla}(\psi_j \circ D_d), \tilde{A}\tilde{\nabla}(\psi_i \circ D_d)\right)\Big|_{(\eta^{(1)},\ldots,\eta^{(d)})}$$

für alle $1 \leq i, j \leq N$.

Wie wir sehen, beruht die Berechnung von A_K auf dem separaten und elementweisen Anwenden der Formel (3.14), wobei wir die in den Formeln auftretenden $\det |D_d'|$-Terme in die Quadraturregel integrieren. Algorithmus 3.6.12 ist leicht zu implementieren und stellt keine weiteren Forderungen an die Struktur der Formfunktionen. Nachteilig ist jedoch seine Komplexität mit $O(p_K^{3d})$ Operationen bei lediglich $O(p_K^d)$ Formfunktionen und Quadraturregeln $\mathrm{QR}^{(i)}$ der Ordnung $O(p_K)$. Es stellt sich daher die Frage, ob und wie wir den Aufwand für das Berechnen der Elementsteifigkeitsmatrix reduzieren können.

Summenfaktorisierung

Betrachten wir die Definitionen 3.6.2 und 3.6.3, so sehen wir, dass unsere auf den Referenzwürfel transformierten Formfunktionen $\phi = \psi \circ D_d$ von spezieller Tensorproduktstruktur sind. Sowohl für $\Psi^{(KS)}$ als auch $\Psi^{(Lag)}$ gilt:

$$\Psi \circ D_2 = \Phi = \left\{\phi_{(B,k_1,k_2)}(\eta_1, \eta_2) = g_{B,k_1}^{(1)}(\eta_1) g_{B,k_1,k_2}^{(2)}(\eta_2)\right\},$$

mit

$$0 \leq B \leq 5,\ 1 \leq k_1 \leq K_1(B),\ 1 \leq k_2 \leq K_2(B,k_1)$$

für $d = 2$. Beziehungsweise

$$\Psi \circ D_3 = \Phi = \left\{ \phi_{(B,k_1,k_2,k_3)}(\eta_1,\eta_2,\eta_3) = g^{(1)}_{B,k_1}(\eta_1) g^{(2)}_{B,k_1,k_2}(\eta_2) g^{(3)}_{B,k_1,k_2,k_3}(\eta_3) \right\},$$

mit

$$0 \leq B \leq 13,\ 1 \leq k_1 \leq K_1(B), 1 \leq k_2 \leq K_2(B,k_1), 1 \leq k_2 \leq K_3(B,k_1,k_2)$$

für $d = 3$. Folglich, da auch die Komponenten $(\tilde{\nabla}\phi_{(\cdot)})_r$ von $\tilde{\nabla}\phi_{(B,k_1,k_2)}$ und $\tilde{\nabla}\phi_{(B,k_1,k_2,k_3)}$ von gleicher Struktur sind:

$$\begin{aligned}
\left(\tilde{\nabla}\phi_{(B,k_1,k_2)}\right)_r &= \tilde{g}^{(1)}_{(B,r,k_1)}(\eta_1)\tilde{g}^{(2)}_{(B,r,k_1,k_2)}(\eta_2) && \text{für } d = 2,\\
\left(\tilde{\nabla}\phi_{(B,k_1,k_2,k_3)}\right)_r &= \tilde{g}^{(1)}_{(B,r,k_1)}(\eta_1)\tilde{g}^{(2)}_{(B,r,k_1,k_2)}(\eta_2)\tilde{g}^{(3)}_{(B,r,k_1,k_2,k_3)}(\eta_3) && \text{für } d = 3,
\end{aligned}$$

können wir uns diese Struktur zu Nutze machen und mittels Summenfaktorisierung einen Algorithmus von $O(p_K^{2d+1})$ Komplexität generieren. Hierzu betrachten wir den 3-dimensionalen Fall und Quadraturregeln

$$\mathrm{QR}^{(i)} = \{(\eta^{(i)}_{l_i}, \omega^{(i)}_{l_i}) \mid l_i = 0,\ldots,q_i)\}, \quad i = 1,\ldots,3,$$

welche die $\det|D_3'|$-Terme aus (3.15) als Gewichtsfunktionen enthalten, so bedeutet dies, dass wir mit den Abkürzungen

$$I := (B,k_1,k_2,k_3) \quad I' := (B',k_1',k_2',k_3')$$

und

$$\begin{aligned}
&J_1 := (B,r,k_1), \quad J_2 := (B,r,k_1,k_2), \quad J_3 := (B,r,k_1,k_2,k_3),\\
&J_1' := (B',r',k_1'), \quad J_2' := (B',r',k_1',k_2'), \quad J_3' := (B',r',k_1',k_2',k_3'),
\end{aligned}$$

die Berechnung von $A_K[I][I']$ als

$$A_K[I][I'] = \sum_{r,r'=1}^{3} \sum_{B,B'=0}^{13} \sum_{l_3=0}^{q_3} \omega^{(3)}_{l_3}\, \tilde{g}^{(3)}_{J_3}\tilde{g}^{(3)}_{J_3'} H^{(2)}_{r,r'}[k_1,k_1',k_2,k_2',l_3]\Big|_{\eta^{(3)}_{l_3}} \tag{3.16}$$

mit den vorab berechneten Hilfsfeldern

$$\begin{aligned}
H^{(2)}_{B,B',r,r'}[k_1,k_1',k_2,k_2',l_3] &= \sum_{l_2=0}^{q_2} \omega^{(2)}_{l_2}\tilde{g}^{(2)}_{J_2}\tilde{g}^{(2)}_{J_2'} H^{(1)}[k_1,k_1',l_3,l_2]\Big|_{(\eta^{(2)}_{l_2},\eta^{(3)}_{l_3})}\\
H^{(1)}_{B,B',r,r'}[k_1,k_1',l_3,l_2] &= \sum_{l_1=0}^{q_1} \tilde{g}^{(1)}_{J_1}\tilde{g}^{(1)}_{J_1'}\hat{C}_{r',r}\Big|_{(\eta^{(1)}_{l_1},\eta^{(2)}_{l_2},\eta^{(3)}_{l_3})}
\end{aligned} \tag{3.17}$$

durchführen. Hier der vollständige Algorithmus:

Algorithmus 3.6.13 (Summenfaktorisierung in 3D)**.**

1. *Wähle für* $i = 1, \ldots, 3$ *geeignete Quadraturregeln*

$$\mathrm{QR}^{(i)} = \{(\eta_{l_i}^{(i)}, \omega_{l_i}^{(i)}) \mid l_i = 0, \ldots, q_i)\},$$

welche die $\det|D_3'|$*-Terme aus (3.15) enthalten.*

2. *Für* $1 \le r \le 3$ *und* $0 \le B \le 13$ *sei*

$$\tilde{\nabla}_r \Phi_B = \Big\{ \tilde{g}^{(1)}_{(B,r,k_1)}(\eta_1) \tilde{g}^{(2)}_{(B,r,k_1,k_2)}(\eta_2) \tilde{g}^{(3)}_{(B,r,k_1,k_2,k_3)}(\eta_3) \;\Big|\; 1 \le k_1 \le K_1(B), 1 \le k_2 \le K_2(B,k_1), 1 \le k_3 \le K_3(B,k_1,k_2) \Big\}.$$

3. *Berechne für* $1 \le r \le 3$, $0 \le B \le 13$, $0 \le l_i \le q_i$, $1 \le k_1 \le K_1(B)$, $1 \le k_2 \le K_2(B,k_1)$, $1 \le k_3 \le K_3(B,k_1,k_2)$ *die Hilfsfelder*

$$\begin{aligned} G^{(1)}(B,r,k_1,l_1) &= \tilde{g}^{(1)}_{B,r,k_1}(\eta^{(1)}_{l_1}), \\ G^{(2)}(B,r,k_1,k_2,l_2) &= \tilde{g}^{(2)}_{B,r,k_1,k_2}(\eta^{(2)}_{l_2}), \\ G^{(3)}(B,r,k_1,k_2,k_3,l_3) &= \tilde{g}^{(3)}_{B,r,k_1,k_2,k_3}(\eta^{(3)}_{l_3}). \end{aligned}$$

4. *Berechne für* $1 \le r, r' \le 3$ *und* $0 \le l_i \le q_i$ *das Hilfsfeld*

$$C(r',r,l_1,l_2,l_3) = \tilde{C}_{(r',r)}(\eta^{(1)}_{l_1}, \eta^{(2)}_{l_2}, \eta^{(3)}_{l_3}).$$

5. *Initialisiere* $A_K = 0$.

6. *Setze*

$$\begin{aligned} &I_1 := (B,r,k_1,l_1), \quad I_2 := (B,r,k_1,k_2,l_2), \quad I_3 := (B,r,k_1,k_2,k_3,l_3), \\ &I_1' := (B',r',k_1',l_1), \; I_2' := (B',r',k_1',k_2',l_2), \; I_3' := (B',r',k_1',k_2',k_3',l_3). \end{aligned}$$

Berechne für $1 \le r, r' \le 3$ *und* $0 \le B, B' \le 13$*:*

$$H^{(1)}[k_1,k_1',l_3,l_2] = \sum_{l_1=0}^{q_1} G^{(1)}(I_1) G^{(1)}(I_1') C(r',r,l_1,l_2,l_3) \omega^{(1)}_{l_1},$$

$$H^{(2)}[k_1,k_1',k_2,k_2',l_3] = \sum_{l_2=0}^{q_2} G^{(2)}(I_2) G^{(2)}(I_2') H^{(1)}[k_1,k_1',l_3,l_2] \omega^{(2)}_{l_2},$$

wobei

$$\begin{aligned} &1 \le k_1 \le K_1(B), \; 1 \le k_2 \le K_2(B,k_1), \; 1 \le k_3 \le K_3(B,k_1,k_2), \\ &1 \le k_1' \le K_1(B'), 1 \le k_2' \le K_2(B',k_1'), 1 \le k_3' \le K_3(B',k_1',k_2'), \\ &0 \le l_i \le q_i. \end{aligned}$$

$$A_K[I][I'] += \sum_{l_3=0}^{q_3} G^{(3)}(I_3) G^{(3)}(I_3') H^{(2)}[k_1,k_1',k_2,k_2',l_3] \omega^{(3)}_{l_3}$$

für

$$\begin{aligned} &1 \le k_1 \le K_1(B), \; 1 \le k_2 \le K_2(B,k_1), \; 1 \le k_3 \le K_3(B,k_1,k_2), \\ &1 \le k_1' \le K_1(B'), 1 \le k_2' \le K_2(B',k_1'), 1 \le k_3' \le K_3(B',k_1',k_2'). \end{aligned}$$

Bemerkung 3.6.14. Für den 2-dimensionalen Fall berechnen wir mit den Abkürzungen $I := (B, k_1, k_2)$ und $I' := (B', k_1', k_2')$

$$A_K[I][I'] = \sum_{r,r'=1}^{2} \sum_{B,B'=0}^{5} \sum_{l_2}^{q_2} \omega_{l_2}^{(2)} \; \tilde{g}_{J_2}^{(2)} \tilde{g}_{J_2'}^{(2)} H_{B,B',r,r'}[k_1, k_1', l_2]\Big|_{\eta_{l_2}^{(2)}},$$

wobei

$$H_{B,B',r,r'}[k_1, k_1', l_2] \;=\; \sum_{l_1}^{q_1} \tilde{g}_{J_1}^{(1)} \tilde{g}_{J_1'}^{(1)} \hat{C}_{r',r}\Big|_{(\eta_{l_1}^{(1)}, \eta_{l_2}^{(2)})}$$

und erhalten das 2-dimensionale Analogon zu Algorithmus 3.6.13.

Bemerkung 3.6.15. Um weitere Rechenzeit einzusparen, könnte man für jedes Paar (B, B') eine separate Quadraturregel wählen.

Bemerkung 3.6.16. Das Auswerten und Zwischenspeichern der Formfunktionen und Koeffizientenmatrix in den Schritten 3 und 4 verhindert ein mehrfaches Berechnen dieser Größen. Insbesondere für höhere Polynomgrade und Koeffizienten von komplizierter Struktur kann damit deutlich Rechenzeit eingespart werden.

Bemerkung 3.6.17. Das Auswerten und Zwischenspeichern der Formfunktionen in Schritt 3 ist nicht für jedes Element $K \in \mathcal{T}(\mathcal{N})$ notwendig. Gehen wir davon aus, dass die benutzte Quadraturregel lediglich von dem dem Element K zugeordneten Polynomgrad p_K abhängt, so ist es, da $p_e, p_f \leq p_K$, ausreichend die Felder aus Schritt 3 für jeden Polynomgrad $p_K \in \{1, \ldots, p_{max}\}$ und angenommener uniformer Polynomgradverteilung (d.h. $p_e = p_f = p_k$ für alle Kanten und Seitenflächen) aufzustellen. Bei der Angabe von Rechenzeiten für das Generieren der Matrix A_K werden wir folglich auch die für die Schritte 1 bis 3 notwendige Rechenzeit vernachlässigen.

Spektral-Galerkin-Algorithmus

Im letzten Abschnitt haben wir bereits die Tensorproduktstruktur der Formfunktionen $\Phi^{(KS)}$ und $\Phi^{(Lag)}$ ausgenutzt. Unsere modifizierten Formfunktionen $\Phi^{(Lag)}$ bieten jedoch noch weitere Möglichkeiten, Rechenzeit einzusparen. Zum einen besitzt $\Phi^{(Lag)}$ gegenüber $\Phi^{(KS)}$ eine teilweise einfachere Struktur:

$$\Phi^{(Lag)} = \left\{ \phi_{(B,k_1,k_2)}(\eta) = g_{B,k_1}^{(1)}(\eta_1) g_{B,k_2}^{(2)}(\eta_2) \;\mid 1 \leq k_i \leq K_i(B) \right\} \tag{3.18}$$

für $d = 2$ und

$$\Phi_{13}^{(Lag)} = \left\{ \phi_{(13,k_1,k_2,k_3)}(\eta) = g_{13,k_1}^{(1)}(\eta_1) g_{13,k_2}^{(2)}(\eta_2) g_{13,k_3}^{(3)}(\eta_3) \mid 1 \leq k_i \leq K \right\} \tag{3.19}$$

für $d = 3$, wobei $g^{(i)}(\eta_i)$ nicht mehr von k_j mit $j \neq i$ abhängt. Zum anderen sind die inneren Formfunktionen von $\Phi^{(Lag)}$ an die Quadraturregel anpassbar, womit eine Verallgemeinerung der in [MGS01] beschriebenen Spektral-Galerkin-Methode auf Dreiecks- und Tetraederelemente möglich wird. Im Folgenden betrachten wir Quadraturregeln

$$\mathrm{QR}^i = S^{(i)} \times W^{(i)} = \{(\eta_0^{(i)}, \omega_0^{(i)}), \ldots, (\eta_{q_i}^{(i)}, \omega_{q_i}^{(i)})\}, \; \mathrm{QR} = \mathrm{QR}^1 \times \ldots \times \mathrm{QR}^d,$$

welche die $\det|D'_d|$-Terme enthalten, in Verbindung mit den Formfunktionen $\Phi^{(Lag)}$, bei denen die Stützstellenmengen $N^{(i)}$ aus Definition 3.6.2 bzw. Definition 3.6.3 Teilmengen der Quadraturstützstellen sind. D.h.

$$N^{(i)} \subset S^{(i)}.$$

Werten wir nun die Gradienten der inneren Formfunktionen in den Quadraturpunkten aus, so erhalten wir, auf Grund der bei der Definition der Formfunktionen verwendeten Lagrange-Interpolationspolynome und den mit den Quadraturen abgestimmten Stützstellen, eine beachtliche Anzahl von Nullen. Da wir in (3.16) und (3.17) natürlich nur die von Null verschiedenen Summanden berücksichtigen müssen, können wir alle Summen in Algorithmus 3.6.13 nach folgendem Muster ersetzen:

$$\sum_{l_1=0}^{q_1} \tilde{g}^{(1)}_{(B,r,k_1)} \tilde{g}^{(1)}_{(B',r',k'_1)} \hat{C}_{r',r}\Big|_{(\eta^{(1)}_{l_1},\eta^{(2)}_{l_2},\eta^{(3)}_{l_3})}$$
$$\rightarrow \sum_{l_i \in NZ^{(1)}_{(r,r')}[k_1,k'_1]} \tilde{g}^{(1)}_{(B,r,k_1)} \tilde{g}^{(1)}_{(B',r',k'_1)} \hat{C}_{r',r}\Big|_{(\eta^{(1)}_{l_1},\eta^{(2)}_{l_2},\eta^{(3)}_{l_3})},$$

wobei die Mengen

$$NZ^{(1)}_{(r,r')}[k_1,k'_1] := \{l_1 \in \{0,\dots,q^{(1)}\} \mid \tilde{g}^{(1)}_{(B,r,k_1)} \tilde{g}^{(1)}_{(B',r',k'_1)}\Big|_{\eta^{(1)}_{l_1}} \neq 0\},$$

mit geringem Aufwand vorab bestimmt werden können (analog für die Summationen über l_2 und l_3).
Während in (3.16) und (3.17) die Summationsreihenfolge bezüglich l_1, l_2, l_3 noch fest vorgegeben ist, kann bei Verwendung der $\Phi^{(Lag)}$ Formfunktionen, bedingt durch die vereinfachte Tensorproduktstruktur (siehe (3.18), (3.19)), diese Summationsreihenfolge speziell für $B = B' = 13$ in 3D sowie für alle Paarungen (B, B') in 2D ohne Schwierigkeiten beliebig permutiert werden.
Aus der Kenntnis der Anzahl von Nicht-Null-Elementen $\#NZ^{(i)}_{(r,r')}[k_1,k'_1]$ lässt sich leicht eine Schätzung für den Aufwand zum Aufstellen der Hilfsfelder H sowie für den Aufwand zum Aufstellen der Elementsteifigkeitsmatrix ermitteln. Wir können daher vorab die jeweils billigste Summationsreihenfolge auswählen. Für den Fall $B = B' = 13$ in 3D sowie für alle Paarungen (B, B') im 2-Dimensionalen kann Punkt 6 aus Algorithmus 3.6.13 somit modifiziert werden zu [2]

Algorithmus 3.6.18 (Spektral-Galerkin-Algorithmus für Tetraeder)**.**

6. Für $B = B' = 13$ und alle $1 \le r, r' \le 3$:

1. Berechne für $i = 1, ..., 3$, $1 \le k_i \le K_i(B)$, $1 \le k'_i \le K_i(B')$ und $0 \le l_i \le q_i$:

$$\begin{aligned} F^{(i)}[k_i,k'_i,l_i] &:= G^{(i)}(B,r,k_i,l_i)G^{(i)}(B',r',k'_i,l_i), \\ NZ^{(i)}[k_i,k'_i] &:= \{l_i \mid F^{(i)}[k_i,k'_i,l_i] \neq 0\}, \\ S_i &:= \sum_{k_i,k'_i} \#NZ^i[k_i,k'_i]. \end{aligned}$$

[2] Wir geben wiederum nur die 3D-Version explizit an. Die 2D-Version ergibt sich jedoch völlig analog.

2. *Schätze für alle Permutationen* (i_1, i_2, i_3) *von* $\{1, 2, 3\}$ *den Arbeitsaufwand für die Summationsreihenfolge* $\sum_{l_{i_1}=0}^{q_{i_1}} \sum_{l_{i_2}=0}^{q_{i_2}} \sum_{l_{i_3}=0}^{q_{i_3}}$*:*

$$\begin{aligned} W_{(i_1,i_2,i_3)} = \; & (q_{i_1}+1)(q_{i_2}+1)S_{i_3} + && \%\textit{Aufstellen von } H^{(1)} \\ & K_{i_3}(B)K_{i_3}(B')(q_{i_1}+1)S_{i_2} + && \%\textit{Aufstellen von } H^{(2)} \\ & K_{i_3}(B)K_{i_3}(B')K_{i_2}(B)K_{i_2}(B')S_{i_1} && \%\textit{Aufstellen von } A_K \end{aligned}$$

3. *Bestimme eine Permutation* (i_1, i_2, i_3) *mit* $W_{(i_1,i_2,i_3)} \leq W_{(i'_1,i'_2,i'_3)}$ *für alle* (i'_1, i'_2, i'_3)*.*

4. *Berechne die Hilfsfelder*

$$\begin{aligned} H^{(1)}[k_{i_3}, k'_{i_3}, l_{i_1}, l_{i_2}] \quad &= \sum_{l_{i_3} \in NZ^{(i_3)}} F^{(i_3)}[k_{i_3}, k'_{i_3}, l_{i_3}] C(r', r, l_1, l_2, l_3) \omega_{l_{i_3}}^{(i_3)}, \\ H^{(2)}[l_{i_1}, k_{i_3}, k'_{i_3}, k_{i_2}, k'_{i_2}] \quad &= \sum_{l_{i_2} \in NZ^{(i_2)}} F^{(i_2)}[k_{i_2}, k'_{i_2}, l_{i_2}] H^{(1)}[l_{i_1}, l_{i_2}, k_{i_3}, k'_{i_3}] \omega_{l_{i_2}}^{(i_2)} \end{aligned}$$

für $1 \leq k_i \leq K_i(B)$*,* $1 \leq k'_i \leq K_i(B')$ *und* $0 \leq l_i \leq q_i$*.*

5. *Addiere*

$$A_K[I][I'] \; += \sum_{l_{i_1} \in NZ^{(i_1)}} F^{(i_1)}[k_{i_1}, k'_{i_1}, l_{i_1}] H^{(2)}[l_{i_1}, k_{i_3}, k'_{i_3}, k_{i_2}, k'_{i_2}] \omega_{l_{i_1}}^{(i_1)}$$

für alle $1 \leq k_i \leq K_i(B)$*,* $1 \leq k'_i \leq K_i(B')$*.*

Für die Komplexitätsbetrachtung von Algorithmus 3.6.18 können wir in völliger Analogie auf die in [MGS01] betrachteten Fälle von Rechtecks- und Hexaederelementen verweisen.

Theorem 3.6.19. *Setzen wir für die Quadraturregeln* $\mathrm{QR}^{(i)}$ *Stützstellenanzahlen von* $p_K + q$ *mit* $q \geq 0$ *und* $q = O(1)$ *voraus, so erhalten wir eine Komplexität von* $O(p_K^{2d})$ *für das Aufstellen der Elementsteifigkeitsmatrix* A_K *mittels Spektral-Galerkin-Algorithmus.*

Beweis. Die Analyse von Algorithmus 3.6.18 erfolgt analog zu [MGS01]. Da in 3D jedoch nur der Fall $B = B' = 13$ mit Algorithmus 3.6.18 behandelt wird, müssen wir noch die Komplexität für die anderen Paarungen $(B, B') \neq (13, 13)$ ansehen. Diese Paarungen werden mittels Algorithmus 3.6.13 behandelt. Betrachten wir Algorithmus 3.6.13, so sehen wir, dass die Hilfsfelder $H^{(1)}$ und $H^{(2)}$ von maximal 5-Parametern mit einem Laufbereich von $O(p_k)$ abhängen. Zusammen mit den auftretenden Summationen über l_* erhalten wir daher eine maximale Komplexität von $O(p_K^6)$ für das Aufstellen der Hilfsfelder. Beim Aufstellen von $A[I][I']$ stecken in I und I' zwar die Parameter $k_1, \ldots, k_3$ und $k'_1, \ldots, k'_3$, jedoch haben, da wir nur Blöcke $(B, B') \neq (13, 13)$ betrachten, höchstens fünf dieser Parameter einen Laufbereich von $O(p_k)$. D.h. auch für das Assemblieren von $A[I][I']$ beträgt der Aufwand maximal $O(p_K^6)$. □

Asymptotisch ist der Spektral-Galerkin-Algorithmus somit der Summenfaktorisierung überlegen. Der kritische Punkt ist jedoch die statische Kondensation (siehe Abschnitt 3.7). Wollen wir in unserer *hp*-FEM-Implementation auf statische Kondensation zurückgreifen, so ist dies ein Prozess, dessen Aufwand mit $O((\#Int)^3)$ pro Element wesentlich von der Anzahl $\#Int$

der inneren Formfunktionen abhängt und asymptotisch den Assemblierungsprozess dominiert. Um zu sehen, ob es trotz vergrößerter Anzahl innerer Formfunktionen einen Bereich $p_K \in \{p_0, \ldots, p_1\}$ gibt, in dem es sich auch bei statischer Kondensation lohnt auf den Spektral-Galerkin-Algorithmus zurückzugreifen, bedarf es konkreter Testrechnungen (siehe Abschnitt 3.11).

3.7 Statische Kondensation

Ein in der *hp*-FEM verbreitetes Verfahren zur Dimensionsreduktion ist die statische Kondensation. Die Einteilung der Formfunktionen in externe $E = \{\text{Vertex, Kanten, Seitenflächen}\}$ und $I =$ innere Formfunktionen impliziert eine Blockstruktur sowohl in den Elementsteifigkeitsmatrizen A_K als auch in der globalen Steifigkeitsmatrix A:

$$A_K = \begin{bmatrix} A_K^{EE} & A_K^{EI} \\ A_K^{IE} & A_K^{II} \end{bmatrix} \qquad A = \begin{bmatrix} A^{EE} & A^{EI} \\ A^{IE} & A^{II} \end{bmatrix}.$$

Da innere Formfunktionen zu verschiedenen Elementen disjunkte Träger haben, ist A^{II} blockdiagonal mit $A^{II} = diag(A_K^{II})$.
Die Idee der statischen Kondensation besteht darin, in dem im Rahmen der Finiten-Element-Methode zu lösenden globalen Gleichungssystem

$$A\underline{u} = \underline{r}$$

die inneren Formfunktionen durch Bilden des Schur-Komplements

$$A^c = A^{EE} - A^{EI}(A^{II})^{-1}A^{IE}, \quad \underline{r}^c = \underline{r}^E - A^{EI}(A^{II})^{-1}\underline{r}^I$$

zu eliminieren und damit ein in der Dimension reduziertes Gleichungssystem

$$A^c\underline{u}_E = \underline{r}^c \tag{3.20}$$

für die externen Freiheitsgrade zu erhalten. Haben wir (3.20) gelöst, so können wir anschließend auf Elementebene und völlig parallel die inneren Freiheitsgrade berechnen.
Analog zum Aufbau der Steifigkeitsmatrix A aus den Elementmatrizen A_K, kann die kondensierte Steifigkeitsmatrix A^c durch Assemblieren der kondensierten Elementsteifigkeitsmatrizen gebildet werden:

$$A^c = \mathcal{A}A_K^c \quad \text{mit} \quad A_K^c = A_K^{EE} - A_K^{EI}(A_K^{II})^{-1}A_K^{IE}.$$

Der größte Nachteil der statische Kondensation ist ihre Komplexität. Die lokale statische Kondensation für A_K ist ein $O((\#Int)^3) \sim O(p_K^{3d})$ Prozess und wird für große Polynomgrade p_K somit unweigerlich das Aufstellen der Matrix A_K dominieren. Durch Rückgriff auf speziell optimierte Lapack-Routinen:

1. „dposv“ um $A_K^{II}X = A_K^{IE}$ zu lösen,
2. „dgemm“ um $A_K^c = A_K^{EE} - A_K^{EI}X$ zu berechnen,

gehen wir jedoch davon aus, dass sich der Einfluss der statischen Kondensation erst bei höheren Polynomgraden bemerkbar machen wird.

3.8 Konstante Koeffizienten - „precomputed arrays“

Bisher haben wir den allgemeinen Fall einer Differentialgleichung mit variablen Koeffizienten und recht allgemeinen Elementtransformationen F_K betrachtet. In diesem Abschnitt wollen wir noch kurz auf Vereinfachungsmöglichkeiten bei stückweise konstanten Koeffizienten in Verbindung mit affinen Elementtransformationen eingehen.
Zu einer beliebigen zulässigen Polynomgradverteilung $p(K) = (p_{e_1}, p_{e_2}, p_{e_3}, p_K)$ für 2D bzw. $p(K) = (p_{e_1}, \ldots, p_{e_6}, p_{f_1}, \ldots, p_{f_4}, p_K)$ für 3D (siehe Definition 3.3.6) sei Ψ die Menge zugehöriger Formfunktionen auf $\mathcal{T}^d$

$$\Psi = \begin{cases} \Psi^V \cup \Psi^{E_1}_{p_{e_1}} \cup \ldots \cup \Psi^{E_3}_{p_{e_3}} \cup \Psi^I_{p_K} & : \ 2D \\ \Psi^V \cup \Psi^{E_1}_{p_{e_1}} \cup \ldots \cup \Psi^{E_6}_{p_{e_6}} \cup \Psi^{F_1}_{p_{f_1}} \cup \ldots \cup \Psi^{F_4}_{p_{f_4}} \cup \Psi^I_{p_K} & : \ 3D \end{cases},$$

die sich in Vertex-, Kanten-, Seitenflächen- (d=3) und innere Formfunktionen unterteilen lässt. Ferner sollen folgende Hierarchien gelten:

$$\Psi^{E_i}_{p_{e_i-1}} \subset \Psi^{E_i}_{p_{e_i}}, \qquad \Psi^{F_i}_{p_{f_i-1}} \subset \Psi^{F_i}_{p_{f_i}},$$

d.h. die Menge der Formfunktion zu einer Kanten- bzw. Seitenfläche mit zugeordnetem Polynomgrad p ist stets eine Teilmenge der Menge der Formfunktion zu dieser Kanten- bzw. Seitenfläche mit Polynomgrad $\tilde{p} \geq p$.
Betrachten wir eine symmetrisch positiv definite Koeffizientenmatrix $\hat{A}$, welche auf $K \in \mathcal{T}(\mathcal{N})$ konstant ist, zusammen mit affinen Elementabbildungen F_K, so gilt

$$A_K = \left[\int_{\mathcal{T}^d} \left(\nabla\psi_i, \tilde{C} \nabla\psi_j \right) \mathrm{d}\Omega \right]_{i,j=1}^{N_{p(K)}}, \quad \tilde{C} = (F_K')^{-1} (\hat{A} \circ F_K)(F_K')^{-T} |\det F_K'|$$

wobei $\tilde{C}$ ebenfalls symmetrisch positiv definit und konstant auf $\mathcal{T}^d$ ist. Seien nun die Mengen $\{M_l\}_{l=1}^{(1/2)d(d+1)}$, $d = 2, 3$, bestehend aus symmetrisch positiv definiten Matrizen, Basen der symmetrischen Matrizen des $\mathbb{R}^{d\times d}$, z.B.

$$\{M_l\}_{l=1}^3 = \left\{ \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \right\}$$

für $d = 2$ und

$$\{M_l\}_{l=1}^6 = \left\{ \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}, \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix} \right\}$$

für $d = 3$, so lassen sich stets Koeffizienten c_l berechnen, so dass

$$\tilde{C} = \sum_{l=1}^{\frac{1}{2}d(d+1)} c_l M_l.$$

Folglich gilt

$$A_K = \sum_{l=1}^{\frac{1}{2}d(d+1)} c_l \left[\int_{T^d} (\nabla\psi_j, M_l \nabla\psi_i) \mathrm{d}\Omega \right]_{i,j=1}^{N_{p(K)}} =: \sum_{l=1}^{\frac{1}{2}d(d+1)} c_l A^{(l)}. \tag{3.21}$$

Berechnen wir die Matrizen $A^{(l)}$ zu Beginn des FE-Programms einmal für jede uniforme Polynomgradverteilung $p(K) = (p, \ldots, p)$ mit $1 \leq p \leq P_{max}$ im Voraus, wobei P_{max} den höchsten in der Vernetzung vorkommenden Polynomgrad bezeichnet, so können wir später die Elementsteifigkeitsmatrix A_K für jedes $K \in \mathcal{T}(\mathcal{N})$ mit konstanten Koeffizienten und affiner Elementtransformation F_K als Linearkombination dieser $A^{(l)}$ sowie dem Streichen von Zeilen und Spalten (falls $p_e, p_f < p$) bestimmen. Wie in Figur 3.4 und Figur 3.6 zu sehen ist, reduziert sich die Rechenzeit dabei drastisch.

Bemerkung 3.8.1. Da wir es mit konstanten Koeffizienten zu tun haben, können wir speziell für die Φ^{KS}- und Φ^{Lag}-Formfunktionen die vorab zu berechnenden Matrizen $A^{(l)}$ sogar analytisch bestimmen bzw. geeignete Quadraturformeln wählen, die $A^{(l)}$ exakt liefern. Im Gegensatz zu variablen Koeffizienten erhalten wir somit auch stets die exakte Steifigkeitsmatrix (im Rahmen der durch Computerarithmetik bedingten Rechengenauigkeit). Der Aufwand für das Vorabberechnen der Matrizen $A^{(l)}$ ist für das späterer Assemblieren der Steifigkeitsmatrix unbedeutend.

3.9 Matrix-Vektor-Multiplikation ohne Aufstellen der Steifigkeitsmatrix

In den letzten Abschnitten haben wir uns mit verschiedenen Methoden zum Generieren der Elementsteifigkeitsmatrizen A_K beschäftigt. Gehen wir jedoch davon aus, dass wir unser globales FE-Gleichungssystem mittels eines iterativen Lösers, wie zum Beispiel dem Verfahren der konjugierten Gradienten, lösen wollen, so müssen wir die globale Steifigkeitsmatrix A nicht explizit kennen. Stattdessen ist es völlig ausreichend, eine Matrix-Vektor-Multiplikation zu realisieren. D.h. wir müssen zu beliebigem Vektor w das Produkt

$$u = Aw = [\mathcal{A}_{K \in \mathcal{T}(\mathcal{N})} A_K] w = \sum_{K \in \mathcal{T}(\mathcal{N})} T_K A_K T_K^T w$$

bilden können. Mit $v := T_K^T w$ erkennen wir, dass wir hierbei vor der Aufgabe stehen für $K \in \mathcal{T}(\mathcal{N})$ eine Matrix-Vektor-Multiplikation

$$b = A_K v$$

mit beliebigem v möglichst effektiv durchzuführen. Im folgenden Abschnitt werden wir daher aufzeigen, wie solch eine Matrix-Vektor-Multiplikation ohne das explizite Generieren der Elementsteifigkeitsmatrix A_K realisiert werden kann.

3.9.1 Summenfaktorisierung

Wir beginnen mit der Idee der Summenfaktorisierung, was zu einem Algorithmus mit Komplexität $O(p_K^{d+1})$ pro Matrix-Vektor-Multiplikation führt. Wie bereits in den vorherigen Abschnitten sind die Fälle $d = 2$ und $d = 3$ hierbei so ähnlich, dass wir lediglich den Fall $d = 3$ im Detail beschreiben.

Bezeichne Ψ entweder die Menge der Formfunktionen $\Psi^{(KS)}$ oder $\Psi^{(Lag)}$, gegeben durch Definition 3.6.3, dann gilt

$$\psi_{(B,k_1,k_2,k_3)} = \phi_{(B,k_1,k_2,k_3)} \circ D_3^{-1}$$

mit

$$\phi_{(B,k_1,k_2,k_3)}(\eta_1,\eta_2,\eta_3) = g^{(1)}_{(B,k_1)}(\eta_1) g^{(2)}_{(B,k_1,k_2)}(\eta_2) g^{(3)}_{(B,k_1,k_2,k_3)}(\eta_3)$$

und

$$0 \le B \le 13, 1 \le k_1 \le K_1(B), 1 \le k_2 \le K_2(B,k_1), 1 \le k_3 \le K_3(B,k_1,k_2).$$

Legen wir für das Berechnen der Einträge von A_K auf dem Referenzwürfel $\mathcal{Q}^3$ die Quadraturregel

$$\mathrm{QR} = \mathrm{QR}^1 \times \mathrm{QR}^2 \times \mathrm{QR}^3$$

mit

$$\mathrm{QR}^i = S^{(i)} \times W^{(i)} = \{(\eta_0^{(i)}, \omega_0^{(i)}), \ldots, (\eta_{q_i}^{(i)}, \omega_{q_i}^{(i)})\}$$

zu Grunde, so erhalten wir mit den Abkürzungen $I = (B,k_1,k_2,k_3)$ und $I' = (B',k_1',k_2',k_3')$ und $\tilde{\nabla}$, $\hat{C}$ aus Korollar 3.6.9:

$$\begin{aligned} A_K &= \left[\sum_{l_1,l_2,l_3} \omega^{(1)}_{l_1} \omega^{(2)}_{l_2} \omega^{(3)}_{l_3} \left\langle \tilde{\nabla}\phi_{I'}, \hat{C}\tilde{\nabla}\phi_I \right\rangle \Big|_{(\eta^{(1)}_{l_1}, \eta^{(2)}_{l_2}, \eta^{(3)}_{l_3})} \right]_{I,I'} \\ &= \left[\sum_{l_1,l_2,l_3} \sum_{r,r'=1}^{3} \omega^{(1)}_{l_1} \omega^{(2)}_{l_2} \omega^{(3)}_{l_3} \left(\tilde{\nabla}_{r'}\phi_{I'} \hat{C}_{r',r} \tilde{\nabla}_r \phi_I \right) \Big|_{(\eta^{(1)}_{l_1}, \eta^{(2)}_{l_2}, \eta^{(3)}_{l_3})} \right]_{I,I'}. \end{aligned}$$

Damit berechnet sich der Vektor $b := A_K v$ zu

$$b_I = \sum_{(r,r',B',k_1',k_2',k_3')} \sum_{(l_1,l_2,l_3)} \omega^{(1)}_{l_1} \omega^{(2)}_{l_2} \omega^{(3)}_{l_3} \left\langle \tilde{\nabla}_{r'}\phi_{I'} \hat{C}_{r',r} \tilde{\nabla}_r \phi_I \right\rangle \Big|_{(\eta^{(1)}_{l_1}, \eta^{(2)}_{l_2}, \eta^{(3)}_{l_3})} v_{I'} \tag{3.22}$$

und für die Summationsreihenfolge

$$b_I = \sum_{l_3} \sum_{l_2} \sum_{r',l_1} \sum_{B',k_1'} \sum_{k_2'} \sum_{k_3'} \omega^{(1)}_{l_1} \omega^{(2)}_{l_2} \omega^{(3)}_{l_3} \; \tilde{\nabla}_{r'}\phi_{I'} \hat{C}_{r',r} \tilde{\nabla}_r \phi_I \Big|_{(\eta^{(1)}_{l_1}, \eta^{(2)}_{l_2}, \eta^{(3)}_{l_3})}$$

kann durch Herausziehen der von den Summationsvariablen unabhängigen Faktoren und Vorabberechnung geeigneter Hilfsfelder (Summenfaktorisierung) folgender Algorithmus mit Komplexität $O(p_K^{d+1})$ pro Matrix-Vektor-Multiplikation generiert werden:

Algorithmus 3.9.1 („on the fly“ Matrix-Vektor-Multiplikation 3D)**.**

1. *Wähle für* $i = 1, \ldots, 3$ *geeignete Quadraturregeln*

$$\mathrm{QR}^{(i)} = \{(\eta_{l_i}^{(i)}, \omega_{l_i}^{(i)}) \mid l_i = 0, \ldots, q_i)\},$$

welche die „$\det|D_3'|$*-Terme“ aus (3.15) enthalten.*

2. *Für* $1 \le r \le 3$ *und* $0 \le B \le 13$ *sei*

$$\tilde{\nabla}_r \Phi_B = \Big\{ \tilde{g}^{(1)}_{(B,r,k_1)}(\eta_1) \tilde{g}^{(2)}_{(B,r,k_1,k_2)}(\eta_2) \tilde{g}^{(3)}_{(B,r,k_1,k_2,k_3)}(\eta_3) \;\Big|\; 1 \le k_1 \le K_1(B), 1 \le k_2 \le K_2(B,k_1), 1 \le k_3 \le K_3(B,k_1,k_2) \Big\}.$$

3. *Berechne für* $1 \le r \le 3$, $0 \le B \le 13$, $0 \le l_i \le q_i$, $1 \le k_1 \le K_1(B)$, $1 \le k_2 \le K_2(B,k_1)$, $1 \le k_3 \le K_3(B,k_1,k_2)$ *die Hilfsfelder*

$$\begin{aligned} G^{(1)}(B,r,k_1,l_1) &= \tilde{g}^{(1)}_{B,r,k_1}(\eta^{(1)}_{l_1}), \\ G^{(2)}(B,r,k_1,k_2,l_2) &= \tilde{g}^{(2)}_{B,r,k_1,k_2}(\eta^{(2)}_{l_2}), \\ G^{(3)}(B,r,k_1,k_2,k_3,l_3) &= \tilde{g}^{(3)}_{B,r,k_1,k_2,k_3}(\eta^{(3)}_{l_3}). \end{aligned}$$

4. *Berechne für* $1 \le r, r' \le 3$ *und* $0 \le l_i \le q_i$ *das Hilfsfeld*

$$C(r',r,l_1,l_2,l_3) = \tilde{C}_{(r',r)}(\eta^{(1)}_{l_1}, \eta^{(2)}_{l_2}, \eta^{(3)}_{l_3}).$$

5. *Initialisiere* $b = 0$.

6. *Setze*

$$\begin{aligned} &I_1 := (B,r,k_1,l_1), \quad I_2 := (B,r,k_1,k_2,l_2), \quad I_3 := (B,r,k_1,k_2,k_3,l_3), \\ &I_1' := (B',r',k_1',l_1), \; I_2' := (B',r',k_1',k_2',l_2), \; I_3' := (B',r',k_1',k_2',k_3',l_3). \end{aligned}$$

Berechne der Reihe nach die Hilfsfelder

$$\begin{aligned} H^{(1)}[r',B',k_1',k_2',l_3] &= \sum_{k_3'} v_{(B',k_1',k_2',k_3')} G^{(3)}(I_3'), \\ H^{(2)}[r',B',k_1',l_2,l_3] &= \sum_{k_2'} H^{(1)}[r',B',k_1',k_2',l_3] G^{(2)}(I_2'), \\ H^{(3)}[r',l_1,l_2,l_3] &= \sum_{B',k_1'} H^{(2)}[r',B',k_1',l_2,l_3] G^{(1)}(I_1'), \\ H^{(4)}[r,B,k_1,l_2,l_3] &= \sum_{r',l_1} \omega^{(1)}_{l_1} H^{(3)}[r',l_1,l_2,l_3] C(r',r,l_1,l_2,l_3) G^{(1)}(I_1), \\ H^{(5)}[r,B,k_1,k_2,l_3] &= \sum_{l_2} \omega^{(2)}_{l_2} H^{(4)}[r,B,k_1,l_2,l_3] G^{(2)}(I_2) \end{aligned}$$

für $1 \le r, r' \le 3$, $0 \le B, B' \le 13$, $0 \le l_i \le q_i$ *und*

$$\begin{aligned} &1 \le k_1 \le K_1(B), \quad 1 \le k_2 \le K_2(B,k_1), \quad 1 \le k_3 \le K_3(B,k_1,k_2), \\ &1 \le k_1' \le K_1(B'), \quad 1 \le k_2' \le K_2(B',k_1'), \quad 1 \le k_3' \le K_3(B',k_1',k_2'). \end{aligned}$$

Berechne

$$b_{(B,k_1,k_2,k_3)} = \sum_{l_3} \omega_{l_3}^{(3)} H^{(5)}[r, B, k_1, k_2, l_3] G^{(3)}(I_3)$$

für $0 \le B \le 13,\ 1 \le k_1 \le K_1(B),\ 1 \le k_2 \le K_2(B, k_1),\ 1 \le k_3 \le K_3(B, k_1, k_2)$.

Bemerkung 3.9.2. Mit seiner Komplexität von $O(p_K^{d+1})$ pro Multiplikation ist Algorithmus 3.9.1 asymptotisch sogar schneller als eine Standard-Matrix-Vektor-Multiplikation mit Aufwand $O(p_K^{2d})$ für eine vollbesetzte $p_K^d \times p_K^d$ Matrix.

Der Algorithmus kann in zwei Abschnitte unterteilt werden. Die Schritte 1-4, welche die Initialisierung darstellen, und die Schritte 5 und 6, welche die eigentliche Matrix-Vektor-Multiplikation realisieren. Unabhängig von der Anzahl der durchzuführenden Multiplikationen muss die Initialisierung (Schritte 1-4) nur einmal durchgeführt werden. Betrachten wir ein FE-Netz $\mathcal{N}$, so brauchen die Hilfsfelder aus Schritt 2 sogar nur einmal für jeden auftretenden Polynomgrad p_K, $K \in \mathcal{T}(\mathcal{N})$, angelegt werden (siehe auch Bemerkung 3.6.17). Lediglich die Berechnung des symmetrischen Hilfsfeldes $C(r', r, l_1, l_2, l_3)$ mit seinen $\approx (1/2)3^2(q_1+1)(q_2+1)(q_3+1)$ Einträgen muss für jedes $K \in \mathcal{T}$ individuell vorgenommen werden. Da jedoch auch für Elemente mit hohem Polynomgrad der Speicheraufwand für $C(r', r, l_1, l_2, l_3)$ recht gering bleibt, lohnt es sich, insbesondere bei rechenintensiven Koeffizienten und mehreren Multiplikationen mit verschiedenen Vektoren, das Feld $C(r', r, l_1, l_2, l_3)$ nur einmal zu berechnen und kurzzeitig zwischenzuspeichern. Nehmen wir zum Beispiel ein Element mit $p_K = q_i = 10$ und 8 Byte pro Eintrag (double precision), so benötigen wir hierfür circa 50 KByte.

Bemerkung 3.9.3. Da, wie soeben dargelegt, die Initialisierungsschritte lediglich einmal, unabhängig von der Anzahl der Matrix-Vektor-Multiplikationen und zum Teil auch unabhängig von der Anzahl der Elemente $K \in \mathcal{T}(\mathcal{N})$ durchzuführen sind, werden wir die Rechenzeit für die Schritte 1-4 in Zukunft vernachlässigen.

Bemerkung 3.9.4. Berücksichtigen wir in Schritt 6 für die Formfunktionen Φ^{Lag} nur die von Null verschiedenen Summanden, so können wir weitere Rechenzeit einsparen. Aus implementatorischer Sicht führt dies jedoch zu nicht Cache-optimalen Sprüngen und Speicherzugriffen, welche einen Großteil der Ersparnisse zunichte machen.

Die Abbildungen 3.8, 3.9 und Tabellen 3.1-3.4 zeigen die Rechenzeit für die verschiedenen Formfunktionen und Methoden in Abhängigkeit vom Polynomgrad p_K. Zu beachten ist, dass hierbei nur die Rechenzeiten für die reine Matrix-Vektor-Multiplikation angegeben sind. Für die Multiplikation mittels „Blas“-Routinen ist daher noch zusätzlich das Aufstellen der Matrix A_K zu berücksichtigen.
Algorithmus 3.9.1 realisiert ein spezielle Summationsreihenfolge und es stellt sich die Frage, ob es nicht noch andere geeignete Reihenfolgen gibt. Für die Formfunktionen Φ^{KS} liefert das folgende Lemma die Antwort:

Lemma 3.9.5. *Seien $\phi^{(KS)}$ die Formfunktionen aus Definition 3.6.2 bzw. Definition 3.6.3. Die einzigen Summationsreihenfolgen, welche mittels Summenfaktorisierung auf eine Komplexität $O(p_K^{d+1})$ pro Matrix-Vektor-Multiplikation führen, sind $(l_3, l_2, l_1, k_1', k_2', k_3')$ für $d = 3$ bzw. (l_2, l_1, k_1', k_2') für $d = 2$. Alle anderen Summationsreihenfolgen führen zu einer Komplexität schlechter als $O(p_K^{d+1})$.*

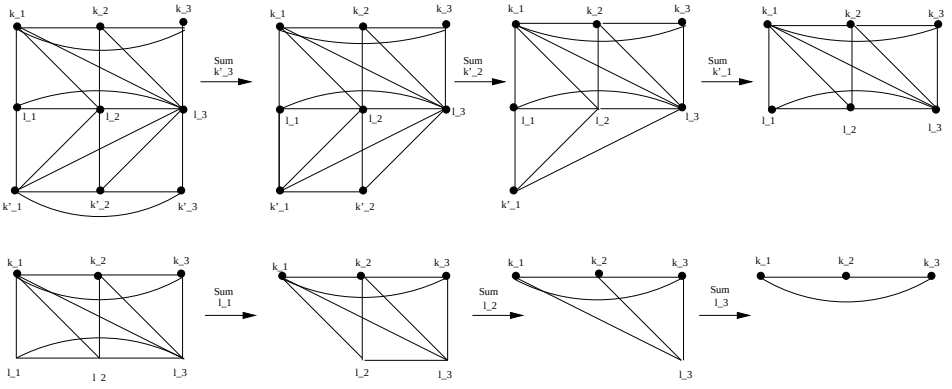
Beweis. Da die Beweise für $d = 2$ und $d = 3$ völlig analog sind, beschränken wir uns darauf, den 3-dimensionalen Fall darzustellen. Betrachten wir lediglich die inneren Formfunktionen,

d.h. B=B'=13, so zeigt der obere linke Graph in Abbildung 3.3 die Abhängigkeiten zwischen den Größen $\{k_1, k_2, k_3, l_1, l_2, l_3, k'_1, k'_2, k'_3\}$. Hierbei repräsentiert jeder Vertex eine dieser Größen und eine Kante zwischen den Vertizes V_1 und V_2 existiert genau dann, wenn in

$$b_{(B,k_1,k_2,k_3)} = \sum_{\{r,r',B'\}} \sum_{\{k'_1,k'_2,k'_3,l_1,l_2,l_3\}} \omega^{(1)}_{l_1}\omega^{(2)}_{l_2}\omega^{(3)}_{l_3} G^{(3)}_{B',r'}(k'_1,k'_2,k'_3,l_3) G^{(2)}_{B',r'}(k'_1,k'_2,l_2)$$
$$G^{(1)}_{B',r'}(k'_1,l_1)\hat{C}_{r',r}(l_1,l_2,l_3) G^{(1)}_{B,r}(k_1,l_1)$$
$$G^{(2)}_{B,r}(k_1,k_2,l_2) G^{(3)}_{B,r}(k_1,k_2,k_3,l_3) v_{(B',k'_1,k'_2,k'_3)}$$

ein Faktor existiert, der von beiden - durch V_1 und V_2 dargestellten - Größen abhängt. Zum Beispiel impliziert $G^{(3)}_{B',r'}(k'_1,k'_2,k'_3,l_3)$ die Kanten $\{k'_1,k'_2\}$, $\{k'_1,k'_3\}$, $\{k'_1,l_3\}$, $\{k'_2,k'_3\}$, $\{k'_2,l_3\}$ und $\{k'_3,l_3\}$. Summieren wir nun über die durch V dargestellte Größe, so bedeutet dies, dass wir ein Hilfsfeld anlegen, welches von allen Größen abhängt, welche zu V benachbart, d.h. mit V durch eine Kante verbunden sind. Da im Falle der inneren Formfunktionen für alle k_i, k'_i und l_i der Laufindex i einen Bereich $O(p)$ durchläuft, beträgt der Arbeitsaufwand für eine solche Summation und Anlegen des Hilfsfeldes jeweils $O(p^{1+N(V)})$, wobei $N(V)$ die Anzahl der Nachbarn von V bezeichnet. Folglich dürfen wir, um eine Komplexität nicht schlechter als $O(p^4)$ zu erhalten, niemals über Größen mit mehr als drei Nachbarknoten summieren. Abbildung 3.3 zeigt somit die eindeutig bestimmte Summationsreihenfolge, welche auf eine Komplexität $O(p^4)$ führt. □

Abbildung 3.3: Beweis zu Lemma 3.9.5



Bemerkung 3.9.6. Zu Lemma 3.9.5 gibt es eine „Ausnahme". Betrachten wir die Summationsreihenfolge

$$b_{(B,k_1,k_2,k_3)} = \sum_{(r,r',B',k'_1,k'_2,k'_3,l_*,l_*)} \sum_{(l_i)} \dots,$$

bei der wir mit der Summation über l_i starten, so sehen wir, dass die innere Summe vom Vektor $v_{(B',k'_1,k'_2,k'_3)}$ unabhängig ist und wir sie theoretisch vorab in einem Hilfsfeld H berechnen könnten. Da in dieser inneren Summe jedoch der Faktor $C(r',r,l_1,l_2,l_3)$ steckt, müssten wir bei einer Vernetzung $\mathcal{N}$ ein Feld $H^{(K)}$ für jedes Element $K \in \mathcal{T}(\mathcal{N})$ berechnen. Starten wir mit

- l_1, führt dies zu

$$H^{(K)}{}_{r,r',B,B'}[k_1,k_1',l_2,l_3] = \sum_{l_1} \omega_{l_1}^{(1)} \hat{C}_{r',r}(l_1,l_2,l_3) G_{B,r}^{(1)}(k_1,l_1) G_{B',r'}^{(1)}(k_1',l_1),$$

- l_2, führt dies zu

$$H^{(K)}{}_{r,r',B,B'}[k_1,k_1',k_2,k_2',l_1,l_3] = \sum_{l_2} \omega_{l_2}^{(2)} \hat{C}_{r',r}(l_1,l_2,l_3) G_{B,r}^{(2)}(k_1,k_2,l_2) G_{B',r'}^{(2)}(k_1',k_2',l_2),$$

- l_3, führt dies zu

$$H^{(K)}{}_{r,r',B,B'}[k_1,k_1',k_2,k_2',k_3,k_3',l_1,l_2] = \sum_{l_3} \omega_{l_3}^{(3)} \hat{C}_{r',r}(l_1,l_2,l_3) G_{B,r}^{(3)}(k_1,k_2,k_3,l_3) G_{B',r'}^{(3)}(k_1',k_2',k_3',l_3).$$

Wie wir sehen können, ist das Aufstellen des Hilfsfeldes in jedem Fall von höherer Komplexität als $O(p^4)$ und muss zudem für jedes Element $K \in \mathcal{T}(\mathcal{N})$ separat ausgeführt werden. Für höhere Polynomgrade und nur einige wenige Matrix-Vektor-Multiplikationen, wie beispielsweise der Fall bei guten Vorkonditionierern, ist es daher fraglich, ob ein Vorabberechnen der Felder $H^{(K)}$ nicht sogar zu höheren Rechenzeiten als beim Anwenden von Algorithmus 3.9.1 führt. Der andere Punkt ist, dass wir bereits für das Abspeichern von $H^{(K)}_{r,r',B,B'}[k_1,k_1',l_2,l_3]$ bei $p_K = l_i = 10$ circa $(1/2)\cdot 3^2 \cdot 6^2 \cdot 10^4 \cdot 8$ Byte ≈ 17 MByte Speicher einplanen müssten. Aus den eben genannten Gründen werden wir daher l_i als innerste Summationsvariablen ausschließen.

Bemerkung 3.9.7. Lemma 3.9.5 besagt, dass $(l_3,l_2,l_1,k_1',k_2',k_3')$ und (l_2,l_1,k_1',k_2') die eindeutig bestimmten besten Summationsreihenfolgen sind. Man könnte noch darüber nachdenken, ob eine variable Summationsreihenfolge, abhängig von B und B', zu Rechenzeiteinsparungen führt. Aufgrund der Faktoren, welche von mehreren k_i bzw. k_i' abhängen, ist dies für $\Phi^{(KS)}$ aber nicht praktikabel und würde allenfalls zu extensiven Fallunterscheidungen führen. Für den Fall unserer modifizierten Formfunktionen $\Phi^{(Lag)}$ werden wir diese Möglichkeit im nächsten Abschnitt näher untersuchen.

3.9.2 Beschleunigte Matrix-Vektor-Multiplikation durch Spektral-Galerkin-Ideen

In diesem Abschnitt wollen wir untersuchen, wie die einfachere Struktur von $\Phi^{(Lag)}$ (siehe Abschnitt 3.6.2) ausgenutzt werden kann, um gegenüber Algorithmus 3.9.1 mit seiner starren Summationsreihenfolge Rechenzeit einzusparen. Da im 3-dimensionalen Fall jedoch lediglich die inneren Formfunktionen eine veränderte, einfachere Struktur haben und deren Anteil am Gesamtaufwand sehr gering ist (siehe Abbildung 3.8), betrachten wir im Folgenden ausschließlich den 2-dimensionalen Fall. D.h. wir haben

$$\psi_{(B,k_1,k_2)} = \phi_{(B,k_1,k_2)} \circ D_2^{-1} \text{ mit } \phi_{(B,k_1,k_2)}(\eta_1,\eta_2) = g^{(1)}_{(B,k_1)}(\eta_1) g^{(2)}_{(B,k_2)}(\eta_2)$$

und

$$0 \leq B \leq 5, \quad 1 \leq k_1 \leq K_1(B), \quad 1 \leq k_2 \leq K_2(B).$$

Die Hauptidee für eine Beschleunigung gegenüber Algorithmus 3.9.1 ist, ähnlich zum Spektral-Galerkin-Algorithmus für das Aufstellen der Elementsteifigkeitsmatrix, mittels variabler Summationsreihenfolge die einfachere Struktur von $\phi \in \Phi^{(Lag)}$ und die Anpassbarkeit der inneren Formfunktionen an die Quadratur auszunutzen. Betrachten wir

$$b_{(B,k_1,k_2)} = \sum_{\{r,r',B'\}} \sum_{\{k_1',k_2',l_1,l_2\}} \omega_{l_1}^{(1)} \omega_{l_2}^{(2)} G_{B',r'}^{(2)}(k_2', l_2) G_{B',r'}^{(1)}(k_1', l_1) \hat{C}_{r',r}(l_1, l_2) \\ G_{B,r}^{(1)}(k_1, l_1) G_{B,r}^{(2)}(k_2, l_2) v_{(B',k_1',k_2')},$$

mit $G_{B,r}^{(i)}(k_i, l_i)$, $\hat{C}_{r',r}(l_1, l_2)$ und $v_{(B',k_1',k_2')}$ analog zum letzten Abschnitt, so gibt es 24 potentielle Summationsreihenfolgen für $\{k_1', k_2', l_1, l_2\}$. Unser Bestreben soll daher sein, in Abhängigkeit von B, B', r, r' die jeweils günstigste (oder zumindest eine gute) Reihenfolge für $\{k_1', k_2', l_1, l_2\}$ zu finden. Einige Summationsreihenfolgen können wir von vornherein ausschließen:

Lemma 3.9.8. *Für alle $i, j \in \{1, 2\}$ ist die Summationsreihenfolge $(l_j, k_i', *, *)$ mindestens genauso effizient wie $(k_i', l_j, *, *)$.*

Beweis. Aus Symmetriegründen betrachten wir o.B.d.A $(k_1', l_j, *, *)$. Für $j = 2$ ergeben sich hierbei die Fälle (k_1', l_2, l_1, k_2') und (k_1', l_2, k_2', l_1). Beide führen zu

$$b_{(B,k_1,k_2)} = \sum_{B',r,r'} \sum_{k_1'} \sum_{l_2} \omega_{l_2}^{(2)} G_{B,r}^{(2)}(k_2, l_2) H_{B',r'}^{(1)}(k_1, k_1', l_2) H_{B,B',r,r'}^{(2)}(k_1', l_2)$$

mit

$$H_{B',r'}^{(1)}(k_1', l_2) = \sum_{k_2'} G_{B',r'}^{(2)}(k_2', l_2) v_{(B',k_1',k_2')}$$

und

$$H_{B,B',r,r'}^{(2)}(k_1, k_1', l_2) = \sum_{l_1} G_{B,r}^{(1)}(k_1, l_1) G_{B',r'}^{(1)}(k_1', l_1) \omega_{l_1}^{(1)} \hat{C}_{r',r}(l_1, l_2).$$

Offensichtlich, da es, abgesehen von $\omega_{l_2}^{(2)}$, nicht möglich ist, einen Faktor vor $\sum_{l_2}$ zu ziehen, können wir die Summation von $\sum_{k_1'} \sum_{l_2}$ auch zu $\sum_{l_2} \sum_{k_1'}$ abändern. Für $j = 1$ erhalten wir

$$b_{(B,k_1,k_2)} = \sum_{B',r,r'} \sum_{k_1'} \sum_{l_1} \omega_{l_1}^{(1)} G_{B,r}^{(1)}(k_1, l_1) G_{B',r'}^{(1)}(k_1', l_1) H_{B,B',r,r'}^{(3)}(k_2, k_1', l_1)$$

und wir können wiederum an Stelle von $\sum_{k_1'} \sum_{l_1}$ auch $\sum_{l_1} \sum_{k_1'}$ setzten, ohne dass die Summation kostspieliger wird. □

Sei $i, \overline{i}, j, \overline{j} \in \{1, 2\}$ mit $i \neq \overline{i}$ und $j \neq \overline{j}$, dann wollen wir zusätzlich folgende Summationsreihenfolgen ausschließen:

- $(k_i', k_{\overline{i}}', l_j, l_{\overline{j}})$ da dies dem Aufstellen eines Blockes der Steifigkeitsmatrix entspricht,

- $(*, *, k'_i, l_j)$ wegen der Überlegungen aus Bemerkung 3.9.6,
- $(l_i, k'_{\bar{i}}, l_{\bar{i}}, k'_i)$ da eine effiziente Umsetzung analog zu $(l_i, k'_{\bar{i}}, k'_i, l_{\bar{i}})$ wäre.

Insgesamt verbleiben somit:

- Vier Permutationen des Typs $(l_j, l_{\bar{j}}, k'_i, k'_{\bar{i}})$,
- Zwei Permutationen des Typs $(l_j, k'_j, l_{\bar{j}}, k'_{\bar{j}})$.

Betrachten wir die Reihenfolge (l_1, k'_1, l_2, k'_2), so erhalten wir

$$b_{(B,k_1,k_2)} = \sum_{l_1} \omega^{(1)}_{l_1} G^{(1)}_{B,r}(k_1, l_1) \sum_{r',B',k'_1} G^{(1)}_{B',r'}(k'_1, l_1) H^{(2)}_{r,r',B,B'}(k_2, k'_1, l_1),$$

wobei

$$\begin{aligned} H^{(2)}_{r,r',B,B'}(k_2, k'_1, l_1) &= \sum_{l_2} H_{r',B'}(k'_1, l_2) \hat{C}_{r',r}(l_1, l_2) G^{(2)}_{B,r}(k_2, l_2) \\ H_{r',B'}(k'_1, l_2) &= \sum_{k'_2} \omega^{(2)}_{l_2} G^{(2)}_{B',r'}(k'_2, l_2) v_{(B',k'_1,k'_2)}. \end{aligned}$$

Da $G^{(1)}_{B,r}(k_1, l_1)$ ausschließlich von Größen abhängt, wie sie bereits im Hilfsfeld $H^{(2)}$ auftauchen, können wir auch auf das Aufstellen von $H^{(2)}$ verzichten und die Summation mit annähernd gleichem Aufwand als

$$b_{(B,k_1,k_2)} = \sum_{l_1} \omega^{(1)}_{l_1} G^{(1)}_{B,r}(k_1, l_1) \sum_{r',B',k'_1} \sum_{l_2} G^{(1)}_{B',r'}(k'_1, l_1) H_{r',B'}(k'_1, l_2) \hat{C}_{r',r}(l_1, l_2) G^{(2)}_{B,r}(k_2, l_2),$$

auswerten. Dies ist vom Aufwand her jedoch wiederum äquivalent zu

$$b_{(B,k_1,k_2)} = \sum_{l_1} \omega^{(1)}_{l_1} G^{(1)}_{B,r}(k_1, l_1) \sum_{l_2} \sum_{r',B',k'_1} G^{(1)}_{B',r'}(k'_1, l_1) H_{r',B'}(k'_1, l_2) \hat{C}_{r',r}(l_1, l_2) G^{(2)}_{B,r}(k_2, l_2).$$

Analoge Überlegungen für (l_2, k'_2, l_1, k'_1) zeigen uns, dass wir unseren Algorithmus auf die Betrachtung der vier verbleibenden Permutationen vom Typ $(l_j, l_{\bar{j}}, k'_i, k'_{\bar{i}})$ einschränken können. Wir präsentieren nun den Algorithmus und geben anschließend noch eine kurz Erläuterung.

Algorithmus 3.9.9 (Matrix-Vektor-Multiplikation (Spektral-Galerkin-Algorithmus))**.**

1. *Wähle für* $i = 1, 2$ *geeignete Quadraturregeln*

   $$\mathrm{QR}^{(i)} = \{(\eta^{(i)}_{l_i}, \omega^{(i)}_{l_i}) \mid l_i = 0, \ldots, q_i)\},$$

   welche die $\det |D'_2|$*-Terme aus (3.15) enthalten.*

2. *Für alle* $1 \leq r \leq 2$ *und* $0 \leq B \leq 5$ *sei*

   $$\tilde{\nabla}_r \Phi_B = \left\{ \tilde{g}^{(1)}_{(B,r,k_1)}(\eta_1) \tilde{g}^{(2)}_{(B,r,k_2)}(\eta_2) \;\middle|\; 1 \leq k_i \leq K_i(B) \right\}.$$

3. *Sei* $I_i = (B, r, k_i, l_i), \quad I_i = (B', r', k_i', l_i)$.
Berechne für $1 \le r \le 2$, $0 \le B \le 5$, $0 \le l_i \le q_i$, $1 \le k_i \le K_i(B)$:

$$G^{(1)}(I_1) = \tilde{g}^{(1)}_{B,r,k_1}(\eta^{(1)}_{l_1}), \qquad NZ^{(1)}(B, r, k_1) = \{l_1 \mid G^{(1)}(I_1) \neq 0\},$$
$$G^{(2)}(I_2) = \tilde{g}^{(2)}_{B,r,k_2}(\eta^{(2)}_{l_2}), \qquad NZ^{(2)}(B, r, k_2) = \{l_2 \mid G^{(2)}(I_2) \neq 0\}.$$

4. *Berechne für* $1 \le r, r' \le 3$ *und* $0 \le l_i \le q_i$ *das Hilfsfeld*

$$C(r', r, l_1, l_2) = \tilde{C}_{(r',r)}(\eta^{(1)}_{l_1}, \eta^{(2)}_{l_2}).$$

5. *Initialisiere* $b = 0$, $H^{(2)}[r', l_2, l_1] = 0$.

6. *Berechne für* $1 \le r' \le 2$, $0 \le l_i \le q_i$

$$H^{(2)}[r', l_2, l_1] = \sum_{B', k_1', k_2'} v_{(B', k_1', k_2')} \omega^{(1)}_{l_1} \omega^{(2)}_{l_2} G^{(1)}(I_1') G^{(2)}(I_2')$$

wie folgt: Für alle $1 \le r' \le 2$, $0 \le B' \le 5$ *berechne*

(a) $S_1 := \sum_{k_1'} \# NZ^{(1)}(B', r', k_1')$, $S_2 := \sum_{k_2'} \# NZ^{(2)}(B', r', k_2')$.

(b) *If* $[(q_2+1)S_1 + K_1(B')S_2] \le [(q_1+1)S_2 + K_2(B')S_1]$

Setze $H^{(1)}[k_1', l_2] = 0$.
Addiere $H^{(1)}[k_1', l_2] += v_{(B', k_1', k_2')} \omega^{(2)}_{l_2} G^{(2)}(B', r', k_2', l_2)$
für alle $1 \le k_1' \le K_1(B')$, $1 \le k_2' \le K_2(B')$, $l_2 \in NZ^{(2)}(B', r', k_2)$.

Addiere $H^{(2)}[r', l_2, l_1] += \omega^{(1)}_{l_1} H^{(1)}[k_1', l_2] G^{(1)}(B', r', k_1', l_1)$
für alle $1 \le k_1' \le K_1(B')$, $l_1 \in NZ^{(1)}(B', r', k_1)$, $0 \le l_2 \le q_2$.

(c) *If* $[(q_2+1)S_1 + K_1(B')S_2] > [(q_1+1)S_2 + K_2(B')S_1]$

Setze $H^{(1)}[k_2', l_1] = 0$.
Addiere $H^{(1)}[k_2', l_1] += v_{(B', k_1', k_2')} \omega^{(1)}_{l_1} G^{(1)}(B', r', k_1', l_1)$
für alle $1 \le k_1' \le K_1(B')$, $1 \le k_2' \le K_2(B')$, $l_1 \in NZ^{(1)}(B', r', k_1)$.
Addiere $H^{(2)}[r', l_2, l_1] += \omega^{(2)}_{l_2} H^{(1)}[k_2', l_1] G^{(2)}(B', r', k_2', l_2)$
für alle $1 \le k_1' \le K_1(B')$, $l_1 \in NZ^{(1)}(B', r', k_1)$, $0 \le l_2 \le q_2$.

7. *Berechne*

$$b_{(B,k_1,k_2)} = \sum_{(r,r',l_1,l_2)} H^{(2)}[r', l_2, l_1] C(r', r, l_1, l_2) G^{(1)}(I_1) G^{(2)}(I_2)$$

wie folgt: Für alle $1 \le r \le 2$, $0 \le B \le 5$ *berechne*

(a) $S_1 := \sum_{k_1} \# NZ^{(1)}(B, r, k_1)$, $S_2 := \sum_{k_2} \# NZ^{(2)}(B, r, k_2)$.

(b) If $[2(q_2+1)S_1 + K_1(B)S_2] \leq [2(q_1+1)S_2 + K_2(B)S_1]$

$$
\begin{aligned}
&\textit{Setze} \quad H^{(3)}[k_1,l_2] = 0.\\
&\textit{Addiere} \quad H^{(3)}[k_1,l_2] += C(r',r,l_1,l_2)H^{(2)}[r',l_2,l_1]G^{(1)}(B,r,k_1,l_1)\\
&\textit{für alle } 1 \leq k_1 \leq K_1(B),\ l_1 \in NZ^{(1)}(B,r,k_1),\ 0 \leq r' \leq 2,\ 0 \leq l_2 \leq q_2.\\
&\textit{Addiere} \quad b_{(B,k_1,k_2)} += H^{(3)}[k_1,l_2]G^{(2)}(B,r,k_2,l_2)\\
&\textit{für alle } 1 \leq k_2 \leq K_2(B),\ l_2 \in NZ^{(2)}(B,r,k_2),\ 1 \leq k_1 \leq K_1(B).
\end{aligned}
$$

(c) If $[2(q_2+1)S_1 + K_1(B)S_2] > [2(q_1+1)S_2 + K_2(B)S_1]$

$$
\begin{aligned}
&\textit{Setze} \quad H^{(3)}[k_2,l_1] = 0.\\
&\textit{Addiere} \quad H^{(3)}[k_2,l_1] += C(r',r,l_1,l_2)H^{(2)}[r',l_2,l_1]G^{(2)}(B,r,k_2,l_2)\\
&\textit{für alle } 1 \leq k_2 \leq K_2(B),\ l_2 \in NZ^{(2)}(B,r,k_2),\ 0 \leq r' \leq 2,\ 0 \leq l_1 \leq q_1.\\
&\textit{Addiere} \quad b_{(B,k_1,k_2)} += H^{(3)}[k_2,l_1]G^{(1)}(B,r,k_1,l_1)\\
&\textit{für alle } 1 \leq k_1 \leq K_1(B),\ l_1 \in NZ^{(1)}(B,r,k_1),\ 1 \leq k_2 \leq K_2(B).
\end{aligned}
$$

Jede der vier möglichen Summationsreihenfolgen führt über das gleiche Zwischenhilfsfeld $H^{(2)}[r',l_2,l_1]$. Die Grundidee des Algorithmus besteht darin, sowohl für das Aufaddieren dieses Zwischenhilfsfeldes

$$H^{(2)}[r',l_2,l_1] = \sum_{B'} \sum_{\{k_1',k_2'\}} v_{(B',k_1',k_2')} \omega_{l_1}^{(1)} \omega_{l_2}^{(2)} G^{(1)}(I_1') G^{(2)}(I_2') \tag{3.23}$$

als auch für die anschließende Berechnung des Vektors

$$b_{(B,k_1,k_2)} = \sum_{\{r,r'\}} \sum_{\{l_1,l_2\}} H^{(2)}[r',l_2,l_1] C(r',r,l_1,l_2) G^{(1)}(I_1) G^{(2)}(I_2) \tag{3.24}$$

die in Abhängigkeit von (B',r') bzw. (B,r) jeweils effizientere Summationsreihenfolge von $\{k_1',k_2'\}$ bzw. $\{l_1,l_2\}$ zu benutzen. Unsere Entscheidung beruht hierbei auf dem vorherigen Abschätzen des Aufwands. Da wir stets nur über die Nicht-Null-Elemente zu summieren brauchen, erhalten wir mit

$$S_i(B,r) = \sum_{k_i} NZ^{(i)}(B,r,k_i), \quad S_i(B',r') = \sum_{k_i'} NZ^{(i)}(B',r',k_i') \tag{3.25}$$

einen Arbeitsaufwand von $W_H(B',r') = (q_2+1)S_1(B',r') + K_1(B')S_2(B',r')$ für

$$H^{(2)}[r',l_2,l_1] \quad += \sum_{k_1'} \omega_{l_1}^{(1)} H^{(1)}[k_1',l_2] \tilde{g}^{(1)}_{B',r',k_1'}(\eta_{l_1}^{(2)})$$

mit

$$H^{(1)}[k_1',l_2] \qquad = \sum_{k_2'} v_{(B',k_1',k_2')} \omega_{l_2}^{(2)} \tilde{g}^{(2)}_{B',r',k_2'}(\eta_{l_2}^{(2)}),$$

bzw. $W_H(B',r') = (q_1+1)S_2(B',r') + K_2(B')S_1(B',r')$ für

$$H^{(2)}[r',l_2,l_1] \quad += \sum_{k_2'} \omega_{l_2}^{(2)} H^{(1)}[k_2',l_1]\tilde{g}^{(1)}_{B',r',k_1'}(\eta_{l_1}^{(2)})$$

mit

$$H^{(1)}[k_1',l_2] \qquad = \sum_{k_1'} v_{(B',k_1',k_2')}\omega_{l_1}^{(1)}\tilde{g}^{(1)}_{B',r',k_1'}(\eta_{l_1}^{(1)}).$$

Analoge Betrachtungen bezüglich (3.24) ergeben daher einen Gesamtaufwand pro Matrix-Vektor-Multiplikation von

$$W_{Mv} = \sum_{B,r} W_b(B,r) + \sum_{B',r'} W_H(B',r'),$$

mit

$$\begin{aligned} W_H(B',r') = \min\{\ & (q_2+1)S_1(B',r') + K_1(B')S_2(B',r'), \\ & (q_1+1)S_2(B',r') + K_2(B')S_1(B',r')\ \}, \\ W_b(B,r) \quad = \min\{\ & 2(q_2+1)S_1(B,r) + K_1(B)S_2(B,r), \\ & 2(q_1+1)S_2(B,r) + K_2(B)S_1(B,r)\ \} \end{aligned}$$

und $S_i(B,r), S_i(B',r')$ aus (3.25). Wie wir sehen, ist die Komplexität von Algorithmus 3.9.9 zwar immer noch $O(p_K^3)$, jedoch erzielen wir eine deutliche Verbesserung der Rechenzeit gegenüber der 2D-Version von Algorithmus 3.9.1 (siehe Abbildung 3.8 und Tabelle 3.1, 3.2).

3.10 Bemerkungen zur Quadraturfehleranalyse

Dadurch, dass wir die Integrale für die Einträge der Elementsteifigkeitsmatrizen und der Elementlastvektoren im Allgemeinen nicht exakt analytisch ausrechnen können, sondern stattdessen auf numerische Quadraturverfahren zurückgreifen, erhalten wir zwangsläufig nur Näherungen der exakten Steifigkeitsmatrix und des exakten Lastvektors. D.h. wir lösen nicht das Problem:

$$\text{Finde } u \in \mathcal{V} \text{ mit:} \qquad a(u,v) = f(v) \quad \forall v \in \mathcal{V},$$

sondern:

$$\text{Finde } \tilde{u} \in \mathcal{V} \text{ mit:} \qquad \tilde{a}(\tilde{u},v) = \tilde{f}(v) \quad \forall v \in \mathcal{V},$$

wobei $\tilde{a}(\cdot,\cdot)$ und $\tilde{f}(\cdot)$ die aus den numerischen Integrationen resultierenden Näherungen für $a(\cdot,\cdot)$ und $f(\cdot)$ sind. Für die Untersuchung, wie sich die numerische Quadratur auf den Fehler $\|u-\tilde{u}\|_{H^1(\Omega)}$ auswirkt, können wir auf das Lemma von Strang [GR92, Lemma 4.14] zurückgreifen. Entscheidend ist hierbei jedoch, dass wir die $\mathcal{V}$-Elliptizität der Bilinearform $\tilde{a}(\cdot,\cdot)$, d.h.

$$\tilde{a}(u,u) \geq C\|u\|^2_{H^1(\Omega)} \qquad \forall u \in \mathcal{V}$$

garantieren müssen.

Theorem 3.10.1 (Elliptizität)**.** *Die von uns betrachteten Algorithmen definieren mittels einer Tensorproduktkonstruktion aus 1D Gauß-Lobatto-Jacobi-Quadraturen* $\mathrm{GLJ}_{\alpha,n}$ *(siehe Abschnitt 2.4) eine Quadraturregel* $\mathrm{GLJ}_{\mathcal{Q}^d,q} := \mathrm{GLJ}_{0,q} \times \ldots \times \mathrm{GLJ}_{(d-1),q}$ *auf dem Einheitswürfel* $\mathcal{Q}^d$*,* $d \in \{2,3\}$*, mit*

$$\begin{aligned} \mathrm{GLJ}_{\mathcal{Q}^d,q}(f(\eta_1,\ldots,\eta_d)) &= \sum_{i_1=0}^{q} \cdots \sum_{i_d=0}^{q} \omega_{i_1}^{(1)} \ldots \omega_{i_d}^{(d)} f(\eta_{i_1}^{(1)},\cdots,\eta_{i_d}^{(d)}) \\ &\approx \int_{\mathcal{Q}^d} f(\eta_1,\ldots,\eta_d) 2^{(2d-3)} |\det D_d'| d\eta_1 \ldots d\eta_d. \end{aligned}$$

Ferner gilt mit $\tilde{u} := u \circ F_K \circ D_d$ *und* $\tilde{A} := 2^{3-2d}(F_K')^{-1}(\hat{A} \circ F_k \circ D_d)(F_K')^{-T} |\det F_K'|$

$$\tilde{a}(u,u)|_{K \in \mathcal{T}(\mathcal{N})} := \mathrm{GLJ}_{\mathcal{Q}^d,q}\left(\langle \nabla\tilde{u}, (D_d')^{-1}\tilde{A}(D_d')^{-T}\nabla\tilde{u}\rangle\right) \geq C\|\nabla u\|^2_{L^2(K)}$$

für alle $u \circ F_K \in \Pi_{p(K)}(\mathcal{T}^d)$ *bzw.* $u \circ F_K \in \tilde{\Pi}_{p(K)}(\mathcal{T}^d)$ *(siehe Lemma 3.6.7 für die Def. von* $\Pi_{p(K)}(\mathcal{T}^d)$ *und* $\tilde{\Pi}_{p(K)}(\mathcal{T}^d)$*), falls* $q \geq p_K$*. Die Konstante* $C > 0$ *hängt hierbei nur von den Koeffizienten* $\hat{A}$ *sowie den Produkten* $\|F_K'\|^{-2}_{L^\infty(\mathcal{T}^d)} \|(F_K')^{-1}\|^{-2}_{L^\infty(\mathcal{T}^d)}$ *beziehungsweise* $\|\det F'\|_{L^\infty(\mathcal{T}^d)} \|\det(F')^{-1}\|_{L^\infty(\mathcal{T}^d)}$ *ab.*

Beweis. Der Beweis von

$$\mathrm{GLJ}_{\mathcal{Q}^d,q}\left(\langle \nabla\tilde{u}, (D_d')^{-1}\tilde{A}(D_d')^{-T}\nabla\tilde{u}\rangle\right) \geq C\|\nabla u\|^2_{L^2(K)} \quad \forall u \circ F_K \in \Pi_{p(K)}(\mathcal{T}^d) \cup \tilde{\Pi}_{p(K)}(\mathcal{T}^d)$$

beruht zum einen auf den Eigenschaften der 1D-GLJ-Quadratur, speziell auf Lemma 2.4.1 und zum anderen auf der Spektralabschätzung aus Lemma 3.10.2. Ein detailierter Beweis ist in [EM06b] zu finden. □

Lemma 3.10.2 (Spektraläquivalenz)**.** *Für* $d \in \{2,3\}$ *definieren wir die Matrizen*

$$B^d := (D_d')^{-1}(D_d')^{-T}$$

Dann gilt: Es existiert $C > 0$*, so dass*

$$C^{-1}\mathrm{diag}B^d \leq B^d \leq C\mathrm{diag}B^d,$$

wobei $\mathrm{diag}B^d$ *die Diagonale von* B^d *bezeichnet. Die Konstante* $C > 0$ *ist unabhängig von den Koordinaten* η*.*

Beweis. Siehe [EM06b]. □

Summieren wir über alle Dreiecke $K \in \mathcal{T}(\mathcal{N})$, so liefert Theorem 3.10.1 die Elliptizität der Bilinearform $\tilde{a}(\cdot,\cdot)$ für $\mathcal{V} = S^{\mathbf{p}}(\Omega,\mathcal{N})$, vorausgesetzt wir verwenden auf jedem Element Quadraturformeln mit $q_i \geq p_K$.

Bemerkung 3.10.3. Für hinreichend glatte Daten und Elliptizität der Bilinearform $a_h(\cdot,\cdot)$ beeinflusst die Verwendung von Quadraturregeln nicht die Konvergenzrate der p-FEM. Siehe hierzu unter anderem [BS92, Mn90].

3.11 Numerische Ergebnisse

Dieser Abschnitt beinhaltet alle numerischen Ergebnisse zu den Algorithmen der letzten Abschnitte, sowohl für den 2-dimensionalen als auch den 3-dimensionalen Fall. In all unseren Rechnungen setzen wir $\hat{A}(x) = I$, verfahren jedoch so, als ob wir es mit variable Koeffizienten zu tun hätten. Lediglich in den explizit mit „const. coeff." gekennzeichneten Kurven der Abbildungen 3.4 und 3.6 machen wir Gebrauch von dem Wissen, dass wir es mit konstanten Koeffizienten zu tun haben. Mit diesen speziellen Rechnungen wollen wir aufzeigen, welch enormer Geschwindigkeitsgewinn sich erzielen lässt, wenn man die Möglichkeit hat, die Konstanz der Koeffizienten ausnutzen zu können.
Als Quadraturregeln verwenden wir stets Gauß-Lobatto-Jacobi-Regeln $\mathrm{QR} = \mathrm{QR}^1 \times \ldots \times \mathrm{QR}^d$ mit

$$\mathrm{QR}^i = S^{(i)} \times W^{(i)} = \{(\eta_0^{(i)}, \omega_0^{(i)}), \ldots, (\eta_{q_i}^{(i)}, \omega_{q_i}^{(i)})\}$$

bezüglich der Gewichtsfunktionen

$$\omega = 1 \text{ für } \mathrm{QR}^1, \quad \omega = (1-\eta_2) \text{ für } \mathrm{QR}^2, \quad \omega = (1-\eta_3)^2 \text{ für } \mathrm{QR}^3 .$$

Für den 2-dimensionalen Fall benutzen wir hierbei Stützstellenzahlen $q_1 = q_2 = p_K$ und für den 3-dimensionalen Fall definieren wir folgende drei Quadraturtypen:

$$\begin{aligned} &\text{Typ-1:} && q_1 = q_2 = q_3 = p_K, \\ &\text{Typ-2:} && q_1 = q_2 = q_3 = p_K + 1, \\ &\text{Typ-3:} && q_1 = q_2 = q_3 = p_K + 2. \end{aligned}$$

Bemerkung 3.11.1. Wie in Abschnitt 3.10 dargelegt, besitzt Quadraturtyp-1 die minimal notwendige Stützstellenanzahl. Unter bestimmten Umständen, z.B auf gestörten Gittern, kann es jedoch sinnvoll sein, auf höhere Quadraturordnungen zurückzugreifen (z.B. auf Typ-2, Typ-3).

In den nachfolgenden Rechnungen sind die Lagrange-Formfunktionen $\Phi^{(Lag)}$ stets an die Quadratur angepasst und wir nehmen eine uniforme Polynomgradverteilung für das Element K an. D.h. wir betrachten unsere Algorithmen für $p(K) = (p, \ldots, p)$.
Alle Rechnungen wurden auf einem Pentium IV mit 2400 MHz und 1GB Hauptspeicher ausgeführt. Die Tabellen 3.1- 3.4 enthalten die Rechenzeiten für das Aufstellen der Steifigkeitsmatrix (gen), für das Durchführen der statischen Kondensation (sc) und für eine Matrix-Vektor-Multiplikation (Av). Des Weiteren benutzen wir folgende Abkürzungen:

KS	-	die Berechnung wurde für $\Phi^{(KS)}$ Formfunktionen durchgeführt,
Lag	-	die Berechnung wurde für $\Phi^{(Lag)}$ Formfunktionen durchgeführt,
blas	-	die Berechnung beruht auf BLAS- bzw. LAPACK-Routinen,
simple	-	die Berechnung wurde mit Algorithmus 3.6.12 durchgeführt,
sum fact.	-	die Berechnung wurden mittels Summenfaktorisierung durchgeführt,
spect Gal.	-	die Berechnung beruht auf dem Spektral-Galerkin-Algorithmus.

Mit DOF bezeichnen wir die Gesamtzahl der Freiheitsgrade und mit INT die Anzahl innerer Freiheitsgrade.

Abbildung 3.4: Aufstellen der Elementsteifigkeitsmatrix - Rechenzeit - 2D

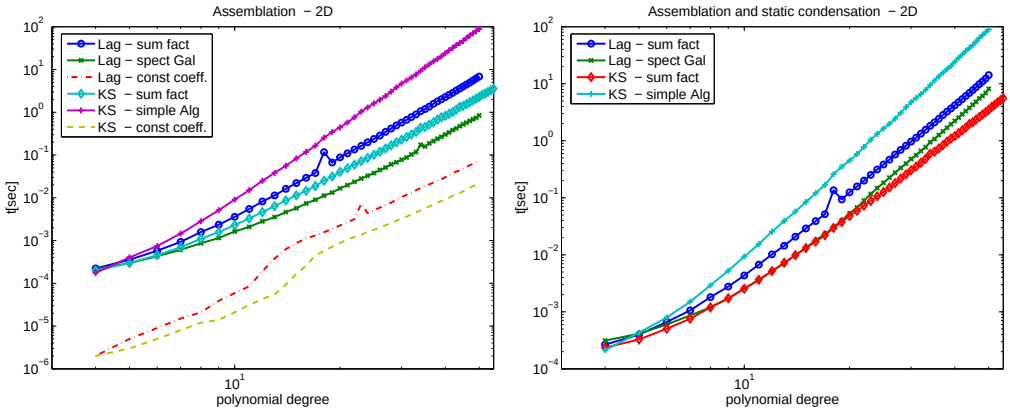


Abbildung 3.5: Aufstellen der Elementsteifigkeitsmatrix - Rechenzeit - Blockweise - 2D

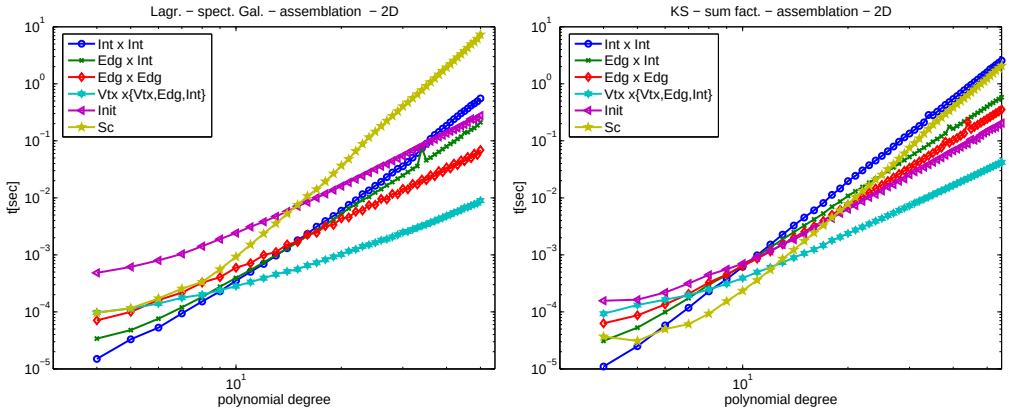


Abbildung 3.6: Aufstellen der Elementsteifigkeitsmatrix - Rechenzeit - 3D

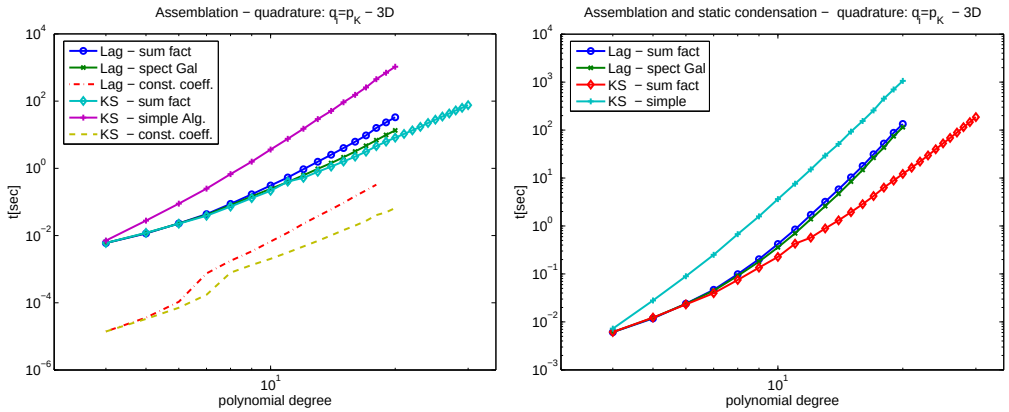


Tabelle 3.1: KS-Formfunktionen - Rechenzeit - Quadratur: $q_i = p_K$ - 2D

p_K	DOF	INT	gen (sum fact)	sc	Av (blas)	Av (sum fact)
4	15	3	1.98e-04	3.70e-05	3.00e-06	3.50e-05
5	21	6	2.97e-04	3.10e-05	4.00e-06	4.00e-05
6	28	10	4.56e-04	5.00e-05	4.00e-06	5.60e-05
7	36	15	6.95e-04	6.10e-05	8.00e-06	7.00e-05
8	45	21	1.10e-03	9.30e-05	9.00e-06	9.40e-05
9	55	28	1.58e-03	1.53e-04	1.20e-05	1.23e-04
10	66	36	2.29e-03	2.35e-04	1.40e-05	1.57e-04
11	78	45	3.29e-03	3.60e-04	1.90e-05	1.97e-04
12	91	55	4.62e-03	5.48e-04	2.50e-05	2.36e-04
13	105	66	6.40e-03	8.55e-04	3.70e-05	2.88e-04
14	120	78	8.64e-03	1.22e-03	4.20e-05	3.44e-04
15	136	91	1.13e-02	1.75e-03	5.90e-05	4.20e-04
20	231	171	4.02e-02	7.79e-03	2.10e-04	9.10e-04
25	351	276	1.00e-01	2.54e-02	4.04e-04	1.69e-03
30	496	406	2.27e-01	7.57e-02	6.89e-04	2.75e-03
50	1326	1176	2.27e+00	1.23e+00	4.49e-03	1.20e-02

Tabelle 3.2: Lag-Formfunktionen - Rechenzeit - Quadratur: $q_i = p_K$ - 2D

p_K	DOF	INT	gen (spect Gal)	sc	Av (blas)	Av (sum fact)	Av (spect Gal)
4	18	6	2.18e-04	9.60e-05	2.00e-06	3.40e-05	2.30e-05
5	27	12	2.96e-04	1.16e-04	4.00e-06	4.10e-05	2.80e-05
6	38	20	4.30e-04	1.72e-04	4.00e-06	5.30e-05	4.90e-05
7	51	30	6.09e-04	2.55e-04	8.00e-06	6.70e-05	4.40e-05
8	66	42	8.62e-04	3.35e-04	1.80e-05	8.80e-05	5.30e-05
9	83	56	1.14e-03	5.58e-04	2.70e-05	1.14e-04	6.30e-05
10	102	72	1.62e-03	9.30e-04	3.60e-05	1.54e-04	7.70e-05
11	123	90	2.09e-03	1.51e-03	4.60e-05	2.03e-04	9.40e-05
12	146	110	2.81e-03	2.36e-03	6.40e-05	2.42e-04	1.11e-04
13	171	132	3.54e-03	3.54e-03	9.10e-05	3.01e-04	1.31e-04
14	198	156	4.63e-03	5.26e-03	1.25e-04	3.58e-04	1.53e-04
15	227	182	5.70e-03	7.51e-03	1.75e-04	4.27e-04	1.81e-04
20	402	342	1.66e-02	3.67e-02	5.32e-04	9.00e-04	3.61e-04
30	902	812	7.74e-02	4.03e-01	2.08e-03	2.69e-03	9.70e-04
40	1602	1482	3.02e-01	1.93e+00	7.00e-03	6.09e-03	2.11e-03
50	2502	2352	8.43e-01	7.31e+00	1.51e-02	1.15e-02	3.98e-03

Tabelle 3.3: KS-Formfunktionen - Rechenzeit - Quadratur: $q_i = p_K$ - 3D

p_K	DOF	INT	gen (sum fact)	sc	Av (blas)	Av (sum fact)
4	35	1	6.00e-03	1.43e-04	5.00e-06	7.65e-04
5	56	4	1.20e-02	2.76e-04	9.00e-06	1.20e-03
6	84	10	2.25e-02	6.24e-04	1.90e-05	1.81e-03
7	120	20	3.87e-02	9.92e-04	3.60e-05	3.23e-03
8	165	35	7.25e-02	2.45e-03	6.90e-05	4.35e-03
9	220	56	1.29e-01	5.57e-03	1.25e-04	6.92e-03
10	286	84	2.14e-01	1.16e-02	2.34e-04	9.47e-03
11	364	120	4.02e-01	2.48e-02	4.11e-04	1.28e-02
12	455	165	5.22e-01	5.00e-02	6.36e-04	1.63e-02
13	560	220	7.90e-01	9.71e-02	9.23e-04	2.22e-02
14	680	286	1.12e+00	1.83e-01	1.34e-03	2.78e-02
15	816	364	1.60e+00	3.37e-01	1.88e-03	3.61e-02
20	1771	969	8.21e+00	3.98e+00	9.07e-03	9.78e-02
25	3276	2024	2.80e+01	2.47e+01	3.08e-02	2.17e-01
30	5456	3654	7.59e+01	1.10e+02	8.52e-02	4.19e-01

Tabelle 3.4: Lag-Formfunktionen - Rechenzeit - Quadratur: $q_i = p_K$ - 3D

p_K	DOF	INT	gen (spect Gal)	sc	Av (blas)	Av (sum fact)
4	35	1	5.99e-03	1.35e-04	4.00e-06	7.69e-04
5	60	8	1.18e-02	3.30e-04	1.00e-05	1.21e-03
6	101	27	2.27e-02	1.03e-03	2.70e-05	1.97e-03
7	164	64	4.13e-02	3.29e-03	7.10e-05	3.03e-03
8	255	125	8.02e-02	1.12e-02	1.81e-04	4.37e-03
9	380	216	1.45e-01	3.44e-02	4.42e-04	6.75e-03
10	545	343	2.46e-01	1.12e-01	8.27e-04	9.15e-03
11	756	512	3.90e-01	3.12e-01	1.56e-03	1.23e-02
12	1019	729	6.21e-01	7.70e-01	2.85e-03	1.59e-02
13	1340	1000	9.65e-01	1.63e+00	5.18e-03	2.13e-02
14	1725	1331	1.43e+00	3.25e+00	8.56e-03	2.73e-02
15	2180	1728	2.13e+00	6.29e+00	1.37e-02	3.52e-02
20	5715	4913	1.34e+01	1.01e+02	9.30e-02	9.49e-02

Abbildung 3.7: Aufstellen der Elementsteifigkeitsmatrix - Rechenzeit - Blockweise - 3D

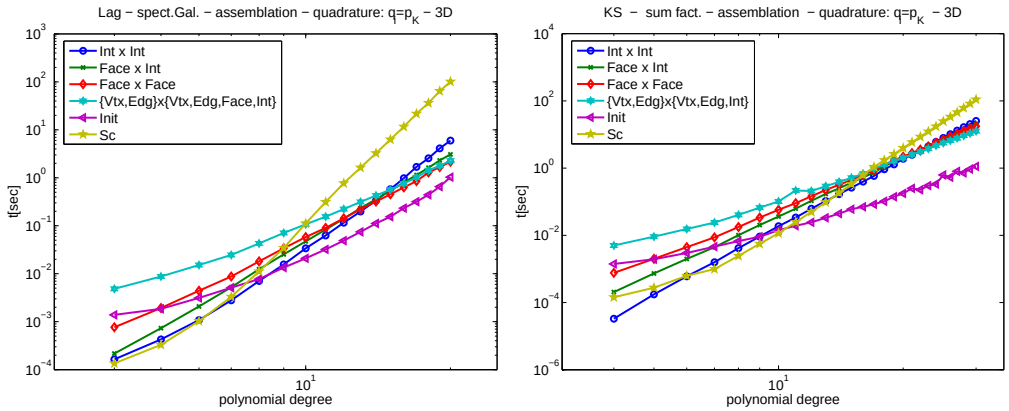


Abbildung 3.8: Matrix-Vektor-Multiplikation - Rechenzeit

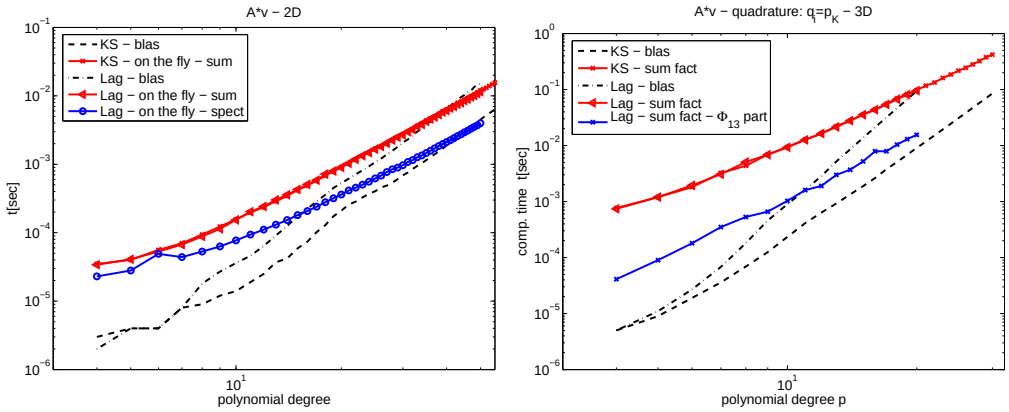
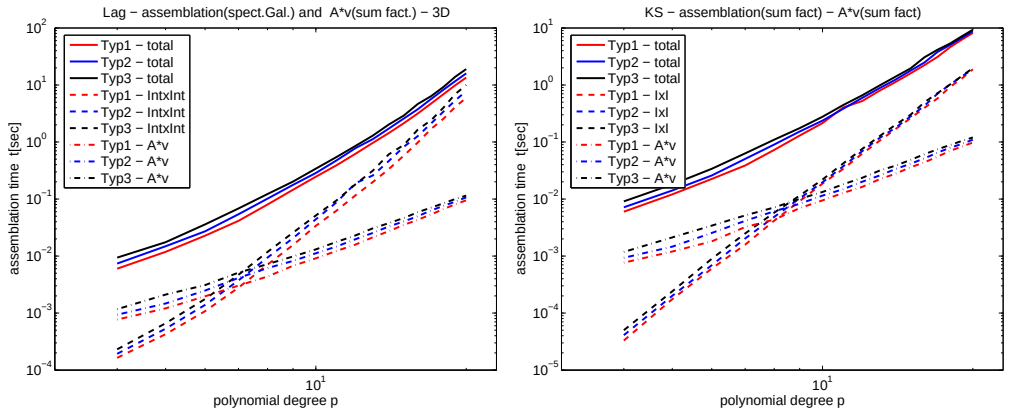


Abbildung 3.9: Verschiedene Quadraturen - Assemblieren und Matrix-Vektor-Multiplikation - Rechenzeit - 3D



3.12 Auswertung der numerischen Ergebnisse

In den letzten Abschnitten haben wir verschiedene Algorithmen für das Aufstellen der Elementsteifigkeitsmatrix sowie einer „on the fly“ Matrix-Vektor-Multiplikation untersucht. Wir betrachteten einerseits die Karniadakis & Sherwin-Formfunktionen $\Phi^{(KS)}$ in Verbindung mit dem Standardalgorithmus und der Summenfaktorisierung, anderseits führten wir die modifizierten Formfunktionen $\Phi^{(Lag)}$ ein und ermöglichten damit eine Verallgemeinerung der Spektral-Galerkin-Idee aus [MGS01] auf Dreiecks- und Tetraederelemente. Um die verschiedenen Algorithmen numerisch vergleichen zu können, haben wir die Verfahren sowohl für den 2-dimensionalen als auch den 3-dimensionalen Fall implementiert und sind zu folgenden Feststellungen gelangt:

- Das Aufstellen der Elementsteifigkeitsmatrizen mittels Standardalgorithmus 3.6.12 ist in jedem Fall die langsamste Methode, ist jedoch am einfachsten zu implementieren und stellt keine speziellen Bedingungen an die Struktur der Formfunktionen.

- Aufgrund der vergrößerten Anzahl innerer Formfunktionen erhalten wir mit $\Phi^{(Lag)}$ eine bessere Approximation der Lösung. Des Weiteren ist in 2D das Aufstellen der Steifigkeitsmatrix mittels Spektral-Galerkin-Algorithmus in Verbindung mit den angepassten Formfunktionen $\Phi^{(Lag)}$ deutlich schneller als das Aufstellen der Steifigkeitsmatrix für $\Phi^{(KS)}$-Formfunktionen mit Summenfaktorisierung. In 3D besitzen die angepassten Formfunktionen $\Phi^{(Lag)}$ die circa 6-fache Anzahl innerer Formfunktionen gegenüber $\Phi^{(KS)}$. Trotzdem ist die Rechenzeit für das Aufstellen der Steifigkeitsmatrix mit dem Spektral-Galerkin-Algorithmus und $\Phi^{(Lag)}$ für praktisch relevante Polynomgrade ($p_K \leq 15$) annähernd gleich zur Summenfaktorisierung für $\Phi^{(KS)}$-Formfunktionen.

- Die statische Kondensation ist wegen der vergrößerten Anzahl innerer Formfunktionen für $\Phi^{(Lag)}$-basierte Steifigkeitsmatrizen langsamer als bei Verwendung von $\Phi^{(KS)}$. Für 2D sind die Rechenzeiten für das Aufstellen der kondensierten $\Phi^{(Lag)}$-basierten Steifigkeitsmatrix mittels Spektral-Galerkin-Algorithmus und die Rechenzeit für das Aufstellen der kondensierten $\Phi^{(KS)}$-basierten Steifigkeitsmatrix mittels Summenfaktorisierung bis zu einem Polynomgrad $p_k \leq 20$ nahezu identisch. In 3D gilt dies nur für Polynomgrade $p_k \leq 8$.

- Für Elemente mit konstanten Koeffizienten und affiner Elementtransformation F_K lässt sich der Prozess des Aufstellens der Elementsteifigkeitsmatrix deutlich beschleunigen.

- Betrachten wir eine *hp*-FEM-Implementation, welche an Stelle des expliziten Aufstellens der Steifigkeitsmatrix eine „on the fly“ Matrix-Vektor-Multiplikation benutzt (insbesondere in Verbindung mit dem PCG-Algorithmus für das Lösen des Gleichungssystems weit verbreitet), so erhalten wir in 2D eine signifikante Rechenzeitersparnis bei Verwendung von $\Phi^{(Lag)}$-Formfunktionen in Verbindung mit Algorithmus 3.9.9. Asymptotisch ist die „on the fly“ Matrix-Vektor-Multiplikation sogar schneller als eine auf Blas-Routinen basierende Matrix-Vektor-Multiplikation, wobei hier sogar noch der Aufwand für das Aufstellen der Matrix hinzukommt.

Kapitel 4

Randkonzentrierte hp-FEM

Die randkonzentrierte Finite-Element-Methode stellt eine spezielle Variante der hp-FEM dar und wurde erstmals in [KM03] vorgestellt. Mit ihrer a priori Vorgabe von Netzstruktur und Polynomgradverteilung vereint sie FEM-typische Eigenschaften mit den Vorteilen der BEM zu einer überaus effektiven und leistungsfähigen Methode. Das Haupteinsatzgebiet der randkonzentrierten FEM sind hierbei elliptische Randwertaufgaben mit analytischen oder stückweise analytischen Koeffizienten, deren Lösungen, bedingt durch Randeffekte, wie zum Beispiel Randbedingungen von niedriger Regularität oder komplizierten Geometrien, jedoch geringe globale Regularität aufweisen. Eine der wesentlichen Eigenschaften der randkonzentrierten FEM ist die Reduktion der Anzahl von Freiheitsgraden auf eine Größe proportional zur Menge der Knoten am Rand. Gemessen in Fehler gegen Freiheitsgrade besitzt die randkonzentrierte FEM damit gleiche Konvergenzeigenschaften wie eine h-BEM, benötigt jedoch keine Fundamentallösung und führt zudem auf FEM-typische schwach besetzte Steifigkeitsmatrizen. Speziell für Kontakt- und Steuerungsprobleme wäre eine Verwendung der randkonzentrierten FEM denkbar.
Nach einer kurzen Zusammenstellung der aus [KM03] stammenden Grundlagen der randkonzentrierten FEM werden wir in diesem Kapitel

- eine verbesserte Konvergenzrate der randkonzentrierten FEM im Gebietsinneren beweisen (Abschnitt 4.2) sowie
- einen Vorkonditionierer konstruieren, um speziell auch Probleme mit Neumann- oder gemischten Randbedingungen effektiv lösen zu können (Abschnitt 4.3).

All unsere theoretischen Ergebnisse werden stets durch numerische Beispiele verifiziert. [1]

4.1 Grundlegende Idee und Eigenschaften

Für die nachfolgenden Abschnitte über die lokale Fehleranalyse und schnelle Löser ist ein Verständnis der grundlegenden Ideen der randkonzentrierten FEM zwingend notwendig. Wir beginnen daher mit einer kurzen Zusammenstellung der für uns wichtigsten Resultate aus [KM03]. Die Darstellungen in [KM03] beschränken sich auf 2-dimensionale polygonale Lipschitz-Gebiete mit affinen Elementtransformationen sowie reine Dirichlet- bzw. reine Neumann-Randbedingungen. Obwohl sich die Resultate aus [KM03] auch auf 3D Probleme, gemischte

[1] Kapitel 4 beinhaltet im Wesentlichen die Ergebnisse aus [EM06a, EM05b].

Randbedingungen und nichtaffine Elementtransformationen übertragen lassen, wollen wir vorerst an den in [KM03] getätigten Annahmen festhalten. Erst im späteren Kapitel 4.3 werden wir explizit auch den 3D Fall und gemischte Randbedingungen mit einbeziehen.
Für diesen und den nachfolgenden Abschnitt 4.2 sei daher das Gebiet Ω vorerst ein 2-dimensionales polygonales Lipschitz-Gebiet, auf dem wir Problem 3.1.2 mit reinen Dirichlet- oder reinen Neumann-Randbedingungen betrachten.

4.1.1 Regularität der Lösung, Voraussetzungen an die Daten

Der Einsatz der randkonzentrierten Finite-Element-Methode verlangt eine Verschärfung der Annahme 3.1.3. Für dieses Kapitel soll daher gelten:

Annahme 4.1.1 (randkonzentrierte FEM - Voraussetzungen an die Daten)**.**

- *Die Koeffizienten $\hat{A}$, b, a_0 und die rechte Seite f sind analytisch auf $\overline{\Omega}$.*
- *Es existiert ein $\delta \in (0,1]$, so dass für die Lösung u von Problem 3.1.2 die Regularität $u \in H^{1+\delta}(\Omega)$ gegeben ist.*

Obwohl durch die Voraussetzungen an die Daten die Analytizität der Lösung u auf Ω bereits gesichert ist (siehe [Mor66]), bleiben, wenn wir uns dem Rand $\partial\Omega$ nähern, höhere Ableitungen von u nicht notwendigerweise beschränkt. Mit Hilfe der in [KM03] eingeführten Räume $\tilde{\mathcal{B}}^2_{1-\delta}$ und des Lemmas 4.1.3 (siehe [KM03, Thm. 1.4]) über die Regularität der Lösung gelingt es jedoch eine Kontrolle über das Anwachsen dieser Ableitungen zu erhalten:

Definition 4.1.2. Für $\delta \in (0,1)$ und $C, \gamma > 0$ sei

$$\tilde{\mathcal{B}}^2_{1-\delta}(C,\gamma) = \{u \in H^{1+\delta}(\Omega) \cap C^\infty(\Omega) \mid \ u \text{ genügt (4.1)}\},$$

wobei mit $r = r(x) = \operatorname{dist}(x, \partial\Omega)$ Bedingung (4.1) gegeben ist durch

$$\|u\|_{H^{1+\delta}(\Omega)} \leq C \quad \text{und} \quad \left\| r^{1-\delta+n} \nabla^{n+2} u \right\|_{L^2(\Omega)} \leq C\gamma^n n! \quad \forall n \in \mathbb{N}_0. \tag{4.1}$$

Lemma 4.1.3 (Regularität der Lösung)**.** *Sei Ω ein Lipschitz-Gebiet. Seien $\hat{A}$, b, a_0, f analytisch auf $\overline{\Omega}$ und $u \in H^{1+\delta}(\Omega)$, $\delta \in (0,1]$, die Lösung von Problem 3.1.2. Dann gilt:*

1. *u ist analytisch auf Ω.*
2. *Es existieren Konstanten $C_u, \gamma_u > 0$, welche nur vom Gebiet Ω, den Koeffizienten $\hat{A}$, b, a_0 und δ, $\|u\|_{H^{1+\delta}(\Omega)}$ abhängen, so dass $u \in \tilde{\mathcal{B}}^2_{1-\delta}(C_u, \gamma_u)$.*

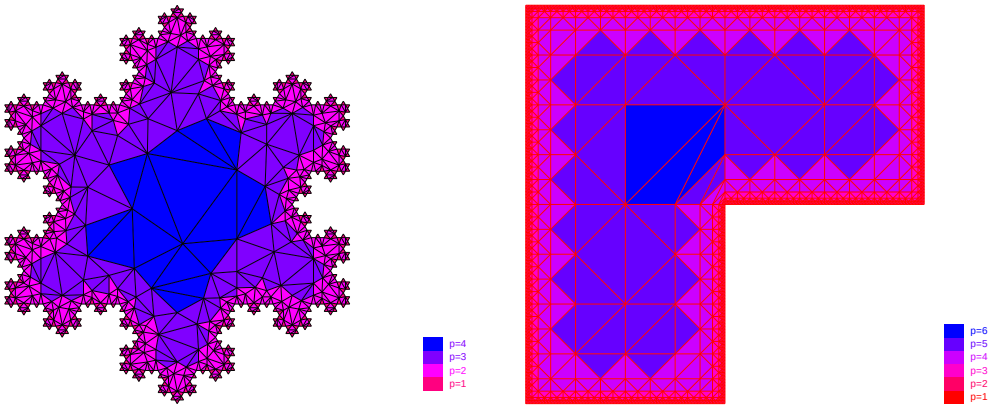
Beweis. [KM03, Theorem 1.4]. □

4.1.2 Netze, Polynomgradverteilungen, FE-Räume, Approximation

Die Hauptidee der randkonzentrierten FEM besteht darin, durch eine geeignete a priori Vorgabe von Netzstruktur und Polynomgradverteilung die Regularitätseigenschaften der Lösung bestmöglich auszunutzen.
Wir beschränken uns hierbei auf γ-formreguläre Netze $\mathcal{N} = \{(K, F_K)\}$, d.h.

$$h_K^{-1} \left\|F_K'\right\|_{L^\infty(K)} + h_K \left\| \left(F_K'\right)^{-1} \right\|_{L^\infty(K)} \leq \gamma, \quad h_K = \operatorname{diam}(K) \quad \forall\, K \in \mathcal{T}(\mathcal{N})$$

Abbildung 4.1: Beispiele einer randkonzentrierten Vernetzung



bestehend aus Dreiecken $K = F_K(\mathcal{T}^2)$ mit $F_K : \mathbb{R}^2 \mapsto \mathbb{R}^2$ affin für alle $K \in \mathcal{T}(\mathcal{N})$.
Der für die randkonzentrierte Finite-Element-Methode geeignete Netztyp ist das geometrische Netz mit Verfeinerung zum gesamten Rand hin:

Definition 4.1.4 (geometrisches Netz)**.** Ein γ-formreguläres Netz $\mathcal{N} = \{(K, F_K)\}$ heißt geometrisches Netz mit Randgitterweite h, falls Konstanten $C_1, C_2 > 0$ existieren, so dass für alle $K \in \mathcal{T}(\mathcal{N})$ gilt:

1. $h \leq h_K \leq C_2 h$ für alle $\overline{K} \cap \partial\Omega \neq \emptyset$
2. $C_1 \inf\limits_{x \in K} r(x) \leq h_K \leq C_2 \sup\limits_{x \in K} r(x)$ für alle $\overline{K} \cap \partial\Omega = \emptyset$.

Abbildung 4.1 zeigt typische Vertreter randkonzentrierter Netze. Wie wir hierbei und anhand von Definition 4.1.4 sehen können, ist die Einschränkung des Netzes auf den Rand stets ein quasi-uniformes Gitter mit Randgitterweite h. Als direkte Folgerung aus Definition 4.1.4 ergibt sich außerdem:

Lemma 4.1.5. *Sei $\mathcal{N}$ ein geometrisches Netz mit Randgitterweite h im Sinne von Definition 4.1.4. Dann existieren Konstanten $\tilde{C}_1$ und $\tilde{C}_2$, die nur vom Formregularitätsparamter γ sowie den Konstanten C_1 und C_2 aus Definition 4.1.4 abhängen, so dass*

1. $\inf\limits_{x \in K} r(x) \geq \tilde{C}_1 h_K \qquad \forall\, K \in \mathcal{T}(\mathcal{N}) \;\; mit \;\; \overline{K} \cap \partial\Omega = \emptyset,$
2. $\sup\limits_{x \in K} r(x) \leq \tilde{C}_2 h_K \qquad \forall\, K \in \mathcal{T}(\mathcal{N}).$

Beweis. (1) Sei $K \in \mathcal{T}(\mathcal{N})$ mit $\overline{K} \cap \partial\Omega = \emptyset$. Wegen der Formregularität gilt für den Durchmesser eines jeden Dreiecks K' mit $K \cap K' \neq \emptyset$ die Abschätzung $h_{K'} \geq C_\gamma h_K$. Da der Abstand von K zum Rand jedoch nicht kleiner sein kann als die kleinste Höhe seiner Nachbarelemente, folgt die erste Behauptung allein aus der Formregularität.
(2) Für $K \in \mathcal{T}(\mathcal{N})$ mit $\overline{K} \cap \partial\Omega \neq \emptyset$ ist die Behauptung offensichtlich. Sei nun $\overline{K} \cap \partial\Omega = \emptyset$, dann gilt

$$\sup_{x \in K} r(x) \leq \inf_{x \in K} r(x) + h_K$$

und aus Definition 4.1.4 folgt die Behauptung

$$\sup_{x\in K} r(x) \leq \inf_{x\in K} r(x) + h_K \leq \left(1 + C_1^{-1}\right) h_K.$$

□

Die in Verbindung mit geometrischen Netzen geeignete Polynomgradverteilung ist die so genannte lineare Polynomgradverteilung:

Definition 4.1.6 (lineare Polynomgradverteilung)**.** Sei $\mathcal{N}$ ein geometrisches Netz mit Randgitterweite h. Die Polynomgradverteilung $p(\mathcal{N})$ heißt linear mit Anstieg $\alpha > 0$, falls für Konstanten $C_1, C_2 > 0$ gilt:

$$1 + \alpha C_1 \log \frac{h_K}{h} \leq p_K \leq 1 + \alpha C_2 \log \frac{h_K}{h} \qquad \forall K \in \mathcal{T}(\mathcal{N}). \tag{4.2}$$

Korollar 4.1.7. *Sei $\mathcal{N}$ ein geometrisches Netz mit Randgitterweite h und $p(\mathcal{N})$ eine lineare Polynomgradverteilung mit Anstieg $\alpha > 0$. Dann gilt für die FE-Räume aus Definition 3.3.7:*

- $\dim S^{\mathbf{p}}(\Omega, \mathcal{N}) \leq Ch^{-1}$,
- $\max\limits_{K\in\mathcal{T}(\mathcal{N})} p_K \leq C|\log h|$,
- $C^{-1} p_{K'} \leq p_K \leq C p_{K'} \qquad \forall K, K' \in \mathcal{T}(\mathcal{N})$ *mit* $\overline{K} \cap \overline{K'} \neq \emptyset$

Beweis. Siehe [KM03]. □

Eine der wichtigsten Aussagen über die randkonzentrierte FEM charakterisiert ihre globale Approximationseigenschaft. Das folgende Theorem zeigt, dass sich die randkonzentrierte FEM beim Verhältnis von Fehler gegenüber Freiheitsgraden analog einer h-BEM verhält:

Theorem 4.1.8. *Seien die Annahmen 3.1.3 und 4.1.1 erfüllt und sei $u \in H^{1+\delta}(\Omega)$ die Lösung von Problem 3.1.2. Sei $\mathcal{N}$ ein geometrisches Netz mit Randgitterweite h und $p(\mathcal{N})$ eine gemäß Definition 3.3.6 zulässige lineare Polynomgradverteilung mit Anstieg α hinreichend groß. Dann gilt für die zugehörige FE-Lösung $u_N \in S^{\mathbf{p}}(\Omega, \mathcal{N})$ die globale Fehlerabschätzung*

$$\|u - u_N\|_{H^1(\Omega)} \leq CN^{-\delta},$$

wobei $N = \dim\ S^{\mathbf{p}}(\Omega, \mathcal{N}) = O(h^{-1})$.

Beweis. Siehe [KM03, Thm. 2.13]. □

4.2 Lokale Fehleranalyse

Nachdem wir die Grundzüge der randkonzentrierten FEM aus [KM03] kurz vorgestellt haben, kommen wir nun zu einem neuen Resultat, dem Beweis einer gegenüber der globalen Konvergenz verbesserten Konvergenz im Gebietsinneren. Wir schränken uns hierbei auf Dirichlet-Randbedingungen ein (siehe Bemerkung 4.2.26) und um den technischen Aufwand zu verringern betrachten wir zudem das Poissonproblem. D.h. unser Modellproblem für Abschnitt 4.2 lautet:

Problem 4.2.1 (Poissonproblem - schwache Formulierung)**.** *Finde zu vorgegebenen Randbedingungen g_D und rechter Seite f ein $u \in H^1(\Omega)$, so dass*

$$\gamma_0 u = g_D \quad \textit{und} \quad \int_\Omega \langle \nabla u, \nabla v \rangle \, d\Omega = \int_\Omega f v \, d\Omega \qquad \forall\, v \in H_0^1(\Omega).$$

Wie bereits in Kapitel 3 besprochen, führt Problem 4.2.1 zu folgender *hp*-FE-Diskretisierung:

Problem 4.2.2 (Poissonproblem - *hp*-FE-Formulierung)**.** *Finde ein $u \in S^{\mathbf{p}}(\Omega, \mathcal{N})$, so dass*

$$\gamma_0 u = g_h \quad \textit{und} \quad \int_\Omega \langle \nabla u, \nabla v \rangle \, d\Omega = \int_\Omega f v \, d\Omega \qquad \forall\, v \in S_0^{\mathbf{p}}(\Omega, \mathcal{N}),$$

wobei g_h die L^2-Approximation von g_D bezeichnet (siehe Problem 3.3.9).

$\mathcal{N}$ bezeichnet hierbei wie üblich eine Vernetzung des Gebiets Ω und $S^{\mathbf{p}}(\Omega, \mathcal{N})$ den FE-Raum bezüglich $\mathcal{N}$ und einer zugehörigen Polynomgradverteilung $p(\mathcal{N})$.
Der Beweis der lokalen Fehleranalyse wird die Lösung des dualen Problems einbeziehen. Hierbei ist folgende Forderung bezüglich Lösbarkeit und Regularität wesentlich:

Annahme 4.2.3. *Es existiert ein $\delta_0 \in (0, 1]$, so dass für jede fixierte kompakte Teilmenge ein $C > 0$ existiert und für beliebiges $K \in \mathcal{T}(\mathcal{N})$ mit $K \subset \Omega'$ das Problem: Finde ein $z \in H_0^1(\Omega)$ mit*

$$\int_\Omega \nabla z \nabla v \, d\Omega = \int_K e v \, d\Omega \qquad \forall\, v \in H_0^1(\Omega),$$

eine eindeutige Lösung $z \in H^{1+\delta_0}(\Omega)$ besitzt, für welche gilt

$$\|z\|_{H^{1+\delta_0}(\Omega)} + \|\gamma_1 z\|_{H^{\delta_0 - \frac{1}{2}}(\partial\Omega)} \leq C \|e\|_{L^2(K)}.$$

Bemerkung 4.2.4. Für den Fall eines polygonalen Gebietes bezeichne $\alpha_{max} \in (0, 2\pi)$ den größten Innenwinkel von Ω. Dann gilt (siehe [Gri85]) für jedes $s \in [0, 1] \cap [0, \pi/\alpha_{max})$ $z \in H^{1+s}(\Omega)$ und $\gamma_1 z \in H^{-1/2+s}(\partial\Omega)$ zusammen mit der a priori Abschätzung

$$\|z\|_{H^{1+s}(\Omega)} + \|\gamma_1 z\|_{H^{s-\frac{1}{2}}(\partial\Omega)} \leq C_s \|e\|_{L^2(K)}.$$

Für allgemeine Lipschitz-Gebiete gilt die Annahme 4.2.3 für jedes $\delta_0 < \frac{1}{2}$. [Neč64] zeigt die Abschätzung $\|z\|_{H^{1+\delta_0}(\Omega)} \leq C\|e\|_{L^2(K)}$ für jedes $\delta_0 < 1/2$ und [Neč67, Thm. 3.1, Chap. 5] zeigt die Abschätzung $\|\gamma_1 u\|_{H^{\delta_0 - 1/2}(\partial\Omega)} \leq C\|e\|_{L^2(K)}$ für den Fall $\delta_0 = 1/2$.

Kommen wir nun zur Formulierung der Hauptaussage bezüglich lokaler Konvergenz im Gebietsinneren:

Theorem 4.2.5 (lokale Fehlerabschätzung)**.** *Sei $\Omega \subset \mathbb{R}^2$ ein polygonales Lipschitz-Gebiet und $\Omega' \subset\subset \Omega$ eine kompakte Teilmenge. Es sei $u \in H^{1+\delta}(\Omega)$, $\delta \in (0, 1]$, Lösung von Problem 4.2.1 und es gelte die Annahme 4.2.3. Sei u_h die FE-Lösung des Problems 4.2.2 bezüglich eines geometrischen Netzes $\mathcal{N}$ mit Randgitterweite h und zulässiger linearer Polynomgradverteilung*

$p(\mathcal{N})$ mit Anstieg α. Dann existiert ein $\beta \in (0, \delta_0]$, so dass für α hinreichend groß, abhängig von u, und alle Elemente $\dot{K} \in \mathcal{T}(\mathcal{N})$ mit $\dot{K} \subset \Omega'$ gilt:

$$\|u - u_h\|_{L^2(\dot{K})} \leq Ch^{\delta+\beta} \leq CN^{-\delta-\beta}, \tag{4.3}$$

$$|u - u_h|_{W^{k,2}(\dot{K})} \leq Cp_{\dot{K}}^{2k} h^{\delta+\beta} \leq C(\log N)^{2k} N^{-\delta-\beta}, \tag{4.4}$$

$$|u - u_h|_{W^{k,\infty}(\dot{K})} \leq Cp_{\dot{K}}^{2k+2} h^{\delta+\beta} \leq C(\log N)^{2k+2} N^{-\delta-\beta}. \tag{4.5}$$

$N = O(h^{-1})$ bezeichnet hierbei die Dimension des Raumes $S^{\mathbf{p}}(\Omega, \mathcal{N})$. Die Konstanten α, β, C sind unabhängig von h und somit unabhängig von N, hängen jedoch von den Netzparametern aus den Definitionen 4.1.4, 4.1.6 ab. Die Größe β hängt zusätzlich von dem Gebiet Ω' und die Konstante C von Ω', k ab.

Bemerkung 4.2.6. Für Anmerkungen bezüglich typischer Werte von β verweisen wir auf den Abschnitt 4.2.3.

Der Beweis vom Theorem 4.2.5 erfordert zahlreiche und oft recht technische Hilfsaussagen. Der besseren Übersicht wegen haben wir all diese Hilfsaussagen in den Abschnitt 4.2.1 verschoben. Eine zentrale Rolle im Beweis von Theorem 4.2.5 spielt die Gewichtsfunktion $\omega_{\beta,\mathcal{T}}$:

Definition 4.2.7 (Gewichtsfunktion)**.** Sei $\mathcal{N}$ eine geometrische Vernetzung mit Randgitterweite h. Für einen Parameter $\beta \in (0, 2]$ definieren wir die Gewichtsfunktion $\omega_{\beta,\mathcal{T}}$ durch:

$$\omega_{\beta,\mathcal{T}}(x) := I\left[\left(\frac{h}{h + r(x)}\right)^{\beta}\right],$$

wobei $r(x) = \operatorname{dist}(x, \partial\Omega)$ und I die Interpolation in den Raum der auf $\mathcal{N}$ stückweise linearen Funktionen, mit $[Iu](x) = u(x)$ in allen Netzknoten, bezeichne.

Neben der Gewichtsfunktion $\omega_{\beta,\mathcal{T}}$ benötigen wir noch folgendes Hilfsproblem:

Definition 4.2.8 (Hilfsproblem)**.** Unter den Voraussetzungen von Theorem 4.2.5 seien $z \in H_0^1(\Omega)$ und $z_h \in S_0^{\mathbf{p}}(\Omega, \mathcal{N})$ gegeben durch

- Finde $z \in H_0^1(\Omega)$, so dass

  $$-\Delta z = \chi_{\dot{K}}(u - u_h) \text{ auf } \Omega \quad \text{und} \quad \gamma_0 z = 0 \text{ auf } \partial\Omega,$$

  bzw. in schwacher Formulierung

  $$\int_{\Omega} \nabla z \cdot \nabla v \mathrm{d}\Omega = \int_{\dot{K}} (u - u_h)\, v \mathrm{d}\Omega \quad \forall\, v \in H_0^1(\Omega). \tag{4.6}$$

- Finde $z_h \in S_0^{\mathbf{p}}(\Omega, \mathcal{T})$, so dass

  $$\int_{\Omega} \nabla z_h \cdot \nabla v \mathrm{d}\Omega = \int_{\dot{K}} (u - u_h)\, v \mathrm{d}\Omega \quad \forall\, v \in S_0^{\mathbf{p}}(\Omega, \mathcal{N}). \tag{4.7}$$

Für Aussagen und Beweise zu den Eigenschaften der Gewichtsfunktion $\omega_{\beta,\mathcal{T}}$ sowie den Eigenschaften der Hilfsfunktionen z und z_h verweisen wir abermals auf Abschnitt 4.2.1 und wenden uns nun dem Beweis von Theorem 4.2.5 zu.

Beweis. (Theorem 4.2.5) Wir beginnen mit der Abschätzung (4.3): Sei $Y := H^{\frac{1}{2}}(\partial\Omega)$ und $Y^* := H^{-\frac{1}{2}}(\partial\Omega)$, so folgt aus der Greenschen Formel (siehe [Gri85, Lemma 1.5.3.7, Lemma 1.5.3.9])

$$\int_{\dot{K}} (u-u_h)^2 \mathrm{d}\Omega \;=\; -\int_{\Omega} \Delta z(u-u_h)\mathrm{d}\Omega = \int_{\Omega} \nabla z \cdot \nabla(u-u_h)\mathrm{d}\Omega - \langle \gamma_1 z, u-u_h\rangle_{Y^*\times Y}\,.$$

Einsetzen der Randbedingungen und Ausnutzen von Galerkin-Orthogonalitäten bezüglich u_h und z_h sowie Ausnutzen der $L^2(\partial\Omega)$-Orthogonalität von g_h liefert für beliebiges $Iu \in S^{\mathbf{p}}(\Omega,\mathcal{N})$ mit $u_h|_{\partial\Omega} = Iu|_{\partial\Omega}$ und beliebiges $q \in (Y^{\mathbf{p}}(\Omega,\mathcal{N}))^*$

$$\begin{aligned}\int_{\dot{K}} (u-u_h)^2 \mathrm{d}\Omega \;&=\; \int_{\Omega} \nabla(z-z_h)\cdot\nabla(u-u_h)\mathrm{d}\Omega - \langle \gamma_1 z, g-g_h\rangle_{Y^*\times Y} \\ &=\; \int_{\Omega} \nabla(z-z_h)\cdot\nabla(u-Iu)\mathrm{d}\Omega - \langle \gamma_1 z - q, g-g_h\rangle_{Y^*\times Y}\,.\end{aligned}$$

Bringen wir nun die Gewichtsfunktion $\omega_{2\beta,\mathcal{T}}$ als produktive Eins ein und wenden die Cauchy-Schwarz-Ungleichung an, so erhalten wir

$$\begin{aligned}\|u-u_h\|^2_{L^2(\dot{K})} \;\leq\; &\left\|\sqrt{\omega_{2\beta,\mathcal{T}}}\nabla(z-z_h)\right\|_{L^2(\Omega)} \left\|\frac{1}{\sqrt{\omega_{2\beta,\mathcal{T}}}}\nabla(u-Iu)\right\|_{L^2(\Omega)} + \\ &\|\gamma_1 z - q\|_{Y^*}\|g-g_h\|_Y.\end{aligned}$$

Für genügend großes α folgt nun die Behauptung aus $g-g_h = \gamma_0(u-u_h)$, dem Spursatz

$$\|g-g_h\|_{H^{\frac{1}{2}}(\partial\Omega)} = \|\gamma_0(u-u_h)\|_{H^{\frac{1}{2}}(\partial\Omega)} \leq C\|u-u_h\|_{H^1(\Omega)} \leq Ch^{\delta}$$

sowie Lemma 4.2.14, Lemma 4.2.17 und Lemma 4.2.20.
Betrachten wir als Nächstes die Abschätzung (4.4): Sei $\mathcal{T}^2$ das Referenzdreieck und die Transformation einer Funktion auf das Referenzelement mit einem Dach markiert. Aus der Formregularität der Vernetzung und da aus $\dot{K} \subset \Omega' \subset\subset \Omega$ die Abschätzung $h_{\dot{K}} \geq C_{\Omega'}$ folgt, erhalten wir

$$|u-u_h|_{W^{k,2}(\dot{K})} \;\leq\; Ch_{\dot{K}}^{1-k}|\hat{u}-\hat{u}_h|_{W^{k,2}(\mathcal{T}^2)} \leq C_{k,\Omega'}|\hat{u}-\hat{q}|_{W^{k,2}(\mathcal{T}^2)} + C_{k,\Omega'}|\hat{q}-\hat{u}_h|_{W^{k,2}(\mathcal{T}^2)}$$

für beliebiges $\hat{q} \in \Pi_{p(\dot{K})}(\mathcal{T}^2)$. Mittels inverser Ungleichungen (siehe [Sch98, (4.6.5)]) ergibt sich daraus

$$\begin{aligned}|u-u_h|_{W^{k,2}(\dot{K})} \;&\leq\; C_{k,\Omega'}|\hat{u}-\hat{q}|_{W^{k,2}(\hat{K})} + C_{k,\Omega'}\overline{p}^{2k}\|\hat{q}-\hat{u}_h\|_{L^2(\hat{K})} \\ &\leq\; C_{k,\Omega'}|\hat{u}-\hat{q}|_{W^{k,2}(\hat{K})} + C_{k,\Omega'}\overline{p}^{2k}\|\hat{u}-\hat{u}_h\|_{L^2(\hat{K})} + C_{k,\Omega'}\overline{p}^{2k}\|\hat{u}-\hat{q}\|_{L^2(\hat{K})} \\ &\leq\; C_{k,\Omega'}\overline{p}^{2k}\|\hat{u}-\hat{q}\|_{W^{k,\infty}(\hat{K})} + C_{k,\Omega'}\overline{p}^{2k}\|\hat{u}-\hat{u}_h\|_{L^2(\hat{K})} \\ &\leq\; C_{k,\Omega'}\overline{p}^{2k}\|\hat{u}-\hat{q}\|_{W^{k,\infty}(\hat{K})} + C_{k,\Omega'}h_{\dot{K}}^{-1}\overline{p}^{2k}\|u-u_h\|_{L^2(\dot{K})},\end{aligned}$$

wobei $\overline{p}$ den Wert des maximalen und $\underline{p}$ den Wert des minimalen Eintrags von $p(\dot{K})$ bezeichne. Schließlich, durch Ausnutzen von $\overline{p} \leq C\underline{p}$ (Korollar 4.1.7), $h_{\dot{K}} \geq C_{\Omega'}$, der Analytizität von u auf $\overline{\Omega'}$, dem Korollar [Mel02, Corollary 3.2.17] und Abschätzung (4.3) erhalten wir für α hinreichend groß und $\beta \in (0,\delta_0]$ (wie in Abschätzung (4.3)):

$$|u-u_h|_{W^{k,2}(\dot{K})} \leq C_{k,\Omega'}\overline{p}^{2k}e^{-b\underline{p}} + C_{k,\Omega'}\overline{p}^{2k}h^{\delta+\beta} \leq C_{k,\Omega'}p_{\dot{K}}^{2k}h^{\delta+\beta}.$$

Für die Abschätzung (4.5) gehen wir wie im Beweis zu Abschätzung (4.4) vor: Als erstes transformieren wir $\dot{K}$ auf das Referenzdreieck $\mathcal{T}^2$ und schieben anschließend ein beliebiges Element $\hat{q} \in \Pi_{\mathbf{p}(\dot{K})}(\hat{K})$ ein:

$$|u - u_h|_{W^{k,\infty}(\dot{K})} \leq C_{k,\Omega'} |\hat{u} - \hat{q}|_{W^{k,\infty}(\hat{K})} + C_{k,\Omega'} |\hat{u}_h - \hat{q}|_{W^{k,\infty}(\hat{K})}.$$

Nutzen wir abermals inverse Ungleichungen (siehe [Sch98, (4.6.1), (4.6.5)]), so erhalten wir

$$\begin{aligned} |u - u_h|_{W^{k,\infty}(\dot{K})} &\leq C_{k,\Omega'} |\hat{u} - \hat{q}|_{W^{k,\infty}(\hat{K})} + C_{k,\Omega'} \overline{p}^{2k} \|\hat{u}_h - \hat{q}\|_{L^\infty(\hat{K})} \\ &\leq C_{k,\Omega'} |\hat{u} - \hat{q}|_{W^{k,\infty}(\hat{K})} + C_{k,\Omega'} \overline{p}^{2k+2} \|\hat{u}_h - \hat{q}\|_{L^2(\hat{K})}. \end{aligned}$$

Da

$$\|\hat{u}_h - \hat{q}\|_{L^2(\hat{K})} \leq \|\hat{u} - \hat{u}_h\|_{L^2(\hat{K})} + \|\hat{u} - \hat{q}\|_{L^2(\hat{K})} \leq \|\hat{u} - \hat{u}_h\|_{L^2(\hat{K})} + |\hat{K}| \|\hat{u} - \hat{q}\|_{L^\infty(\hat{K})},$$

gelangen wir schließlich zu

$$|u - u_h|_{W^{k,\infty}(\dot{K})} \leq C_{k,\Omega'} \overline{p}^{2k+2} \|\hat{u} - \hat{q}\|_{W^{k,\infty}(\hat{K})} + C_{k,\Omega'} \overline{p}^{2k+2} \|\hat{u} - \hat{u}_h\|_{L^2(\hat{K})}$$

und die gewünschte Abschätzung folgt analog zum Beweis von Abschätzung (4.4). □

4.2.1 Hilfsaussagen

In diesem recht technischen Abschnitt stellen wir alle notwendigen Lemmata, die für den Beweis von Theorem 4.2.5 eine Rolle spielen, bereit. Beginnen wollen wir dabei mit dem Vorstellen eines Gauß-Lobatto-*hp*-Interpolationsoperators für nicht uniforme Polynomgradverteilungen.

Ein *hp*-Interpolationsoperator

Theorem 4.2.9. *Sei $\mathcal{T}^2$ das Referenzdreieck mit den Kanten $e_1, \ldots, e_3$. Sei $k > 3/2$ und $p(K) = (p_1, p_2, p_3, p_K)$ eine Polynomgradverteilung mit*

$$\underline{p} := \min_{i=1,\ldots,3} p_i, \qquad \overline{p} := \max_{i=1,\ldots,3} p_i \leq p_K.$$

Ferner bezeichne

$$i_p : C([-1,1]) \mapsto \mathcal{P}_p, \qquad u \mapsto i_p u(x) = \sum_{i=0}^{p} u(\xi_i) l_i(x)$$

den eindimensionalen Gauß-Lobatto-Interpolationsoperator, wobei die Gauß-Lobatto-Punkte ξ_i, $i = 0, \ldots, p$, die Nullstellen des Polynoms $x \mapsto (1 - x^2) L_p'(x)$ sind (siehe Abschnitt 2.4) und

$$l_i = \prod_{\substack{j=0 \\ j \neq i}}^{p} \frac{x - \xi_i}{\xi_j - \xi_i}$$

die zugehörigen Lagrange-Interpolationspolynome. Dann existiert eine Konstante $C > 0$ und ein linearer Operator $I : H^k(\mathcal{T}^2) \to \Pi_{p(K)}(\mathcal{T}^2)$, so dass

1. $(Iu)|_{e_i} = i_{p_i}(u|_{e_i})$ *für* $i \in \{1, \ldots, 3\}$,

2. $Iu = u$ *für alle* $u \in \Pi_{p(K)}(\mathcal{T}^2)$,

3. $\|Iu\|_{H^1(\mathcal{T}^2)} \leq C(1 + p'/\underline{p})\|u\|_{H^1(\mathcal{T}^2)}$ *für alle* $u \in \mathcal{P}_{p'}(\mathcal{T}^2)$,

4. $|Iu|_{H^1(\mathcal{T}^2)} \leq C(1 + p'/\underline{p})|u|_{H^1(\mathcal{T}^2)}$ *für alle* $u \in \mathcal{P}_{p'}(\mathcal{T}^2)$.

Weiterhin gelten folgende Approximationseigenschaften, wobei die Konstante $C_k > 0$ *einzig von* k *abhängt:*

$$\begin{aligned} \|u - Iu\|_{H^1(\mathcal{T}^2)} &\leq C_k \underline{p}^{-(k-1)} \|u\|_{H^k(\mathcal{T}^2)}, \\ \|u - Iu\|_{H^{1/2}_{00}(e_i)} &\leq C_k \underline{p}^{-(k-1)} \|u\|_{H^k(\mathcal{T}^2)}, \qquad i = 1, \ldots, 3. \end{aligned}$$

Beweis. Siehe [EM06a]. □

Die Gewichtsfunktion $\omega_{\beta,\mathcal{T}}$

Als Nächstes beschäftigen wir uns mit den Eigenschaften der in Definition 4.2.7 eingeführten Gewichtsfunktion $\omega_{\beta,\mathcal{T}}$.

Lemma 4.2.10 (Eigenschaften von $\omega_{\beta,\mathcal{T}}$)**.** *Sei* $\mathcal{N}$ *eine geometrische Vernetzung mit Randgitterweite* h *und* $\omega_{\beta,\mathcal{T}}$ *gegeben durch Definition 4.2.7. Dann existieren Konstanten* $C_1, \ldots, C_4 > 0$*, die nur von der Formregularitätskonstanten* γ *und den Konstanten aus Definition 4.1.4 abhängen, so dass für alle* $K \in \mathcal{T}(\mathcal{N})$ *und beliebiges* $\beta \in (0, 2]$ *gilt:*

1. $\inf\limits_{x \in K} |\omega_{\beta,\mathcal{T}}(x)| \geq C_2 \left(\dfrac{h}{h_K}\right)^{\beta}$,

2. $\sup\limits_{x \in K} |\omega_{\beta,\mathcal{T}}(x)| \leq C_1 \left(\dfrac{h}{h_K}\right)^{\beta}$,

3. $|\nabla \omega_{\beta,\mathcal{T}}(x)| = C_{\beta,\mathcal{T},K} \leq C_3 \beta \dfrac{\omega_{\beta,\mathcal{T}}(x)}{h_K} \leq C_4 \beta \dfrac{\omega_{\beta,\mathcal{T}}(x)}{r(x)} \qquad \forall\, x \in K.$

Beweis.

1. Die Einschränkung von $\omega_{\beta,\mathcal{T}}(x)$ auf das Dreieck K ist eine affine Funktion, die in den Eckpunkten von K mit der nicht interpolierten Gewichtsfunktion $\tilde{\omega}_{\beta,\mathcal{T}}(x) = \left(\frac{h}{h + r(x)}\right)^{\beta}$ übereinstimmt. Folglich gilt

   $$\inf_{x \in K} |\omega_{\beta,\mathcal{T}}(x)| \geq \inf_{x \in K} |\tilde{\omega}_{\beta,\mathcal{T}}(x)| = \inf_{x \in K} \left(\frac{h}{h + r(x)}\right)^{\beta}$$

   und die für das geometrische Netz geltenden Abschätzungen $r(x) \leq C h_K$ und $h \leq C h_K$ liefern

   $$\inf_{x \in K} \left(\frac{h}{h + r(x)}\right)^{\beta} \geq C^{\beta} \left(\frac{h}{h_K}\right)^{\beta} \geq \min\{C^0, C^2\} \left(\frac{h}{h_K}\right)^{\beta}.$$

2. Analog zu Punkt (1) gilt

   $$\sup_{x \in K} |\omega_{\beta,\mathcal{T}}(x)| \leq \sup_{x \in K} |\tilde{\omega}_{\beta,\mathcal{T}}(x)| = \sup_{x \in K} \left(\frac{h}{h + r(x)}\right)^{\beta}$$

und aus $r(x) \geq Ch_K$ für alle $K \in \mathcal{T}(\mathcal{N})$ mit $\overline{K} \cap \Omega = \emptyset$ bzw. $h \geq Ch_K$ für alle $K \in \mathcal{T}(\mathcal{N})$ mit $\overline{K} \cap \Omega \neq \emptyset$ folgt

$$\sup_{x \in K} \left(\frac{h}{h + r(x)} \right)^{\beta} \leq C^{\beta} \left(\frac{h}{h_K} \right)^{\beta} \leq \max\{C^0, C^2\} \left(\frac{h}{h_K} \right)^{\beta}.$$

3. Betrachten wir die Funktion

$$x \mapsto \left(\frac{h}{h + x h_K} \right)^{\beta}, \qquad x \in \mathbb{R},$$

so existiert laut Mittelwertsatz für $x_2 > x_1 \geq 0$ ein $\xi \in (x_1, x_2)$, so dass

$$\left(\frac{h}{h + x_1 h_K} \right)^{\beta} - \left(\frac{h}{h + x_2 h_K} \right)^{\beta} = \frac{\beta(x_2 - x_1) h_k}{h + \xi h_k} \left(\frac{h}{h + \xi h_K} \right)^{\beta}. \tag{4.8}$$

Ferner, da $\omega_{\beta,\mathcal{T}}(x)$ auf K affin ist, gilt $|\nabla \omega_{\beta,\mathcal{T}}(x)| = C_{\beta,\mathcal{T},K}$ für alle $x \in K$. Mit der γ-Formregularität des Netzes erhalten wir

$$C_{\beta,\mathcal{T},K} \leq C h_K^{-1} \left| \sup_{x \in K} \omega_{\beta,\mathcal{T}}(x) - \inf_{x \in K} \omega_{\beta,\mathcal{T}}(x) \right|,$$

wobei die Konstante C nur von γ abhängt. Folglich, aus der Definition von $\omega_{\beta,\mathcal{T}}(x)$ in Verbindung mit Lemma 4.1.5, ergibt sich:

$$C_{\beta,\mathcal{T},K} \leq C h_K^{-1} \begin{cases} \left| 1 - \left(\frac{h}{h + c_2 h_K} \right)^{\beta} \right| & : \ \overline{K} \cap \partial\Omega \neq \emptyset \\ \left| \left(\frac{h}{h + C_1 h_K} \right)^{\beta} - \left(\frac{h}{h + C_2 h_K} \right)^{\beta} \right| & : \ \overline{K} \cap \partial\Omega = \emptyset \end{cases},$$

wobei die Konstanten $C_1, C_2 > 0$ wiederum nur von der Formregularitätskonstanten γ abhängen. Wenden wir nun Ungleichung (4.8) an, so führt dies auf

$$C_{\beta,\mathcal{T},K} \leq C\beta h_K^{-1} \begin{cases} C_2 \frac{h_K}{h + \xi_1 h_K} \left(\frac{h}{h + \xi_1 h_K} \right)^{\beta} & : \ \overline{K} \cap \partial\Omega \neq \emptyset \\ |C_1 - C_2| \frac{h_K}{h + \xi_2 h_K} \left(\frac{h}{h + \xi_2 h_K} \right)^{\beta} & : \ \overline{K} \cap \partial\Omega = \emptyset \end{cases}$$

mit $\xi_1 \in (0, C_2)$ bzw. $\xi_2 \in (C_1, C_2)$. Da $h_K \leq Ch$ und $\xi_2 \geq C_1$ erhalten wir

$$C_{\beta,\mathcal{T},K} \leq C\beta h_K^{-1} \begin{cases} 1 & \text{für alle } K \in \mathcal{T}(\mathcal{N}) \mid \overline{K} \cap \partial\Omega \neq \emptyset \\ \left(\frac{h}{h + c h_K} \right)^{\beta} & \text{für alle } K \in \mathcal{T} \mid \overline{K} \cap \partial\Omega = \emptyset \end{cases}$$

und die Behauptung folgt aus $h_K \geq Cr(x)$ für alle $K \in \mathcal{T}(\mathcal{N})$ und $h \geq Ch_K$ falls $\overline{K} \cap \partial\Omega \neq \emptyset$.

□

Lemma 4.2.11. *Sei $\Omega \in \mathbb{R}^2$ ein polygonales Gebiet und $\mathcal{N}$ eine geometrische Vernetzung mit Randgitterweite h. Ferner sei die Gewichtsfunktion $\omega_{\beta,\mathcal{T}}$ gegeben durch Definition 4.2.7. Dann existiert ein $\beta' \in (0, 2]$ sowie ein zugehöriges $C_{\beta'} > 0$, wobei β' nur vom Gebiet Ω, der*

Formregularitätskonstanten γ und den Konstanten aus Definition 4.1.4 abhängt, so dass für alle $\beta \in (0, \beta']$ und alle $f \in H_0^1(\Omega)$ gilt:

$$\begin{aligned} \left\| r^{-1} \sqrt{\omega_{\beta,\mathcal{T}}} f \right\|_{L^2(\Omega)} &\leq C_{\beta'} \left\| \sqrt{\omega_{\beta,\mathcal{T}}} \nabla f \right\|_{L^2(\Omega)}, & (4.9) \\ \left\| r^{-1} \omega_{\beta,\mathcal{T}} f \right\|_{L^2(\Omega)} &\leq C_{\beta'} \left\| \omega_{\beta,\mathcal{T}} \nabla f \right\|_{L^2(\Omega)}. & (4.10) \end{aligned}$$

Beweis. Da $f \in H_0^1(\Omega)$ können wir Hardys Ungleichung (siehe [Gri85, Thm. 1.4.4.3]) anwenden:

$$\begin{aligned} \left\| \sqrt{\omega_{\beta,\mathcal{T}}} \frac{f}{r} \right\|_{L^2(\Omega)} &\leq C_1 \left\| \nabla (\sqrt{\omega_{\beta,\mathcal{T}}} f) \right\|_{L^2(\Omega)} \\ &\leq C_1 \left\| \sqrt{\omega_{\beta,\mathcal{T}}} \nabla f \right\|_{L^2(\Omega)} + \frac{C_1}{2} \left\| \frac{f}{\sqrt{\omega_{\beta,\mathcal{T}}}} \nabla \omega_{\beta,\mathcal{T}} \right\|_{L^2(\Omega)}. \end{aligned} \tag{4.11}$$

Aus Lemma 4.2.10 - Eigenschaft 3 folgt:

$$\left\| \frac{f}{\sqrt{\omega_{\beta,\mathcal{T}}}} \nabla \omega_{\beta,\mathcal{T}} \right\|_{L^2(\Omega)}^2 \leq (C_2 \beta)^2 \int_\Omega \left(\frac{f}{r} \sqrt{\omega_{\beta,\mathcal{T}}} \right)^2 \mathrm{d}\Omega = (C_2 \beta)^2 \left\| \sqrt{\omega_{\beta,\mathcal{T}}} \frac{f}{r} \right\|_{L^2(\Omega)}^2. \tag{4.12}$$

Kombinieren wir nun die beiden Ungleichungen (4.11) und (4.12), so erhalten wir für $2 - C_1 C_2 \beta > 0$

$$\left\| \sqrt{\omega_{\beta,\mathcal{T}}} \frac{f}{d} \right\|_{L^2(\Omega)} \leq \frac{2C_1}{2 - C_1 C_2 \beta} \left\| \sqrt{\omega_{\beta,\mathcal{T}}} \nabla f \right\|_{L^2(\Omega)},$$

woraus sich für beliebiges $0 < \beta' < 2(C_1 C_2)^{-1}$ und $C_{\beta'} := \frac{2C_1}{2 - C_1 C_2 \beta'}$ die Abschätzung (4.9) ergibt. Für Abschätzung 4.10 gehen wir völlig analog vor. Die Hardy-Ungleichung liefert

$$\left\| \omega_{\beta,\mathcal{T}} \frac{f}{r} \right\|_{L^2(\Omega)} \leq C_1 \left\| \omega_{\beta,\mathcal{T}} \nabla f \right\|_{L^2(\Omega)} + C_1 \left\| f \nabla \omega_{\beta,\mathcal{T}} \right\|_{L^2(\Omega)} \tag{4.13}$$

und aus Lemma 4.2.10 folgt

$$\left\| f \nabla \omega_{\beta,\mathcal{T}} \right\|_{L^2(\Omega)}^2 \leq (C_2 \beta)^2 \left\| \omega_{\beta,\mathcal{T}} \frac{f}{r} \right\|_{L^2(\Omega)}^2. \tag{4.14}$$

Die Kombination beider Ungleichungen ergibt auch hier wieder die gewünschte Abschätzung. □

Interpolationsoperator $S_0^{2\mathbf{p}}(\Omega, \mathcal{N}) \to S_0^{\mathbf{p}}(\Omega, \mathcal{N})$

Lemma 4.2.12. *Sei $\Omega \in \mathbb{R}^2$ ein polygonales Gebiet und $\mathcal{N} = \{(K, F_K)\}$ eine geometrische Vernetzung mit linearer Polynomgradverteilung $p(\mathcal{N})$. Ferner sei der lineare Operator $I : S_0^{2\mathbf{p}}(\Omega, \mathcal{N}) \to S_0^{\mathbf{p}}(\Omega, \mathcal{N})$ gegeben durch:*

$$[Iu](x) := \left[\hat{I} (u \circ F_K) \right] (F_K^{-1} x) \quad \forall x \in \overline{K} \ \forall K \in \mathcal{T}(\mathcal{N}),$$

wobei $\hat{I}$ den Operator aus Theorem 4.2.9 bezeichne. Dann existiert ein $\beta' \in (0, 2]$ sowie ein zugehöriges $C_{\beta'} > 0$, wobei β' nur vom Gebiet Ω, der Formregularitätskonstanten γ und den

Konstanten aus den Definitionen 4.1.4, 4.1.6 abhängt, so dass für alle $\beta \in (0, \beta']$ und alle $g \in S_0^{\mathbf{P}}(\Omega, \mathcal{N})$ gilt:

$$\left\| \frac{1}{\sqrt{\omega_{\beta,\mathcal{T}}}} \nabla \left(\omega_{\beta,\mathcal{T}} g - I(\omega_{\beta,\mathcal{T}} g) \right) \right\|_{L^2(\Omega)} \leq C_{\beta'} \beta \left\| \sqrt{\omega_{\beta,\mathcal{T}}} \nabla g \right\|_{L^2(\Omega)}.$$

Beweis. Mit den Abkürzungen

$$\overline{\omega}_{\beta,\mathcal{T},K} := \sup_{x \in K} \omega_{\beta,\mathcal{T}}(x), \qquad \underline{\omega}_{\beta,\mathcal{T},K} := \inf_{x \in K} \omega_{\beta,\mathcal{T}}(x)$$

erhalten wir

$$\begin{aligned} &\left\| \frac{1}{\sqrt{\omega_{\beta,\mathcal{T}}}} \nabla \left(\omega_{\beta,\mathcal{T}} g - I(\omega_{\beta,\mathcal{T}} g) \right) \right\|^2_{L^2(\Omega)} \\ &\qquad \leq \sum_{K \in \mathcal{T}(\mathcal{N})} \frac{1}{\underline{\omega}_{\beta,\mathcal{T},K}} \left\| \nabla \left(\omega_{\beta,\mathcal{T}} g - q_K - I(\omega_{\beta,\mathcal{T}} g - q_K) \right) \right\|^2_{L^2(K)}, \end{aligned}$$

wobei für jedes $K \in \mathcal{T}(\mathcal{N})$ ein beliebiges q_K mit $q_K \circ F_K \in \Pi_{p(K)}(\mathcal{T}^2)$ gewählt werden kann. Eigenschaft (4) des Theorems 4.2.9 impliziert

$$\begin{aligned} &\left\| \frac{1}{\sqrt{\omega_{\beta,\mathcal{T}}}} \nabla \left(\omega_{\beta,\mathcal{T}} g - I(\omega_{\beta,\mathcal{T}} g) \right) \right\|^2_{L^2(\Omega)} \\ &\qquad \leq 2 \sum_{K \in \mathcal{T}(\mathcal{N})} \frac{1}{\underline{\omega}_{\beta,\mathcal{T},K}} \left(|\omega_{\beta,\mathcal{T}} g - q_K|^2_{H^1(K)} + |I(\omega_{\beta,\mathcal{T}} g - q_K)|^2_{H^1(K)} \right) \\ &\qquad \leq C \sum_{K \in \mathcal{T}(\mathcal{N})} \frac{1}{\underline{\omega}_{\beta,\mathcal{T},K}} |\omega_{\beta,\mathcal{T}} g - q_K|^2_{H^1(K)}. \end{aligned}$$

Mit der speziellen Wahl von

$$q_K := \omega_{\beta,\mathcal{T},K}\, g := \omega_{\beta,\mathcal{T}}(x_K) g \quad \text{für einen beliebigen Punkt} \quad x_K \in K$$

gelangen wir zu

$$\begin{aligned} &\left\| \frac{1}{\sqrt{\omega_{\beta,\mathcal{T}}}} \nabla \left(\omega_{\beta,\mathcal{T}} g - I(\omega_{\beta,\mathcal{T}} g) \right) \right\|^2_{L^2(\Omega)} \\ &\qquad \leq C \sum_{K \in \mathcal{T}} \frac{1}{\underline{\omega}_{\beta,\mathcal{T},K}} |(\omega_{\beta,\mathcal{T}} - \omega_{\beta,\mathcal{T},K}) g|^2_{H^1(K)} \\ &\qquad \leq C \sum_{K \in \mathcal{T}} \frac{1}{\underline{\omega}_{\beta,\mathcal{T},K}} \left(\|g \nabla \omega_{\beta,\mathcal{T}}\|^2_{L^2(K)} + \|(\omega_{\beta,\mathcal{T}} - \omega_{\beta,\mathcal{T},K}) \nabla g\|^2_{L^2(K)} \right). \end{aligned}$$

Aus Lemma 4.2.10 folgt $\overline{\omega}_{\beta,\mathcal{T},K} \leq C \underline{\omega}_{\beta,\mathcal{T},K}$ für alle $K \in \mathcal{T}(\mathcal{N})$ und durch mehrmaliges

Ausnutzen der Eigenschaften der Gewichtsfunktion (Lemma 4.2.10) erhalten wir schließlich

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{\omega_{\beta,\mathcal{T}}}} \nabla \left(\omega_{\beta,\mathcal{T}} g - I(\omega_{\beta,\mathcal{T}} g) \right) \right\|^2_{L^2(\Omega)} \\
&\qquad \leq C \sum_{K \in \mathcal{T}} \frac{1}{\underline{\omega}_{\beta,\mathcal{T},K}} \left(\beta^2 \left\| g \frac{\omega_{\beta,\mathcal{T}}}{r} \right\|^2_{L^2(K)} + h_K^2 |\nabla \omega_{\beta,\mathcal{T}}|^2_K \|\nabla g\|^2_{L^2(K)} \right) \\
&\qquad \leq C\beta^2 \sum_{K \in \mathcal{T}} \frac{1}{\underline{\omega}_{\beta,\mathcal{T},K}} \left(\left\| g \frac{\omega_{\beta,\mathcal{T}}}{r} \right\|^2_{L^2(K)} + \overline{\omega}^2_{\beta,\mathcal{T},K} \|\nabla g\|^2_{L^2(K)} \right) \\
&\qquad \leq C\beta^2 \sum_{K \in \mathcal{T}} \left(\left\| g \frac{\sqrt{\omega_{\beta,\mathcal{T}}}}{r} \right\|^2_{L^2(K)} + \|\sqrt{\omega_{\beta,\mathcal{T}}} \nabla g\|^2_{L^2(K)} \right),
\end{aligned}$$

woraus sich mittels Lemma 4.2.11 die Behauptung ergibt. □

Approximation von $\tilde{\mathcal{B}}^2_{1-\delta}$-Funktionen in einer ω-gewichteten Norm

In diesem Teilabschnitt leiten wir mittels der Resultate aus [KM03, Section 2.3.2] eine Approximationsaussage für die $\omega_{\beta,\mathcal{T}}$ - gewichtete H^1-Seminorm her.

Lemma 4.2.13. *Sei $\mathcal{N}$ eine geometrische Vernetzung mit Randgitterweite h und sei $p(\mathcal{N})$ eine lineare Polynomgradverteilung mit Anstieg $\alpha > 0$. Sei $u \in \tilde{\mathcal{B}}^2_{1-\delta}(C_u, \gamma_u)$, $C_u, \gamma_u > 0$. Ferner sei g_h die $L^2(\partial\Omega)$-Projektion von $\gamma_0 u$ auf $Y^{\mathbf{p}}(\Omega, \mathcal{N})$. Dann existiert ein Element $Iu \in S^{\mathbf{p}}(\Omega, \mathcal{N})$, so dass*

$$\begin{aligned}
Iu|_{\partial\Omega} &= g_h, \\
\|u - Iu\|_{H^1(K)} &\leq C C_K h_K^{\delta - \alpha b'} h^{\alpha b'} \qquad \forall K \in \mathcal{T}(\mathcal{N}) \quad mit \quad K \cap \partial\Omega = \emptyset, \\
\sum_{\substack{K \in \mathcal{T} \\ \overline{K} \cap \partial\Omega \neq \emptyset}} \|u - Iu\|^2_{H^1(K)} &\leq C C_u h^{2\delta},
\end{aligned}$$

wobei C, $b' > 0$ von der Formregularitätskonstanten γ, den Konstanten aus den Definitionen 4.1.4, 4.1.6 und γ_u abhängen. C ist zusätzlich von δ und α abhängig. Die Konstanten C_K sind gegeben durch

$$C_K^2 := \sum_{n=0}^{\infty} \frac{1}{(2\gamma_u)^{2n} (n!)^2} \left\| r^{n+1-\delta} \nabla^{n+2} u \right\|^2_{L^2(K)}$$

und es gilt

$$\sum_{K \in \mathcal{T}(\mathcal{N})} C_K^2 \leq \frac{4}{3} C_u^2.$$

Beweis. Siehe [EM06a]. □

Mit Hilfe von Lemma 4.2.13 sind wir nun in der Lage eine Aussage über die Approximierbarkeit von $\tilde{\mathcal{B}}^2_{1-\delta}(C_u, \gamma_u)$-Funktionen durch Funktionen aus $S^{\mathbf{p}}(\Omega, \mathcal{T})$ zu beweisen.

Lemma 4.2.14. *Sei $\mathcal{N}$ eine geometrische Vernetzung mit Randgitterweite h und $p(\mathcal{N})$ eine linearen Polynomgradverteilung mit Anstieg $\alpha > 0$. Sei $\beta \in (0,2]$, $u \in \tilde{\mathcal{B}}^2_{1-\delta}(C_u, \gamma_u)$ und $\omega_{\beta,\mathcal{T}}$ gegeben durch Definition 4.2.7. Ferner sei g_h die $L^2(\partial\Omega)$-Projektion von $\gamma_0 u$ auf $Y^{\mathbf{p}}(\Omega, \mathcal{N})$. Dann existiert ein Element $Iu \in S^{\mathbf{p}}(\Omega, \mathcal{T})$, so dass für α, C hinreichend groß, abhängig von der Formregularitätskonstanten γ, den Konstanten aus den Definitionen 4.1.4, 4.1.6 und den Werten von γ_u, δ, gilt:*

$$Iu|_{\partial\Omega} = g_h, \qquad \left[\int_\Omega \left(\frac{1}{\sqrt{\omega_{\beta,\mathcal{T}}}}\nabla\,(u - Iu)\right)^2 d\Omega\right]^{\frac{1}{2}} \leq CC_u h^\delta.$$

Beweis. Wir werden zeigen, dass das Element Iu aus Lemma 4.2.13 die gewünschte Eigenschaft besitzt.

1. Da $\omega_{\beta,\mathcal{T}}|_K \sim 1$ für alle $K \in \mathcal{T}(\mathcal{N})$ mit $\overline{K} \cap \partial\Omega \neq \emptyset$ folgt aus Lemma 4.2.13 unmittelbar

$$\sum_{\substack{K\in\mathcal{T}\\ \overline{K}\cap\partial\Omega\neq\emptyset}} \left\|\frac{1}{\sqrt{\omega_{\beta,\mathcal{T}}}}\nabla(u - Iu)\right\|^2_{L^2(K)} \leq C \sum_{\substack{K\in\mathcal{T}\\ \overline{K}\cap\partial\Omega\neq\emptyset}} \|u - Iu\|^2_{H^1(K)} \leq CC_u^2 h^{2\delta}.$$

2. Für $K \in \mathcal{T}(\mathcal{N})$ mit $\overline{K} \cap \partial\Omega = \emptyset$ gilt

$$\left\|\frac{1}{\sqrt{\omega_{\beta,\mathcal{T}}}}\nabla\,(u - Iu)\right\|_{L^2(K)} \leq CC_K \left\|\frac{1}{\sqrt{\omega_{\beta,\mathcal{T}}}}\right\|_{L^\infty(K)} h_K^{\delta-\alpha b'} h^{\alpha b'}$$

und aus

$$\omega_{\beta,\mathcal{T}}(x) \geq C\left(\frac{h}{h_K}\right)^\beta \quad \text{für alle} \quad x \in K$$

folgt

$$\left\|\frac{1}{\sqrt{\omega_{\beta,\mathcal{T}}}}\nabla\,(u - Iu)\right\|_{L^2(K)} \leq CC_K h_K^{\delta-\alpha b'+\frac{\beta}{2}} h^{\alpha b'-\frac{\beta}{2}}.$$

Für α hinreichend groß, d.h. $\alpha > b'^{-1}(\delta + \beta/2)$, gelangen wir mit der Abschätzung $h_K \geq Ch$ schließlich zu:

$$\left\|\frac{1}{\sqrt{\omega_{\beta,\mathcal{T}}}}\nabla\,(u - Iu)\right\|_{L^2(K)} \leq CC_K h^\delta$$

und die Summation über alle Elemente im Inneren liefert:

$$\sum_{\substack{K\in\mathcal{T}\\ \overline{K}\cap\partial\Omega=\emptyset}} \left\|\frac{1}{\sqrt{\omega_{\beta,\mathcal{T}}}}\nabla(u - Iu)\right\|^2_{L^2(K)} \leq Ch^{2\delta} \sum_{\substack{K\in\mathcal{T}\\ \overline{K}\cap\partial\Omega=\emptyset}} C_K^2 \leq CC_u^2 h^{2\delta}.$$

□

Bemerkung 4.2.15. Die Forderung $\alpha > b'^{-1}(\delta + \beta/2)$ aus Lemma 4.2.14 ist unabhängig von β und δ stets erfüllt, falls $\alpha > 2/b'$.

Eigenschaften der Hilfsgrößen z und z_h

Lemma 4.2.16 (Eigenschaften von z und z_h)**.** *Seien die Voraussetzungen von Theorem 4.2.5 erfüllt. Ferner sei $\dot{K} \subset \Omega'' \subset\subset \Omega' \subset\subset \Omega$ und z sowie z_h gegeben durch Definition 4.2.8. Dann existieren Konstanten C_Ω, $C_{\Omega'}$, γ_z, abhängig von Ω, Ω', Ω'', δ_0, so dass*

1. $\|z\|_{H^1(\Omega)} \leq \|z\|_{H^{1+\delta_0}(\Omega)} \leq C_\Omega \|u - u_h\|_{L^2(\dot{K})}$,
2. $z \in H^2(\Omega')$ *und* $\|z\|_{H^2(\Omega')} \leq C_{\Omega'} \|u - u_h\|_{L^2(\dot{K})}$,
3. $z|_{\Omega\setminus\Omega'} \in \tilde{\mathcal{B}}^2_{1-\delta_0}(C_{\Omega'}\|u - u_h\|_{L^2(\dot{K})}, \gamma_z)$,
4. $\|z - z_h\|_{H^1(\Omega)} \leq c\|u - u_h\|_{L^2(\dot{K})}$.

Beweis.

1. Entspricht Annahme 4.2.3.
2. Entspricht der inneren Regularität elliptischer Probleme. Aus [Hac96, Thm. 9.1.26] erhalten wir $z \in H^2(\Omega')$ zusammen mit

   $$\|z\|_{H^2(\Omega')} \leq C_{\Omega'} \left(\|u - u_h\|_{L^2(\dot{K})} + \|z\|_{H^1(\Omega)} \right)$$

   und die gewünschte Abschätzung folgt nun aus der vorherigen.
3. Folgt aus [KM03, Thm. A.1]: o.B.d.A. sei Ω'' ein glattes Gebiet. Für $z \in H^{1+\delta_0}(\Omega\setminus\Omega'')$ gilt $-\Delta z = 0$ auf $\Omega\setminus\Omega''$ und da $\|z\|_{H^{1+\delta_0}(\Omega\setminus\Omega'')} \leq \|z\|_{H^{1+\delta_0}(\Omega)} \leq C\|u - u_h\|_{L^2(\dot{K})}$, folgt die Behauptung aus [KM03, Thm. A.1].
4. Folgt aus dem Lax-Milgram-Lemma in Verbindung mit der ersten Abschätzung:

   $$\|z - z_h\|_{H^1(\Omega)} \leq C \inf_{v \in S^{\mathbf{p}}_0(\Omega,\mathcal{T})} \|z - v\|_{H^1(\Omega)} \quad \leq C\|z\|_{H^1(\Omega)} \leq C\|u - u_h\|_{L^2(\dot{K})}.$$

□

Lemma 4.2.17. *Seien die Voraussetzungen von Theorem 4.2.5 erfüllt und z sowie z_h gegeben durch Definition 4.2.8. Dann existiert ein Element $q \in Y^{\mathbf{p}}(\Omega, \mathcal{T})$, so dass*

$$\|\gamma_1 z - q^*\|_{H^{-\frac{1}{2}}(\partial\Omega)} \leq Ch^{\delta_0}\|u - u_h\|_{L^2(\dot{K})}, \tag{4.15}$$

wobei $q^* \in Y^{\mathbf{p}}(\Omega, \mathcal{T})^*$ und $q \in Y^{\mathbf{p}}(\Omega, \mathcal{T})$ durch den Rieszschen-Darstellungssatz in Beziehung stehen.

Beweis. Annahme 4.2.3 gibt uns $\gamma_1 z \in H^{-1/2+\delta_0}(\partial\Omega)$ mit

$$\|\gamma_1 z\|_{H^{\delta_0-\frac{1}{2}}(\partial\Omega)} \leq C\|\chi_{\dot{K}}(u - u_h)\|_{L^2(\Omega)} = C\|u - u_h\|_{L^2(\dot{K})}$$

und [KM03, Lemma 2.8] sichert die Existenz eines Elements $q \in Y^{\mathbf{p}}(\Omega, \mathcal{T})$ mit

$$\|\gamma_1 z - q^*\|_{H^{-\frac{1}{2}}(\partial\Omega)} \leq Ch^{\delta_0}\|\gamma_1 z\|_{H^{\delta_0-\frac{1}{2}}(\partial\Omega)}.$$

Die Behauptung (4.15) folgt aus der Kombination beider Ungleichungen. □

Lemma 4.2.18. *Seien die Voraussetzungen von Theorem 4.2.5 erfüllt, z sowie z_h gegeben durch Definition 4.2.8 und $\omega_{\beta,\mathcal{T}}$ gegeben durch Definition 4.2.7. Sei die Randgitterweite h hinreichend klein und α hinreichend groß, abhängig nur von der Formregularitätskonstanten γ, den Konstanten aus den Definitionen 4.1.4, 4.1.6 und u. Dann existiert ein Element $Iz \in S_0^{\mathbf{p}}(\Omega, \mathcal{T})$, so dass*

$$\left\| \sqrt{\omega_{2\beta,\mathcal{T}}} \nabla (z - Iz) \right\|_{L^2(\Omega)} \leq C h^{\beta} \left\| u - u_h \right\|_{L^2(\dot{K})}$$

für alle $\beta \in (0, \delta_0]$ und C unabhängig von h und β ist.

Beweis. Wir konstruieren das Element Iz wie folgt:

$$[Iz](x) := \left[\hat{I} (z \circ F_K) \right] \left(F_K^{-1} x \right) \quad \forall x \in \overline{K} \ \ \forall K \in \mathcal{T}(\mathcal{N})$$

und unterscheiden zwei Fälle:

- Für $K \in \mathcal{T}_1 := \{ K \in \mathcal{T} \,|\, \overline{K} \cap \overline{\dot{K}} = \emptyset \}$ bezeichne $\hat{I}$ den Operator aus [KM03, Lemma 2.9].
- Für $K \in \mathcal{T}_2 := \{ K \in \mathcal{T} \,|\, \overline{K} \cap \overline{\dot{K}} \neq \emptyset \}$ bezeichne $\hat{I}$ den in Theorem 4.2.9 konstruierten Operator.

Man beachte, dass die Operatoren für beide Fälle auf den Kanten übereinstimmen. Gehen wir davon aus, dass kein $K \in \mathcal{T}_2$ existiert mit $\overline{K} \cap \partial\Omega \neq \emptyset$ (siehe Bemerkung 4.2.19), so existiert eine Umgebung

$$\tilde{\Omega} \ : \ \bigcup_{K \in \mathcal{T}_2} \overline{K} \subset \tilde{\Omega} \subset\subset \Omega,$$

unabhängig von h. Mit $\omega_{2\beta,\mathcal{T}}(x) \in (0, 1]$ und den Eigenschaften aus Lemma 4.2.10 ergibt sich

$$\begin{aligned} \left\| \sqrt{\omega_{2\beta,\mathcal{T}}} \nabla (z - Iz) \right\|_{L^2(\Omega)}^2 &\leq \sum_{K \in \mathcal{T}} \left\| \sqrt{\omega_{2\beta,\mathcal{T}}} \right\|_{L^\infty(K)}^2 |z - Iz|_{H^1(K)}^2 \\ &\leq \sum_{K \in \mathcal{T}_1} |z - Iz|_{H^1(K)}^2 + C \sum_{K \in \mathcal{T}_2} \left(\frac{h}{h_K} \right)^{2\beta} |z - Iz|_{H^1(K)}^2 . \end{aligned}$$

Da nach Lemma 4.2.16 $z \in H^2(\tilde{\Omega})$ mit $\|z\|_{H^2(\tilde{\Omega})} \leq C_{\tilde{\Omega}} \|u - u_h\|_{L^2(\dot{K})}$, können wir durch eine Transformation auf das Referenzelement und Theorem 4.2.9 die zweite Summe folgendermaßen abschätzen:

$$\begin{aligned} \sum_{K \in \mathcal{T}_2} \left(\frac{h}{h_K} \right)^{2\beta} |z - Iz|_{H^1(K)}^2 &\leq C \sum_{K \in \mathcal{T}_2} \left(\frac{h}{h_K} \right)^{2\beta} |\hat{z}|_{H^2(\hat{K})}^2 \\ &\leq C \sum_{K \in \mathcal{T}_2} \left(\frac{h}{h_K} \right)^{2\beta} h_K^2 \, |z|_{H^2(K)}^2 \\ &\leq C h^{2\beta} \sum_{K \in \mathcal{T}_2} |z|_{H^2(K)}^2 \leq C h^{2\beta} \|u - u_h\|_{L^2(\dot{K})}^2, \end{aligned} \tag{4.16}$$

wobei C unabhängig von h und β ist. Für die erste Summe nutzen wir $z|_{\Omega\setminus\tilde{\Omega}} \in \tilde{\mathcal{B}}^2_{1-\delta_0}(C_{\tilde{\Omega}}\|u-u_h\|_{L^2(\dot{K})}, \gamma_z)$. Nach Lemma 4.2.13 und da kein $K \in \mathcal{T}_1$ einen kleineren Abstand als $ch_{\dot{K}}$ von $\dot{K}$ besitzt, erhalten wir

$$\sum_{K\in\mathcal{T}_1} |z-Iz|^2_{H^1(K)} \leq \sum_{K\in\mathcal{T}_1|K\cap\partial\Omega\neq\emptyset} C^2_{K,z}h^{2\delta_0} + \sum_{K\in\mathcal{T}_1|K\cap\partial\Omega\,=\emptyset} C^2_{K,z}h_K^{2(\delta_0-\alpha b')}h^{2\alpha b'},$$

wobei $C_{K,z}$ die Konstanten aus Lemma 4.2.13 bezüglich z bezeichnen. Für α hinreichend groß folgt somit

$$\sum_{K\in\mathcal{T}_1} |z-Iz|^2_{H^1(K)} \leq h^{2\delta_0}C\sum_{K\in\mathcal{T}_1} C^2_{K,z} \leq h^{2\delta_0}C\|u-u_h\|^2_{L^2(\dot{K})} \tag{4.17}$$

und die Kombination von (4.16) und (4.17) liefert die Behauptung. □

Bemerkung 4.2.19. Im Beweis zu Lemma 4.2.18 haben wir gefordert, dass kein $K \in \mathcal{T}_2$ existiert mit $\overline{K}\cap\partial\Omega \neq \emptyset$. Diese Forderung stellt jedoch keine Einschränkung dar, da sie sich automatisch einstellt, sobald h hinreichend klein ist.

Lemma 4.2.20. *Seien die Voraussetzungen von Theorem 4.2.5 erfüllt. Ferner seien z sowie z_h gegeben durch Definition 4.2.8 und $\omega_{\beta,\mathcal{T}}$ gegeben durch Definition 4.2.7. Dann existiert ein $\beta' \in (0,\delta_0]$, nur abhängig von Ω, Ω', der Formregularitätskonstanten γ und den Konstanten aus den Definitionen 4.1.4, 4.1.6, so dass*

$$\left\|\sqrt{\omega_{2\beta,\mathcal{T}}}\nabla(z-z_h)\right\|_{L^2(\Omega)} \leq C_{\beta'}h^{\beta}\|u-u_h\|_{L^2(\dot{K})} \tag{4.18}$$

für alle $\beta \in (0,\beta']$. Des Weiteren ist $C_{\beta'}$ unabhängig von h und β.

Beweis. Wir setzen $e = z - z_h$ und es gilt

$$\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla e\right\|^2_{L^2(\Omega)} = \int_\Omega \nabla e\cdot\nabla(\omega_{\beta,\mathcal{T}}e)\mathrm{d}\Omega - \int_\Omega e\nabla e\cdot\nabla\omega\mathrm{d}\Omega.$$

Lemma 4.2.11 sichert die Existenz von $\tilde{\beta}\in(0,2]$ und $C'_{\tilde{\beta}}>0$, so dass

$$\begin{aligned}\left|\int_\Omega e\nabla e\cdot\nabla w\mathrm{d}\Omega\right| &\leq C\beta\int_\Omega\left|\sqrt{\omega_{\beta,\mathcal{T}}}\frac{e}{r}\right|\left|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla e\right|\mathrm{d}\Omega\\ &\leq C\beta\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\frac{e}{r}\right\|_{L^2(\Omega)}\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla e\right\|_{L^2(\Omega)}\\ &\leq CC'_{\tilde{\beta}}\beta\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla e\right\|^2_{L^2(\Omega)}\end{aligned}$$

für alle $\beta\in(0,\tilde{\beta}]$. $C_{\tilde{\beta}}$ wächst in $\tilde{\beta}$ monoton (siehe Beweis zu Lemma 4.2.11) und wir fordern $\tilde{\beta}$ so klein, dass $CC'_{\tilde{\beta}}\tilde{\beta}<1$ und erhalten somit

$$\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla e\|^2_{L^2(\Omega)} \leq C_{\tilde{\beta}}\left|\int_\Omega\nabla e\cdot\nabla(\omega_{\beta,\mathcal{T}}e)\mathrm{d}\Omega\right|$$

für alle $\beta \in (0, \tilde{\beta}]$ und $C_{\tilde{\beta}} := \frac{1}{1-CC'_{\tilde{\beta}}\tilde{\beta}}$. Als Nächstes fügen wir das Element $Iz \in S_0^{\mathbf{p}}(\Omega, \mathcal{T})$ aus Lemma 4.2.18 ein und wenden die Dreiecksungleichung an:

$$\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla e\right\|^2_{L^2(\Omega)} \leq C_{\beta'}\left|\int_\Omega \nabla e \cdot \nabla(\omega_{\beta,\mathcal{T}}(z - Iz))\mathrm{d}\Omega\right| + C_{\tilde{\beta}}\left|\int_\Omega \nabla e \cdot \nabla(\omega_{\beta,\mathcal{T}}(Iz - z_h)\mathrm{d}\Omega\right|.$$

Mittels Cauchy-Schwarz-Ungleichung, Lemma 4.2.10 und Lemma 4.2.11 kann der erste Term dieser Summe für alle $\beta \in (0, \tilde{\beta}]$ wie folgt abgeschätzt werden:

$$\begin{aligned}\left|\int_\Omega \nabla e \cdot \nabla(\omega_{\beta,\mathcal{T}}(z - Iz))\mathrm{d}\Omega\right| &\leq \left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla e\right\|_{L^2(\Omega)} \left\|\frac{1}{\sqrt{\omega_{\beta,\mathcal{T}}}}\nabla\left(\omega_{\beta,\mathcal{T}}(z - Iz)\right)\right\|_{L^2(\Omega)} \\ &\leq (1 + CC_{\tilde{\beta}}\beta)\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla e\right\|_{L^2(\Omega)}\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla(z - Iz)\right\|_{L^2(\Omega)}.\end{aligned}$$

Für den zweiten Term nutzen wir Galerkin-Orthogonalitäten und erhalten:

$$\begin{aligned}\left|\int_\Omega \nabla e \cdot \nabla(\omega_{\beta,\mathcal{T}}(Iz - z_h))\mathrm{d}\Omega\right| &= \left|\int_\Omega \nabla e \cdot \nabla(\omega_{\beta,\mathcal{T}}(Iz - z_h) - \overline{I}(\omega_{\beta,\mathcal{T}}(Iz - z_h)))\mathrm{d}\Omega\right| \\ &\leq CC_{\tilde{\beta}}\beta\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla e\right\|_{L^2(\Omega)}\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla(Iz - z_h)\right\|_{L^2(\Omega)},\end{aligned}$$

für alle $\beta \in (0, \tilde{\beta}]$, wobei $\overline{I} : S_0^{2\mathbf{p}}(\Omega, \mathcal{T}) \to S_0^{\mathbf{p}}$ den Operator aus Lemma 4.2.12 bezeichnet. Kombinieren wir schließlich die Abschätzungen des ersten und zweiten Terms, so erhalten wir

$$\begin{aligned}\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla e\right\|^2_{L^2(\Omega)} \quad\leq\quad & (1 + CC_{\tilde{\beta}}\beta)\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla e\right\|_{L^2(\Omega)}\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla(z - Iz)\right\|_{L^2(\Omega)} + \\ & CC_{\tilde{\beta}}\beta\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla e\right\|_{L^2(\Omega)}\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla(z - Iz)\right\|_{L^2(\Omega)} + \\ & CC_{\tilde{\beta}}\beta\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla e\right\|_{L^2(\Omega)}\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla e\right\|_{L^2(\Omega)}.\end{aligned}$$

Für $\tilde{\beta}$ hinreichend klein haben wir somit

$$\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla e\right\|^2_{L^2(\Omega)} \leq C_{\tilde{\beta}}\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla e\right\|_{L^2(\Omega)}\left\|\sqrt{\omega_{\beta,\mathcal{T}}}\nabla(z - Iz)\right\|_{L^2(\Omega)}$$

für alle $\beta \in (0, \tilde{\beta}]$. Folglich, mit $\beta' := \frac{\tilde{\beta}}{2} \in (0, 1]$, erhalten wir

$$\left\|\sqrt{\omega_{2\beta,\mathcal{T}}}\nabla e\right\|_{L^2(\Omega)} \leq C_{\beta'}\left\|\sqrt{\omega_{2\beta,\mathcal{T}}}\nabla(z - Iz)\right\|_{L^2(\Omega)}$$

für alle $\beta \in (0, \beta']$ und Lemma 4.2.18 liefert die Behauptung (4.18). □

4.2.2 Numerische Beispiele

In diesem Abschnitt betrachten wir einige Beispiele, welche unsere theoretische Aussage bezüglich der verbesserten lokalen Konvergenz im Gebietsinneren verifizieren. Alle hier angegebenen Resultate wurden mit dem von uns implementierten 2D-*hp*-FE-Programmpaket ADURAKON erzielt. In allen Rechnungen starten wir mit einem Grobnetz $\mathcal{N}_0$ und erzeugen durch rot-Verfeinerung aller Randdreiecke und anschließender Beseitigung der hängenden Knoten mittels grünem Abschluss eine Folge hierarchisch geschachtelter geometrischer Netze $\{\mathcal{N}_l\}_{l=0,1,\ldots}$ mit

$$\mathcal{T}(\mathcal{N}_0) \subset \mathcal{T}(\mathcal{N}_1) \subset \mathcal{T}(\mathcal{N}_2) \subset \ldots$$

und Randgitterweite $h_l = 2^{-l}h_0$.

Algorithmus 4.2.21 (Randkonzentrierte Verfeinerung)**.**

- *Input: Netz* $\mathcal{N}_i$
- *Output: Verfeinertes Netz* $\mathcal{N}_{i+1}$
- *Algorithmus:*
  1. *Zerlege alle* $K \in \mathcal{T}(\mathcal{N}_i)$ *mit* $\overline{K} \cap \partial\Omega \neq \emptyset$ *in vier kongruente Dreiecke.*
  2. *Solange ein Dreieck mit mehr als einem hängenden Knoten existiert, zerlege dies in vier kongruente Dreiecke.*
  3. *Zerlege alle Dreiecke mit einem hängenden Knoten in zwei Dreiecke (grüner Abschluss).*

Lemma 4.2.22 (Randkonzentrierte Verfeinerung)**.** *Sei* $\mathcal{N}_0$ *ein beliebiges Grobnetz und* $\{\mathcal{N}_i\}_{i=0,1,2,\ldots}$*, die Folge der mittels Algorithmus 4.2.21 erzeugten verfeinerten Netze. Zu* $e \in e(\mathcal{N}_i)$ *bezeichnen* $v_a(e), v_e(e) \in v(\mathcal{N}_i)$ *die beiden Endpunkte der Kante und zu* $K \in \mathcal{T}(\mathcal{N}_i)$ *bezeichnen* $v_1(K), v_2(K), v_3(K) \in v(\mathcal{N}_i)$ *die drei Eckknoten von* K*. Mit*

$$W^l(\mathcal{N}_i) := \Big\{ (v_0, \ldots, v_l) \in (v(\mathcal{N}_i))^{l+1} \;\Big|\; \exists e_j \in e(\mathcal{N}_i) \text{ mit } \{v_a(e_j), v_e(e_j)\} = \{v_j, v_{j+1}\} \; \forall\, 0 \leq j < l \Big\}$$

bezeichnen wir die Menge aller Wege der Länge $l \geq 0$ *im Netz* $\mathcal{N}_i$ *und mit* $v_j(w)$*,* $0 \leq j \leq l$*, den* j*-ten Vertex des Weges* $w = (v_0, \ldots, v_l) \in W^l$*. Sei*

$$D(v, \mathcal{N}_i) := \min\{l \mid \exists w \in W^l(\mathcal{N}_i) \text{ mit } v_0(w) \in \partial\Omega,\ v_l(w) = v\}$$

der diskrete Abstand von $v \in v(\mathcal{N}_i)$ *zum Rand* $\partial\Omega$*. Dann gilt:*

- *Aus* $K \in \mathcal{T}(\mathcal{N}_i)$ *mit* $\sum_{k=1}^{3} D(v_k(K), \mathcal{N}_i) \geq 5$ *folgt* $K \in \mathcal{T}(\mathcal{N}_{i+1})$*.*
- *Wird im* i*-ten Netz ein Element* $K \in \mathcal{T}(\mathcal{N}_i)$ *grün verfeinert, so wird bei fortgesetzter Anwendung von Algorithmus 4.2.21 keiner der beiden Söhne nochmals verfeinert.*

Beweis. Wir betrachten die Anwendung von Algorithmus 4.2.21 auf das Netz $\mathcal{N}_i$:

1. Es ist leicht einzusehen, dass bei einem Durchlauf von Algorithmus 4.2.21, d.h. beim Generieren von $\mathcal{N}_{i+1}$ aus $\mathcal{N}_i$, jede Kante höchstens einmal geteilt wird und von keinem Dreieck $K \in \mathcal{T}(\mathcal{N}_i)$ Enkel oder höhere Generationen entstehen.

2. Definieren wir für $(k, l) \in \{(0,0), (0,1), (1,1)\}$ die Mengen

   $$E_i^{kl} := \{e \in e(\mathcal{N}_i) \mid \{D(v_a(e), \mathcal{N}_i), D(v_e(e), \mathcal{N}_i)\} = \{k, l\}\},$$

   so folgt unmittelbar aus Algorithmus 4.2.21, dass alle Kanten $e \in E_i^{00} \cup E_i^{01}$ geteilt werden (man beachte, dass laut Definition $E_i^{kl} = E_i^{lk}$ gilt). Umgekehrt können wir jedoch auch zeigen, dass im Verlauf von Algorithmus 4.2.21 keine Kante aus

   $$E^I := e(\mathcal{N}_i) \setminus \left(E_i^{00} \cup E_i^{01} \cup E_i^{11}\right)$$

geteilt wird: In Schritt 1 des Algorithmus werden ausschließlich Dreiecke $K \in \mathcal{T}(\mathcal{N}_i)$ mit $\max\{D(v_k(K), N_i) \mid k = 1,2,3\} \leq 1$ geteilt. Folglich kann zu Beginn von Schritt 2 noch keine Kante $e \in E^I$ geteilt worden sein. Angenommen im Verlauf von Schritt 2 werden Kanten $e \in E^I$ geteilt, so betrachten wir das erstmalige Auftreten dieses Ereignisses. Sei K das betroffene Dreieck, so muss K zwei hängende Knoten besitzen und die dritte noch ungeteilte Kante aus der Menge E^I stammen. Da wir die erstmalige Teilung einer Kante aus E^I betrachten, müssen die beiden Kanten mit den hängenden Knoten in der Vereinigung $E_i^{00} \cup E_i^{01} \cup E_i^{11}$ liegen. Daraus folgt jedoch $\max\{D(v_k(K), \mathcal{N}_i) \mid k = 1,2,3\} \leq 1$ und wir erhalten einen Widerspruch zu der Annahme, dass die dritte Kante zur Menge E^I gehört. In Schritt 3 werden keine weiteren Kanten geteilt.

3. Betrachten wir die Menge aller in Schritt 3 grün zu zerteilenden Dreiecke. Offensichtlich, da alle Dreiecke, die direkt am Rand liegen, bereits im ersten Schritt in vier kongruente Dreiecke zerlegt werden, muss für ein grün zu zerteilendes Dreieck $K \in \mathcal{T}(\mathcal{N}_i)$ gelten: $\min\{D(v_k(K), \mathcal{N}_i) \mid k = 1,2,3\} \geq 1$. Da nach Punkt 2 mindestens eine Kante von K aus $E_i^{00} \cup E_i^{01} \cup E_i^{11}$ stammen muss, kommen somit nur Dreiecke K mit

$$\Big\{D(v_1(K), \mathcal{N}_i), D(v_2(K), \mathcal{N}_i), D(v_3(K), \mathcal{N}_i)\Big\} \in \Big\{\{1,1,1\}, \{1,1,2\}\Big\}$$

für die grüne Teilung in Betracht.

4. Aus Punkt 2 folgt, dass in den ersten beiden Schritten von Algorithmus 4.2.21 nur Dreiecke $K \in \mathcal{T}(\mathcal{N}_i)$ mit $\sum_{k=1}^3 D(v_k(K), \mathcal{N}_i) \leq 3$ rot geteilt werden. Zusammen mit Punkt 3 folgt somit die Behauptung, dass alle Dreiecke $K \in \mathcal{T}(\mathcal{N}_I)$ mit $\sum_{k=1}^3 D(v_k(K), \mathcal{N}_i) \geq 5$ unverändert bleiben.

5. Die Tatsache, dass ein aus der grünen Verfeinerung von $K \in \mathcal{T}(\mathcal{N}_i)$ stammendes Sohndreieck $K_S \in \mathcal{T}(\mathcal{N}_{i+1})$ nicht weiter verfeinert wird, folgt unmittelbar aus den Feststellungen:

$$D(v, \mathcal{N}_i) = 1 \;\Rightarrow\; D(v, \mathcal{N}_{i+1}) = 2 \qquad \forall\, v \in v(\mathcal{N}_i),$$

$$D(v, \mathcal{N}_{i+1}) \geq D(v, \mathcal{N}_i) \qquad \forall\, v \in v(\mathcal{N}_i),$$

$$D(v, \mathcal{N}_{i+1}) \in \{2,3\} \quad \text{falls } v \in v(\mathcal{N}_{i+1}) \backslash v(\mathcal{N}_i) \text{ Mittelpunkt von } \in e(\mathcal{N}_i) \in E_i^{11}.$$

Zusammen mit Punkt 3 ergibt sich für K_S damit nämlich

$$\sum_{k=1}^{3} D(v_k(K_S), \mathcal{N}_j) \geq 5 \qquad \forall\, j > i.$$

□

Bemerkung 4.2.23. Lemma 4.2.22 garantiert, dass trotz grüner Verfeinerungen die Netzverfeinerung zu keiner Entartung der Netze führt und nur Dreiecke in unmittelbarer Randnähe zerteilt werden.

Die zum Netz N_l gehörige Polynomgradverteilung $p(\mathcal{N}_l)$ definieren wir mittels

$$p_{K,l} := \left\lfloor \frac{3}{2} + \alpha \ln\left(\frac{h_K}{\underline{h_l}}\right) \right\rfloor \;\; \forall\, K \in \mathcal{T}(\mathcal{N}_l), \;\; \underline{h_l} := \min\{\text{length}(e) \mid e \in e(\mathcal{N}_l)\},$$

wobei der Anstieg α jeweils für alle Level eines Beispiels gleich ist. Um unsere Aussage über die Konvergenz im Gebietsinneren zu überprüfen, wählen wir einen Punkt $\dot{x} \in \Omega \backslash \{e \mid e \in e(\mathcal{N}_l) \text{ für ein } l \leq 0\}$ im Gebietsinneren und betrachten die Folge $\{\dot{K}_l\}_{l=0,1,\ldots}$, wobei $\dot{K}_l$ das eindeutig bestimmte Dreieck mit $\dot{K}_l \in \mathcal{T}(\mathcal{N}_l)$ und $\dot{x} \in \dot{K}_l$ ist. Aus Lemma 4.2.22 folgt die Existenz von $L > 0$ mit $\dot{K}_m = \dot{K}_n$ für alle $n, m \geq L$ und somit ist $\{\dot{K}_l\}_{l=0,1,\ldots}$ eine geeignete Folge zur Berechnung des lokalen Fehlers $\{\|u - u_l\|_{W^{(\ldots)}(\dot{K}_l)}\}_{l=0,1,\ldots}$ in Abhängigkeit von der Randgitterweite h.

Beispiel 4.2.24. *Wir betrachten auf* $\Omega_L = (0,1)^2 \backslash ([0,1] \times [-1,0])$ *das Randwertproblem*

$$-\Delta u = f \text{ auf } \Omega_L, \qquad u = 0 \text{ auf } \partial\Omega$$

mit einer rechten Seite f*, so dass die exakte Lösung* $u \in H^{5/3-\epsilon}(\Omega_L)$*,* $\epsilon > 0$*, gegeben ist durch*

$$u = r^{\frac{2}{3}} \sin\left(\frac{2}{3}\varphi\right) \left(1 - r^2 \cos^2 \varphi\right) \left(1 - r^2 \sin^2 \varphi\right).$$

Da $u \in \cap_{\epsilon>0} H^{5/3-\epsilon}(\Omega_L)$, erwarten wir eine globale H^1-Konvergenz von $O(N^{-2/3})$ sowie nach Theorem 4.2.5 eine verbesserte lokale Konvergenz im Gebietsinneren. Für unsere Rechnung wählen wir $\alpha = 1$, KS-Formfunktionen (siehe Kapitel 3) sowie für die lokale Konvergenz die beiden inneren Punkte $\dot{x}_1 = (0.4, 0.3)$, $\dot{x}_2 = (0.1, 0.2)$. Die Ergebnisse sind in Tabelle 4.1 und Abbildung 4.2 dargestellt. Wie wir sehen können, ist die lokale Konvergenzrate in den beiden markierten Elementen in etwa doppelt so groß wie die globale Konvergenzordnung.
Mit dem zweiten Beispiel wollen wir Theorem 4.2.5 auch für ein Gebiet mit etwas komplizierterem Rand verifizieren.

Beispiel 4.2.25. *Wir betrachten auf* Ω_S *aus Abbildung 4.3 das Randwertproblem*

$$-\Delta u = 1 \text{ auf } \Omega_S, \qquad u = 0 \text{ auf } \partial\Omega.$$

Diesmal benutzen wir die Lagrange-Formfunktionen (siehe Kapitel 3) und da wir die exakte Lösung nicht kennen, ziehen wir die Lösungen vom Verfeinerungslevel 9 als „exakte Lösung“ heran, um eine Approximation der Fehler auf den Levels 1 bis 7 zu bestimmen. Die in Abbildung 4.3 dargestellten Ergebnisse belegen auch diesmal eine verbesserte lokale Konvergenz auf dem markierten Dreieck im Inneren.

4.2.3 Bemerkungen

Bemerkung 4.2.26 (gemischte Randbedingungen). Der Beweis von Theorem 4.2.5 beruht auf der Hardy-Ungleichung und ist damit auf Dirichlet-Randbedingungen beschränkt. Numerische Experimente zeigen jedoch, dass ähnliche Resultate auch für andere Randbedingungen gelten.

Beispiel 4.2.27 (gemischte Randbedingungen)**.** *Wir betrachten auf* $\Omega_L = (0,1)^2 \backslash ([0,1] \times [-1,0])$ *das Randwertproblem*

$$\begin{aligned} -\Delta u &= f \quad \text{auf} \quad \Omega = (-1,1)^2 \backslash ([0,1] \times [-1,0]) \\ \gamma_1 u &= 0 \quad \text{auf} \quad \partial\Gamma_N = (\{-1\} \times [-1,1]) \cup ([-1,1] \times \{1\}) \\ u &= 0 \quad \text{auf} \quad \partial\Gamma_D = \partial\Omega \backslash \Gamma_N, \end{aligned}$$

mit einer rechten Seite f*, so dass die exakte Lösung* $u \in H^{5/3-\epsilon}(\Omega_L)$*,* $\epsilon > 0$*, gegeben ist durch*

$$u = r^{\frac{2}{3}} \sin\left(\frac{2}{3}\varphi\right) \left(1 - r^2 \cos^2 \varphi\right) (1 + r \cos\varphi) \left(1 - r^2 \sin^2 \varphi\right) (1 - r \sin\varphi).$$

Tabelle 4.1: Beispiel 4.2.24 mit $e = u - u_h$ und KS-Formfunktionen

				$x_1 = (0.4, 0.3)$		$x_2 = (0.1, 0.2)$	
Level	h	N	p_{max}	$\|e\|_{L^2(\dot{K})}$	$\|\nabla e\|_{L^2(\dot{K})}$	$\|e\|_{L^2(\dot{K})}$	$\|\nabla e\|_{L^2(\dot{K})}$
1	5.000e-01	17	1	—	—	—	—
2	2.500e-01	69	1	2.4e-03	2.9e-02	2.7e-03	2.7e-02
3	1.250e-01	165	1	1.9e-03	2.8e-02	1.7e-03	2.5e-02
4	6.250e-02	462	2	2.8e-04	1.7e-03	5.0e-04	3.6e-03
5	3.125e-02	1054	3	5.6e-05	7.6e-04	1.5e-04	1.4e-03
6	1.562e-02	2402	3	1.9e-05	1.5e-04	5.0e-05	5.1e-04
7	7.812e-03	5307	4	7.0e-06	3.7e-05	1.8e-05	1.0e-04
8	3.906e-03	11091	5	2.7e-06	1.3e-05	7.1e-06	4.0e-05
9	1.953e-03	23023	6	1.0e-06	5.0e-06	2.8e-06	1.5e-05
10	9.766e-04	47107	6	4.3e-07	2.0e-06	1.1e-06	5.4e-06
11	4.883e-04	95652	7	1.7e-07	7.9e-07	4.4e-07	2.0e-06
12	2.441e-04	192620	8	6.9e-08	3.1e-07	1.7e-07	8.7e-07
13	1.221e-04	387040	8	2.7e-08	1.2e-07	7.0e-08	3.3e-07
14	6.104e-05	776369	9	1.0e-08	4.9e-08	2.7e-08	1.2e-07
15	3.052e-05	1554857	10	4.3e-09	1.9e-08	1.1e-08	5.2e-08

Wie bereits in Beispiel 4.2.24 betrachten wir auch diesmal die Dreiecke um die Punkte $x_1 = (0.4, 0.3)$, $x_2 = (0.1, 0.2)$ und wählen $\alpha = 1$. Die Ergebnisse unsere Rechnung sind in Abbildung 4.4 dargestellt. Wie wir deutlich sehen, erhalten wir auch hier eine verbesserte Konvergenzrate auf den Elementen zu $x_1 = (0.4, 0.3)$ bzw. $x_2 = (0.1, 0.2)$.

Bemerkung 4.2.28 (Größenordnung von β). In allen unseren numerischen Experimenten erhalten wir $\beta \approx \delta$, falls $u \in H^{1+\delta}(\Omega)$.

Abbildung 4.2: Beispiel 4.2.24: Gitter Level 6 - Ergebnisse

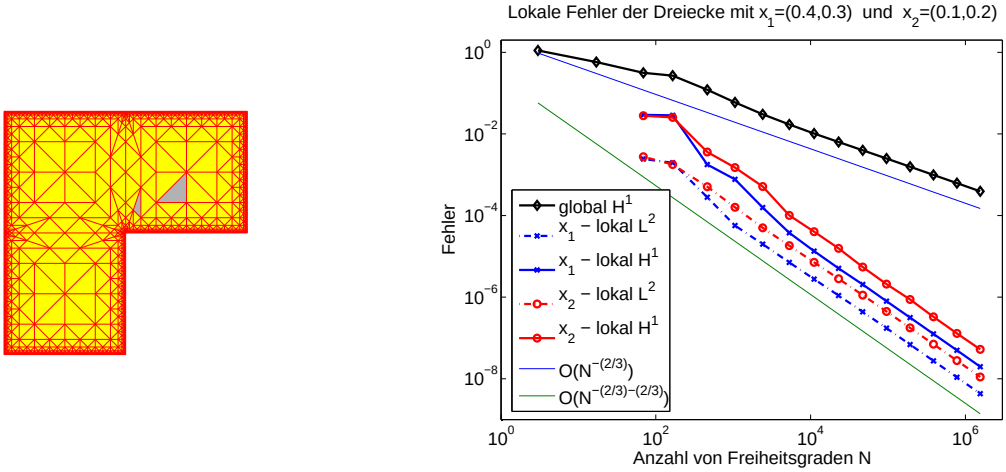


Abbildung 4.3: Beispiel 4.2.25: Grobgitter - Ergebnisse

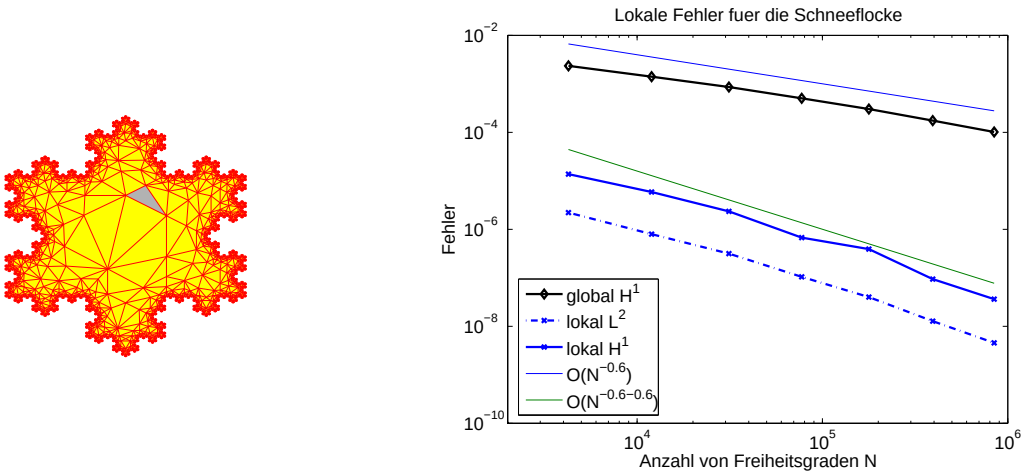
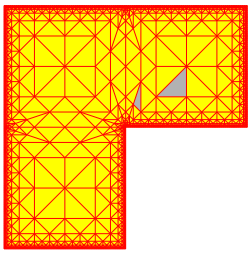
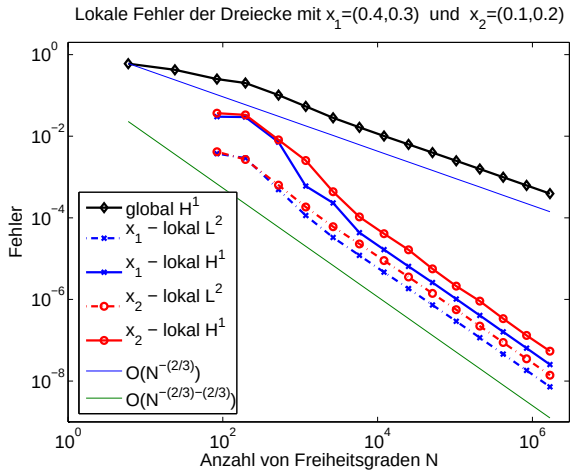


Abbildung 4.4: Beispiel 4.2.27: Gitter Level6 - Ergebnisse





4.3 Multilevel-Vorkonditionierer für die randkonzentrierte *hp*-FEM

Wollen wir die Lösung eines Randwertproblems mittels randkonzentrierter Finiter-Element-Methode bestimmen, so führt dies letztendlich auf ein großdimensionales lineares Gleichungssystem, wobei die Anzahl N der Unbekannten durchaus eine Größenordnung von einigen Millionen annehmen kann. Speziell für den 2D-Fall wurde bereits in [KM02] ein auf der LU-Zerlegung der Steifigkeitsmatrix beruhender Lösungsalgorithmus vorgestellt, der uns die gesuchte Lösung nach $O(N \log^8 N)$ Rechenoperationen liefert. Wollen wir jedoch 3D-Probleme lösen, die schwache Besetztheit der Steifigkeitsmatrix ausnutzen, oder haben wir, um Speicherplatz zu sparen, die Steifigkeitsmatrix gar nicht explizit aufgestellt, sondern stattdessen nur eine Matrix-Vektor-Multiplikation implementiert (siehe Kapitel 3), so kommen eigentlich nur iterative Lösungsverfahren in Frage. Das wohl bekannteste und am weitesten verbreitete Iterationsverfahren ist der PCG-Algorithmus (siehe z.B. [Hac94]). Betrachten wir das zu lösende lineare Gleichungssystem

$$A\underline{u} = b \tag{4.19}$$

mit symmetrisch positiv definiter Matrix $A \in \mathbb{R}^{N\times N}$, $b \in \mathbb{R}^N$ und geeignetem Vorkonditionierer $B^{-1} \in \mathbb{R}^{N\times N}$, ebenfalls symmetrisch positiv definit, so bestimmt der PCG-Algorithmus, ausgehend von einer Startnäherung $\underline{u}_0 \in \mathbb{R}^N$, eine Folge $\{\underline{u}^k\}_{k=0,1,\ldots} \subset \mathbb{R}^N$, die mit

$$\langle \underline{u}, \underline{v} \rangle_A := \langle A\underline{u}, \underline{v} \rangle , \qquad \|\underline{u}\|_A^2 := \langle \underline{u}, \underline{u} \rangle_A$$

und exakte Lösung $\underline{u}^*$, der Abschätzung

$$\|\underline{u}^* - \underline{u}^k\|_A \leq 2 \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^k \|\underline{u}^* - \underline{u}^0\|_A, \qquad \kappa = \mathrm{cond}_2(AB^{-1}) = \frac{\lambda_{max}(AB^{-1})}{\lambda_{min}(AB^{-1})}$$

genügt.
Wie wir an obiger Fehlerabschätzung sehen, ist die Konditionszahl $\kappa = \mathrm{cond}_2(AB^{-1})$ von entscheidender Bedeutung für die Konvergenzgeschwindigkeit des PCG-Algorithmus. Je näher κ an 1 liegt, um so weniger Iterationsschritte sind notwendig, um eine vorgegebene Genauigkeit zu erzielen. Leider ist bei der randkonzentrierten FEM aber ausschließlich das reine Dirichlet-Problem von Haus aus gut konditioniert. Aus der hierfür in [KM03] gezeigten Schranke von $O(\log^\beta N)$ für die Konditionszahl der diagonalskalierten Steifigkeitsmatrix, wobei $\beta \geq 0$ von der Wahl der Formfunktionen abhängt, folgt, dass sich im Falle reiner Dirichlet-Randbedingungen selbst mit simpler Diagonalskalierung noch akzeptable Konvergenzraten im PCG erzielen lassen.
Ganz anders sieht die Sache jedoch aus, sobald wir es mit Neumann- bzw. gemischten Randbedingungen zu tun haben. Um auch diese Probleme effektiv lösen zu können, ist ein guter Vorkonditionierer unentbehrlich (siehe auch Abbildung 4.5, 4.6 bzw. Tabelle 4.2).
Ein erster, auf der schnellen Realisierung von $H^{1/2}(\partial\Omega)$-ähnlichen Normen basierender Vorkonditionierer, welcher auf polylogarithmisch in N anwachsende Konditionszahlen für (AB^{-1}) führt, wurde in [KM03] kurz vorgestellt.
Wir werden im nun Folgenden zwei Vorkonditionierer für die randkonzentrierten FEM vorstellen, die sowohl für 2- als auch 3-dimensionale Probleme mit Dirichlet-, Neumann- oder gemischten Randbedingungen anwendbar sind und dabei zu optimalen, von der Problemgröße N unabhängigen Konditionszahlen $O(1)$ führen. Beide Vorkonditionierer basieren auf

der Additiv-Schwarz-Methode (siehe zum Beispiel [Nep86, Osw94, TW05]) und lassen sich zudem mit optimaler Komplexität $O(N)$ realisieren. [2]

4.3.1 Modellproblem

Es sei $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, ein polygonales Lipschitz-Gebiet mit geeigneter Skalierung, so dass $\operatorname{diam}\Omega \leq 1$. Der Dirichlet-Rand $\Gamma_D \subset \partial\Omega$ sei eine Vereinigung von Kanten (für $d=2$) oder Flächen (für $d=3$), wobei $\Gamma_D = \emptyset$ zugelassen ist. Wir schreiben

$$H^1_D(\Omega) := \{u \in H^1(\Omega) \,|\, u|_{\Gamma_D} = 0\}$$

und betrachten für den Rest des Abschnittes 4.3 das in schwacher Formulierung gegebene Randwertproblem 4.3.1 sowie das im Rahmen der randkonzentrierten FEM daraus resultierende diskretisierte Problem 4.3.2.

Problem 4.3.1 (Modellproblem-Vorkonditionierung)**.** *Finde $u \in H^1_D(\Omega)$, so dass*

$$a(u,v) := \int_\Omega \left\langle \nabla u, \hat{A}(x)\nabla v \right\rangle + a_0(x)uv d\Omega = f(v) \qquad \forall\, v \in H^1_D(\Omega). \tag{4.20}$$

Problem 4.3.2 (BC-FE-Diskretisierung)**.** *Finde $u \in S^{\mathbf{p}}_D(\Omega,\mathcal{N}) := S^{\mathbf{p}}(\Omega,\mathcal{N}) \cap H^1_D(\Omega)$, so dass*

$$a(u,v) = f(v) \qquad \forall\, v \in S^{\mathbf{p}}_D(\Omega,\mathcal{N}).$$

Ferner gehen wir von affinen Elementtransformationen F_K aus und nehmen wieder an, dass $a_0 \in L^\infty(\Omega)$, $\hat{A} \in L^\infty(\Omega, \mathbb{R}^{d\times d})$ punktweise symmetrisch positiv definit ist (siehe auch Annahme 3.1.3) und eine Konstante $C > 0$ existiert mit

$$C^{-1}\|u\|^2_{H^1(\Omega)} \leq a(u,u) \leq C\|u\|^2_{H^1\Omega} \qquad \forall u \in H^1(\Omega).$$

4.3.2 Die Additiv-Schwarz-Methode (ASM)

Eine Methode zur Konstruktion und Analyse von Vorkonditionierern ist die so genannte Additiv-Schwarz-Methode (kurz ASM), die auf einer nicht notwendigerweise disjunkten Zerlegung des Finiten-Element-Raums

$$\mathcal{V} := S^{\mathbf{p}}_D(\Omega,\mathcal{N}) = \sum_{i=0}^{K} \mathcal{V}_i, \qquad N := \dim\mathcal{V}, \qquad N_i := \dim\mathcal{V}_i, \tag{4.21}$$

basiert. Haben wir eine solche Zerlegung vorliegen, so ist der zugehörige ASM-Vorkonditionierer $\mathcal{B}^{-1} : \mathcal{V}^* \to \mathcal{V}$ bereits vollständig bestimmt und lässt sich am einfachsten anhand seiner Wirkung auf das für alle $w \in \mathcal{V}$ definierte Residuenfunktional $r_w := a(w,\cdot) - f(\cdot) \in \mathcal{V}^*$ darstellen:

$$\mathcal{B}^{-1} r_w := \sum_{i=0}^{K} u_i,$$

wobei die Funktionen $u_i \in \mathcal{V}_i$ die Teilraumlösungen von

$$a(u_i, v_i) = \langle r_w, v_i\rangle \qquad \forall v_i \in \mathcal{V}_i \tag{4.22}$$

[2] Abschnitt 4.3 fasst die wesentlichen Ergebnisse des Papers [EM05b] zusammen.

sind. Der Vorkonditionierer $\mathcal{B}^{-1}$ ist invertierbar und seine Inverse $\mathcal{B} : \mathcal{V} \to \mathcal{V}^*$ induziert eine symmetrisch positiv definite Bilinearform $b : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, welche auch durch

$$b(u,u) = \inf \left\{ a(u_i, u_i) \;\middle|\; u = \sum_{i=0}^{K} u_i,\ u_i \in \mathcal{V}_i \right\} \tag{4.23}$$

charakterisiert werden kann (siehe [TW05, Lemma 2.5]).

Bemerkung 4.3.3. [Matrixdarstellung $B^{-1} \in \mathbb{R}^{N\times N}$ des Vorkonditionierers $\mathcal{B}^{-1}$] Speziell für computerbasierte Rechnungen ist es notwendig, die Räume $\mathcal{V}$, $\mathcal{V}_i$ für $i \in \{0,\dots,K\}$ mit Basen auszustatten. Es sei daher

$$\Phi = \{\phi_j\}_{j=1}^{N} \text{ eine Basis von } \mathcal{V}, \qquad \Phi^i = \{\phi_j^i\}_{j=1}^{N_i} \text{ eine Basis von } \mathcal{V}_i.$$

Wir schreiben

$$[\Phi] := [\phi_1,\dots,\phi_N], \qquad [\Phi^i] := [\phi_1^i,\dots,\phi_{N_i}^i]$$

und es sei $I^i \in \mathbb{R}^{N\times N_i}$, gegeben durch:

$$[\Phi^i] = [\Phi] I^i, \quad \text{d.h.} \quad \phi_j^i = \sum_{k=1}^{N} \phi_k I_{kj}^i \qquad \forall 1 \le j \le N_i, \quad i \in \{0,\dots,K\}.$$

Ferner bezeichnen wir mit

$$A = [a(\phi_k,\phi_j)]_{j,k=1}^{N} \in \mathbb{R}^{N\times N}, \qquad \underline{r}_w = [r_w(\phi_i)]_{i=1}^{N} \in \mathbb{R}^{N\times 1}$$

die globale Steifigkeitsmatrix zu Problem 4.3.2 sowie den Residuenvektor zu $r_w \in \mathcal{V}^*$. Setzen wir $u = [\Phi]\underline{u} \in \mathcal{V}$ mit $\underline{u} \in \mathbb{R}^N$ und $u_i = [\Phi^i]\underline{u}_i \in \mathcal{V}_i$ mit $\underline{u}_i \in \mathbb{R}^{N_i}$, so folgt aus (4.22)

$$\begin{aligned}
& a\left([\Phi^i]\underline{u}_i, [\Phi^i]\underline{v}_i\right) &&= r_w\left([\Phi^i]\underline{v}_i\right) && \forall \underline{v}_i \in \mathbb{R}^{N_i} \\
\Leftrightarrow\ & a\left([\Phi] I^i \underline{u}_i, [\Phi] I^i \underline{v}_i\right) &&= r_w\left([\Phi] I^i \underline{v}_i\right) && \forall \underline{v}_i \in \mathbb{R}^{N_i} \\
\Leftrightarrow\ & \left(I^i \underline{v}_i\right)^T A \left(I^i \underline{u}_i\right) &&= \left(I^i \underline{v}_i\right)^T \underline{r}_w && \forall \underline{v}_i \in \mathbb{R}^{N_i} \\
\Leftrightarrow\ & \left(I^i\right)^T A \left(I^i\right) \underline{u}^i &&= \left(I^i\right)^T \underline{r}_w
\end{aligned}$$

und wir erhalten

$$u_i = [\Phi^i] \left[\left(I^i\right)^T A \left(I^i\right)\right]^{-1} \left(I^i\right)^T \underline{r}_w = [\Phi]\, I^i \left[\left(I^i\right)^T A \left(I^i\right)\right]^{-1} \left(I^i\right)^T \underline{r}_w. \tag{4.24}$$

Damit gilt

$$q_w := [\Phi]\underline{q}_w := \mathcal{B}^{-1} r_w \;=\; \sum_{i=0}^{K} [\Phi]\, I^i \left[\left(I^i\right)^T A \left(I^i\right)\right]^{-1} \left(I^i\right)^T \underline{r}_w$$

und wir erhalten eine Matrixrepräsentation $B^{-1} \in \mathbb{R}^{N\times N}$ des Vorkonditionierers $\mathcal{B}^{-1}$ mit $\underline{q}_w := B^{-1}\underline{r}_w$ und

$$B^{-1} = \sum_{i=0}^{K} \underbrace{I^i}_{Prolongation} \underbrace{\left[\left(I^i\right)^T A \left(I^i\right)\right]^{-1}}_{Teilraumlösung} \underbrace{\left(I^i\right)^T}_{Restriktion} .$$

4.3.3 Zwei Vorkonditionierer für die randkonzentrierte *hp*-FEM

Um für das aus Problem 4.3.2 resultierende Gleichungssystem geeignete ASM-Vorkonditionierer zu konstruieren, gehen wir von einer Folge geschachtelter geometrischer Netze

$$\mathcal{N}_0 \subset \mathcal{N}_1 \subset \ldots \subset \mathcal{N}_M = \mathcal{N} \tag{4.25}$$

mit Elementgrößen $h_K = \operatorname{diam}(K)$ für $K \in \mathcal{T}(\mathcal{N}_m)$, gegeben durch

$$h_K \sim \begin{cases} h_0(1/2)^m & : \quad \overline{K} \cap \partial\Omega \neq \emptyset \\ \operatorname{dist}(K, \partial\Omega) & : \quad \text{sonst} \end{cases} \tag{4.26}$$

aus. Ferner nehmen wir an, dass die Verfeinerung $\mathcal{N}_m \mapsto \mathcal{N}_{m+1}$ alle Elemente am Rand regulär rot verfeinert. D.h.

$$K \in \mathcal{T}(\mathcal{N}_m) \text{ und } \overline{K} \cap \partial\Omega \neq \emptyset \;\Rightarrow \mathcal{T}(\mathcal{N}_{m+1}) \text{ enthält alle } 2^d \text{ Söhne von } K. \tag{4.27}$$

Mit V_m bezeichnen wir die Menge der freien Knoten von $\mathcal{N}_m$, d.h. die Menge aller Knoten, die nicht auf dem Dirichlet-Rand liegen. Jedem Knoten $v \in V_m$ ordnen wir ein Patch

$$\overline{w}_v^m := \cup\{\overline{K} \mid K \in \mathcal{T}(\mathcal{N}_m) \mid v \in \overline{K}\}, \quad \omega_v^m := \overline{w}_v^m \backslash \partial\overline{w}_v^m \tag{4.28}$$

zu und führen die Räume der Hutfunktionen

$$\mathcal{V}_v^m = \{u \in S_D^1(\Omega, \mathcal{N}_m) \mid u(v') = 0 \; \forall v' \in V_m \backslash \{v\}\}, \quad \forall m \in \{0, \ldots, M\}, \; \forall v \in V_m \tag{4.29}$$

sowie die Räume der höherpolynomialen Patch-Funktionen

$$\mathcal{S}_v = \{u \in S_D^{\mathbf{p}}(\Omega, \mathcal{N}_M) \mid \operatorname{supp} u \subset \overline{\omega_v^M}\} \quad \forall v \in V_M \tag{4.30}$$

ein.
Mit diesen Definitionen sind wir nun bereits in der Lage, unseren ersten Vorkonditionierer für die randkonzentrierte *hp*-FEM vorzustellen:

Theorem 4.3.4 (Erster Vorkonditionierer)**.** *Sei $\{\mathcal{N}_i\}_{i=0,1,\ldots,M}$ eine Folge geschachtelter geometrischer Netze, welche den Bedingungen (4.25)–(4.27) genügen. Sei $p(\mathcal{N}_M)$ eine lineare Polynomgradverteilung und seien die Patches ω_v^m sowie die Räume $\mathcal{V}_v^m$, $\mathcal{S}_v$ wie in (4.28)–(4.30) definiert. Sei $I_m^B := \{v \in V_m \mid v \in \partial\Omega\}$. Dann bestimmt die Zerlegung*

$$S_D^{\mathbf{p}}(\Omega, \mathcal{N}_M) = \sum_{v \in V_M} \mathcal{S}_v + \sum_{m=0}^{M} \sum_{v \in I_m^B} \mathcal{V}_v^m \tag{4.31}$$

einen ASM-Vorkonditionierer $\mathcal{B}^{-1}$, dessen zugehörige Bilinearform $b(\cdot,\cdot)$ für eine von der Problemgröße N unabhängige Konstante $C > 0$ der Abschätzung

$$C^{-1} a(u,u) \leq b(u,u) \leq C a(u,u) \qquad \forall u \in S_D^{\mathbf{p}}(\Omega, \mathcal{N}_M)$$

genügt. Die Kosten für das Anwenden des Vorkonditionierers betragen $O(N)$.

Beweis. Siehe Abschnitt 4.3.6. □

Für die Definition unseres zweiten Vorkonditionierers versehen wir jeden Patch noch mit zwei Zahlen:

$$\begin{aligned} l(\omega_v^m) &:= \lceil -\log_2 \operatorname{diam} \omega_v^m \rceil, \\ g(\omega_v^m) &:= \min \left\{ k \in \{0, \dots, M\} \mid \exists \omega_{v'}^k \text{ mit } \omega_{v'}^k = \omega_v^m \right\}. \end{aligned} \tag{4.32}$$

Die erste Zahl, genannt das Level des Patches ω_v^m, ist ein diskretes Maß der Größe des Patches. Die zweite Zahl spezifiziert das Netz $\mathcal{N}_{g(\omega_v^m)}$, in dem der Patch ω_v^m zum ersten Mal auftritt. Offensichtlich gilt für beliebiges ω_i^m

$$0 \le g(\omega_i^m) \le m.$$

Aus unserer Skalierungsannahme $\operatorname{diam} \Omega \le 1$ und den Annahmen an die Elementgrößen h_K ergibt sich außerdem:

$$0 \le l(\omega_v^m) \le L := \max\{l(\omega_v^m) \mid m = 0, \dots, M, \;\; v \in V_m\} \le CM. \tag{4.33}$$

Theorem 4.3.5 (Zweiter Vorkonditionierer)**.** *Sei $\{\mathcal{N}_i\}_{i=0,1,\dots,M}$ eine Folge geschachtelter geometrischer Netze, welche den Bedingungen (4.25)–(4.27) genügen. Sei $p(\mathcal{N}_M)$ eine lineare Polynomgradverteilung und seien die Patches ω_v^m, die Räume $\mathcal{V}_v^m$, $\mathcal{S}_v$ sowie die Zahlen $l(\cdot)$, $g(\cdot)$ wie in (4.28)-(4.30), (4.32) definiert. Es sei $I_m := \{v \in V_m \mid g(\omega_v^m) = m\}$ und es existiere eine Konstante $C_1 > 0$, so dass*

$$g(\omega_v^m) \le l(\omega_v^m) + C_1 \quad \forall m \in \{0, \dots, M\}, \; \forall v \in I_m. \tag{4.34}$$

Dann bestimmt die Zerlegung

$$S_D^{\mathfrak{p}}(\Omega, \mathcal{N}_M) = \sum_{v \in V_M} \mathcal{S}_v + \sum_{m=0}^{M} \sum_{v \in I_m} \mathcal{V}_v^m \tag{4.35}$$

einen ASM-Vorkonditionierer $\mathcal{B}^{-1}$, dessen zugehörige Bilinearform $b(\cdot,\cdot)$ für eine von der Problemgröße N unabhängige Konstante $C > 0$ der Abschätzung

$$C^{-1} a(u,u) \le b(u,u) \le C a(u,u) \qquad \forall u \in S_D^{\mathfrak{p}}(\Omega, \mathcal{N}_M)$$

genügt. Die Kosten für das Anwenden des Vorkonditionierers betragen $O(N)$.

Beweis. Siehe Abschnitt 4.3.6. □

Bemerkung 4.3.6. Nach Definition von $g(\cdot)$ kann die Zerlegung (4.35) auch geschrieben werden als

$$S_D^{\mathfrak{p}}(\Omega, \mathcal{T}_M) = \sum_{v \in V_M} \mathcal{S}_v + \left(\sum_{m=0}^{M} \sum_{v \in V_m} \mathcal{V}_v^m \right)',$$

wobei der Strich bedeutet, dass in der Doppelsumme mehrfach auftretende Räume nur einmal berücksichtigt werden.

Bemerkung 4.3.7. Annahme (4.34) ist erfüllt, falls bei der Verfeinerung $\mathcal{N}_k \mapsto \mathcal{N}_{k+1}$ jeweils nur Dreiecke am Rand oder in einer begrenzten Anzahl von Schichten um den Rand zerteilt werden. Für die von uns benutze Verfeinerungsstrategie 4.2.21 zeigt Lemma 4.2.22, dass diese Bedingung erfüllt ist. Eine analoge Strategie kann auch für 3D konstruiert werden.

Bemerkung 4.3.8. Bezeichne wiederum $A \in \mathbb{R}^{N\times N}$ die Steifigkeitsmatrix zu Problem 4.3.2 bezüglich einer Basis Φ von $\mathcal{V} = S_D^{\mathbf{p}}(\Omega,\mathcal{N})$ und $B^{-1} \in \mathbb{R}^{N\times N}$ die Matrixdarstellung des Vorkonditionierers (siehe Bemerkung 4.3.3), so bedeuten die in den Theoremen 4.3.4 und 4.3.5 gezeigten Aussagen

$$\mathrm{cond}_2(B^{-1}A) \le C. \tag{4.36}$$

4.3.4 Numerische Beispiele

In diesem Abschnitt betrachten wir einige numerische Beispielrechnungen, welche die theoretischen Resultate bestätigen und zudem die Effizienz unserer Vorkonditionierer demonstrieren. Alle hier angegebenen Resultate wurden mit dem von uns implementierten 2D-*hp*-FE-Programmpaket ADURAKON erzielt.
Unser erstes Beispiel ist ein Randwertproblem auf dem L-Gebiet:

Beispiel 4.3.9. *Wir betrachten*

$$\begin{aligned} -\Delta u &= f \quad auf \quad \Omega = (0,1)^2\backslash([0,1]\times[-1,0]) \\ \frac{\partial u}{\partial n} &= 0 \quad auf \quad \Gamma_N = (\{-1\}\times[-1,1]) \cup ([-1,1]\times\{1\}) \\ u &= 0 \quad auf \quad \Gamma_D = \partial\Omega\backslash\Gamma_N \end{aligned}$$

mit einer rechten Seite f, so dass die exakte Lösung gegeben ist durch

$$u = r^{\frac{2}{3}}\sin\left(\frac{2}{3}\varphi\right)\left(1-r^2\cos^2\varphi\right)\left(1-r^2\sin^2\varphi\right)(1+r\cos\phi)(1-r\sin\phi).$$

Mit dem zweiten Beispiel wollen wir unsere theoretischen Ergebnisse auch für ein Gebiet mit kompliziertem Rand verifizieren:

Beispiel 4.3.10. *Auf dem in Abbildung (4.6) dargestellten Gebiet betrachten wir das Randwertproblem*

$$\begin{aligned} -\Delta u &= 1 \quad auf \quad \Omega \\ \frac{\partial u}{\partial n} &= 0 \quad auf \quad \Gamma_N = \{(x,y)\in\partial\Omega \mid y<0\} \\ u &= 0 \quad auf \quad \Gamma_D = \partial\Omega\backslash\Gamma_N. \end{aligned}$$

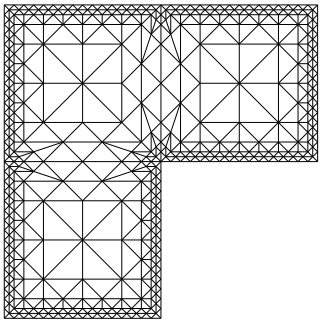
In beiden Beispielen starten wir mit einem Grobgitter $\mathcal{N}_0$ und erzeugen mittels Algorithmus 4.2.21 eine Folge hierarchisch geschachtelter geometrischer Netze $\{\mathcal{N}_j\}_{j=0,1,\dots,M}$ mit Randgitterweiten $h_j \sim 2^{-j}h_0$. Die zum Netz $\mathcal{N}_j$ gehörige Polynomgradverteilung definieren wir wie bisher zu

$$p_{K,j} := \left\lfloor \frac{3}{2} + \alpha\ln\left(\frac{h_K}{\underline{h_j}}\right)\right\rfloor \ \forall\, K\in\mathcal{T}(\mathcal{N}_j),\ \ \underline{h_j} := \min\{\mathrm{length}(e) \mid e\in e(\mathcal{N}_j)\},$$

wobei wir konkret $\alpha = 1$ wählen.
In unseren Beispielrechnungen betrachten wir für die Beispiele 4.3.9, 4.3.10 den nichtvorkonditionierten CG-Algorithmus sowie die Vorkonditionierer der Theoreme 4.3.4 und 4.3.5. Die Ergebnisse unserer Rechnungen sind in den Abbildungen 4.5 und 4.6 dargestellt. CG steht

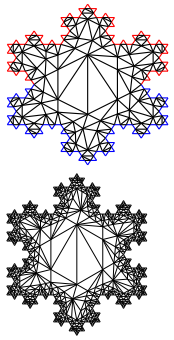
hierbei für den nicht vorkonditionierten CG-Algorithmus, PCG-1 bezeichnet den Vorkonditionierer von Theorem 4.3.4 und PCG-2 den Vorkonditionierer von Theorem 4.3.5. Alle Iterationen wurden mit $u^0 \equiv 0$ gestartet. Die angegebenen Iterationszahlen beziehen sich auf die Anzahl von Iterationsschritten, die notwendig ist, um das Residuum um einen vorgegebenen Faktor zu reduzieren. Wie zu erwarten war, bleiben die Iterationszahlen der PCG-Algorithmen beschränkt und liegen deutlich unter denen des CG-Algorithmus.

Abbildung 4.5: (Beispiel 4.3.9) Links: Netz von Level 4. Rechts: Iterationszahlen, um $\|\underline{r}^k\|^2/\|\underline{r}^0\|^2 \leq 10^{-12}$ zu erreichen



L	#Elemente	p_{max}	N	CG	PCG-1	PCG-2
0	12	1	6	4	4	4
1	48	1	24	14	14	11
2	168	1	84	29	25	18
3	392	1	196	46	36	22
4	840	2	525	87	44	24
5	1736	3	1181	195	51	25
6	3528	3	2657	263	55	27
7	7112	4	5818	368	60	27
8	14280	5	12114	474	63	28
9	28616	6	25070	581	66	29
10	57288	6	51202	691	69	30
11	114632	7	103843	884	71	31
12	229320	8	209003	1120	74	31
13	458696	8	419807	1316	77	31
14	917448	9	841904	1596	79	33
15	1834952	10	1685928	—	81	33

Abbildung 4.6: (Beispiel 4.3.10) Links: Grobgitter und Netz von Level 1 Rechts: Iterationszahlen, um $\|\underline{r}^k\|^2/\|\underline{r}^0\|^2 \leq 10^{-10}$ zu erreichen



L	#Elemente	p_{max}	N	CG	PCG-1	PCG-2
0	298	1	149	63	44	44
1	1054	3	619	185	78	57
2	3346	4	2096	318	79	62
3	8398	5	5227	477	104	65
4	18826	5	12963	659	106	70
5	40006	6	30024	927	112	78
6	82690	7	66228	1156	120	79
7	168382	7	141311	1644	128	80
8	340090	8	296938	2321	133	81
9	683830	9	613767	3282	138	81
10	1371634	10	1253012	4638	142	81

4.3.5 Komplexität der Vorkonditionierer

Die Wirkung unserer Vorkonditionierer aus den Theoremen 4.3.4 und 4.3.5 beruht auf dem Lösen von Teilproblemen bezüglich der Räume $\mathcal{S}_v$ und $\mathcal{V}_v^m$. Da die Räume $\mathcal{V}_v^m$ eindimensional sind, kann der Gesamtaufwand $W_{\mathcal{V}}$ für das Lösen der zugehörigen Teilprobleme durch die Gesamtzahl aller Knoten bzw. die Gesamtzahl aller Elemente auf allen Netzen abgeschätzt werden:

$$W_{\mathcal{V}} \leq C \sum_{m=0}^{M} \sum_{K \in \mathcal{T}(\mathcal{N}_m)} 1.$$

Analog zu Lemma 4.3.11 erhalten wir $\sum_{K \in \mathcal{T}(\mathcal{N}_m)} 1 \leq C h_m^{-(d-1)}$ und aus $h_m \sim h_0 2^{-m}$ ergibt sich

$$W_{\mathcal{V}} \leq C h_M^{-(d-1)}.$$

Da für die Anzahl der Elemente des feinsten Gitters $\#\mathcal{T}(\mathcal{N}_M) \geq C h_M^{-(d-1)}$ gilt, folgt

$$W_{\mathcal{V}} \leq CN, \qquad N = \dim S_D^{\mathbf{p}}(\mathcal{N}, \Omega).$$

Etwas komplizierter sind Abschätzungen für die Teilprobleme zu den $\mathcal{S}_v$-Räumen. Das folgende Lemma zeigt, dass der Aufwand, um diese Probleme mittels Cholesky-Zerlegung zu lösen, ebenfalls nur $O(N)$ beträgt.

Lemma 4.3.11. *Sei $\mathcal{N}_M$ ein geometrisches Netz mit Randgitterweite h_M und linearer Polynomgradverteilung $p(\mathcal{N}_M)$. Dann ist sowohl der Arbeitsaufwand $W_{\mathcal{S}}$ um die Cholesky-Faktoren der Teilraumsteifigkeitsmatrizen bezüglich $\mathcal{S}_v$ für alle $v \in V_M$ zu berechnen als auch der Speicheraufwand* Mem *um diese Faktoren zu speichern $O(N)$.*

Beweis. Aus der linearen Algebra ist bekannt, dass die Cholesky-Zerlegung einer $n \times n$ Matrix mit $O(n^3)$ Rechenoperationen bewerkstelligt werden kann und dass der notwendige Speicher um diese Zerlegung zu speichern $O(n^2)$ beträgt.
Da wir von formregulären Netzen ausgehen, ist die Anzahl der in einem beliebigen Knoten $v \in V_M$ aufeinander treffenden Elemente durch eine Konstante beschränkt. Des Weiteren ist nach Korollar 4.1.7 der Polynomgrad für alle Elemente $K \subset \omega_v^M$ von vergleichbarer Größe. Folglich kann die Dimension der Teilraumsteifigkeitsmatrix bezüglich $\mathcal{S}_v$ durch $n_v \leq C p_K^d$ mit beliebig gewähltem $K \subset \omega_v^M$ abgeschätzt werden. Somit ergibt sich:

$$\mathrm{Mem} \leq \sum_{v \in V_M} (C p_K^d)^2 \leq C \sum_{K \in \mathcal{T}(\mathcal{N}_M)} p_K^{2d}, \qquad \mathrm{W}_{\mathcal{S}} \leq \sum_{v \in V_M} (C p_K^d)^3 \leq C \sum_{K \in \mathcal{T}(\mathcal{N}_M)} p_K^{3d}.$$

Um die Summen $\sum_{K \in \mathcal{T}_M} p_K^{2d}$ und $\sum_{K \in \mathcal{T}_M} p_K^{3d}$ weiter abzuschätzen, verfahren wir wie in [KM03, Prop. 2.7]. Wir betrachten nur $W_{\mathcal{S}}$:

$$\sum_{K \in \mathcal{T}(\mathcal{N}_M)} p_K^{3d} = \sum_{K \in \mathcal{T}(\mathcal{N}_M), \overline{K} \cap \partial\Omega \neq \emptyset} p_K^{3d} + \sum_{K \in \mathcal{T}(\mathcal{N}_M), \overline{K} \cap \partial\Omega = \emptyset} p_K^{3d}.$$

Da $p_K \leq C$ für alle $K \in \mathcal{T}(\mathcal{N}_M)$ mit $\overline{K} \cap \partial\Omega \neq \emptyset$ und $\mathcal{N}_M|_{\partial\Omega}$ ein quasi-uniformes Netz mit Randgitterweite h_M ist, erhalten wir für die erste Summe

$$\sum_{K \in \mathcal{T}(\mathcal{N}_M), \overline{K} \cap \partial\Omega \neq \emptyset} p_K^{3d} \leq C \sum_{K \in \mathcal{T}(\mathcal{N}_M), \overline{K} \cap \partial\Omega \neq \emptyset} 1 \leq C h_M^{1-d} = O(N).$$

Die zweite Summe läuft über alle Elemente K mit $\mathrm{dist}(K, \partial\Omega) > 0$. Für diese Elemente können wir $p_K \leq 1 + C\log(h_K/h_M)$ zusammen mit $h_K \sim r(x) := \mathrm{dist}(x, \partial\Omega)$, gleichmäßig für alle $x \in K$, ausnutzen. Wir erhalten

$$p_K^{3d} \leq \int_K \frac{(1 + C\log(h_K/h_M))^{3d}}{\mathrm{vol}(K)} \mathrm{d}\Omega \leq C \int_K \frac{(1 + C\log(r(x)/h_M))^{3d}}{(r(x))^d} \mathrm{d}\Omega$$

für alle $K \in \mathcal{T}(\mathcal{N}_M)$ mit $\overline{K} \cap \partial\Omega = \emptyset$. Analog zu [KM03, Prop. 2.7] ergibt sich somit

$$\sum_{K \in \mathcal{T}(\mathcal{N}_M), \overline{K} \cap \partial\Omega = \emptyset} p_K^{3d} \leq C \int_{x \in \Omega, r(x) \geq C h_M} \frac{(1 + C\log(r(x)/h_M))^{3d}}{(r(x))^d} \mathrm{d}\Omega \leq C h_M^{1-d} = O(N).$$

□

Wie uns Lemma 4.3.11 zeigt, ist die Anzahl der großdimensionalen $\mathcal{S}_v$-Teilprobleme so klein, dass wir die Cholesky-Zerlegungen in jedem PCG-Schritt aufs Neue berechnen können und unsere Vorkonditionierer trotzdem mit optimalem Aufwand $O(N)$ berechenbar sind.
Alternativ können wir jedoch unter Verwendung von zusätzlichem Speicher die Cholesky-Zerlegungen auch vorab bestimmen und für die Dauer der PCG-Iteration zwischenspeichern. Da sich das in jedem PCG-Schritt notwendige Lösen der $\mathcal{S}_v$-Teilraumprobleme somit auf Vorwärts- und Rückwärtseinsetzen reduziert, verringert sich natürlich die Rechenzeit. Aus Lemma 4.3.11 folgt, dass der hierfür zusätzlich benötigte Speicherbedarf wie $O(N)$ anwächst.
Für das Beispiel 4.3.10 haben wir diese zwei Möglichkeiten anhand einer Testrechnung einmal gegenübergestellt. Die Ergebnisse sind in Tabelle 4.2 zu sehen. Die Spalte „assemb“ zeigt für Vergleichszwecke die notwendige Rechenzeit für das Aufstellen der globalen Steifigkeitsmatrix unter Berücksichtigung konstanter Koeffizienten (siehe Abschnitt 3.8). Die Spalte „CG“ enthält die Rechenzeiten für den nicht vorkonditionierten CG-Algorithmus, „PCG–memory opt.“ steht für den PCG-Algorithmus bezüglich der Zerlegung (4.35), bei dem die Cholesky-Faktoren in jedem PCG-Schritt von Neuem berechnet werden. „PCG–runtime opt.“ steht schließlich für den PCG-Algorithmus bezüglich der Zerlegung (4.35), bei dem die Cholesky-Faktoren nur einmal berechnet und anschließend abgespeichert werden.
Die Spalten „total“ zeigen die Gesamtrechenzeit für das Lösen des Gleichungssystems, inklusive Vorkonditionierung. Die Spalten „precond“ zeigen die dabei für das Vorkonditionieren (inklusive Cholesky-Faktorisierungen) verwandte Rechenzeit. Für die laufzeitoptimale Strategie ist mit „Mem“ der zusätzlich benötigte Speicherbedarf angegeben.
Wie wir am Beispiel des Netzes $\mathcal{N}_{15}$ mit $N = 1685928$ Unbekannten sehen, reduziert sich die notwendige Rechenzeit für das Lösen des linearen Gleichungssystems erheblich. Während der reine CG-Algorithmus noch 4400 Sekunden benötigt, kommt die speicheroptimale PCG-2 Variante bereits mit 373 Sekunden aus. Der laufzeitoptimalen PCG-2 Algorithmus benötigt sogar nur noch 149 Sekunden.

Bemerkung 4.3.12. An Stelle des direkten Lösens der $\mathcal{S}_v$-Teilprobleme könnten diese Räume, wie in [SMPZ05] gezeigt, noch weiter zerlegt werden.

4.3.6 Analyse der Vorkonditionierer

Für die Analyse unserer ASM-Vorkonditionierer werden wir insbesondere auf folgendem Theorem der abstrakten Additiv-Schwarz-Theorie aufbauen:

Tabelle 4.2: Rechenzeit[sek] und zusätzlicher Speicher[kB] für Beispiel 4.3.9 und Abbruchkriterium $\|\underline{r}^k\|^2/\|\underline{r}^0\|^2 \leq 10^{-12}$.

				PCG - memory opt.		PCG - runtime opt.		
L	DOF	assemb.	CG	total	precond.	total	precond.	Mem
0	6	7.74e-04	4.10e-05	1.45e-04	1.08e-04	1.38e-04	1.01e-04	<1
1	24	2.85e-03	1.25e-04	4.32e-04	3.33e-04	3.81e-04	2.80e-04	<1
2	84	9.92e-03	3.35e-04	1.18e-03	9.62e-04	8.64e-04	6.38e-04	<1
3	196	2.33e-02	8.29e-04	2.59e-03	2.06e-03	1.74e-03	1.21e-03	<1
4	525	5.27e-02	3.59e-03	1.50e-02	1.35e-02	7.72e-03	6.24e-03	7
5	1181	1.12e-01	1.99e-02	6.10e-02	5.32e-02	2.01e-02	1.35e-02	31
6	2657	2.34e-01	2.65e-01	2.31e-01	1.92e-01	9.35e-02	5.37e-02	123
7	5818	4.86e-01	1.63e+00	6.48e-01	5.08e-01	2.78e-01	1.38e-01	457
8	12114	9.93e-01	5.66e+00	1.58e+00	1.21e+00	7.15e-01	3.40e-01	1174
9	25070	2.02e+00	1.62e+01	3.74e+00	2.84e+00	1.71e+00	7.99e-01	3105
10	51202	4.06e+00	4.17e+01	8.42e+00	6.42e+00	3.77e+00	1.75e+00	7435
11	103843	8.22e+00	1.12e+02	1.89e+01	1.46e+01	8.19e+00	3.82e+00	17198
12	209003	1.64e+01	2.92e+02	3.98e+01	3.09e+01	1.69e+01	7.88e+00	36577
13	419807	3.29e+01	6.98e+02	8.38e+01	6.55e+01	3.45e+01	1.64e+01	77033
14	841904	6.60e+01	1.73e+03	1.85e+02	1.46e+02	7.39e+01	3.50e+01	160464
15	1685928	1.32e+02	4.40e+03	3.73e+02	2.95e+02	1.49e+02	7.13e+01	326792

Theorem 4.3.13. *Sei $\mathcal{V}$ ein Hilbert-Raum und $a(\cdot,\cdot) : \mathcal{V}\times\mathcal{V} \mapsto \mathbb{R}$ symmetrisch positiv definit. Sei*

$$\mathcal{V} = \sum_{i=0}^{K} \mathcal{V}_i \tag{4.37}$$

eine nicht notwendigerweise direkte Zerlegung von $\mathcal{V}$ in Teilräume $\mathcal{V}_i$ und $\rho(E)$ der Spektralradius der symmetrischen Matrix $E \in \mathbb{R}^{K\times K}$, deren Einträge e_{ij}, $1 \leq i,j \leq K$, gegeben sind durch

$$e_{ij} = \sup_{u\in\mathcal{V}_i}\sup_{v\in\mathcal{V}_j} \frac{|a(u,v)|}{\sqrt{a(u,u)}\sqrt{a(v,v)}} \in [0,1].$$

Ferner existiere eine Konstante $C_0 > 0$, so dass

$$\min\left\{\sum_{i=0}^{K} a(u_i,u_i) \;\Big|\; u = \sum_{i=0}^{K} u_i,\;\; u_i \in \mathcal{V}_i\right\} \leq C_0^2 a(u,u) \qquad \forall\, u \in \mathcal{V}.$$

Dann definiert die Zerlegung (4.37) einen ASM-Vorkonditionierer $\mathcal{B}^{-1} : \mathcal{V}^* \to \mathcal{V}$ und für die von der Inverse $\mathcal{B} : \mathcal{V} \to \mathcal{V}^*$ induzierten Bilinearform $b(\cdot,\cdot) : \mathcal{V}\times\mathcal{V} \mapsto \mathbb{R}$ gilt:

$$\frac{a(u,u)}{(1+\rho(E))} \leq b(u,u) \leq C_0^2 a(u,u) \qquad \forall u \in \mathcal{V}. \tag{4.38}$$

Wir erkennen, dass sich die Untersuchungen der Konditionszahlen unserer vorkonditionierten Systeme auf Abschätzungen für $\rho(E)$ und C_0 reduzieren lassen. Im Folgenden wollen wir solche Abschätzungen für unsere Vorkonditionierer herleiten.

Da aus Annahme (4.27) für die in den Theoremen 4.3.4, 4.3.5 definierten Mengen I_m^B und I_m

$$I_m^B \subset I_m \qquad \text{für} \quad m = 0, \ldots, M, \tag{4.39}$$

folgt, können wir die Beweise von Theorem 4.3.4 und Theorem 4.3.5 dahingehend vereinheitlichen, als dass es genügt, eine Abschätzung für den Spektralradius $\rho(E)$ für die Zerlegung aus Theorem 4.3.5 zu beweisen (siehe Theorem 4.3.25, Korollar 4.3.26) sowie eine stabile Zerlegung für die Zerlegung aus Theorem 4.3.4 zu konstruieren (siehe Theorem 4.3.30).

Bemerkung 4.3.14. In 2D garantiert Algorithmus 4.2.21, dass Annahme (4.27) erfüllt ist. Eine ähnliche Verfeinerungsstrategie ist auch in 3D möglich.

Einige Hilfssätze

In diesem Unterabschnitt werden wir die wichtigsten Hilfsaussagen für unsere späteren Spektralradiusabschätzungen bereitstellen. Diese Aussagen umfassen bekannte Sätze der linearen Algebra sowie Aussagen über die Anzahl überlappender Patches ω_v^m.

Lemma 4.3.15 (Aussagen über den Spektralradius)**.**

(i) Für $A \in \mathbb{R}^{n\times m}$ und orthogonale Matrizen $P \in \mathbb{R}^{n\times n}, Q \in \mathbb{R}^{m\times m}$ gilt:

$$\|A\|_2 = \|PAQ\|_2 = \|A^T\|_2.$$

(ii) Sei $A \in \mathbb{R}^{n\times m}$ und N_c die maximale Anzahl der von Null verschiedenen Einträge einer Spalte von A. Dann gilt: $\|A\|_2 \leq \sqrt{N_c} \max_{i=1,\ldots,n} \|A_{i,\cdot}\|_2$.

(iii) Sei $A = [A_{ij}]_{i,j=1}^{N,M}$ mit $A_{ij} \in \mathbb{R}^{n_i \times m_j}$. Dann gilt

$$\|A\|_2 \leq \left\| \left[\|A_{ij}\|_2 \right]_{i,j=1}^{N,M} \right\|_2.$$

Beweis. Bekannte Fakten der linearen Algebra. □

Lemma 4.3.16. *Sei $\{\mathcal{N}_m\}_{m=0}^M$ eine Folge geschachtelter geometrischer Netze, welche den Bedingungen (4.25)–(4.27) genügen. Seien die Patches ω_v^m und die Zahlen $l(\cdot)$, $g(\cdot)$ gegeben durch (4.28), (4.32). Dann existiert eine Konstante $C \geq 0$, die nur von den Formregularitätskonstanten der Gitter sowie der Randgitterweite h_0 des Grobgitters $\mathcal{N}_0$ abhängt, so dass für alle ω_v^m, $0 \leq m \leq M$, $v \in V_m$*

$$l(\omega_v^m) \leq g(\omega_v^m) + C.$$

Beweis. Aus der Definition von $g(\omega_v^m)$ folgt in Verbindung mit (4.26) die Existenz eines $C > 0$, welches nur von den Formregularitätskonstanten abhängt, so dass $\operatorname{diam} \omega_v^m \geq C h_0 2^{-g(\omega_v^m)}$. Somit ergibt sich

$$-\log_2 (\operatorname{diam} \omega_v^m) \leq -\log_2 \left(C h_0 2^{-g(\omega_v^m)} \right) = g(\omega_v^m) - \log_2(C h_0).$$

□

Lemma 4.3.17. *Sei $\{\mathcal{N}_m\}_{m=0}^M$ eine Folge geschachtelter geometrischer Netze, welche den Bedingungen (4.25)–(4.27) genügen. Seien die Patches ω_v^m und die Zahlen $l(\cdot)$, $g(\cdot)$ gegeben durch (4.28), (4.32) und es gelte Annahme (4.34). Dann existiert eine Konstante $C > 0$, welche nur von der Dimension d, den Formregularitätskonstanten der Gitter, der Randgitterweite h_0 des Grobgitters $\mathcal{N}_0$ sowie der Konstanten aus Annahme (4.34) abhängt, so dass für gegebene l, l' mit $0 \le l \le l' \le L$ und beliebiges $\omega_{v'}^{m'}$ mit $l(\omega_{v'}^{m'}) = l'$, $g(\omega_{v'}^{m'}) = m'$ gilt:*

$$\#\{\omega_v^m \mid l(\omega_v^m) = l,\ \ g(\omega_v^m) = m,\ \ \omega_v^m \cap \omega_{v'}^{m'} \neq \emptyset,\ m \in \{0,\dots,M\},\ v \in V_m\} \le C.$$

Beweis. Wir betrachten einen festen Patch $w_{v'}^{m'}$ und bezeichnen mit P_m die Menge aller Patches des Netzes $\mathcal{N}_m$, welche von Level l sind und mit $\omega_{v'}^{m'}$ einen nicht-leeren Durchschnitt haben:

$$P_m := \{\omega_v^m \mid v \in V_m,\ l(\omega_v^m) = l, \omega_v^m \cap \omega_{v'}^{m'} \neq \emptyset\}, \qquad m \in \{0,\dots,M\}.$$

Aus der Definition von $l(\cdot)$ folgt

$$2^{-(l'+1)} \le \operatorname{diam}(\omega_{v'}^{m'}) \le 2^{-l'} \quad \text{und} \qquad 2^{-(l+1)} \le \operatorname{diam}(\omega_i^m) \le 2^{-l}.$$

Somit gilt

$$\bigcup_{\omega_v^m \in P_m} \omega_v^m \subset B_l,$$

wobei B_l eine geeignet gewählte Kugel mit Durchmesser

$$2^{-l} \le \operatorname{diam}(B_l) \le 2^{-l'} + 2 \cdot 2^{-l} \le 3 \cdot 2^{-l}$$

bezeichnet. Sei χ_E die charakteristische Funktion zur Menge E, so folgt aus der Formregularität unserer Netze:

$$\#P_m = \sum_{\omega_v^m \in P_m} 1 = \sum_{\omega_v^m \in P_m} \frac{1}{\operatorname{vol}(\omega_v^m)} \int_{B_l} \chi_{\omega_v^m} \le \sum_{\omega_v^m \in P_m} C 2^{dl} \int_{B_l} \chi_{\omega_v^m} \le C 2^{dl} \int_{B_l} \sum_{\omega_v^m \in P_m} \chi_{\omega_v^m},$$

wobei $C > 0$ nur von der Dimension d abhängt. Da alle Patches $\omega_v^m \in P_m$ Patches auf dem gleichen Netz sind und wir es mit Dreiecks- bzw. Tetraedervernetzungen zu tun haben, kann ein beliebiger Punkt $x \in \Omega$ in maximal $d+1$ Patches enthalten sein. Somit erhalten wir

$$\#P_m \le C 2^{dl} \int_{B_l} (d+1) \le C 2^{dl} (2^{-l})^d \le C$$

und folglich auch

$$\#\widehat{P}_m \le C, \quad \text{mit } \widehat{P}_m := \{\omega_v^m \in P_m \mid g(\omega_v^m) = m\} \quad \forall m \in \{0,\dots,M\}.$$

Aus Lemma 4.3.16 und Annahme (4.34) ergibt sich zudem

$$g(\omega_v^m) - C \le l(\omega_v^m) \le g(\omega_v^m) + C \qquad \forall \omega_v^m, \tag{4.40}$$

was $\#\widehat{P}_m = 0$ für $|m - l(\omega_v^m)| > C$ zur Folge hat. Die Behauptung folgt schließlich aus

$$\{\omega_v^m \mid l(\omega_v^m) = l,\ \ g(\omega_v^m) = m,\ \ \omega_v^m \cap \omega_{v'}^{m'} \neq \emptyset,\ m \in \{0,\dots,M\},\ v \in V_m\} = \bigcup_{m=0}^{M} \widehat{P}_m.$$

□

Lemma 4.3.18. *Sei $\{\mathcal{N}_m\}_{m=0}^M$ eine Folge geschachtelter geometrischer Netze, welche den Bedingungen (4.25)–(4.27) genügen. Seien die Patches ω_v^m und die Zahlen $l(\cdot)$, $g(\cdot)$ gegeben durch (4.28), (4.32). Es gelte Annahme (4.34) und es sei $l \in \{0,\dots,L\}$ fest. Dann existiert eine Konstante $C > 0$, welche nur von den Formregularitätskonstanten der Netze, der Randgitterweite h_0 des Grobgitters sowie der Konstanten aus Annahme (4.34) abhängt, so dass für beliebiges $\omega_{v'}^M$ gilt:*

$$\#\{\omega_v^m \mid 0 \le m \le M,\ v \in V_m,\ l(\omega_v^m) = l,\ g(\omega_v^m) = m,\ \omega_v^m \cap \omega_{v'}^M \neq \emptyset\} \le C.$$

Beweis. Wir setzen

$$P_m := \{\omega_v^m \mid v \in V_m,\ l(\omega_v^m) = l, g(\omega_v^m) = m, \omega_v^m \cap \omega_{v'}^M \neq \emptyset\}, \qquad m \in \{0,\dots,M\}.$$

Auf Grund der Formregularität unserer Netze besteht $\omega_{v'}^M$ aus höchstens C' Elementen, wobei C' ausschließlich von der Formregularitätskonstanten des Netzes $\mathcal{N}_M$ abhängt. Da $\omega_{v'}^M$ ein Patch des feinsten Netzes ist und wir von einer Folge geschachtelter Netze ausgehen, können wir folgern, dass für jedes Netz $\mathcal{N}_m$ und jedes $K \subset \omega_{v'}^M$ ein eindeutig bestimmtes $K' \in \mathcal{T}(\mathcal{N}_m)$ mit $K \subset K'$ existiert. Somit gilt

$$\#\{K' \in \mathcal{T}(\mathcal{N}_m) \mid K' \cap \omega_{v'}^M \neq \emptyset\} \le C'$$

und da jedes $K' \in \mathcal{T}(\mathcal{N}_m)$ höchstens zu $d+1$ Patches des Netzes $\mathcal{N}_m$ gehört, ergibt sich

$$\#P_m \le (d+1)C' \quad \text{für alle } m \in \{0,\dots,M\}.$$

Des Weiteren folgt aus (4.40) $\#P_m = 0$ für $|m - l(\omega_i^m)| > \tilde{C}$ und wir erhalten die Behauptung aus

$$\#\{\omega_v^m \mid 0 \le m \le M,\ v \in V_m,\ l(\omega_v^m) = l,\ g(\omega_v^m) = m,\ \omega_v^m \cap \omega_{v'}^M \neq \emptyset\} = \sum_{m=0}^{M} \#P_m.$$

□

Lemma 4.3.19. *Sei $\{\mathcal{N}_m\}_{m=0}^M$ eine Folge geschachtelter geometrischer Netze, welche den Bedingungen (4.25)–(4.27) genügen. Seien die Patches ω_v^m und die Zahlen $l(\cdot)$, $g(\cdot)$ gegeben durch (4.28), (4.32). Dann existiert eine Konstante $C > 0$, welche nur von den Netzkonstanten abhängt, so dass für alle ω_v^m, $0 \le m \le M$, $v \in V_m$ und $w_{v'}^M$, $v' \in V_M$ gilt:*

$$l(\omega_v^m) - l(\omega_{v'}^M) > C \quad \Rightarrow \quad \omega_v^m \cap \omega_{v'}^M = \emptyset.$$

Beweis. Wir betrachten einen festen Patch $\omega_{v'}^M$ und unterscheiden zwei Fälle:

- Sei $\overline{\omega_{v'}^M} \cap \partial\Omega = \emptyset$. Aus $l(\omega_{v'}^M) = l' \in \{0,\dots,L\}$ folgt

  $$2^{-(l'+1)} \le \operatorname{diam}(\omega_{v'}^M) \le 2^{-l'}$$

  und auf Grund der Formregularität von $\mathcal{N}_M$ in Verbindung mit (4.26) ergibt sich

  $$\inf_{x \in \omega_{v'}^M} \operatorname{dist}(x, \partial\Omega) \ge C_1 2^{-l'}.$$

Andererseits impliziert $l(\omega_v^m) = l \in \{0, \dots, L\}$ jedoch auch

$$\sup_{x \in \omega_v^m} \operatorname{dist}(x, \partial\Omega) \leq C_2 2^{-l}.$$

Beide Ungleichungen gemeinsam ergeben, dass $\omega_{v'}^m \cap \omega_v^M \neq \emptyset$ nur erfüllt ist, falls

$$C_1 2^{-l'} \leq C_2 2^{-l} \quad \Leftrightarrow \quad 2^{l-l'} \leq C_2 C_1^{-1} \quad \Leftrightarrow \quad l - l' \leq \log_2(C_2) - \log_2(C_1) =: C'.$$

- Sei $\overline{\omega_{v'}^M} \cap \partial\Omega \neq \emptyset$. Da alle Elemente am Rand von gleicher Größe sind, gilt $l(\omega_{v'}^M) = l' \in \{L - \widetilde{C}, \dots, L\}$ und wir haben stets $l(\omega_v^m) - l(\omega_{v'}^M) \leq L - (L - \widetilde{C}) = \widetilde{C}$.

Mit $C := \max\{C', \widetilde{C}\}$ folgt die Behauptung. □

Eine Abschätzung für den Winkel zwischen den Teilräumen

Lemma 4.3.20. *Sei $\{\mathcal{N}_m\}_{m=0}^M$ eine Folge geschachtelter geometrischer Netze, welche den Bedingungen (4.25)–(4.27) genügen. Seien die Patches ω_v^m wie in (4.28) definiert. Dann existiert eine Konstante $C > 0$, welche nur von den Formregularitätskonstanten und den Koeffizienten $\hat{A}$, a_0 der Bilinearform $a(\cdot,\cdot)$ abhängt, so dass für beliebige $U, U' \in \{\mathcal{S}_v \mid v \in V_M\} \cup \{\mathcal{V}_v^m \mid m = 0, \dots, M,\ v \in V_m\}$ mit zugehörigen Patches ω_v^m und $\omega_{v'}^{m'}$, d.h. supp$\{u \in U\} = \omega_v^m$, supp$\{u' \in U'\} = \omega_{v'}^{m'}$, gilt:*

$$e_{UU'}^2 := \sup_{u \in U} \sup_{u' \in U'} \frac{|a(u,u')|^2}{a(u,u)a(u',u')} \leq \min\left\{1, C \frac{\operatorname{vol}(\omega_v^m \cap \omega_{v'}^{m'})}{\operatorname{vol}(w_v^m)} p^{2d}\right\},$$

wobei p den maximalen Polynomgrad von $u \in U$ bezeichnet.

Beweis. Die Cauchy-Schwarz-Ungleichung liefert

$$|a(u,u')|^2 \leq a_{\omega_v^m \cap \omega_{v'}^{m'}}(u,u) a(u',u'), \tag{4.41}$$

wobei

$$\begin{aligned} a_{\omega_v^m \cap \omega_{v'}^{m'}}(u,u) &= \int_{\omega_v^m \cap \omega^{m'}} \left\langle \nabla u, \hat{A} \nabla u \right\rangle + a_0 uu \, \mathrm{d}\Omega \\ &\leq \max\left\{ \|a_0\|_{L^\infty(\Omega)}, \|\hat{A}\|_{L^\infty(\Omega)} \right\} \left(\|\nabla u\|^2_{L^2(\omega_v^m \cap \omega_{v'}^{m'})} + \|u\|^2_{L^2(\omega_v^m \cap \omega_{v'}^{m'})} \right) \\ &\leq C_{\hat{A}, a_0} \operatorname{vol}\left(\omega_v^m \cap \omega_{v'}^{m'}\right) \left(\|\nabla u\|^2_{L^\infty(\omega_v^m)} + \|u\|^2_{L^\infty(\omega_v^m)} \right). \end{aligned}$$

Da u stückweise polynomiell ist und unsere Netze formregulär sind, liefert das Anwenden einer inversen Ungleichung

$$\|u\|^2_{L^\infty(\omega_v^m)} \leq C p^{2d} \left(\operatorname{vol}(\omega_v^m)\right)^{-1} \|u\|^2_{L^2(\omega_v^m)},$$

was wiederum zu

$$a_{\omega_v^m \cap \omega_{v'}^{m'}}(u,u) \leq C_{\hat{A}, a_0} p^{2d} \frac{\operatorname{vol}\left(\omega_v^m \cap \omega_{v'}^{m'}\right)}{\operatorname{vol}(\omega_v^m)} \|u\|^2_{H^1(\omega_v^m)} \leq C_{\hat{A}, a_0} p^{2d} \frac{\operatorname{vol}\left(\omega_v^m \cap \omega_{v'}^{m'}\right)}{\operatorname{vol}(\omega_v^m)} a(u,u)$$

führt. Setzen wir dies in (4.41) ein, so erhalten wir die Behauptung. □

Abschätzungen für den Spektralradius

In diesem Unterabschnitt wollen wir Abschätzungen für die Spektralradien $\rho(E)$ und $\rho(\tilde{E})$ angeben. E bezeichnet hierbei die Matrix, die die Winkel zwischen den Teilräumen der Zerlegung (4.35) enthält und $\tilde{E}$ bezeichnet die analoge Matrix bezüglich der Zerlegung (4.31).
Um zu den gewünschten Abschätzungen zu gelangen, werden wir folgendermaßen vorgehen: Wir ordnen die Einträge der Matrix E in kleinere Teilblöcke, schätzen die Spektralnormen dieser Teilblöcke ab und erhalten daraus, mittels Lemma 4.3.15, eine Abschätzung für die Spektralnorm von E. Die Abschätzung für $\rho(\tilde{E})$ ergibt sich aus der Tatsache, dass $\tilde{E}$ eine Teilmatrix von E ist.
Die in den nachfolgenden Beweisen auftretenden generischen Konstanten $C, C', ...$ sind allesamt unabhängig von M, der Anzahl der Netze, und hängen lediglich von den Formregularitätskonstanten der Netze $\mathcal{N}_0, \ldots, \mathcal{N}_M$, der Randgitterweite h_0 des Grobgitters, der Konstanten aus Annahme (4.34), den Netzparametern aus (4.26), der Dimension d von $\Omega \subset \mathbb{R}^d$ sowie der Bilinearform $a(\cdot,\cdot)$ ab.

Lemma 4.3.21. *Es gelten die Voraussetzungen von Theorem 4.3.5. Sei $l \leq l'$ und bezeichne $E^{ll'}_{\mathcal{SS}} = \left[e_{\mathcal{S}_v \mathcal{S}_{v'}}\right]$ diejenige Matrix, welche die Winkel zwischen den Teilräumen $\mathcal{S}_v \in \{\mathcal{S}_v \mid v \in V_M,\ l(\mathcal{S}_v) = l\}$ und $\mathcal{S}_{v'} \in \{\mathcal{S}_{v'} \mid v' \in V_M,\ l(\mathcal{S}_{v'}) = l'\}$ enthält. Sei $\tilde{C}$ die Konstante aus Lemma 4.3.19. Dann existiert eine Konstante $C > 0$, so dass*

$$\left\| E^{ll'}_{SS} \right\|_2 \leq \begin{cases} C & : \ l' - l \leq \tilde{C} \\ 0 & : \ l' - l > \tilde{C} \end{cases} . \tag{4.42}$$

Insbesondere folgt aus (4.42) die Existenz einer Konstanten $C' > 0$, so dass

$$\left\| E^{ll'}_{SS} \right\|_2 \leq C' 2^{(l-l')/2}.$$

Beweis. Für $l' - l > \tilde{C}$ folgt die Behauptung direkt aus Lemma 4.3.19. Betrachten wir nun den Fall $l' - l \leq \tilde{C}$. Jede Zeile von $E^{ll'}_{\mathcal{SS}}$ gehört zu einem Teilraum $\mathcal{S}_v$ mit $\operatorname{supp}(\mathcal{S}_v) \subset \omega^M_v$ und jede Spalte von $E^{ll'}_{\mathcal{SS}}$ gehört zu einem Teilraum $\mathcal{S}_{v'}$ mit $\operatorname{supp}(\mathcal{S}_{v'}) \subset \omega^M_{v'}$. Da ω^M_v und $\omega^M_{v'}$ Patches auf dem gleichen Netz sind, kann wegen der Formregularität der Netze jede Zeile und jede Spalte von $E^{ll'}_{\mathcal{SS}}$ höchstens $O(1)$ Nicht-Null-Einträge haben, wobei die Konstante $O(1)$ einzig von der Formregularitätskonstanten des Netzes $\mathcal{N}_M$ abhängt. Somit folgt die Behauptung aus Lemma 4.3.15. □

Lemma 4.3.22. *Es gelten die Voraussetzungen von Theorem 4.3.5. Sei $l \leq l'$ und bezeichne $E^{ll'}_{S\mathcal{V}} = [e_{\mathcal{S}_v \mathcal{V}^{m'}_{v'}}]$ diejenige Matrix, welche die Winkel zwischen den Teilräumen $\mathcal{S}_v \in \{\mathcal{S}_v \mid v \in V_M,\ l(\mathcal{S}_v) = l\}$ und $\mathcal{V}^{m'}_{v'} \in \{\mathcal{V}^{m'}_{v'} \mid 0 \leq m' \leq M,\ v \in V_m,\ l(\mathcal{V}^{m'}_{v'}) = l',\ g(\mathcal{V}^{m'}_{v'}) = m'\}$ enthält. Sei $\tilde{C}$ die Konstante aus Lemma 4.3.19. Dann existieren Konstanten $C, C' > 0$, so dass*

$$\left\| E^{ll'}_{S\mathcal{V}} \right\|_2 \leq \begin{cases} C & : \ l' - l \leq \tilde{C} \\ 0 & : \ l' - l > \tilde{C} \end{cases} \leq C' 2^{(l-l')/2}.$$

Beweis. Die Träger der Teilräume $\mathcal{S}_v$ und $\mathcal{V}^{m'}_{v'}$ sind die Patches ω^M_v und $\omega^{m'}_{v'}$ mit $l(\omega^M_v) = l$, $l(\omega^{m'}_{v'}) = l'$. Die Behauptung für $l' - l > \tilde{C}$ folgt somit direkt aus Lemma 4.3.19. Betrachten wir nun den Fall $l' - l \leq \tilde{C}$. Jede Zeile von $E^{ll'}_{S\mathcal{V}}$ gehört zu einem Teilraum $\mathcal{S}_v$ mit $\operatorname{supp}(\mathcal{S}_v) \subset \omega^M_v$ und jede Spalte von $E^{ll'}_{S\mathcal{V}}$ gehört zu einem Teilraum $\mathcal{V}^{m'}_{v'}$ mit $\operatorname{supp}(\mathcal{V}^{m'}_{v'}) \subset \omega^{m'}_{v'}$. Betrachten wir

eine feste Zeile von $E^{ll'}_{\mathcal{SV}}$, d.h. einen festen Teilraum $\mathcal{S}_v$, so folgt aus Lemma 4.3.18, dass jede Zeile von $E^{ll'}_{\mathcal{SV}}$ nur $O(1)$ Nicht-Null-Einträge haben kann. Betrachten wir andererseits eine feste Spalte, d.h. einen festen Teilraum $\mathcal{V}^{m'}_{v'}$, so ergibt sich aus $l(\mathcal{S}_v) \leq l(\mathcal{V}^{m'}_{v'}) \leq l(\mathcal{S}_v)+\tilde{C}$ und einer Argumentation wie in Lemma 4.3.17, dass jede Spalte ebenfalls nur $O(1)$ Nicht-Null-Einträge besitzen kann. Folglich gilt nach Lemma 4.3.15: $\left\|E^{ll'}_{\mathcal{SV}}\right\|_2 \leq C$. □

Lemma 4.3.23. *Es gelten die Voraussetzungen von Theorem 4.3.5. Sei $l \leq l'$ und bezeichne $E^{ll'}_{\mathcal{VS}} = \left[e_{\mathcal{V}^m_v \mathcal{S}_{v'}}\right]$ diejenige Matrix, welche die Winkel zwischen den Teilräumen $\mathcal{V}^m_v \in \{\mathcal{V}^m_v \mid 0 \leq m \leq M,\ v \in V_m,\ l(\mathcal{V}^m_v) = l,\ g(\mathcal{V}^m_v) = m\}$ und $\mathcal{S}_{v'} \in \{\mathcal{S}_{v'} \mid v' \in V_M,\ l(\mathcal{S}_{v'}) = l'\}$ enthält. Dann existiert eine Konstante $C > 0$, so dass*

$$\left\|E^{ll'}_{\mathcal{VS}}\right\|_2 \leq C 2^{(l-l')/2}.$$

Beweis. Jede Spalte von $E^{ll'}_{\mathcal{VS}}$ gehört zu einem Teilraum $\mathcal{S}_{v'}$ mit $l(\mathcal{S}_{v'}) = l' \geq l$ und Träger $\omega^M_{v'}$. Folglich ist nach Lemma 4.3.17 die Anzahl der Nicht-Null-Elemente in jeder Spalte von $E^{ll'}_{\mathcal{VS}}$ durch eine Konstante C beschränkt.
Als Nächstes, um Lemma 4.3.15 anwenden zu können, schätzen wir die l^2-Norm der zu $\mathcal{V}^m_v$ mit $l(\mathcal{V}^m_v) = l$ gehörigen Zeile ab. Es gilt

$$\|E^{ll'}_{\mathcal{V}^m_v,\cdot}\|_2^2 \leq \sum_{\mathcal{S}_{v'}: l(\mathcal{S}_{v'})=l'} |e_{\mathcal{V}^m_v \mathcal{S}_{v'}}|^2.$$

Träger der Funktionen aus $\mathcal{S}_{v'}$ ist der Patch $\omega^M_{v'}$. Da alle Patches zu $\mathcal{S}_v$-Räumen, $v \in V_M$, Patches auf dem gleichen Netz sind, können sich maximal $d+1$ viele dieser Patches überlappen. Weiterhin befinden sich alle Patches von Level l' in einer $O(2^{-l'})$-Umgebung des Randes $\partial\Omega$. Zusammen mit der Tatsache, dass die Räume $\mathcal{V}^m_v$ Räume von stückweise linearen Funktionen sind, d.h. p=1, liefert Lemma 4.3.20:

$$\sum_{\mathcal{S}_{v'}: l(\mathcal{S}_{v'})=l'} |e_{\mathcal{V}^m_v \mathcal{S}_{v'}}|^2 \leq C \frac{\mathrm{vol}(\omega^m_v \cap \{x \in \Omega \mid \mathrm{dist}(x, \partial\Omega) \leq 2^{-l'}\})}{\mathrm{vol}(\omega^m_v)}.$$

Da $l(\omega^m_v) \leq l'$, folgt aus elementargeometrischen Überlegungen

$$\mathrm{vol}(\omega^m_v \cap \{x \in \Omega \mid \mathrm{dist}(x, \partial\Omega) \leq 2^{-l'}\}) \leq C(2^{-l})^{d-1} 2^{-l'}, \qquad \mathrm{vol}(\omega^m_v) \sim (2^{-l})^d.$$

Es ergibt sich

$$\sum_{\mathcal{S}_{v'}: l(\mathcal{S}_{v'})=l'} |e_{\mathcal{V}^m_v \mathcal{S}_{v'}}|^2 \leq C \frac{\mathrm{vol}(\omega^m_v \cap \{x \in \Omega \mid \mathrm{dist}(x, \partial\Omega) \leq 2^{-l'}\})}{\mathrm{vol}(\omega^m_v)} \leq C 2^{(l-l')}.$$

Anwenden von Lemma 4.3.15 vervollständigt den Beweis. □

Lemma 4.3.24. *Es gelten die Voraussetzungen von Theorem 4.3.5. Sei $l \leq l'$ und bezeichne $E^{ll'}_{\mathcal{VV}} = \left[e_{\mathcal{V}^m_v \mathcal{V}_{v'}}\right]$ diejenige Matrix, welche die Winkel zwischen den Teilräumen $\mathcal{V}^m_v \in \{\mathcal{V}^m_v \mid 0 \leq m \leq M,\ v \in V_m,\ l(\mathcal{V}^m_v) = l,\ g(\mathcal{V}^m_v) = m\}$ und $\mathcal{V}^{m'}_{v'} \in \{\mathcal{V}^{m'}_{v'} \mid 0 \leq m' \leq M,\ v' \in V_{m'},\ l(\mathcal{V}^{m'}_{v'}) = l,\ g(\mathcal{V}^{m'}_{v'}) = m'\}$ enthält. Dann existiert eine Konstante $C > 0$, so dass*

$$\left\|E^{ll'}_{\mathcal{VV}}\right\|_2 \leq C 2^{(l-l')/2}.$$

Beweis. Jede Spalte von $E^{ll'}_{\mathcal{VV}}$ gehört zu einem festen Teilraum $\mathcal{V}^{m'}_{v'}$ mit $l(\mathcal{V}^{m'}_{v'}) = l' \geq l$. Folglich ist nach Lemma 4.3.17 die Anzahl der Nicht-Null-Elemente in jeder Spalte von $E^{ll'}_{\mathcal{VV}}$ durch eine Konstante C beschränkt.
Betrachten wir als Nächstes die l^2-Norm der zu $\mathcal{V}^m_v$ mit $l(\mathcal{V}^m_v) = l$ gehörigen Zeile. Es gilt

$$\|E^{ll'}_{\mathcal{V}^m_v,.}\|_2^2 \leq \sum_{m'=0}^{M} \sum_{\left(\mathcal{V}^{m'}_{v'} \,:\, l(\mathcal{V}^{m'}_{v'})=l', g(\mathcal{V}^{m'}_{v'})=m'\right)} |e_{\mathcal{V}^m_v \mathcal{V}^{m'}_{v'}}|^2.$$

Für den $\mathcal{V}^{m'}_{v'}$ zugrunde liegenden Patch $\omega^{m'}_{v'}$ gilt (4.40), d.h.

$$g(\omega^{m'}_{v'}) - C \leq l(\omega^{m'}_{v'}) \leq g(\omega^{m'}_{v'}) + C.$$

Da nach Voraussetzung stets $g(\omega^{m'}_{v'}) = m'$ gelten soll, haben wir $l' - C \leq m' \leq l' + C$ für alle $\omega^{m'}_{v'}$ und die innere Summe ist nur für m' mit $l' - C \leq m' \leq l' + C$ ungleich Null.
Betrachten wir ein festes m' mit $l' - C \leq m' \leq l' + C$, so können wir die innere Summe analog zu Lemma 4.3.23 mit

$$\sum_{\mathcal{V}^{m'}_{v'}: l(\mathcal{V}^{m'}_{v'})=l', g(\mathcal{V}^{m'}_{v'})=m'} |e_{\mathcal{V}^m_v \mathcal{V}^{m'}_{v'}}|^2 \leq C 2^{(l-l')}$$

abschätzen. Daraus folgt

$$\|E^{ll'}_{\mathcal{V}^m_v,.}\|_2^2 \leq \sum_{m'=l'-C}^{l'+C} C 2^{(l-l')} \leq C 2^{(l-l')}.$$

Anwenden von Lemma 4.3.15 vervollständigt den Beweis. □

Theorem 4.3.25 (Spektralradius)**.** *Es gelten die Voraussetzungen von Theorem 4.3.5. E bezeichne diejenige Matrix, die die Winkel zwischen den Teilräumen bezüglich Zerlegung (4.35) enthält. Dann gilt $\rho(E) \leq C$.*

Beweis. Mit den wie in Lemma 4.3.21 bis Lemma 4.3.24 definierten Matrizen $E^{ll'}_{**}$ können wir die Matrix E nach folgendem Muster umordnen:

$$E = \begin{bmatrix} \begin{bmatrix} E^{00}_{\mathcal{VV}} & E^{00}_{\mathcal{VS}} \\ E^{00}_{\mathcal{SV}} & E^{00}_{\mathcal{SS}} \end{bmatrix} & \cdots & \begin{bmatrix} E^{0L}_{\mathcal{VV}} & E^{0L}_{\mathcal{VS}} \\ E^{0L}_{\mathcal{SV}} & E^{0L}_{\mathcal{SS}} \end{bmatrix} \\ \vdots & \ddots & \vdots \\ \begin{bmatrix} E^{L0}_{\mathcal{VV}} & E^{L0}_{\mathcal{VS}} \\ E^{L0}_{\mathcal{SV}} & E^{L0}_{\mathcal{SS}} \end{bmatrix} & \cdots & \begin{bmatrix} E^{LL}_{\mathcal{VV}} & E^{LL}_{\mathcal{VS}} \\ E^{LL}_{\mathcal{SV}} & E^{LL}_{\mathcal{SS}} \end{bmatrix} \end{bmatrix}.$$

Mit den Abkürzungen $e^{ll'}_{**} := \|E^{ll'}_{**}\|_2$ folgt aus Lemma 4.3.15:

$$\rho(E) \leq \left\| \left[\begin{bmatrix} e^{ll'}_{\mathcal{VV}} & e^{ll'}_{\mathcal{VS}} \\ e^{ll'}_{\mathcal{SV}} & e^{ll'}_{\mathcal{SS}} \end{bmatrix} \right]_{l,l'=0}^{L} \right\|_2.$$

Für $l \leq l'$ haben wir in den Lemmatas 4.3.21 - 4.3.24 gezeigt, dass $e^{ll'}_{**} \leq C 2^{(l-l')/2}$ und da unsere Matrix E symmetrisch ist, folgt unmittelbar

$$e^{ll'}_{**} \leq C \left(\sqrt{2}\right)^{-|l-l'|} \qquad \forall\, l, l' \in \{0, \ldots, L\}.$$

Aus $\|E\|_2^2 \leq \|E\|_1 \|E\|_\infty$ und $\|E\|_1 = \|E\|_\infty$ ergibt sich schließlich die Behauptung des Theorems:

$$\rho(E) \leq \max_{l=0}^{L} \sum_{l'=0}^{L} C\sqrt{2}^{-|l-l'|} \leq C.$$

□

Korollar 4.3.26 (Spektralradius)**.** *Es gelten die Voraussetzungen von Theorem 4.3.5. E bezeichne diejenige Matrix, die die Winkel zwischen den Teilräumen bezüglich Zerlegung (4.31) enthält. Dann gilt $\rho(\tilde{E}) \leq C$.*

Beweis. Da (4.39), ist $\tilde{E}$ eine Teilmatrix der nicht-negativen Matrix E aus Theorem 4.3.25. Somit gilt $\rho(\tilde{E}) \leq \rho(E) \leq C$. □

Konstruktion einer stabilen Zerlegung

In diesem Unterabschnitt wollen wir eine stabile Zerlegung für den Vorkonditionierer aus Theorem 4.3.4 beweisen. Nach (4.39) ist dies auch gleichzeitig eine stabile Zerlegung für den Vorkonditionierer aus Theorem 4.3.5.
Unser Vorgehen gliedert sich hierbei in drei Etappen: Als Erstes konstruieren wir eine stabile Zerlegung für $u \in S_0^1(\Omega, \mathcal{N}_M)$. Danach erweitern wir diese zu einer stabilen Zerlegung für $u \in S_D^1(\Omega, \mathcal{N}_M)$ und im letzten Schritt betrachten wir schließlich $u \in S_D^{\mathbf{p}}(\Omega, \mathcal{N}_M)$.

Lemma 4.3.27 (stabile Zerlegung für $u \in S_0^1(\Omega, \mathcal{N}_M)$)**.** *Sei $\Omega \subset \mathbb{R}^d$, $d = 2, 3$. Sei $\mathcal{N}_M$ ein geometrisches Netz mit Randgitterweite h und $u \in S_0^1(\Omega, \mathcal{N}_M) := S^1(\Omega, \mathcal{N}_M) \cap H_0^1(\Omega)$. Dann existiert eine Zerlegung*

$$u = \sum_{v \in V_M} u_v \quad \textit{mit} \quad u_v \in \mathcal{V}_v^M \subset \mathcal{S}_v \quad \textit{und} \quad \sum_{v \in V_M} a(u_v, u_v) \leq C a(u,u),$$

wobei $C > 0$ unabhängig von h ist.

Beweis. Für $v \in V_M$ bezeichne $\phi_v \in \mathcal{V}_v^M$ die Standard-Hutfunktion mit $\phi_v(v) = 1$, womit sich die Funktion $u \in S_0^1(\Omega, \mathcal{N}_M)$ eindeutig als Summe $u = \sum_{v \in V_M} \underline{u}_v \phi_v$ schreiben lässt. Mit $h_v := \operatorname{diam}(\operatorname{supp} \phi_v)$ ist dies äquivalent zu

$$u = \sum_{v \in V_M} h_v^{(d/2-1)} \underline{u}_v h_v^{(1-d/2)} \phi_v$$

und wir setzen

$$\underline{w}_v := h_v^{(d/2-1)} \underline{u}_v, \qquad \psi_v := h_v^{(d/2-1)} \phi_v.$$

Der Rest des Beweises gliedert sich in mehrere Schritte.
1. Schritt: Für ein festes Element $K \in \mathcal{T}(\mathcal{N}_M)$ mit den Ecken $v_{(K,1)}, \ldots, v_{(K,d+1)}$ gilt $u|_K = \sum_{j=1}^{d+1} \underline{w}_{v_{(K,j)}} \psi_{v_{(K,j)}}|_K$. Auf Grund der Formregularität von $\mathcal{N}_M$ erhalten wir

$$\sum_{j=1}^{d+1} |\underline{w}_{v_{(K,j)}}|^2 \leq C h_K^{(d-2)} \sum_{j=1}^{d+1} |u(v_{(K,j)})|^2 \leq C h_K^{(d-2)} (d+1) \|u\|_{L^\infty(K)}^2,$$

wobei $C > 0$ nur von der Formregularitätskonstanten γ abhängt. Bezeichne $\hat{K}$ das Referenzelement und sei die Transformation einer Funktion auf das Referenzelement durch ein Dach markiert, so erhalten wir

$$\begin{aligned}\sum_{j=1}^{d+1} |w_{v_{(K,j)}}|^2 &\leq C h_K^{(d-2)} (d+1) \|\hat{u}\|^2_{L^\infty(\hat{K})} \\ &\leq C h_K^{(d-2)} (d+1) \|\hat{u}\|^2_{L^2(\hat{K})} \leq C h_K^{-2} \|u\|^2_{L^2(K)}.\end{aligned} \tag{4.43}$$

2. Schritt: Sei $v \in V_M$ und $K \in \mathcal{T}(\mathcal{N}_M)$ mit $K \subset \operatorname{supp} \psi_v$. Dann gilt

$$\|\nabla \psi_v\|^2_{L^2(K)} \leq C h_K^{(d-2)} \|\nabla \hat{\psi}_v\|^2_{L^2(\hat{K})} \leq C h_K^{(d-2)} h_K^{(2-d)} \leq C$$

und

$$\|\psi_v\|^2_{L^2(K)} \leq C h_K^{d} \|\hat{\psi}_v\|^2_{L^2(\hat{K})} \leq C h_K^{d} h_K^{(2-d)} \leq C h_K^2 \leq C,$$

wobei $C > 0$ nur von der Formregularitätskonstanten abhängt.
3. Schritt: Aus der Hardy-Ungleichung

$$\left\| \frac{v}{\operatorname{dist}(x, \partial\Omega)} \right\|_{L^2(\Omega)} \leq C \|\nabla v\|_{L^2(\Omega)} \qquad \forall v \in H^1_0(\Omega),$$

der Abschätzung (4.43) und den Annahmen an die Elementgrößen h_K (4.26) folgt

$$\begin{aligned}\sum_{v \in V_M} |\underline{w}_v|^2 &\leq \sum_{K \in \mathcal{T}(\mathcal{N}_M)} \sum_{j=1}^{d+1} |\underline{w}_{v_{(K,j)}}|^2 \\ &\leq C \sum_{K \in \mathcal{T}} h_K^{-2} \|u\|^2_{L^2(K)} \leq \left\| \frac{u}{\operatorname{dist}(\cdot, \partial\Omega)} \right\|^2_{L^2(\Omega)} \leq \|\nabla u\|^2_{L^2(\Omega)}.\end{aligned}$$

Die Behauptung des Lemmas ergibt sich damit aus $a(\cdot,\cdot) \sim \|\cdot\|^2_{H^1(\Omega)}$ und den Abschätzungen aus Schritt 2:

$$\sum_{v \in V_M} a(u_v, u_v) = \sum_{v \in V_M} \underline{w}_v^2 a(\psi_v, \psi_v) \leq C \sum_{v \in V_M} \underline{w}_v^2 \leq C \|\nabla u\|^2_{L^2(\Omega)} \leq C a(u,u).$$

□

Bemerkung 4.3.28. Ein Beweis für den 2-dimensionalen Fall des Lemmas 4.3.27 ist ebenfalls in [Yse99, Mel01] zu finden.

Lemma 4.3.29 (stabile Zerlegung für $u \in S^1_D(\Omega, \mathcal{N}_M)$)**.** *Sei $\Omega \subset \mathbb{R}^d$, $d = 2, 3$. Sei $\{\mathcal{N}_m\}_{m=0}^M$ eine Folge geschachtelter geometrischer Netze, welche den Bedingungen (4.25)–(4.27) genügen. Sei $I_m^B := \{v \in V_m \mid v \in \partial\Omega\}$. Dann existiert für jedes $u \in S^1_D(\Omega, \mathcal{N}_M)$ eine Zerlegung*

$$u = \sum_{v \in V_M} u_v^{\mathcal{S}} + \sum_{m=0}^{M} \sum_{v \in I_m^B} u_v^m, \quad \textit{mit } u_v^m \in \mathcal{V}_v^m \quad \textit{und } u_v^{\mathcal{S}} \in \mathcal{S}_v \cap S^1_D(\Omega, \mathcal{N}_M), \tag{4.44}$$

so dass

$$\sum_{v\in V_M} a(u_v^{\mathcal{S}}, u_v^{\mathcal{S}}) + \sum_{m=0}^{M}\sum_{v\in I_m^B} a(u_v^m, u_v^m) \le Ca(u,u),$$

wobei $C>0$ unabhängig von u, h_M und M ist.

Beweis. 1. Schritt: Wir wählen ein quasi-uniformes Netz $\tilde{\mathcal{N}}_0$, das am Rand mit $\mathcal{N}_0$ übereinstimmt. D.h.

$$\{K\in\mathcal{T}(\mathcal{N}_0) \mid \overline{K}\cap\partial\Omega\neq\emptyset\}\subset\mathcal{T}(\tilde{\mathcal{N}}_0). \tag{4.45}$$

Sollte $\mathcal{N}_0$ selbst quasi-uniform sein, so können wir natürlich $\tilde{\mathcal{N}}_0=\mathcal{N}_0$ wählen. Ansonsten betrachten wir die Menge $\Omega' := \Omega\setminus\cup\{\overline{K} \mid K\in\mathcal{T}(\mathcal{N}_0) \text{ und } \overline{K}\cap\partial\Omega\neq\emptyset\}$ und führen die quasi-uniforme Vernetzung von $\partial\Omega'$, gegeben durch die Einschränkung von $\mathcal{N}_0$ auf $\partial\Omega'$, zu einer quasi-uniformem Vernetzung $\mathcal{N}_0'$ von Ω' fort. Die Vereinigung von $\mathcal{N}_0'$ mit der auf $\Omega^B := \cup\{\overline{K} \mid K\in\mathcal{T}(\mathcal{N}_0) \text{ und } \overline{K}\cap\partial\Omega\neq\emptyset\}$ eingeschränkten Vernetzung $\mathcal{N}_0|_{\Omega^B}$ ergibt dann ein quasi-uniformes Netz $\tilde{\mathcal{N}}_0$, welches der Forderung (4.45) genügt.
2. Schritt: Es sei $\tilde{\mathcal{N}}_m$ die aus m uniformen Verfeinerungen von $\tilde{\mathcal{N}}_0$ resultierende Vernetzung von Ω. Aus Annahme (4.27) folgt, dass die Netze $\mathcal{N}_m$ and $\tilde{\mathcal{N}}_m$ in allen Elementen am Rand übereinstimmen. D.h.

$$\{K\in\mathcal{T}(\mathcal{N}_m) \mid \overline{K}\cap\partial\Omega\neq\emptyset\} = \{K\in\mathcal{T}(\tilde{N}_m) \mid \overline{K}\cap\partial\Omega\neq\emptyset\}. \tag{4.46}$$

Mit $\tilde{V}_m$ bezeichnen wir alle Knoten des Netzes $\tilde{\mathcal{N}}_m$ mit Ausnahme der Knoten auf dem Dirichlet-Rand. Ferner seien die $\tilde{\phi}_v^m$ die Standard-Hutfunktionen auf dem Netz $\tilde{\mathcal{T}}_m$ bezüglich der Knoten $v\in\tilde{V}_m$. Wir schreiben

$$\tilde{\mathcal{V}}_v^m := \operatorname{span}\{\tilde{\phi}_v^m\},$$

wobei wir explizit darauf hinweisen, dass nach (4.45) die Gleichheit $\tilde{\mathcal{V}}_v^m=\mathcal{V}_v^m$ für $v\in I_m^B$ gilt.
3. Schritt: Für $u\in S_D^1(\Omega,\mathcal{N}_M)$ kann leicht eine Funktion $\tilde{u}\in S_D^1(\Omega,\tilde{\mathcal{N}}_M) := S^1(\Omega,\tilde{\mathcal{N}}_M)\cap H_D^1(\Omega)$ gefunden werden, so dass

$$\tilde{u}|_{\partial\Omega}=u|_{\partial\Omega}, \qquad \|\tilde{u}\|_{H^1(\Omega)}\le C\|u\|_{H^1(\Omega)} \tag{4.47}$$

und $C>0$ unabhängig von u. Des Weiteren folgt aus der klassischen Theorie (siehe [Zha92], [Osw94]) die Existenz einer Zerlegung

$$\tilde{u}=\sum_{m=0}^{M}\sum_{v\in\tilde{V}_m}\tilde{u}_v^m, \quad \text{mit} \quad \tilde{u}_v^m\in\tilde{\mathcal{V}}_v^m, \quad \sum_{m=0}^{M}\sum_{v\in\tilde{V}_m} a(\tilde{u}_v^m,\tilde{u}_v^m)\le Ca(\tilde{u},\tilde{u}) \tag{4.48}$$

sowie $C>0$ unabhängig von $\tilde{u}$ und h_M.
4. Schritt: Es gilt $I_m^B\subset V_m\cap\tilde{V}_m$ und aus (4.46) folgt $\mathcal{V}_v^m=\tilde{\mathcal{V}}_v^m$ für alle $v\in I_m^B$, $m\in\{0,\dots,M\}$. Damit erhalten wir

$$u^B := \sum_{m=0}^{M}\sum_{v\in I_m^B}\tilde{u}_v^m \subset S_D^1(\Omega,\mathcal{N}_M) \quad \text{mit} \quad u^B|_{\partial\Omega}=u|_{\partial\Omega} \tag{4.49}$$

und nach (4.47) zusammen mit (4.48) gilt

$$\sum_{m=0}^{M} \sum_{v \in I_m^B} a(\tilde{u}_v^m, \tilde{u}_v^m) \leq Ca(\tilde{u}, \tilde{u}) \leq Ca(u, u). \tag{4.50}$$

Des Weiteren gilt

$$\begin{aligned} a(u^B, u^B) &= a(\sum_{m=0}^{M} \sum_{v \in I_m^B} \tilde{u}_v^m, \sum_{m=0}^{M} \sum_{v \in I_m^B} \tilde{u}_v^m) \\ &\leq C \sum_{m=0}^{M} \sum_{v \in I_m^B} \sum_{m'=0}^{M} \sum_{v' \in I_{m'}^B} e_{vv'}^{mm'} \sqrt{a(\tilde{u}_v^m, \tilde{u}_v^m)} \sqrt{a(\tilde{u}_{v'}^{m'}, \tilde{u}_{v'}^{m'})}, \end{aligned}$$

wobei $e_{vv'}^{mm'}$ den Winkel zwischen den Teilräumen $\tilde{\mathcal{V}}_v^m \equiv \mathcal{V}_v^m$ und $\tilde{\mathcal{V}}_{v'}^{m'} \equiv \mathcal{V}_{v'}^{m'}$ bezeichnet. Sei $E := [e_{vv'}^{mm'}]$ die Matrix, die alle Winkel enthält, so wissen wir aus dem vorherigen Unterabschnitt, dass $\rho(E) \leq C$, und es folgt

$$a(u^B, u^B) \leq \rho(E) \sum_{m=0}^{M} \sum_{v \in I_m^B} \left(\sqrt{a(\tilde{u}_v^m, \tilde{u}_v^m)} \right)^2 \leq C \sum_{m=0}^{M} \sum_{v \in I_m^B} \left(\sqrt{a(\tilde{u}_v^m, \tilde{u}_v^m)} \right)^2 \leq Ca(u, u).$$

5. Schritt Wir schreiben $u := u^B + u^H$ mit u^B aus Schritt 4 und $u^H := u - u^B \in S_0^1(\Omega, \mathcal{N}_M)$. Es gilt

$$a(u^H, u^H) = a(u, u) + a(u^B, u^B) - 2a(u, u^B) \leq Ca(u, u) + 2\sqrt{a(u, u)a(u^B, u^B)} \leq Ca(u, u)$$

und die Existenz einer stabilen Zerlegung für u ergibt sich als Kombination der stabilen Zerlegung für u^H (Lemma 4.3.27) und der stabilen Zerlegung von u^B (siehe (4.49)). □

Das nun abschließende Theorem konstruiert eine stabile Zerlegung für $u \in S_D^{\mathbf{p}}(\Omega, \mathcal{N}_M)$.

Theorem 4.3.30 (stabile Zerlegung für $u \in S_D^{\mathbf{p}}(\Omega, \mathcal{N}_M)$)**.** *Sei $\Omega \subset \mathbb{R}^d$, $d = 2, 3$. Sei $\{\mathcal{N}_m\}_{m=0}^M$ eine Folge geschachtelter geometrischer Netze, welche den Bedingungen (4.25)–(4.27) genügen. Sei $p(\mathcal{N}_M)$ eine lineare Polynomgradverteilung. Seien die Patches ω_v^m und die Räume $\mathcal{V}_v^m$, $\mathcal{S}_v$ wie in Abschnitt 4.3.3 definiert. Sei $I_m^B := \{v \in V_m \mid v \in \partial\Omega\}$. Dann existiert $C > 0$, so dass für alle $u \in S_D^{\mathbf{p}}(\Omega, \mathcal{N}_M)$ eine Zerlegung*

$$u = \sum_{v \in V_M} u_{\mathcal{S}_v} + \sum_{m=0}^{M} \sum_{v \in I_m^B} u_v^m$$

mit $u_{\mathcal{S}_v} \in \mathcal{S}_v$, $u_v^m \in \mathcal{V}_v^m$ existiert und

$$\sum_{v \in V_M} a(u_{\mathcal{S}_v}, u_{\mathcal{S}_v}) + \sum_{m=0}^{M} \sum_{v \in I_m^B} a(u_v^m, u_v^m) \leq Ca(u, u).$$

Beweis. Aus [SMPZ05] (siehe Bemerkung 4.3.31) folgt die Existenz einer stabilen Zerlegung

$$u = u_D + \sum_{v \in V_M} u_{\mathcal{S}_v}$$

mit

$$a(u_D, u_D) + \sum_{v \in V_M} a(u_{\mathcal{S}_v}, u_{\mathcal{S}_v}) \leq Ca(u,u) \quad \forall u \in S_D^{\mathbf{P}}(\Omega, \mathcal{N}_M),$$

wobei $u_D \in S_D^1(\mathcal{N}_M, \Omega)$ und $u_{\mathcal{S}_v} \in \mathcal{S}_v$. Aus Lemma 4.3.29 folgt für beliebiges $u_D \in S_D^1(\Omega, \mathcal{N}_M)$

$$u_D = \sum_{v \in V_M} \hat{u}_{\mathcal{S}_v} + \sum_{m=0}^{M} \sum_{v \in I_m^B} u_v^m,$$

mit

$$\sum_{v \in V_M} a(\hat{u}_{\mathcal{S}_v}, \hat{u}_{\mathcal{S}_v}) + \sum_{m=0}^{M} \sum_{v \in I_m^B} a(u_v^m, u_v^m) \leq Ca(u_D, u_D),$$

wobei $\hat{u}_{\mathcal{S}_v} \in \mathcal{S}_v$ und $u_v^m \in \mathcal{V}_v^m$. Folglich können wir schreiben

$$u = \sum_{v \in V_M} (u_{\mathcal{S}_v} + \hat{u}_{\mathcal{S}_v}) + \sum_{m=0}^{M} \sum_{v \in I_m^B} u_v^m$$

mit $(\hat{u}_{\mathcal{S}_v} + u_{\mathcal{S}_v}) \in \mathcal{S}_v$, $u_v^m \in \mathcal{V}_v^m$ und es gilt

$$\begin{aligned}
&\sum_{v \in V_M} a(u_{\mathcal{S}_v} + \hat{u}_{\mathcal{S}_v}, u_{\mathcal{S}_v} + \hat{u}_{\mathcal{S}_v}) + \sum_{m=0}^{M} \sum_{v \in I_m^B} a(u_v^m, u_v^m) \\
&\quad \leq \quad 2 \sum_{v \in V_M} a(u_{\mathcal{S}_v}, u_{\mathcal{S}_v}) + 2 \sum_{v \in V_M} a(\hat{u}_{\mathcal{S}_v}, \hat{u}_{\mathcal{S}_v}) + \sum_{m=0}^{M} \sum_{v \in I_m^B} a(u_v^m, u_v^m) \\
&\quad \leq \quad Ca(u_D, u_D) + 2 \sum_{v \in V_M} a(u_{\mathcal{S}_v}, u_{\mathcal{S}_v}) \leq Ca(u,u).
\end{aligned}$$

□

Bemerkung 4.3.31. In [SMPZ05] werden zwar nur uniforme Polynomgradverteilungen betrachtet, jedoch erkennt man beim Durchsehen der Beweise, dass eine Erweiterung auf nicht uniforme Polynomgradverteilungen möglich ist.

Kapitel 5

Adaptive hp-FEM

Im vorigen Kapitel haben wir uns mit der randkonzentrierten Finiten-Element-Methode beschäftigt, welche mittels a priori vorgegebener Netzstruktur und Polynomgradverteilung auf ein zur h-BEM analoges Verhältnis von Fehler gegenüber Freiheitsgraden führt (siehe [KM03] und Kapitel 4). In diesem Kapitel wollen wir uns nun von der Vorgabe von Netzstruktur und Polynomgradverteilung lösen und uns einer adaptiven Steuerung dieser Größen zuwenden.

Die Grundidee aller adaptiven Strategien ist, ausgehend von einem Grobgitter, bestehend aus nur wenigen Elementen von niedrigem Polynomgrad, mit Hilfe eines Fehlerschätzers sowie einer geeigneten Verfeinerungsstrategie, eine Folge sich stetig verbessernder FE-Näherungen zu generieren, wobei wir hierbei stets bestrebt sind, mit so wenig wie möglich Freiheitsgraden bestmögliche Approximation zu erreichen.

In der h-FEM ist die Verfeinerungsstrategie recht simpel: Wir teilen schlicht und einfach jedes Element, für das der Fehlerschätzer einen zu großen Fehler ausweist, in mehrere kleinere Elemente. Betrachten wir jedoch die hp-FEM, so haben wir neben dem Verfeinern von Elementen auch die Möglichkeit, den den Elementen zugeordneten Polynomgrad zu erhöhen. Wie wichtig es ist hierbei die richtige Entscheidung zu treffen, stellt sich in den Arbeiten von I. Babuška und B. Guo (siehe [BG86c, Sch98]) heraus. In diesen Arbeiten wird nämlich gezeigt, dass für eine große Klasse von Aufgaben die hp-FEM zu exponentieller Konvergenz führt, falls Netz und zugehörige Polynomgradverteilung passend gewählt sind.

Allgemein kann man sagen, dass für Elemente, auf denen die Lösung glatt ist, die p-Verfeinerung die zu bevorzugende Wahl ist, wohingegen für Elemente, auf denen die Lösung weniger glatt ist, eine h-Verfeinerung günstiger ist.

In den meisten adaptiven hp-FE-Algorithmen basiert die Entscheidung zwischen h- und p-Verfeinerung auf einem Schätzen der lokalen Sobolev-Regularität (siehe [AS97, AS98, AS99, BOV01, HS05]). Die Arbeit [HS05] beschreibt hierbei eine Verallgemeinerung des in [Mav94] vorgeschlagenen Weges, dem Schätzen der lokalen Sobolev-Regularität anhand der Zerlegungskoeffizienten der Lösung u bezüglich einer Basis aus Legendre-Polynomen. Während die Arbeiten [Mav94, HS05] sich jedoch auf eindimensionale Probleme bzw. Elemente mit natürlicher Tensorproduktstruktur konzentrieren, werden wir im Folgenden zeigen, dass sich dieser Ansatz ebenso erfolgreich auch auf Dreiecks- und Tetraederelemente übertragen lässt. Die dazu notwendigen theoretischen Grundlagen liefern die Theoreme 5.1.1 bzw. 5.1.8, welche besagen, dass eine Funktion u genau dann analytisch in einer Umgebung eines Dreiecks bzw. Tetraeders ist, falls die Zerlegungskoeffizienten der Entwicklung von u nach einer geeigneten Orthogonalbasis exponentiell abklingen. Natürlich kennen wir die exakte Lösung nicht

und müssen uns damit begnügen, die aktuelle FE-Näherung nach orthogonalen Polynomen zu entwickeln. Unsere für Dreiecksnetze durchgeführten Vergleichsrechnungen zeigen jedoch, dass dieser Ansatz sehr gut funktioniert, vorausgesetzt wir starten mit einer Polynomgradverteilung, die genügend viele Zerlegungskoeffizienten und somit Informationen über deren Abklingen bereitstellt.
Kapitel 5 gliedert sich wie folgt: Wir beginnen mit Betrachtungen zu analytischen Funktionen auf Tetraedern und Dreiecken und bereiten damit die theoretische Grundlage der Verfeinerungsstrategie. Danach stellen wir für Vergleichszwecke zwei weitere Verfeinerungsstrategien vor und führen anschließend mit dem von uns implementierten *hp*-FE-Code ADURAKON Testrechnungen auf Dreiecksnetzen durch, um die einzelnen Strategien einander gegenüberzustellen.[1]

5.1 Analytische Funktionen auf Dreiecken und Tetraedern

Für den 2-dimensionalen Fall erhalten wir aus [Mel02, Prop. 3.2.14] und [Mel02, Lemma 3.2.15], analog zum weiter unten behandelten 3-dimensionalen Fall, die Aussage:

Theorem 5.1.1. *Es bezeichne $\mathcal{T}^2$ das Referenzdreieck aus Definition 3.3.1 und es sei die $L^2(\mathcal{T}^2)$-Orthogonalbasis $\{\psi_{pq} \mid p,q \in \mathbb{N}_0\}$, gegeben durch*

$$\psi_{pq} = \phi_{pq} \circ D_2^{-1} \quad \textit{mit } \phi_{pq} = P_p^{(0,0)}(\eta_1) \left(\frac{1-\eta_2}{2}\right)^p P_q^{(2p+1,0)}(\eta_2),$$

wobei D_2 die Transformationsvorschrift aus Lemma 3.5.2 darstellt. Für eine Funktion $u \in L^2(\mathcal{T}^2)$, dargestellt als $u = \sum_{p,q\in\mathbb{N}_0} u_{pq}\psi_{pq}$, gilt: u ist auf $\overline{\mathcal{T}}^2$ genau dann analytisch, wenn Konstanten C, $b > 0$ existieren, so dass

$$|u_{pq}| \leq Ce^{-b(p+q)} \qquad \textit{für alle } p,\, q \in \mathbb{N}_0.$$

Beweis. Folgt aus [Mel02, Prop. 3.2.14] und [Mel02, Lemma 3.2.15]. Siehe auch den Beweis zu Theorem 5.1.8. □

Bemerkung 5.1.2. Die Funktion $u \in L^2(\mathcal{T}^d)$ ist analytisch auf $\overline{\mathcal{T}}^d$, $d \in \{2,3\}$ bedeutet, dass eine offene Umgebung $\mathcal{T}' \supset \overline{\mathcal{T}}^d$ und eine Fortsetzung u' von u auf $\mathcal{T}'$ existieren, so dass u' auf $\mathcal{T}'$ analytisch ist. Im Folgenden werden wir darauf verzichten zwischen u' und u explizit zu unterscheiden. Ferner seien, wenn wir die Polynome ψ_{pq} bzw. ψ_{pqr} (siehe Definition 5.1.3) auf $\mathcal{T}'$ betrachten, die natürlichen und eindeutig bestimmten analytischen Fortsetzungen als Polynome gemeint.

Wollen wir diese Aussage auf den 3D Fall, d.h. auf das Referenztetraeder, übertragen, so können wir zwar zum Teil ähnlich wie in [Mel02, Abschnitt 3.2] vorgehen, müssen jedoch an einigen Stellen auch Zusatzüberlegungen anstellen. Insbesondere die Herleitung der notwendigen Hilfsaussagen wird in 3D wesentlich komplexer. Zum Beispiel muss hier nun das Verhalten von auf $\overline{\mathcal{Q}}_3$ holomorphen Funktionen unter der Abbildung $D_3 : \mathcal{Q}_3 \mapsto \mathcal{T}^3$ genauer untersucht werden.
Beginnen wir mit der Definition orthogonaler Polynome $(\psi_{pqr})_{p,q,r\in\mathbb{N}_0}$ auf dem Referenztetraeder $\mathcal{T}^3$.

[1] Kapitel 5 beinhaltet im Wesentlichen die Ergebnisse aus [EM04].

Definition 5.1.3 (orthogonale Polynome auf dem Tetraeder)**.** Es bezeichne D_3 die Transformation aus Lemma 3.5.2. Für $p, q, r \in \mathbb{N}_0$ definieren wir

$$\psi_{pqr} = \phi_{pqr} \circ D_3^{-1} \quad \text{mit} \quad \phi_{pqr}(\eta) = \phi_{pqr}(\eta_1, \eta_2, \eta_3)$$

gegeben durch

$$\phi_{pqr}(\eta) = P_p^{(0,0)}(\eta_1) P_q^{(2p+1,0)}(\eta_2) P_r^{(2p+2q+2,0)}(\eta_3) \left(\frac{1-\eta_2}{2}\right)^p \left(\frac{1-\eta_3}{2}\right)^{p+q}.$$

Lemma 5.1.4. *Sei $\mathcal{T}^3$ das Referenztetraeder aus Definition 3.3.1 und bezeichne ψ_{pqr} die Funktionen aus Definition 5.1.3. Dann gilt:*

- *$\psi_{pqr} \in P_{p+q+r}(\mathcal{T}^3)$, d.h. ψ_{pqr} ist ein Polynom vom Gesamtgrad kleiner gleich $p+q+r$.*
- *Die Funktionen ψ_{pqr} sind orthogonal auf $\mathcal{T}^3$ bezüglich des üblichen $L^2(\mathcal{T}^3)$-Skalarproduktes. Wir haben*

$$\left(\psi_{pqr}, \psi_{p'q'r'}\right)_{L^2(\mathcal{T}^3)} = \frac{2}{(2p+1)} \frac{2}{(2p+2q+2)} \frac{2}{(2r+2p+2q+3)} \delta_{p'p} \delta_{q'q} \delta_{r'r}.$$

Beweis. Transformation von $\left(\psi_{pqr}, \psi_{p'q'r'}\right)_{L^2(\mathcal{T}^3)}$ mittels D_3 auf ein Integral über $\mathcal{Q}^3$. Ausnutzen von Eigenschaften der Jacobi-Polynome. Rest analog zur 2-dimensionalen Version in [Mel02]. □

Bevor wir nun zu dem Theorem über analytische Funktionen auf dem Tetraeder kommen, wollen wir noch die Definition der Ellipse $\mathcal{E}_\rho$ sowie zwei für die späteren Beweise unbedingt notwendigen Lemmata angeben. Alle weiteren Hilfsaussagen, die zum Teil sehr technisch sind, werden wir der besseren Übersicht wegen jedoch in das eigenständige Kapitel 5.1.1 verschieben.

Definition 5.1.5. In der komplexen Ebene sei für $\rho > 1$ die Ellipse $\mathcal{E}_\rho$ gegeben durch (siehe auch Abbildung 5.6):

$$\mathcal{E}_\rho := \left\{ z \in \mathbb{C} \mid |z+1| + |z-1| \leq \rho + \rho^{-1} \right\}.$$

Lemma 5.1.6. *Es sei $\mathcal{E}_\rho \subset \mathbb{C}$ die Ellipse aus Definition 5.1.5. Für den Abstand des Punktes $(1,0) \subset \mathbb{C}$ zum Rand von $\mathcal{E}_\rho$ gilt:*

$$\operatorname{dist}(\partial\mathcal{E}_\rho, 1) = \frac{(\rho-1)^2}{2\rho}.$$

Beweis. Mit den Abkürzungen $a := \frac{1}{2}\left(\rho + \rho^{-1}\right)$ und $b := \frac{1}{2}\left(\rho - \rho^{-1}\right)$ können wir den Rand der Ellipse $\mathcal{E}_\rho$ durch

$$\partial\mathcal{E}_\rho = \{a\cos\phi + \mathbf{i}b\sin\phi \mid \phi \in [0, 2\pi)\}$$

beschreiben. Für unser Abstandsproblem erhalten wir somit:

$$\left(\operatorname{dist}(\partial\mathcal{E}_\rho, 1)\right)^2 = \min_{\phi \in [0,2\pi)} |a\cos\phi - 1 + \mathbf{i}b\sin\phi|^2 = \min_{\phi \in [0,2\pi)} \left((a\cos\phi - 1)^2 + b^2\sin^2\phi\right).$$

Bezeichne $f(\phi) := (a\cos\phi - 1)^2 + b^2\sin^2\phi$, so lautet die für ein Minimum notwendige Bedingung:

$$\frac{\partial f}{\partial \phi} = (2\sin\phi)\left(a + (b^2 - a^2)\cos\phi\right) = (2\sin\phi)\left(\frac{1}{2}\left(\rho + \rho^{-1}\right) - \cos\phi\right) = 0. \tag{5.1}$$

Da $\frac{1}{2}\left(\rho + \rho^{-1}\right) - \cos\phi > 0$ für alle $\phi \in [0, 2\pi)$ und $\rho > 1$ folgt, dass $\phi_{1;2} = \{0, \pi\}$ die einzigen Lösungen von (5.1) sind. Ferner zeigt sich, dass $f(0)$ das globale Minimum darstellt, womit die Behauptung $\operatorname{dist}(\partial\mathcal{E}_\rho, 1) = f^{1/2}(0) = 1/2(\rho-1)^2\rho^{-1}$ gezeigt ist. □

Lemma 5.1.7. *Für $\rho > 1$ sei $\mathcal{E}_\rho \subset \mathbb{C}$ die Ellipse aus Definition 5.1.5. Die Funktion*

$$w \mapsto \tilde{Q}_q^{(\alpha,\beta)}(w) = \int_{-1}^{1} (1-t)^\alpha (1+t)^\beta \frac{P_q^{(\alpha,\beta)}(t)}{w - t}\, dt$$

ist für alle $q \in \mathbb{N}_0$ holomorph auf $\mathbb{C}\backslash[-1,1]$ und es gilt

$$\begin{aligned}
\left|\tilde{Q}_q^{(0,0)}(w)\right| &\leq \frac{2\pi}{1 - 1/\rho}\rho^{-(q+1)} \quad \forall\, w \in \partial\mathcal{E}_\rho, \\
\left|\tilde{Q}_q^{(\alpha,0)}(w)\right| &\leq \frac{2^{\alpha+2}}{\alpha+1}\frac{q+2}{(1-1/\rho)^2}\rho^{-(q+1)} \quad \forall\, w \in \partial\mathcal{E}_\rho.
\end{aligned}$$

Beweis. [Mel02, Lemma 3.2.10, Corollary 3.2.11] □

Kommen wir nun zu dem Theorem über analytische Funktionen auf dem Tetraeder:

Theorem 5.1.8. *Es bezeichne $\mathcal{T}^3$ den Referenztetraeder aus Definition 3.3.1 und es sei $\{\psi_{pqr} \mid p,q,r \in \mathbb{N}_0\}$ die $L^2(\mathcal{T}^3)$-Orthogonalbasis aus Definition 5.1.3. Für eine Funktion $u \in L^2(\mathcal{T}^3)$, dargestellt als $u = \sum_{p,q,r\in\mathbb{N}_0} u_{pqr}\psi_{pqr}$, gilt: u ist auf $\overline{\mathcal{T}}^3$ genau dann analytisch, wenn Konstanten C, $b > 0$ existieren, so dass*

$$|u_{pqr}| \leq Ce^{-b(p+q+r)} \qquad \textit{für alle } p,q,r \in \mathbb{N}_0.$$

Beweis. Als Erstes zeigen wir, dass aus der Existenz von $C, b > 0$ und $|u_{pqr}| \leq Ce^{-b(p+q+r)}$ für alle $p,q,r \in \mathbb{N}_0$ folgt, dass u auf $\overline{\mathcal{T}}^3$ eine analytische Funktion darstellt. Wählen wir $\rho > 1$ so klein, dass $\ln\rho \leq b/2$, dann sichert Lemma 5.1.12 die Existenz einer komplexen offenen Umgebung $\mathcal{T}'$ von $\overline{\mathcal{T}}^3$ mit

$$\|\psi_{pqr}\|_{L^\infty(\mathcal{T}')} \leq (p+q+r)^3 e^{(b/2)(p+q+r)}. \tag{5.2}$$

Setzen wir nun $u_k := \sum_{p+q+r\leq k} u_{pqr}\psi_{pqr}$, so folgt aus (5.2)

$$\begin{aligned}
\|u - u_k\|_{L^\infty(\mathcal{T}')} &= \Big\| \sum_{p+q+r>k} u_{pqr}\psi_{pqr} \Big\|_{L^\infty(\mathcal{T}')} \leq \sum_{p+q+r>k} |u_{pqr}| \|\psi_{pqr}\|_{L^\infty(\mathcal{T}')} \\
&\leq C \sum_{p+q+r>k} (p+q+r)^3 e^{-(b/2)(p+q+r)}.
\end{aligned}$$

Für k hinreichend groß gilt

$$(p+q+r)^3 \leq e^{(b/4)(p+q+r)} =: \gamma^{(p+q+r)} \qquad \forall\, (p+q+r) > k.$$

Aus

$$\begin{aligned}\|u-u_k\|_{L^\infty(\mathcal{T}')} &\leq C\sum_{p+q+r>k}\gamma^{-(p+q+r)} = \frac{\gamma^{2-k}}{(\gamma-1)^3}+\frac{(k+1)\gamma^{1-k}}{(\gamma-1)^2}+\frac{(k+1)(k+2)\gamma^{-k}}{2(\gamma-1)}\\ &\leq C_\gamma(k+2)^2\gamma^{-k}\end{aligned}$$

mit $\gamma = e^{b/4} > 1$ ergibt sich die gleichmäßige Konvergenz der Folge $(u_k)_{k=0}^\infty$. Da alle Funktionen u_k analytisch auf $\mathcal{T}'$ sind, ist auch die Grenzfunktion u auf $\mathcal{T}'$ analytisch (siehe z.B. [Hör90, Cor. 2.2.4]).
Widmen wir uns nun der umgekehrten Richtung des Beweises. Da die Polynome ψ_{pqr} orthogonal bezüglich des $L^2(\mathcal{T}^3)$-Skalarproduktes sind, gilt

$$u_{pqr} = \frac{(\psi_{pqr},u)_{L^2(\mathcal{T}^3)}}{\|\psi_{pqr}\|^2_{L^2(\mathcal{T}^3)}} = \frac{1}{\|\psi_{pqr}\|^2_{L^2(\mathcal{T}^3)}}\int\limits_{\mathcal{T}^3} u\ \psi_{pqr}\ \mathrm{d}\Omega$$

und wir müssen die Existenz von C, $b>0$ zeigen, so dass

$$\left|(\psi_{pqr},u)_{L^2(\mathcal{T}^3)}\right| \leq Ce^{-b(p+q+r)}. \tag{5.3}$$

Bezeichne U_{pq} die in Lemma 5.1.11 definierte Funktion. Dann liefert die Transformation von $\mathcal{T}^3$ auf $\mathcal{Q}^3$ mittels D_3:

$$\left|(\psi_{pqr},u)_{L^2(\mathcal{T}^3)}\right| = \left|\int_{-1}^1 P_r^{(2p+2q+2,0)}(\eta_3)\left(\frac{1-\eta_3}{2}\right)^{(p+q+2)} U_{pq}(\eta_3)\mathrm{d}\eta_3\right|.$$

Lemma 5.1.11 besagt, dass für ein $\rho>1$ die Funktion U_{pq} auf $\mathcal{E}_\rho$ holomorph ist und eine Nullstelle der Ordnung $p+q$ bei $\eta_3=1$ besitzt. Folglich können wir den Cauchyschen Integralsatz auf die holomorphe Funktion $\eta_3 \mapsto U_{pq}(\eta_3)/(1-\eta_3)^{(p+q)}$ anwenden:

$$\begin{aligned}\left|(\psi_{pqr},u)_{L^2(\mathcal{T}^3)}\right| &= \left|2^{-p-q-2}\int_{-1}^1(1-\eta_3)^{(2p+2q+2)}\ \frac{U_{pq}(\eta_3)}{(1-\eta_3)^{p+q}}P_r^{(2p+2q+2,0)}(\eta_3)\mathrm{d}\eta_3\right|\\ &= \left|\frac{2^{-(p+q+2)}}{2\pi\mathbf{i}}\oint\limits_{\zeta_3\in\partial\mathcal{E}_\rho}\frac{U_{pq}(\zeta_3)}{(1-\zeta_3)^{p+q}}\tilde{Q}_r^{(2p+2q+2,0)}(\zeta_3)\mathrm{d}\zeta_3\right|.\end{aligned}$$

Mit $\mathrm{length}(\partial\mathcal{E}_\rho)\leq 4\rho$ und den Lemmata 5.1.6, 5.1.7, 5.1.11 erhalten wir somit eine erste Abschätzung:

$$\begin{aligned}\left|(\psi_{pqr},u)_{L^2(\mathcal{T}^3)}\right| &\leq C\frac{2^{-(p+q)}\mathrm{length}(\partial\mathcal{E}_\rho)}{(\mathrm{dist}(\partial\mathcal{E}_\rho,1))^{(p+q)}}\,\|U_{pq}\|_{L^\infty(\partial\mathcal{E}_\rho)}\,\|\tilde{Q}_r^{(2p+2q+2,0)}\|_{L^\infty(\partial\mathcal{E}_\rho)}\\ &\leq C\rho\left(\frac{2\rho}{(\rho-1)^2}\right)^{p+q}e^{-b(p+q)}\frac{2^{(p+q)}}{(2p+2q+3)}\,\frac{(r+2)}{(1-1/\rho)^2}\rho^{-(r+1)}\\ &\leq C\gamma^{p+q}\delta^{-r},\end{aligned} \tag{5.4}$$

wobei $\delta > 1$ und C, γ, δ unabhängig von p, q, r sind. Eine zweite Abschätzung erhalten wir mittels Cauchy-Schwarz-Ungleichung, Eigenschaften der Jacobi-Polynome und Lemma 5.1.11:

$$\begin{aligned}\left|(\psi_{pqr}, u)_{L^2(\mathcal{T}^3)}\right| &= \left|\int_{-1}^{1} P_r^{(2p+2q+2,0)}(\eta_3)\left(\frac{1-\eta_3}{2}\right)^{(p+q+2)} U_{pq}(\eta_3)\mathrm{d}\eta_3\right| \\ &\leq \left(\int_{-1}^{1}\left(\frac{1-\eta_3}{2}\right)^{(2p+2q+2)}\left(P_r^{(2p+2q+2,0)}(\eta_3)\right)^2\mathrm{d}\eta_3\right)^{\frac{1}{2}} \times \\ &\quad \left(\int_{-1}^{1}\left(\frac{1-\eta_3}{2}\right)^{2}|U_{pq}(\eta_3)|^2\mathrm{d}\eta_3\right)^{\frac{1}{2}} \\ &\leq C\left(\frac{2}{2p+2q+2r+3}\right)^{\frac{1}{2}} e^{-b(p+q)} \leq Ce^{-b(p+q)}.\end{aligned} \tag{5.5}$$

Insgesamt erhalten wir aus (5.4) und (5.5) damit

$$\left|(\psi_{pqr}, u)_{L^2(\mathcal{T}^3)}\right| \leq C\min\{e^{-b(p+q)}, \gamma^{p+q}\delta^{-r}\}. \tag{5.6}$$

Da für $\gamma < 1$ sofort die Behauptung (5.3) folgt, müssen wir lediglich den Fall $\gamma \geq 1$ genauer untersuchen. Wir wählen $\lambda > 0$, so dass $\gamma^\lambda/\delta =: q < 1$. Betrachten wir den Fall $(p+q) \leq \lambda r$, so folgt (5.3) aus

$$\left|(\psi_{pqr}, u)_{L^2(\mathcal{T}^3)}\right| \leq C\gamma^{p+q}\delta^{-r} \leq C(\gamma^\lambda/\delta)^r = Cq^{r/2+r/2} \leq Cq^{\frac{1}{2}\min\{1,1/\lambda\}(r+p+q)}.$$

Für $(p+q) > \lambda r$ folgt (5.3) aus

$$\left|(\psi_{pqr}, u)_{L^2(\mathcal{T}^3)}\right| \leq Ce^{-b(p+q)} \leq Ce^{-b\frac{\lambda}{2}(p+q+r)} \leq Ce^{-b'(p+q+r)}.$$

□

5.1.1 Hilfsaussagen

In diesem recht technischen Abschnitt stellen wir alle notwendigen Lemmata, die für den Beweis von Theorem 5.1.8 eine Rolle spielen, bereit.

Lemma 5.1.9. *Sei u analytisch auf $\overline{\mathcal{T}}^3$ und D_3 die Transformation aus Lemma 3.5.2. Dann existieren nur von u abhängige Konstanten $C > 0$, $\delta > 0$, $\rho > 1$, so dass gilt:*

1. *Die Funktion $u \circ D_3$ ist holomorph auf $\overline{\mathcal{Q}}^3$ und kann zu einer auf $\mathcal{E}_\rho \times \mathcal{E}_\rho \times \mathcal{E}_\rho$ holomorphen Funktion $\tilde{u}$ mit*

   $$\|\tilde{u}\|_{L^\infty(\mathcal{E}_\rho^3)} \leq C.$$

   fortgesetzt werden.

2. *Die Funktion $\eta_1 \mapsto \tilde{u}(\eta_1, \eta_2, \eta_3)$ ist für alle $\eta_2, \eta_3 \in (-1, 1)$ holomorph auf $\mathcal{E}_{1+\delta/((1-\eta_2)(1-\eta_3))}$ und*

   $$\sup_{(\eta_2,\eta_3)\in\mathcal{Q}^2} \|\tilde{u}(\cdot, \eta_2, \eta_3)\|_{L^\infty(\mathcal{E}_{1+\delta/((1-\eta_2)(1-\eta_3))})} \leq C.$$

3. *Die Funktion* $\eta_2 \mapsto \tilde{u}(\eta_1, \eta_2, \eta_3)$ *ist für alle* $\eta_1, \eta_3 \in (-1, 1)$ *holomorph auf* $\mathcal{E}_{1+\delta/(1-\eta_3)}$ *und*

$$\sup_{(\eta_1,\eta_3)\in\mathcal{Q}^2} \|\tilde{u}(\eta_1, \cdot, \eta_3)\|_{L^\infty(\mathcal{E}_{1+\delta/(1-\eta_3)})} \leq C.$$

Beweis. Da die Funktion u analytisch auf $\overline{\mathcal{T}}^3$ ist, existiert eine offene komplexe Umgebung $\mathcal{T}' \subset \mathbb{C}^3$ von $\overline{\mathcal{T}}^3$, auf der u zu einer holomorphen und beschränkten Funktion $\tilde{u}$ fortgesetzt werden kann. Auf Grund der Stetigkeit von D_3 existiert ein $\rho > 1$, so dass $D_3(\mathcal{E}_\rho^3) \subset \mathcal{T}'$. Um die zweite Aussage zu beweisen, müssen wir zeigen, dass zu beliebigem $\epsilon > 0$ ein $\delta(\epsilon) > 0$ existiert, so dass aus

$$(\eta_1, \eta_2, \eta_3) \in G_\delta := \left\{(\eta_1, \eta_2, \eta_3) \mid \eta_1 \in \mathcal{E}_{\rho_1}, (\eta_2, \eta_3) \in \mathcal{Q}^2\right\}$$

mit

$$\rho_1 = 1 + \frac{\delta}{(1-\eta_2)(1-\eta_3)} \tag{5.7}$$

folgt

$$\inf_{\mathbf{x}\in\overline{\mathcal{T}}^3} |D_3(\eta_1, \eta_2, \eta_3) - \mathbf{x}| \leq \epsilon. \tag{5.8}$$

Dazu setzen wir $\eta_1 = a + b\mathbf{i}$, wählen $\delta < \frac{8}{\sqrt{5}}\epsilon$ und unterscheiden drei Fälle (siehe auch Abbildung 5.6):

1. Für $a \leq -1$ gilt $A_1 := D_3(-1, \eta_2, \eta_3) \in \overline{\mathcal{T}}^3$ mit:

$$\begin{aligned} &|D_3(\eta_1, \eta_2, \eta_3) - A_1| \\ &\quad = \frac{1}{4}(1-\eta_2)(1-\eta_3)|\eta_1 + 1| \leq \frac{\sqrt{5}}{4}(1-\eta_2)(1-\eta_3)\left|\frac{1}{2}(\rho_1 + \rho_1^{-1}) - 1\right| \\ &\quad = \frac{1}{4}(1-\eta_2)(1-\eta_3)\frac{\sqrt{5}}{2}\frac{\delta}{(1-\eta_2)(1-\eta_3)} \cdot \frac{\delta}{(1-\eta_2)(1-\eta_3)+\delta} < \epsilon. \end{aligned}$$

2. Für $|a| < 1$ gilt $A_2 := D_3(a, \eta_2, \eta_3) \in \overline{\mathcal{T}}^3$ mit:

$$\begin{aligned} &|D_3(\eta_1, \eta_2, \eta_3) - A_2| \\ &\quad = \frac{1}{4}(1-\eta_2)(1-\eta_3)|b| \leq \frac{1}{8}(1-\eta_2)(1-\eta_3)|\rho_1 - \rho_1^{-1}| \\ &\quad = \frac{1}{8}(1-\eta_2)(1-\eta_3)\left|\frac{\delta}{(1-\eta_2)(1-\eta_3)} \cdot \frac{2(1-\eta_2)(1-\eta_3)+\delta}{(1-\eta_2)(1-\eta_3)+\delta}\right| < \epsilon. \end{aligned}$$

3. Für $a \geq 1$ gilt $A_3 := D_3(1, \eta_2, \eta_3) \in \overline{\mathcal{T}}^3$ und analog zum Fall $a \leq -1$:

$$|D_3(\eta_1, \eta_2, \eta_3) - A_3| \quad \leq \quad \frac{\sqrt{5}}{8}\delta < \epsilon.$$

Damit ist die zweite Aussage bewiesen und die dritte kann völlig analog gezeigt werden. □

Lemma 5.1.10. *Sei u analytisch auf $\overline{\mathcal{T}}^3$ und D_3 die Transformation aus Lemma 3.5.2. Für $p \in \mathbb{N}_0$ sei die Funktion $(\eta_2, \eta_3) \mapsto U_p(\eta_2, \eta_3)$ definiert durch:*

$$U_p(\eta_2, \eta_3) := \int_{-1}^{1} P_p^{(0,0)}(\eta_1)[u \circ D_3](\eta_1, \eta_2, \eta_3)\, d\eta_1.$$

Dann existieren nur von u abhängige Konstanten $\rho > 1$, $\delta > 0$, $C > 0$, so dass gilt:

1. *Die Funktion U_p ist holomorph und beschränkt auf $\mathcal{E}_\rho \times \mathcal{E}_\rho$ mit*

$$\|U_p\|_{L^\infty(\mathcal{E}_\rho \times \mathcal{E}_\rho)} \leq C\rho^{-p}.$$

2. *Die Funktion $\eta_2 \mapsto U_p(\eta_2, \eta_3)$ ist für jedes $\eta_3 \in (-1,1)$ holomorph auf $\mathcal{E}_{1+\delta/(1-\eta_3)}$ mit*

$$\sup_{\eta_3 \in (-1,1)} \|U_p(\cdot, \eta_3)\|_{L^\infty(\mathcal{E}_{1+\delta/(1-\eta_3)})} \leq C.$$

3. *Die Funktion U_p besitzt Nullstellen der Vielfachheit p bei $\eta_2 = 1$ und $\eta_3 = 1$.*

Beweis. Die zweite Aussage folgt aus Lemma 5.1.9-3 und die Holomorphie von U_p auf $\mathcal{E}_\rho \times \mathcal{E}_\rho$ aus Lemma 5.1.9-1. Um zu zeigen, dass $\|U_p\|_{L^\infty(\mathcal{E}_\rho \times \mathcal{E}_\rho)} \leq C\rho^{-p}$, benutzen wir die Cauchysche Integralformel zusammen mit Lemma 5.1.7:

$$\begin{aligned}
|U_p(\zeta_2, \zeta_3)| &= \left| \frac{1}{2\pi \mathbf{i}} \int_{-1}^{1} \oint_{\zeta_1 \in \partial\mathcal{E}_\rho} \frac{\tilde{u}(\zeta_1, \zeta_2, \zeta_3)}{\zeta_1 - \eta_1} P_p^{(0,0)}(\eta_1) \mathrm{d}\zeta_1 \mathrm{d}\eta_1 \right| \\
&= \left| \frac{1}{2\pi \mathbf{i}} \oint_{\zeta_1 \in \partial\mathcal{E}_\rho} \tilde{u}(\zeta_1, \zeta_2, \zeta_3) \tilde{Q}_p^{(0,0)}(\zeta_1) \mathrm{d}\zeta_1 \right| \\
&\leq \frac{\mathrm{length}(\mathcal{E}_\rho)}{2\pi} \|\tilde{u}\|_{L^\infty(\mathcal{E}_\rho^3)} \|\tilde{Q}_p^{(0,0)}\|_{L^\infty(\partial\mathcal{E}_\rho)} \leq C\rho^{-p}.
\end{aligned}$$

Da U_p auf $\mathcal{E}_\rho \times \mathcal{E}_\rho$ holomorph ist, reicht es für die Nullstellen der Vielfachheit p bei $\eta_2 = 1$ und $\eta_3 = 1$ zu zeigen, dass ein $C > 0$, unabhängig von η_2, η_3, existiert mit

$$|U_p(\eta_2, \eta_3)| \leq C(1 - \eta_2)^p (1 - \eta_3)^p \qquad \forall \eta_2, \eta_3 \in (-1, 1).$$

Lemma 5.1.9-2 zusammen mit der Cauchyschen Integralformel liefert:

$$|U_p(\eta_2, \eta_3)| = \left| \frac{1}{2\pi \mathbf{i}} \int_{-1}^{1} \oint_{\zeta_1 \in \partial\mathcal{E}_{\rho_1}} \frac{\tilde{u}(\zeta_1, \eta_2, \eta_3)}{\zeta_1 - \eta_1} P_p^{(0,0)}(\eta_1) \mathrm{d}\zeta_1 \mathrm{d}\eta_1 \right|$$

mit ρ_1 gegeben durch (5.7). Mittels Lemma 5.1.7 erhalten wir für $\eta_2, \eta_3 \in (-1, 1)$ und G_δ aus Lemma 5.1.9

$$|U_p(\eta_2, \eta_3)| \leq \frac{\mathrm{length}(\partial\mathcal{E}_{\rho_1})}{2\pi} \|\tilde{u}\|_{L^\infty(G_\delta)} \|\tilde{Q}_p^{(0,0)}\|_{L^\infty(\partial\mathcal{E}_{\rho_1})} \leq C \left(\frac{(1 - \eta_2)(1 - \eta_3)}{\delta + (1 - \eta_2)(1 - \eta_3)} \right)^p,$$

wobei C nur von u abhängt. Da $\delta > 0$ und $\eta_2, \eta_3 \in (-1,1)$, gelangen wir zur gewünschten Abschätzung

$$|U_p(\eta_2,\eta_3)| \leq C\delta^{-p}(1-\eta_2)^p(1-\eta_3)^p.$$

□

Lemma 5.1.11. *Sei u analytisch auf $\overline{\mathcal{T}}^3$ und D_3 die Transformation aus Lemma 3.5.2. Für $p,q \in \mathbb{N}_0$ sei die Funktion $\eta_3 \mapsto U_{pq}(\eta_3)$ definiert durch:*

$$U_{pq}(\eta_3) \;=\; \int_{-1}^{1} U_p(\eta_2,\eta_3) P_q^{(2p+1,0)}(\eta_2) \left(\frac{1-\eta_2}{2}\right)^{p+1} d\eta_2,$$

wobei U_p die Funktion aus Lemma 5.1.10 bezeichnet. Dann existieren nur von u abhängige Konstanten $\rho > 1$, $b > 0$, $C > 0$, so dass gilt:

1. *Die Funktion U_{pq} ist auf $\mathcal{E}_\rho$ holomorph und besitzt bei $\eta_3 = 1$ eine Nullstelle der Ordnung $(p+q)$.*
2. $|U_{pq}(\zeta_3)| \leq Ce^{-b(p+q)} \quad \forall\, \zeta_3 \in \mathcal{E}_\rho$.

Beweis. U_{pq} holomorph folgt aus Lemma 5.1.10-1. Um zu zeigen, dass U_{pq} eine Nullstelle der Ordnung $(p+q)$ bei $\eta_3 = 1$ hat, genügt es wiederum die Existenz von C, unabhängig von η_3, zu zeigen, so dass

$$|U_{pq}(\eta_3)| \leq C(1-\eta_3)^{p+q} \qquad \forall\, \eta_3 \in (-1,1).$$

Aus Lemma 5.1.10 wissen wir, dass für jedes $\eta_3 \in (-1,1)$ die Funktion $U_p(\cdot,\eta_3)$ holomorph auf $\mathcal{E}_{\rho_2} = \mathcal{E}_{1+\delta/(1-\eta_3)}$ ist und eine Nullstelle der Ordnung p bei $\eta_2 = 1$ besitzt. Folglich können wir den Cauchyschen Integralsatz auf die holomorphe Funktion $\eta_2 \mapsto U_p(\eta_2,\eta_3)/(1-\eta_2)^p$ anwenden und erhalten

$$\begin{aligned} U_{pq}(\eta_3) &= \frac{1}{2\pi\mathbf{i}}\left(\frac{1}{2}\right)^{p+1} \oint_{\zeta_2 \in \partial\mathcal{E}_{\rho_2}} \frac{U_p(\zeta_2,\eta_3)}{(1-\zeta_2)^p} \int_{-1}^{1} \frac{P_q^{(2p+1,0)}(\eta_2)}{\zeta_2-\eta_2}(1-\eta_2)^{2p+1} \mathrm{d}\eta_2 \mathrm{d}\zeta_2 \\ &= C \oint_{\zeta_2 \in \partial\mathcal{E}_{\rho_2}} \frac{U_p(\zeta_2,\eta_3)}{(1-\zeta_2)^p} \tilde{Q}_q^{(2p+1,0)}(\zeta_2)\mathrm{d}\zeta_2. \end{aligned} \tag{5.9}$$

Somit gilt

$$\begin{aligned} |U_{pq}(\eta_3)| &\leq C\frac{\mathrm{length}(\partial\mathcal{E}_{\rho_2})}{(\mathrm{dist}(\partial\mathcal{E}_{\rho_2},1))^p} \|\tilde{Q}_q^{(2p+1,0)}\|_{L^\infty(\partial\mathcal{E}_{\rho_2})} \|U_p(\cdot,\eta_3)\|_{L^\infty(\mathcal{E}_{\rho_2})} \\ &\leq C\rho_2 \left(\frac{(\rho_2-1)^2}{2\rho_2}\right)^{-p} \left(\frac{2^{2p+3}}{2p+2}\frac{q+2}{(1-1/\rho_2)^2}\rho_2^{-(q+1)}\right) \\ &\leq C\frac{\rho_2^{p-q+2}}{(\rho_2-1)^{2p+2}} \leq C\frac{(1-\eta_3+\delta)^{p-q+2}}{\delta^{2p+2}}(1-\eta_3)^{p+q} \end{aligned}$$

mit C unabhängig von η_3. Da $\delta > 0$ und $\eta_3 \in (-1,1)$, erhalten wir die gewünschte Abschätzung

$$|U_{pq}(\eta_3)| \leq C_{\delta,p,q}(1-\eta_3)^{p+q}.$$

Für eine erste Abschätzung für die zweite Behauptung verfahren wir ähnlich. Wir benutzen (5.9), integrieren aber über $\partial\mathcal{E}_\rho$. D.h.

$$\begin{aligned}
|U_{pq}(\zeta_3)| &= C\left|\oint_{\zeta_2\in\partial\mathcal{E}_\rho}\frac{U_p(\zeta_2,\zeta_3)}{(1-\zeta_2)^p}\tilde{Q}_q^{(2p+1,0)}(\zeta_2)\mathrm{d}\zeta_2\right| \\
&\leq C\frac{\mathrm{length}(\partial\mathcal{E}_\rho)}{(\mathrm{dist}(\partial\mathcal{E}_\rho,1))^p}\|\tilde{Q}_q^{(2p+1,0)}\|_{L^\infty(\partial\mathcal{E}_\rho)}\,\|U_p(\zeta_2,\zeta_3)\|_{L^\infty(\mathcal{E}_\rho^2)} \\
&\leq C\rho\left(\frac{2\rho}{(\rho-1)^2}\right)^p\frac{2^{2p+3}}{2p+2}\frac{q+2}{(1-1/\rho)^2}\rho^{-(q+1)} \leq C\rho^{-q}\gamma^p,
\end{aligned}$$

mit C, γ unabhängig von p,q, und η_3. Eine zweite Abschätzung für $|U_{pq}(\zeta_3)|$ erhalten wir aus der Cauchy-Schwarz-Ungleichung zusammen mit Lemma 5.1.10-1 und Eigenschaften der Jacobi-Polynome:

$$\begin{aligned}
|U_{pq}(\zeta_3)| &= \int_{-1}^1 U_p(\eta_2,\zeta_3)P_q^{(2p+1,0)}(\eta_2)\left(\frac{1-\eta_2}{2}\right)^{p+1}\mathrm{d}\eta_2 \\
&\leq \left\{\int_{-1}^1|U_p(\eta_2,\zeta_3)|^2\left(\frac{1-\eta_2}{2}\right)\mathrm{d}\eta_2\right\}^{\frac{1}{2}}\times \\
&\quad\left\{\int_{-1}^1\left(P_q^{(2p+1,0)}(\eta_2)\right)^2\left(\frac{1-\eta_2}{2}\right)^{2p+1}\mathrm{d}\eta_2\right\}^{\frac{1}{2}} \\
&\leq C\rho^{-p}\frac{2}{2p+2q+2} \leq C\rho^{-p}.
\end{aligned}$$

Insgesamt erhalten wir somit $|U_{pq}(\zeta_3)| \leq C\min\{\rho^{-p},\gamma^p\rho^{-q}\}$ und mit einer Überlegung wie im Beweis von Theorem 5.1.8 gelangen wir zu

$$|U_{pq}(\zeta_3)| \leq Ce^{-b(p+q)} \qquad \forall\zeta_3\in\mathcal{E}_\rho.$$

□

Lemma 5.1.12. *Seien die orthogonalen Polynome ψ_{pqr} gegeben durch Definition 5.1.3. Dann existiert zu beliebigem $\rho>1$ ein $C>0$ und eine offene komplexe Umgebung $\mathcal{T}'\subset\mathbb{C}^3$ mit $\overline{\mathcal{T}}^3\subset\mathcal{T}'$, so dass*

$$\|\psi_{pqr}\|_{L^\infty(\mathcal{T}')} \leq C(p+q+r)^3\rho^{p+q+r} \quad \forall p,q,r\in\mathbb{N}_0.$$

Beweis. Für Polynome einer Veränderlichen gilt (siehe [DL93, Chap 4, Thm. 2.2])

$$\|u\|_{L^\infty(\mathcal{E}_\rho)} \leq \rho^p\|u\|_{L^\infty(-1,1)} \qquad \forall\rho>1 \quad \forall u\in P_p.$$

Mittels einer Tensorproduktargumentation erhalten wir daraus

$$\|u\|_{L^\infty(\mathcal{E}^3_{\rho^{1/3}})} \leq \rho^p\|u\|_{L^\infty(\mathcal{Q}^3)} \qquad \forall\rho>1 \quad \forall u\in Q_p(\mathcal{Q}^3)$$

und mit Hilfe einer affinen Variablentransformation kann dies auf ein beliebiges Parallelepiped P verallgemeinert werden. D.h. für beliebiges $\rho>1$ existiert eine offene komplexe Umgebung P' von $\overline{P}$, so dass

$$\|u\|_{L^\infty(P')} \leq \rho^p\|u\|_{L^\infty(\overline{P})} \qquad \forall u\in Q_p(\mathcal{Q}^3). \tag{5.10}$$

Da es möglich ist, Parallelepipede $P^1,\ldots,P^{10}$ zu finden, so dass

$$\overline{\mathcal{T}}^3 = \bigcup_{i=1}^{10} \overline{P}^i,$$

wobei nicht notwendigerweise $P^i \cap P^j = \emptyset$ für $i \neq j$ gelten muss, erhalten wir für beliebiges $\rho > 1$ eine komplexe Umgebung $\mathcal{T}' := \cup_{i=1}^{10} P'^i$ von $\overline{\mathcal{T}}^3$ mit P'^i gegeben durch (5.10) und

$$\|u\|_{L^\infty(\mathcal{T}')} = \max_{i=1}^{10} \|u\|_{L^\infty(P'^i)} \leq \rho^p \max_{i=1}^{10} \|u\|_{L^\infty(\overline{P}^i)} \leq \rho^p \|u\|_{L^\infty(\overline{\mathcal{T}}^3)} \tag{5.11}$$

für alle $u \in P_p(\mathcal{T}^3)$. Auf die L^∞-Norm der rechten Seite wenden wir die inverse Ungleichung

$$\|u\|_{L^\infty(\overline{\mathcal{T}}^3)} \leq Cp^3 \|u\|_{L^2(\overline{\mathcal{T}}^3)} \qquad \forall\, u \in P_p(\mathcal{T}^3) \tag{5.12}$$

an (diese ist analog zum 2D Fall aus [Sch98, Thm. 4.76]) und erhalten für das Polynom ψ_{pqr}

$$\|\psi_{pqr}\|_{L^\infty(\mathcal{T}')} \leq C(p+q+r)^3 \rho^{p+q+r} \|\psi_{pqr}\|_{L^2(\mathcal{T}^3)}.$$

Lemma 5.1.4 liefert $\|\psi_{pqr}\|_{L^2(\mathcal{T}^3)} \leq 2/\sqrt{3}$ für alle $p,q,r \in \mathbb{N}_0$ und es folgt die Behauptung. □

5.2 Adaptive *hp*-Strategien

Im letzten Abschnitt haben wir bewiesen, dass eine Funktion genau dann analytisch auf einer Umgebung des Dreiecks bzw. Tetraeders ist, wenn ihre Zerlegungskoeffizienten bezüglich geeigneter Orthogonalbasen exponentiell abklingen. Diese Aussage bildet die theoretische Grundlage für eine adaptive *hp*-Strategie auf Dreiecks- und Tetraedernetzen, welche anhand der Abklingrate der Zerlegungskoeffizienten zwischen *h*- und *p*-Verfeinerung differenziert. Ziel dieses Abschnittes soll sein, die Wirksamkeit dieser Methode für den 2-dimensionalen Fall der Dreiecksnetze numerisch zu untersuchen und anderen, weiter unten vorgestellten, adaptiven *hp*-Verfahren gegenüberzustellen. Mit dem von uns implementierten Programmpaket ADURAKON werden wir dazu verschiedene Testrechnungen durchführen und die Ergebnisse anschließend auswerten.

5.2.1 Modellproblem

Für ein polygonal berandetes Gebiet Ω mit affinen Elementtransformationen $F_K : \mathcal{T}^2 \mapsto K$ betrachten wir folgendes Dirichlet-Modellproblem:

Problem 5.2.1. (Modellproblem für adaptive Strategien) *Finde $u \in H_0^1(\Omega)$ mit*

$$\int_\Omega \nabla u \cdot \nabla v d\Omega = \int_\Omega f v d\Omega \qquad \forall\, v \in H_0^1(\Omega). \tag{5.13}$$

5.2.2 Fehlerschätzer

Eine wesentliche Grundlage eines jeden adaptiven Verfahrens ist, zu gegebenem Netz $\mathcal{N}$, gegebener Polynomgradverteilung $p(\mathcal{N})$ und gegebenem FE-Raum $S_0^{\mathbf{p}}(\Omega,\mathcal{N})$ auf Basis der zugehörigen FE-Lösung u_{FE} eine Schätzung für die Verteilung des Fehlers $\|u - u_{FE}\|_{H^1(\Omega)}$ zu bestimmen. Diese Aufgabe erledigt der in [MW01] vorgestellte Residuenfehlerschätzer.

Definition 5.2.2 (Fehlerschätzer)**.** Sei $\mathcal{N}$ eine Vernetzung von Ω, $p(\mathcal{N})$ eine Polynomgradverteilung auf $\mathcal{N}$ und $u_{FE} \in S_0^{\mathrm{P}}(\Omega, \mathcal{N})$ die zugehörige FE-Lösung von Problem 5.2.1. Für $K \in \mathcal{T}(\mathcal{N})$ bezeichne p_K den dem Element zugeordneten Polynomgrad, h_K den Durchmesser und f_{p_K} die $L^2(K)$-Projektion von f in den Raum der Polynome vom Grad $p_K - 1$. Für eine Kante e im Inneren von Ω bezeichne p_e den zugehörigen Polynomgrad, h_e die Länge und $[\frac{\partial u_{FE}}{\partial n_e}]$ den Sprung der Normalenableitung von u_{FE} über die Kante e, wobei die Richtung der Normalen frei gewählt werden kann. Dann ist der für das Element $K \in \mathcal{T}(\mathcal{N})$ geschätzte lokale Fehler η_K gegeben durch

$$\eta_K^2 := \eta_{B_K}^2 + \eta_{E_K}^2,$$

wobei

$$\eta_{B_K}^2 := \frac{h_K^2}{p_K^2} \|f_{p_K} + \Delta u_{FE}\|_{L^2(K)}^2 \text{ und } \eta_{E_K}^2 := \sum_{e \subset \partial K \cap \Omega} \frac{h_e}{2p_e} \left\| [\frac{\partial u_{FE}}{\partial n_e}] \right\|_{L^2(e)}^2.$$

Der geschätzte Gesamtfehler η ergibt sich als

$$\eta^2 := \sum_{K \in \mathcal{T}(\mathcal{N})} \eta_K^2.$$

Folgendes Theorem fasst die wichtigsten Eigenschaften des obigen Fehlerschätzers zusammen.

Theorem 5.2.3. *Sei $\epsilon > 0$ und $\mathcal{N}$ eine γ-formreguläre Vernetzung. Ferner gelte für die Polynomgradverteilung $p(\mathcal{N})$*

$$\frac{1}{\gamma} p_K \leq p_{K'} \leq \gamma p_K \qquad \forall K, K' \in \mathcal{T}(\mathcal{N}) \text{ mit } \overline{K} \cap \overline{K'} \neq \emptyset.$$

Dann existieren Konstanten $C_1, C_2 > 0$, unabhängig von den Elementgrößen und unabhängig von $p(\mathcal{N})$, so dass

$$\|u - u_{FE}\|_{H^1(\Omega)}^2 \leq C_1 \sum_{K \in \mathcal{T}(\mathcal{N})} \eta_K^2 + \frac{h_K^2}{p_K^2} \|f - f_{p_K}\|_{L^2(K)}^2,$$

$$\eta_K^2 \leq C_2(\epsilon) p_K^{1+2\epsilon} \left(p_K \|u - u_{FE}\|_{H^1(\omega_K)}^2 + p_K^{2\epsilon} \frac{h_K^2}{p_K^2} \|f_{p_K} - f\|_{L^2(\omega_K)}^2 \right).$$

Beweis. Siehe [MW01]. □

Verhalten des Fehlerschätzers

In [MW01] wurden bereits numerische Untersuchungen zum Verhalten des oben vorgestellten Fehlerschätzers durchgeführt, jedoch bezogen sich diese Untersuchungen auf Netze, bestehend aus Rechteckselementen mit hängenden Knoten. Bevor wir uns den adaptiven Strategien widmen, wollen wir daher noch kurz das Verhalten des Fehlerschätzers auf Dreiecksvernetzungen anhand der folgenden drei unterschiedlichen Beispiele betrachten:

Beispiel 5.2.4 (ganze Lösung)**.** *Wir betrachten das Modellproblem (5.2.1) auf dem Quadrat $\Omega_S = (0,1)^2$ mit einer rechten Seite f, so dass die exakte Lösung gegeben ist durch*

$$u = x(1-x)y(1-y)(1-2y)e^{-\frac{5}{2}(2x-1)^2}.$$

Beispiel 5.2.5 (Lösung mit Singularität auf einem Knoten)**.** *Wir betrachten das Modellproblem (5.2.1) auf dem L-Gebiet* $\Omega_L = (0,1)^2 \backslash ([0,1] \times [-1,0])$ *mit einer rechten Seite* f*, so dass die exakte Lösung* $u \in \cap_{\epsilon>0} H^{5/3-\epsilon}(\Omega_L)$ *gegeben ist durch*

$$u = r^{\frac{2}{3}} \sin\left(\frac{2}{3}\varphi\right)\left(1 - r^2\cos^2\varphi\right)\left(1 - r^2\sin^2\varphi\right).$$

Beispiel 5.2.6 (Lösung mit Singularität nicht auf einem Knoten)**.** *Sei* $\Omega = (-1,1) \times (0,1)$*,* $\Gamma_N = (-1,1) \times \{0\}$ *und* $\Gamma_D = \partial\Omega \backslash \Gamma_N$*. Für*

$$g_N(x,y) = -\frac{2}{3} g\left(r\cos\phi\right) \cos\left(-\frac{1}{3}\phi\right) r^{-\frac{1}{3}}$$

betrachten wir das gemischte Problem: Finde $u \in H^1_D := \{u \in H^1(\Omega) \mid u|_{\Gamma_D} = 0\}$*, so dass*

$$\int_\Omega \nabla u \cdot \nabla v d\Omega = \int_\Omega f v d\Omega + \int_{\Gamma_N} g_n v d\Gamma \qquad \forall\, v \in H^1_D,$$

wobei f *so gewählt ist, dass mit*

$$g(t) = \begin{cases} -250t^3 - 675t^2 - 600t - 175 & : \quad t < -0.8 \\ 1 & : \quad |t| \leq 0.8 \\ 250t^3 - 675t^2 + 600t - 175 & : \quad t > 0.8 \end{cases}.$$

die exakte Lösung $u \in \cap_{\epsilon>0} H^{5/3-\epsilon}(\Omega)$ *gegeben ist durch*

$$u(r,\phi) = g(r\cos\phi) g(r\sin\phi) r^{\frac{2}{3}} \sin\left(\frac{2}{3}\phi\right).$$

Bemerkung 5.2.7. Beispiel 5.2.6 weicht etwas von unserem Modellproblem ab und soll im Wesentlichen auch nur der Illustration des Verhaltens des Fehlerschätzers bei Singularitäten, die nicht mit Netzknoten zusammenfallen, dienen. Um die im Beispiel 5.2.6 auftretenden Neumann-Bedingungen in den Fehlerschätzer zu integrieren, muss, wie in der Herleitung in [MW01] zu erkennen ist, die Berechnung von η_K^2 lediglich um die Sprünge an der Neumannkante ergänzt werden. D.h. wir erhalten

$$\eta_K^2 := \eta_{B_K}^2 + \eta_{E_K}^2 + \eta_{N_K}^2,$$

wobei sich die Terme $\eta_{B_K}^2$ und $\eta_{E_K}^2$ wie in Definition 5.2.2 angegeben berechnen und

$$\eta_{N_K}^2 := \sum_{e \subset \partial K \cap \Gamma_N} \frac{h_e}{p_e} \left\| g_N - \frac{\partial u_{FE}}{\partial n_e} \right\|^2_{L^2(\Gamma_N)}.$$

Da für festen Polynomgrad p der Fehlerschätzer in einen Standard-h-Methoden-Fehlerschätzer übergeht, interessieren wir uns vor allem für die p-Abhängigkeit von η und betrachten daher für die Beispiele 5.2.4 bis 5.2.6 reine p-Methoden auf festen Netzen. Als Formfunktionen verwenden wir die Karniadakis und Sherwin Formfunktionen Ψ^{KS} aus Definition 3.6.2. Die Ergebnisse, zusammen mit den verwendeten Netzen, sind in den Abbildungen 5.1-5.3 und Tabelle 5.1 dargestellt.

Alle Grafiken zeigen sowohl den globalen Fehler in der H^1-Norm als auch den geschätzten Gesamtfehler η. Wie wir erkennen, wird der tatsächliche Fehler in allen Beispielen überschätzt. Speziell für Beispiel 5.2.5 erkennen wir zudem, dass mit wachsendem Polynomgrad p der tatsächliche Fehler schneller abnimmt als der geschätzte Fehler. Dieses Überschätzen des Fehlers ist jedoch konform mit Theorem 5.2.3, da die Effizienzabschätzung von η nur suboptimal ist. Des Weiteren liegt mit Beispiel 5.2.5 ein spezieller Fall vor, nämlich dass die Lösung zwar nur $H^k(\Omega)$, mit $k < 5/3$, Regularität besitzt, die Singularität aber mit einem Knoten des Netzes zusammenfällt. Für Beispiel 5.2.5 kann die Funktion u somit durch stückweise Polynome vom Grad p mit einem Fehler $O(p^{-2(k-1)})$ approximiert werden, wohingegen für allgemeine H^k-Funktionen lediglich $O(p^{-(k-1)})$ gilt (siehe [BS87, Sch98]).
In Beispiel 5.2.6 haben wir ebenfalls $u \in \cap_{\epsilon>0} H^{5/3-\epsilon}(\Omega)$, jedoch fällt hier die Singularität nicht mit einem Knoten des Netzes zusammen. Wir sehen, dass tatsächlicher Fehler und Fehlerschätzer nun einen nahezu identischen Verlauf haben.

Tabelle 5.1: Effizienz $\eta/\|u - u_{FE}\|_{H^1(\Omega)}$ für die Beispiele 5.2.4, 5.2.5 und 5.2.6.

p	5.2.4	5.2.5	5.2.6	p	5.2.4	5.2.5	5.2.6
1	4.74	4.26	15.71	11	7.73	8.82	3.46
2	6.82	5.43	15.97	12	6.18	9.37	3.60
3	5.43	4.85	7.18	13	—	9.93	3.49
4	6.49	5.02	4.07	14	—	10.49	3.61
5	5.85	5.54	3.56	15	—	11.04	3.50
6	7.29	6.08	3.70	16	—	11.60	3.61
7	6.20	6.62	3.45	17	—	12.15	3.52
8	8.14	7.16	3.64	18	—	12.71	3.62
9	7.03	7.72	3.45	19	—	13.26	3.54
10	8.96	8.27	3.60	20	—	13.83	3.63

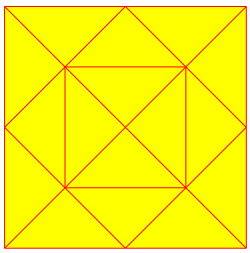
5.2.3 Der Basisalgorithmus zur adaptiven *hp*-FEM

Obwohl die Art und Weise, in der zwischen h- oder p-Verfeinerung unterschieden wird, für die drei nachfolgenden Verfahren völlig verschieden ist, so ist der grundlegende Algorithmus jedoch gleich:

Algorithmus 5.2.8 (Basisalgorithmus zur adaptiven *hp*-FEM)**.**

- *Input: Zulässiges Netz $\mathcal{N}$, Polynomgradverteilung $p(\mathcal{N})$, zugehörige FE-Lösung $u_{FE} \in S_0^{\mathbf{p}}(\Omega, \mathcal{N})$.*
- *Output: Verfeinertes Netz $\mathcal{N}_{ref}$, Polynomgradverteilung $p(\mathcal{N}_{ref})$.*
- *Algorithmus:*
  1. *Berechne η_K^2 für alle $K \in \mathcal{T}(\mathcal{N})$.*
  2. *Bestimme die Elemente $\mathcal{T}_{h\_ref} \subset \mathcal{T}(\mathcal{N})$ für die h-Verfeinerung.*
  3. *Bestimme die Elemente $\mathcal{T}_{p\_ref} \subset \mathcal{T}(\mathcal{N})$ für die p-Verfeinerung.*

Abbildung 5.1: Fehlerschätzer - Beispiel 5.2.4



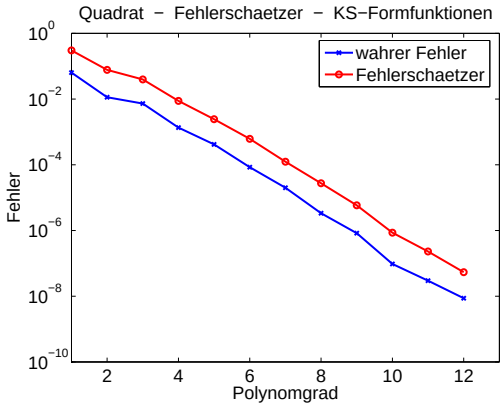
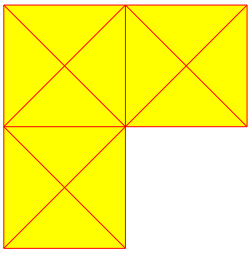


Abbildung 5.2: Fehlerschätzer - Beispiel 5.2.5



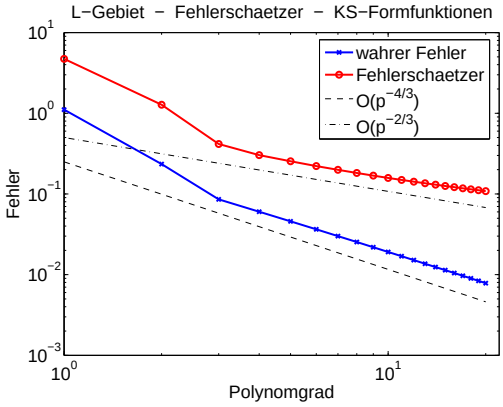
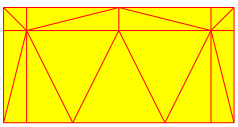
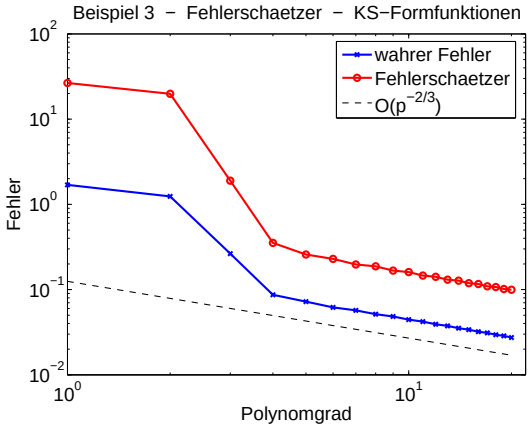


Abbildung 5.3: Fehlerschätzer - Beispiel 5.2.6





4. *Zerlege alle Dreiecke $K \in \mathcal{T}_{h\_ref}$ in vier kongruente Söhne (rote Verfeinerung).*
5. *Bestimme die verfeinerte Vernetzung $\mathcal{N}_{ref}$ durch Elimination aller hängender Knoten.*
6. *Erhöhe den Polynomgrad $p_K := p_K + 1$ für alle $K \in \mathcal{T}(\mathcal{N}_{ref}) \cap \mathcal{T}_{p\_ref}$. D.h. für Elemente, die bereits h-verfeinert wurden, entfällt die etwaige p-Verfeinerung.*

Algorithmus 5.2.9 (Elimination hängender Knoten)**.**

1. *Solange es Elemente mit mehr als einem hängenden Knoten gibt, zerlege diese in vier kongruente Söhne (rote Verfeinerung).*
2. *Zerlege alle Elemente mit einem hängenden Knoten in zwei Söhne (grüne Verfeinerung).*

Bemerkung 5.2.10. Da die grüne Verfeinerung einen Innenwinkel des Vaterelements in zwei Winkel von etwa halber Größe zerlegt, kann eine wiederholte Anwendung der grünen Teilung auf Dreiecke, die bereits einer grünen Verfeinerung entstammen, zur Entartung des Netzes führen. Um dies zu vermeiden, verbieten wir jede weitere Teilung der beiden einer grünen Verfeinerung entstammenden Söhne. Sollte es notwendig werden ein solches Dreieck weiter zu verfeinern, so machen wir erst die grüne Teilung rückgängig und teilen das Vaterelement anschließend rot (siehe Abbildung 5.4). Falls notwendig, werden die Dreiecke Sohn*_1 und Sohn*_2 noch grün geteilt (für weitere Informationen bezüglich Netzverfeinerung siehe zum Beispiel [GRT93] und enthaltene Referenzen).

Bemerkung 5.2.11 (Polynomgrad beim Auflösen einer grünen Verfeinerung). Im Falle des Auflösens einer grünen Verfeinerung stellt sich noch die Frage: Was passiert, wenn sich der Polynomgrad der beiden Söhne inzwischen unterschiedlich entwickelt hat, d.h. Sohn_1 einen anderen Polynomgrad als Sohn_2 besitzt? Um den nach dem Wiedervereinigen mit anschließender roten Teilung entstehenden vier Dreiecken einen jeweils eindeutig bestimmten Polynomgrad zuzuweisen, können wir nach einer der folgenden Methoden verfahren (für die Bezeichnungen siehe Abbildung 5.4):

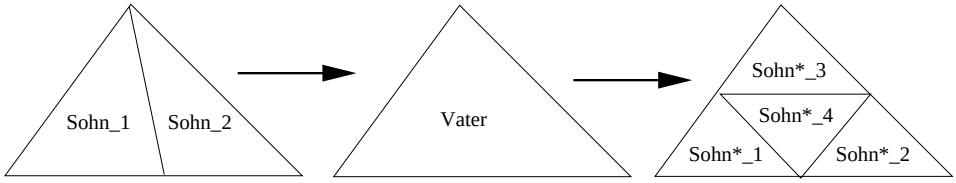
1. Die beiden Söhne einer grünen Teilung werden ausschließlich simultan p-verfeinert, d.h. wird eines der Dreiecke p-verfeinert, so wird auch das andere Dreieck p-verfeinert, egal ob es zur Verfeinerung vorgesehen war oder nicht.
2. Sohn*_1 erbt den Polynomgrad von Sohn_1 und Sohn*_2 den von Sohn_2. Den Dreiecken Sohn*_3 und Sohn*_4 weisen wir den niedrigeren (höheren) Polynomgrad von Sohn_1 und Sohn_2 zu.
3. Sohn*_1 bis Sohn*_4 bekommen den niedrigeren (höheren) Polynomgrad von Sohn_1 und Sohn_2 zugewiesen.

In unseren Rechnungen verwenden wir Variante 2.
Nachdem wir den Basisalgorithmus vorgestellt haben, kommen wir nun zu den verschiedenen Strategien für die Unterteilung zwischen h- und p-Verfeinerung.

5.2.4 Strategie I - Vergleich von geschätztem und vorhergesagtem Fehler

Die erste Strategie, die wir vorstellen, stammt aus [MW01] und die Entscheidung, ob ein Element geteilt oder aber sein Polynomgrad erhöht wird, beruht auf einem Vergleich zwischen

Abbildung 5.4: Auflösen der grünen Verfeinerung



geschätztem Fehler η_K und vorhergesagtem Fehler $\eta_K^{(pred)}$. Dieser Vergleich liefert indirekt eine Aussage über die lokale Regularität der Lösung: Die Vorhersagen (siehe (5.14) und (5.15)) gehen von maximaler Glattheit der Lösung aus, d.h. wir nehmen unter anderem Analytizität der Lösung an, um in (5.15) exponentielles Abklingen zu erreichen. Stellt sich später heraus, dass der geschätzte Fehler in der vorhergesagten Größenordnung liegt, so führen wir eine p-Verfeinerung durch, da die Annahme der Glattheit anscheinend erfüllt ist. Andernfalls, wenn der geschätzte Fehler deutlich größer als der vorhergesagte ist, führen wir eine h-Verfeinerung durch, da unsere Regularitätsannahme vermutlich nicht zutreffend war.
Um die Strategie an unsere Gegebenheiten anzupassen, müssen wir den in [MW01] angegebenen Algorithmus lediglich um den Fall der grünen Teilung sowie den Fall des Auflösens einer grünen Teilung erweitern.

Algorithmus 5.2.12 (Strategie I)**.** *Für gewählte Parameter σ, γ_h, γ_p und γ_n bestimme:*

1. *Den durchschnittlichen Elementfehler*

$$\overline{\eta}^2 = \frac{1}{\#\mathcal{N}(\mathcal{T})} \sum_{K \in \mathcal{N}(\mathcal{T})} \eta_K^2.$$

2. *Die Mengen*

$$\begin{aligned} \mathcal{T}_{p\_ref} &= \left\{ K \in \mathcal{N}(\mathcal{T}) \mid \eta_K^2 \geq \sigma \overline{\eta}^2 \wedge \eta_K^2 < \left(\eta_K^{pred}\right)^2 \right\}, \\ \mathcal{T}_{h\_ref} &= \left\{ K \in \mathcal{N}(\mathcal{T}) \mid \eta_K^2 \geq \sigma \overline{\eta}^2 \wedge \eta_K^2 \geq \left(\eta_K^{pred}\right)^2 \right\}. \end{aligned}$$

Für die vorhergesagten Fehler auf dem Ausgangsnetz $\mathcal{N}_0$ setzen wir

$$\eta_K^{pred} = \begin{cases} 0 \\ \infty \end{cases} \text{ falls wir eine } \begin{matrix} h\text{-Verfeinerung} \\ p\text{-Verfeinerung} \end{matrix} \text{ im ersten Verfeinerungsschritt bevorzugen.}$$

Jeweils nach Beendigung von Schritt 6 des Basisalgorithmus durchlaufen wir alle Elemente $K \in \mathcal{T}(\mathcal{N})$ und aktualisieren den vorhergesagten Fehler:

Algorithmus 5.2.13 (Aktualisieren der Fehlervorhersage)**.**

- *Wurde K einer h-Verfeinerung unterzogen, so setze für alle Söhne K_s von K*

$$\left(\eta_{K_s}^{pred}\right)^2 := \eta_K^2 \cdot \begin{cases} \frac{1}{4}\gamma_h \left(\frac{1}{2}\right)^{2p_K} & \textit{im Falle einer roten Teilung} \\ \frac{1}{2} & \textit{im Falle einer grünen Teilung} \end{cases}. \tag{5.14}$$

- *Wurde K einer p-Verfeinerung unterzogen, so setze*

$$\left(\eta_K^{pred}\right)^2 := \gamma_p \eta_K^2. \tag{5.15}$$

- *Wurde K keiner Verfeinerung unterzogen, so setze*

$$\left(\eta_K^{pred}\right)^2 := \gamma_n \left(\eta_K^{pred}\right)^2. \tag{5.16}$$

Ein im obigen Algorithmus 5.2.13 noch nicht enthaltener Spezialfall ist das Auflösen einer grünen Verfeinerung. In einem ersten Schritt werden hierbei die beiden grünen Zwillinge wieder zum Vaterelement K_{Vater} vereint. Für den Fehler setzen wir dabei

$$\eta^2_{K_{Vater}} := \eta^2_{K_{Sohn\_1}} + \eta^2_{K_{Sohn\_2}}.$$

In der anschließenden roten Verfeinerung wird das Element K_{Vater} dann in vier kongruente Söhne zerlegt, die wie bei einer ganz gewöhnlichen Teilung gemäß (5.14) behandelt werden.

Bemerkung 5.2.14. Für rote und p-Verfeinerungen berechnen wir η_K^{pred} wie in [MW01] angegeben. Im Falle einer grünen Teilung wird das Dreieck in zwei Teile zerlegt, wobei sich jedoch weder Durchmesser noch Seitenlängen der Söhne signifikant verkleinern. Aus diesem Grund erwarten wir hierbei auch keine zur roten Teilung vergleichbare Fehlerreduktion und verteilen den Fehler des Vaterelements lediglich gleichmäßig auf die beiden Söhne.

5.2.5 Strategie II - der Drei-Klassen-Algorithmus

Die zweite Strategie wurde ursprünglich in [HMS01] für die Behandlung von hyperpersingulären und schwach singulären Integralgleichungen im Kontext der Randelementmethode vorgeschlagen. Im Gegensatz zur vorherigen Strategie ist hier für die Auswahl der zu verfeinernden Elemente nicht der durchschnittliche Fehler, sondern der maximal auftretende Elementfehler $\eta^2_{max} = \max_{K\in\mathcal{N}(\mathcal{T})} \eta^2_K$ ausschlaggebend. Die Idee ist die folgende: Elemente, für die der Fehlerschätzer einen geringen Fehler ausweist, d.h. $\eta^2_K/\eta^2_{max} < \delta_1$, bleiben unverändert. Elemente mit mittlerem Fehler werden p-verfeinert und Elemente mit großem Fehler werden h-verfeinert.

Algorithmus 5.2.15 (Strategie II)**.** *Für gewählte Parameter δ_1, δ_2 mit $0 < \delta_1 < \delta_2 < 1$ bestimme:*

1. *Den maximalen Elementfehler*

$$\eta^2_{max} = \max_{K\in\mathcal{N}(\mathcal{T})} \eta^2_K.$$

2. *Die Mengen*

$$\begin{aligned} \mathcal{T}_{p\_ref} &= \{K \in \mathcal{T} \mid \delta_1 \eta^2_{max} \le \eta^2_K \le \delta_2 \eta^2_{max}\}, \\ \mathcal{T}_{h\_ref} &= \{K \in \mathcal{T} \mid \eta^2_K > \delta_2 \eta^2_{max}\}. \end{aligned}$$

5.2.6 Strategie III - Abklingen der Legendre-Zerlegungskoeffizienten

Die dritte Strategie, die wir nun vorstellen, basiert auf der lokalen Zerlegung der FE-Lösung nach einer geeigneten Orthogonalbasis und wurde erstmals in [Mav94] (siehe auch [HS05, HSS03]) vorgeschlagen. Die für diese Strategie notwendigen theoretischen Grundlagen beschränkten die Anwendung des Algorithmus bisher jedoch auf eindimensionale Gebiete bzw. höherdimensionale Gebiete, bei denen die Elemente der Vernetzung eine natürliche Tensorproduktstruktur aufweisen. Erst die in diesem Kapitel besprochenen Erweiterungen der theoretischen Grundlagen auf Dreieckselemente (siehe Theorem 5.1.1, [Mel02]) und Tetraederelemente (siehe Theorem 5.1.8) versetzen uns in die Lage, den Anwendungsbereich der Strategie des Abklingens der Legendre-Zerlegungskoeffizienten auch auf Dreiecks- und Tetraedervernetzungen auszudehnen [2].
Wie bereits angedeutet, besteht die grundlegende Idee der Strategie darin, die Finite-Element-Lösung u_{FE} nach Einschränkung auf das zur Verfeinerung ausgewählte Element $K \in \mathcal{T}(\mathcal{N})$ und anschließender Transformation auf das Referenzelement $\mathcal{T}^2$ bezüglich einer geeigneten Orthogonalbasis zu zerlegen und anhand des Abklingens der Zerlegungskoeffizienten zwischen h- und p-Verfeinerung zu unterscheiden.
Aus Abschnitt 5.1 wissen wir, dass für eine Funktion $u \in L^2(\mathcal{T}^2)$ die Koeffizienten u_{pq} der Zerlegung

$$u = \sum_{p,q \in \mathbb{N}_0} u_{pq} \psi_{pq}$$

bezüglich der Orthogonalbasis ψ_{pq} (siehe Abschnitt 5.1) genau dann exponentiell mit

$$|u_{pq}| \leq C e^{-b(p+q)} \qquad C, b > 0$$

abklingen, falls u analytisch auf einer Umgebung des Dreiecks $\mathcal{T}^2$ ist. Da wir die exakte Lösung u natürlich nicht kennen, zerlegen wir die aktuelle Finite-Element-Lösung u_{FE} in

$$u_{FE}|_K \circ F_K = \sum_{p,q=0}^{p+q \leq p_K} u_{pq} \psi_{pq}$$

und bestimmen die Abklingrate b mittels Kleinster-Quadrate-Methode für

$$\ln |u_{ij;K}| \sim C_K - b_K (i+j).$$

Klingen die Koeffizienten hinreichend schnell ab, d.h. falls für einen gewählten Parameter δ gilt $b_K \geq \delta$, so können wir von exponentieller Konvergenz in p und damit verbunden von einer auf K lokal glatten Lösung ausgehen. In diesem Fall entscheiden wir uns für eine p-Verfeinerung, andernfalls für eine h-Verfeinerung.

Algorithmus 5.2.16 (Strategie III)**.** *Für gewählte Parameter σ und δ bestimme:*

1. *Den durchschnittlichen Fehler*

$$\overline{\eta}^2 = \frac{1}{\#\mathcal{N}(\mathcal{T})} \sum_{K \in \mathcal{N}(\mathcal{T})} \eta_K^2.$$

[2] Auf Grund des enormen Programmieraufwands für eine adaptive 3D-*hp*-FEM beschränken wir, wie bei allen andere Strategien auch, die numerischen Betrachtungen auf den 2-dimensionalen Fall der Dreiecksvernetzung.

2. *Für alle* $K \in \mathcal{T}(\mathcal{N})$ *mit* $\eta_K^2 \geq \sigma\overline{\eta}^2$ *die Zerlegungskoeffizienten*

$$u_{ij;K} = \|\psi_{ij}\|^{-2}_{L^2(\mathcal{T}^2)}(u_{FE}|_K \circ F_K, \psi_{ij})_{L^2(\mathcal{T}^2)}, \quad 0 \leq i+j \leq p_K,$$

sowie mittels mittels Kleinster-Quadrate-Methode für

$$\ln|u_{ij;K}| \sim C_K - b_K(i+j)$$

die Abklingrate b_K*.*

3. *Die Mengen*

$$\begin{aligned} \mathcal{T}_{p\_ref} &= \{K \in \mathcal{N}(\mathcal{T}) \mid \eta_K^2 \geq \sigma\overline{\eta}^2 \wedge b_K \geq \delta\}, \\ \mathcal{T}_{h\_ref} &= \{K \in \mathcal{N}(\mathcal{T}) \mid \eta_K^2 \geq \sigma\overline{\eta}^2 \wedge b_K < \delta\}. \end{aligned}$$

5.2.7 Numerische Ergebnisse

In diesem Abschnitt wollen wir die drei oben vorgestellten Algorithmen anhand numerischer Testrechnungen miteinander vergleichen. Wir demonstrieren das Verhalten der drei Strategien zunächst an Beispielrechnungen zu zwei sehr unterschiedlichen Problemen mit bekannten Lösungen und testen anschließend, im Unterabschnitt zur Parameterwahl, den von uns auf Dreiecks- und Tetraedervernetzungen erweiterten Algorithmus 5.2.16 noch an einem etwas komplexeren Problem.
Das erste Beispiel, das wir betrachten, ist Beispiel 5.2.4. Da hier eine auf $\overline{\Omega}$ analytische Lösung vorliegt, besteht die optimale Strategie in einer reinen p-Methode auf einem geeigneten Netz. Von einer erfolgreichen adaptiven *hp*-Strategie erwarten wir daher, dass sie die Glattheit der Lösung erkennt und bis auf wenige *h*-Verfeinerung, vorwiegend in den ersten Iterationsleveln, zu einer reinen *p*-Methode übergeht. Auf Grund der für die optimale Strategie zu erwartenden exponentiellen Konvergenz in p tragen wir für dieses Beispiel den Fehler gegen $\sqrt{\text{DOF}}$ ab.
Das zweite Beispiel, welches wir betrachten, ist das klassische L-Gebiet mit einer Singularität an der einspringenden Ecke (Beispiel 5.2.5). Für dieses Beispiel erwarten wir eine starke Netzverfeinerung zur Singularität hin und überwiegend *p*-Verfeinerungen für den Rest des Gebietes. Die für dieses Beispiel beste bekannte *hp*-Strategie (siehe [BG86c, BG86d, Sch98]) liefert eine Fehlerabschätzung von

$$\|u - u_{FE}\|_{H^1(\Omega)} \leq Ce^{-b(\text{DOF})^{1/3}}.$$

In unseren Grafiken tragen wir daher den Fehler gegen $(\text{DOF})^{1/3}$ auf.
Alle Berechnungen wurden mit dem von uns implementierten *hp*-FE-Programmpaket ADURA-KON sowohl für die Karniadakis & Sherwin Formfunktionen Ψ^{KS}, als auch für die angepassten Formfunktionen Ψ^{Lag} (siehe Definition 3.6.2) durchgeführt. Betrachten wir die einzelnen Strategien im Detail:

- **Strategie I (Vergleich von geschätztem und vorhergesagtem Fehler):** Für unsere Berechnungen haben wir $\sigma = 0.75$, $\gamma_p = 0.75$, $\gamma_h = 6.0$ und $\gamma_n = 1.0$ gewählt. Der vorhergesagte Fehler η^{pred} im Ausgangsnetz wurde so initialisiert, dass die erstmalige Verfeinerung eines jeden Dreiecks eine *p*-Verfeinerung ist. Die Ergebnisse der Testrechnungen sind in den Abbildungen 5.7-5.12, 5.25, 5.28, 5.31, 5.34, 5.35 sowie in den Tabellen 5.2, 5.5, 5.8, 5.11 dargestellt. Wie wie sehen, erhalten wir für Beispiel 5.2.4

die erwartete p-Methode und für Beispiel 5.2.5 die starke Netzverfeinerung in der Nähe der Singularität. Wir erkennen auch, dass der zum Teil recht hohe Polynomgrad an der einspringenden Ecke bereits in den ersten Iterationsschritten zustande kommt und im späteren Verlauf des adaptiven Prozesses keine Erhöhung mehr stattfindet (siehe hierzu auch Bemerkung 5.2.19). Eine leichte Tendenz zur h-Verfeinerung in höheren Iterationsleveln kann durch das Erreichen der maximal möglichen Rechengenauigkeit erklärt werden. In diesem Fall kann keine Fehlerreduktion in dem vorhergesagten Ausmaß mehr erzielt werden und folglich wird stets eine h-Verfeinerung vorgeschlagen. Insgesamt beobachten wir für Beispiel 5.2.5 die gewünschte exponentielle Konvergenz.

- **Strategie II (Drei-Klassen-Algorithmus):** Für unsere Berechnungen haben wir $\sigma = 0.75$ und $\delta_1 = 0.07$, $\delta_2 = 0.65$ gewählt, die Ergebnisse sind in den Abbildungen 5.13-5.18, 5.26, 5.29, 5.32, 5.34, 5.35 und den Tabellen 5.3, 5.6, 5.9, 5.12 dargestellt. Wie wir sehen funktioniert Strategie II für das L-Gebiet sehr gut, liefert für Beispiel 5.2.4 jedoch deutlich zu viele h-Verfeinerungen. Die Erklärung für dieses schlechte Abschneiden liegt einerseits darin begründet, dass Algorithmus 5.2.15 keine reine p-Methode vorschlagen kann und zum anderen auch darin, dass Algorithmus 5.2.15 unabhängig von der Glattheit der Lösung bei nahezu gleichmäßig verteiltem Fehler stets in Richtung einer reinen h-Methode degeneriert.

- **Strategie III (Abklingen der Legendre-Zerlegungskoeffizienten):** Für unsere Berechnungen haben wir $\sigma = 0.75$ und $\delta = 0.95$ gewählt. Des Weiteren, um für das Abschätzen der Abklingrate eine hinreichend große Anzahl von Daten zur Verfügung zu haben, starten wir mit einer uniformen Polynomgradverteilung $p_K = 3$ für alle $K \in \mathcal{T}(\mathcal{N})$. Die Ergebnisse der Testrechnungen sind in den Abbildungen 5.19-5.24, 5.27, 5.30, 5.33, 5.34, 5.35 sowie in den Tabellen 5.4, 5.7, 5.10, 5.13 dargestellt. Wie wir sehen, liefert Algorithmus 5.2.16 stets gute Ergebnisse und trotz der Tatsache, dass wir mit einer uniformen Polynomgradverteilung von $p_K = 3$ im Ausgangsnetz gestartet sind, wächst der Polynomgrad an der einspringenden Ecke nicht über vier (bzw. auf wenigen Dreiecke fünf bei Verwendung der angepassten Lagrange-Formfunktionen). Gemessen in Fehler pro Freiheitsgraden ist Algorithmus 5.2.16 konkurrenzfähig zu den beiden anderen Strategien und erfüllt die Erwartungen sowohl für Beispiel 5.2.4 als auch für Beispiel 5.2.5.

Bemerkung 5.2.17. Die Bilder 5.31-5.33 zeigen die Polynomgradverteilung für das L-Gebiet entlang der Strecke von (0,0) nach (-1/2,1).

Bemerkung 5.2.18. Solang nicht explizit etwas anderes angegeben ist, zeigen die nachfolgenden Grafiken den tatsächlichen und nicht den geschätzten Fehler. Angaben zum Fehlerschätzer beziehen sich auf $\eta := \sqrt{\eta^2}$. Normen sind echte Normen, keine Normquadrate.

Bemerkung 5.2.19. Unsere Berechnungen zeigen, dass der Großteil aller Fehlentscheidungen bezüglich der Wahl zwischen h- und p-Verfeinerung in den ersten Iterationsleveln geschieht. Zu diesen Fehlentscheidungen zählt insbesondere das Erhöhen des Polynomgrades in Regionen wo die Lösung nicht glatt ist. In höheren Iterationsleveln sind solche Fehlentscheidungen eher die Ausnahme (siehe einspringende Ecke im L-Gebiet). Wir gehen daher davon aus, dass eine Kombination obiger Strategien mit einer geeigneten Vergröberungsstrategie, welche die anfänglichen Fehlentscheidungen bezüglich überflüssiger p-Verfeinerungen wieder zurücknimmt, zu weiteren Verbesserungen der adaptiven Algorithmen führen könnte.

Bemerkungen zur Parameterwahl

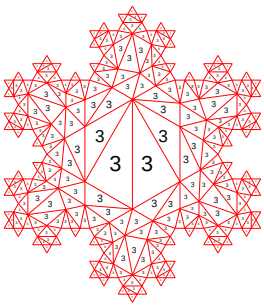
Eines haben alle oben vorgestellten Algorithmen gemein, nämlich dass ihre Entscheidung zwischen h- und p-Verfeinerung von geeignet gewählten Parametern abhängig ist. Die Anzahl der Parameter reicht dabei von einem Parameter für Strategie III (Abklingen der Legendre-Zerlegungskoeffizienten) bis hin zu drei Parametern für Strategie I (Vergleich von geschätztem und vorhergesagtem Fehler). Um zu demonstrieren, dass die Wahl geeigneter Parameter entscheidenden Einfluss auf das Abschneiden des jeweiligen Algorithmus haben kann, betrachten wir speziell für Strategie III folgendes Randwertproblem in schwacher Formulierung:

Beispiel 5.2.20. *Sei Ω das in Abbildung 5.5 dargestellte Gebiet (Schneeflocke). Gesucht ist $u \in H_0^1(\Omega)$ mit*

$$\int_\Omega \nabla u \cdot \nabla v d\Omega = \int_\Omega v d\Omega \qquad \forall v \in H_0^1(\Omega). \tag{5.17}$$

Der für die Entscheidung zwischen h- und p-Verfeinerung wesentliche Parameter ist δ (siehe Algorithmus 5.2.16). Kleine Werte von δ begünstigen die p-Verfeinerung und große Werte von δ die h-Verfeinerung. Abbildung 5.5 zeigt das Verhältnis zwischen Fehler und Freiheitsgraden für verschiedene Parameterwerte. Da wir die exakte Lösung zu Beispiel 5.2.20 nicht kennen, haben wir zum näherungsweisen Berechnen des tatsächlichen Fehlers jeweils eine wesentlich genauere Rechnung als Ersatz zu Grunde gelegt.
Wie wir sehen, schwankt die notwendige Anzahl von Freiheitsgraden, die wir benötigen um einen Fehler $|u - u_{FEM}|_{H^1(\Omega)} \leq 10^{-7}$ zu erreichen, für $\delta \in [0.75, 1.75]$ zwischen $6 * 10^5$ und $12 * 10^5$. Je größer wir δ wählen, umso stärker nähern wir uns einer reinen h-Methode. Speziell für $\delta \geq 1.5$ zeigt Abbildung 5.5, dass nach anfänglich rascher Konvergenz eine Abschwächung hin zu eher algebraischen Konvergenzraten eintritt. Umgekehrt führen kleine Wert von δ zwangsläufig in Richtung einer reinen p-Methode, welche, abgesehen von extrem glatten Lösungen, ebenfalls nur algebraische Konvergenzraten liefert.
Wie aus Abbildung 5.5 recht deutlich zu erkennen ist, führt $0.75 \leq \delta \leq 1.15$ zu exponentieller Konvergenz, wobei wir aber auch sehen, dass $\delta \leq 0.85$ wesentlich schlechtere Ergebnisse liefert, als wie wir zum Beispiel für $\delta = 0.95$ bzw. $\delta = 1.05$ erhalten.
Insgesamt zeigt sich, dass es schon bei nur einem Parameter nicht einfach ist einen möglichst optimalen Wert zu finden und je mehr Parameter im Spiel sind, umso schwieriger wird es. Nichtsdestotrotz haben wir uns für alle drei Verfahren gleichermaßen bemüht, eine möglichst optimale Parameterwahl zu treffen.

Abbildung 5.5: Parameterwahl für Strategie III - Ausgangsnetz - Ergebnisse



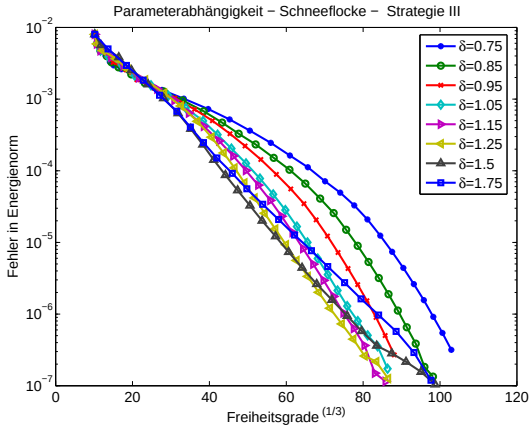


Abbildung 5.6: Ellipse $\mathcal{E}_\rho$

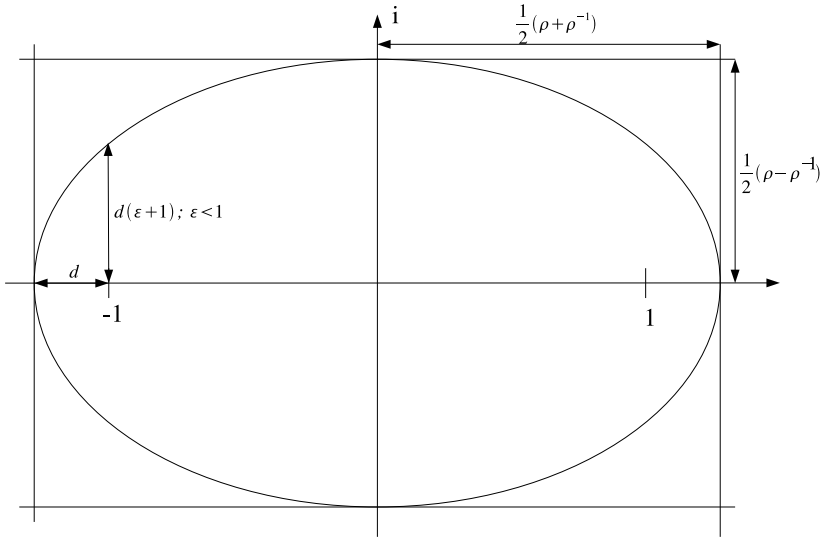


Abbildung 5.7: Strategie I - L-Gebiet - KS-Formfunktionen - Iterationslevel 0, 15, und 25

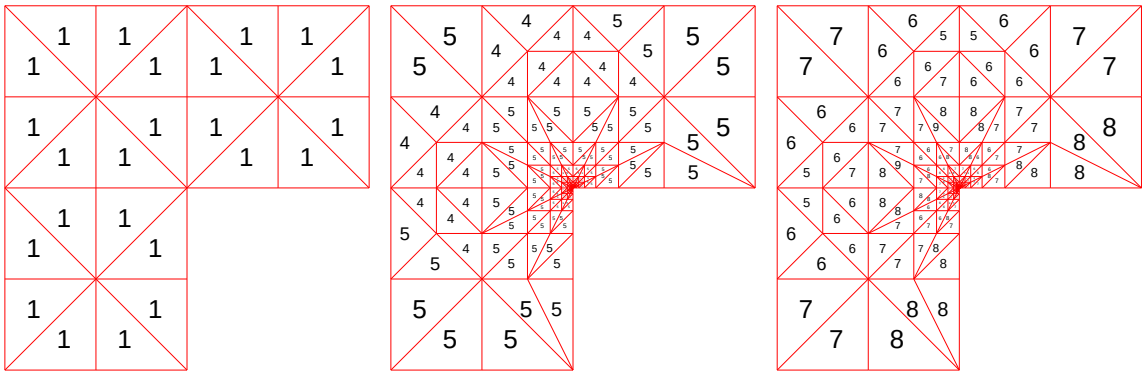


Abbildung 5.8: Zoom zur einspringenden Ecke: Level 15 (Vergrößerungsfaktor 2^9) - Level 25 (Vergrößerungsfaktor 2^{19})

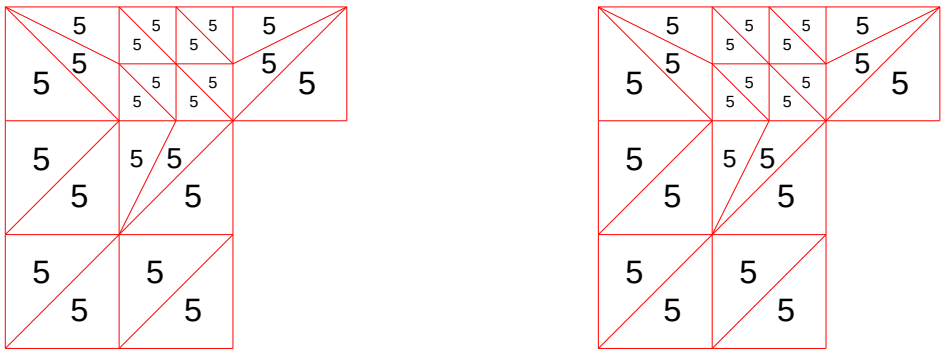


Abbildung 5.9: Strategie I - Quadrat - KS-Formfunktionen - Iterationslevel 0, 10, und 20

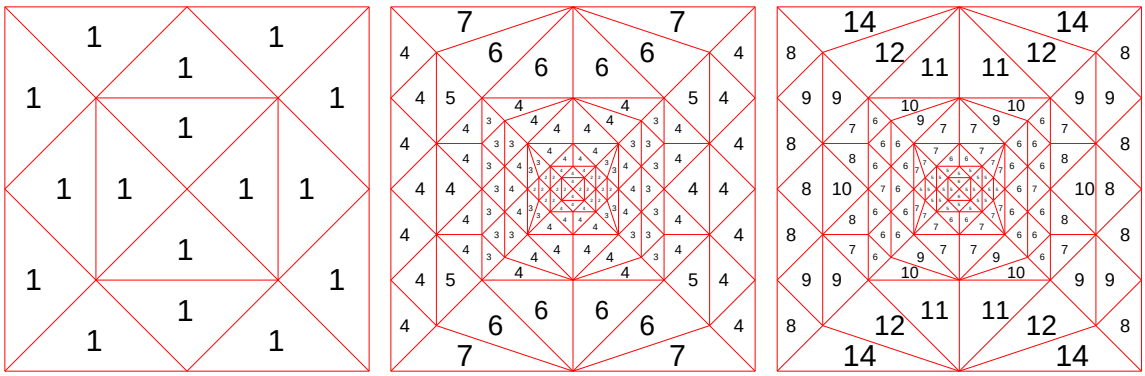


Abbildung 5.10: Strategie I - L-Gebiet - Lag-Formfunktionen - Iterationslevel 0, 15, und 25

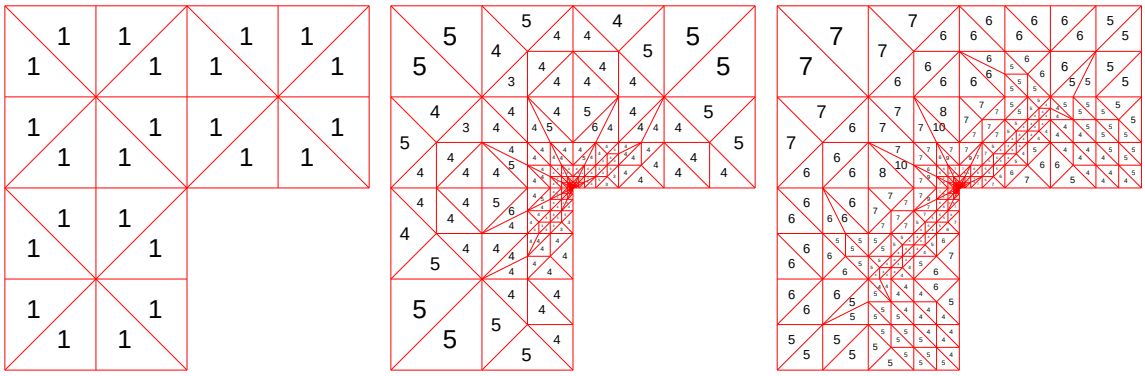


Abbildung 5.11: Zoom zur einspringenden Ecke: Level 15 (Vergrößerungsfaktor 2^9) - Level 25 (Vergrößerungsfaktor 2^{19})

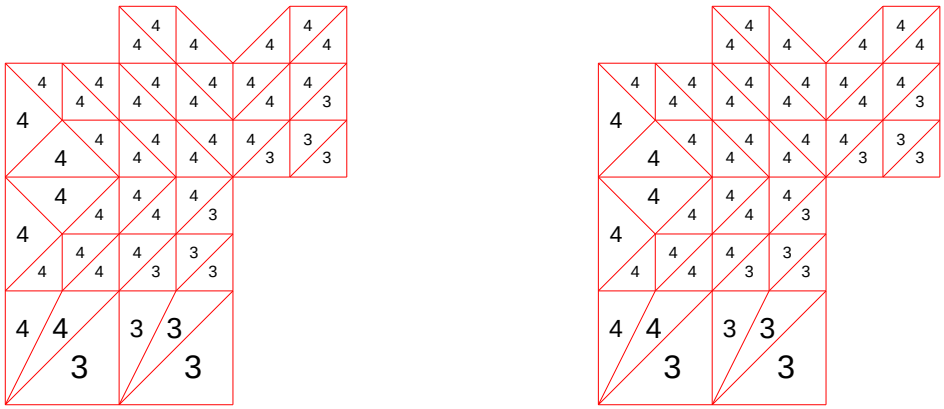


Abbildung 5.12: Strategie I - Quadrat - Lag-Formfunktionen - Iterationslevel 0, 10, und 20

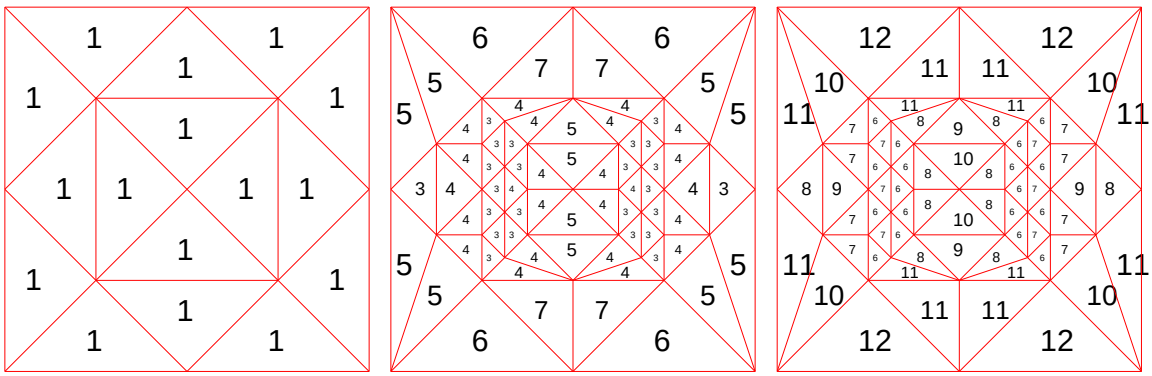


Abbildung 5.13: Strategie II - L-Gebiet - KS-Formfunktionen - Iterationslevel 0, 15, und 25

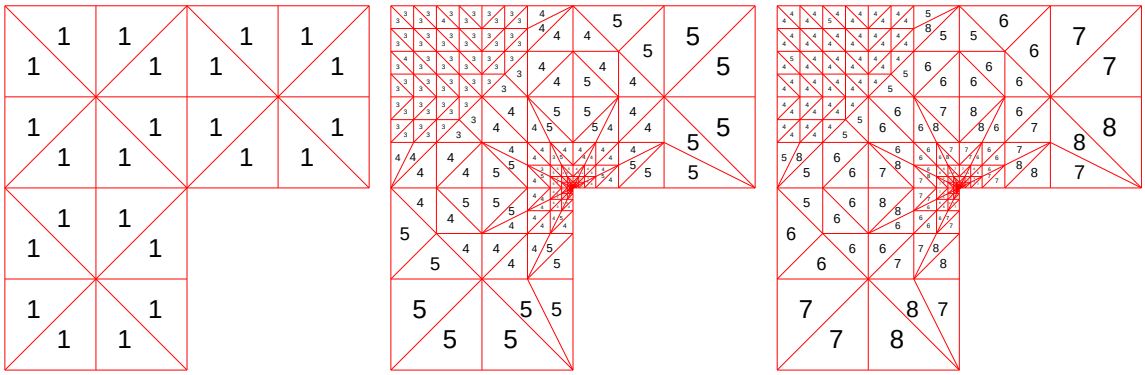


Abbildung 5.14: Zoom zur einspringenden Ecke: Level 15 (Vergrößerungsfaktor 2^9) - Level 25 (Vergrößerungsfaktor 2^{19})

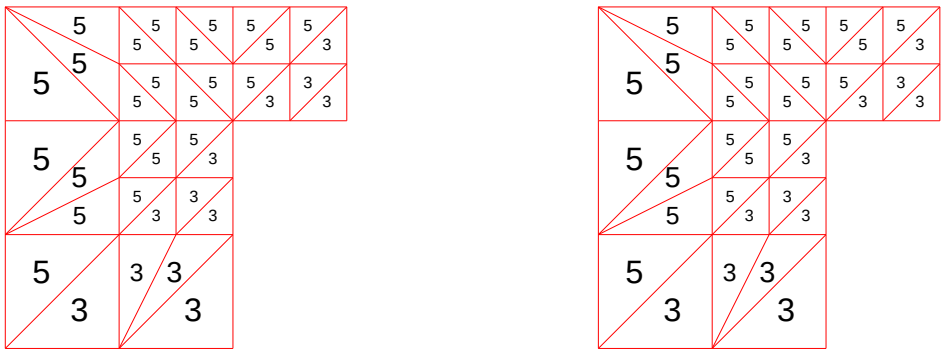


Abbildung 5.15: Strategie II - Quadrat - KS-Formfunktionen - Iterationslevel 0, 10, und 20

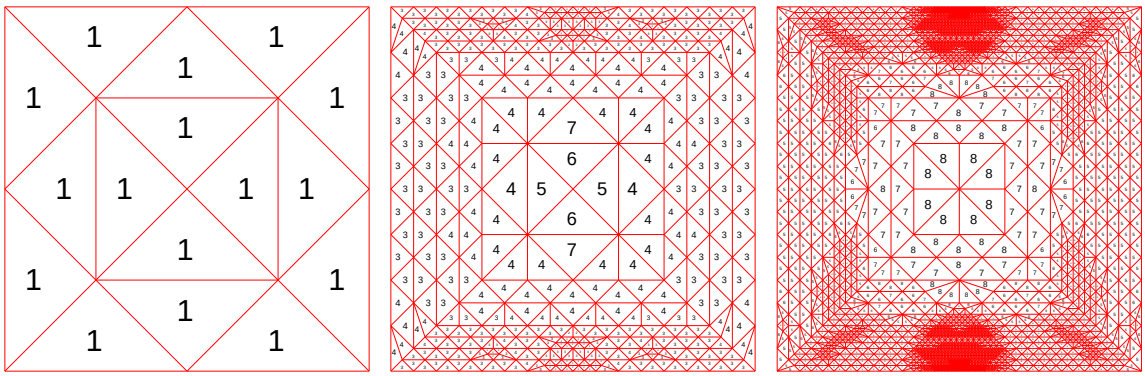


Abbildung 5.16: Strategie II - L-Gebiet - Lag-Formfunktionen - Iterationslevel 0, 15, und 25

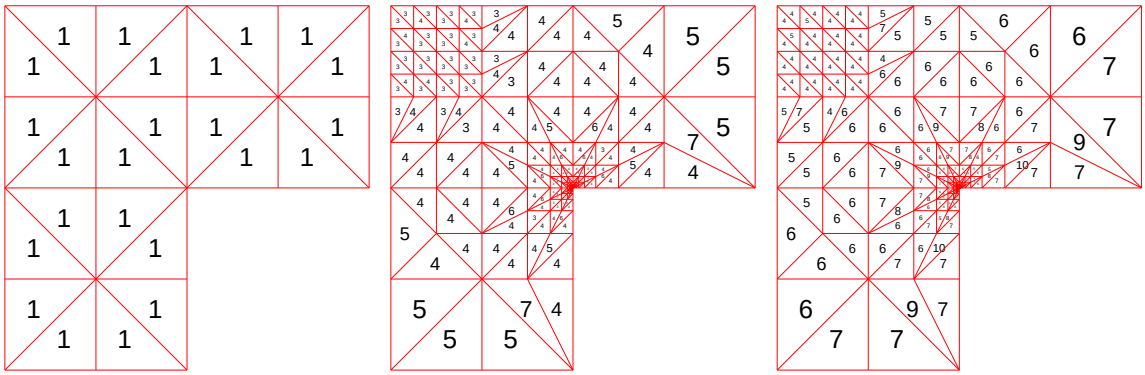


Abbildung 5.17: Zoom zur einspringenden Ecke: Level 15 (Vergrößerungsfaktor 2^{11}) - Level 25 (Vergrößerungsfaktor 2^{21})

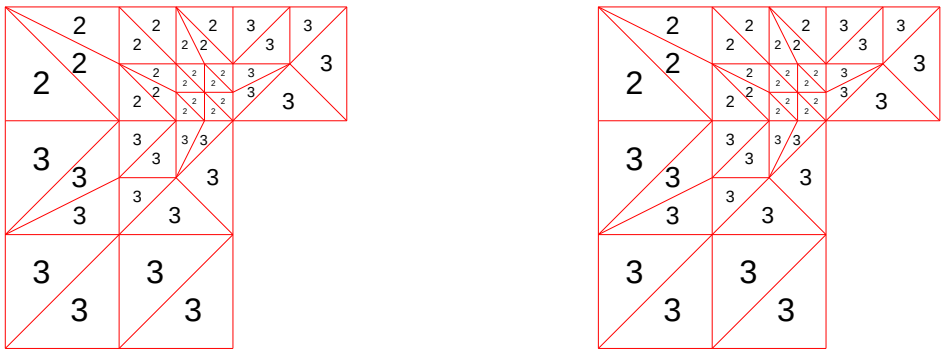


Abbildung 5.18: Strategie II - Quadrat - Lag-Formfunktionen - Iterationslevel 0, 10, und 20

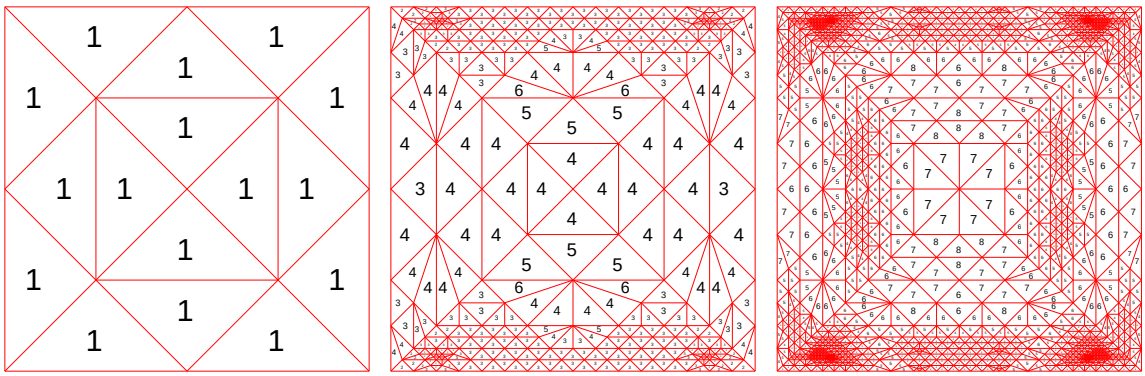


Abbildung 5.19: Strategie III - L-Gebiet - KS-Formfunktionen - Iterationslevel 0, 15, und 25

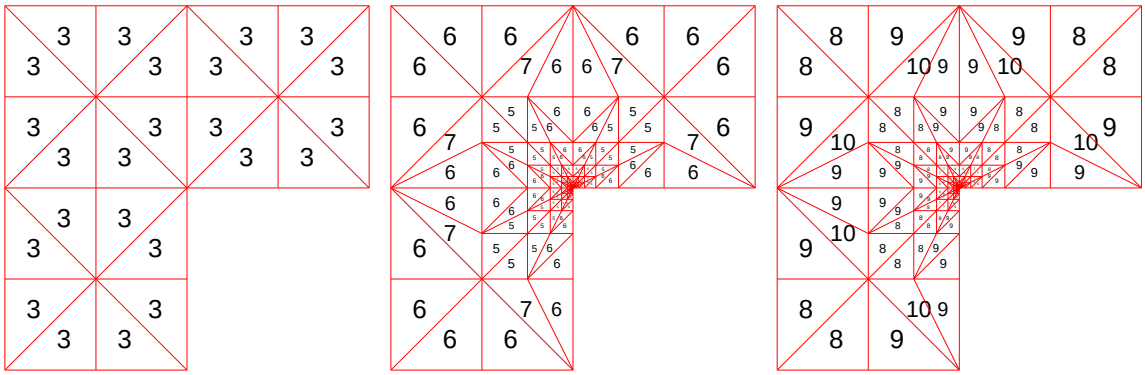


Abbildung 5.20: Zoom zur einspringenden Ecke: Level 15 (Vergrößerungsfaktor 2^{12}) - Level 25 (Vergrößerungsfaktor 2^{23})

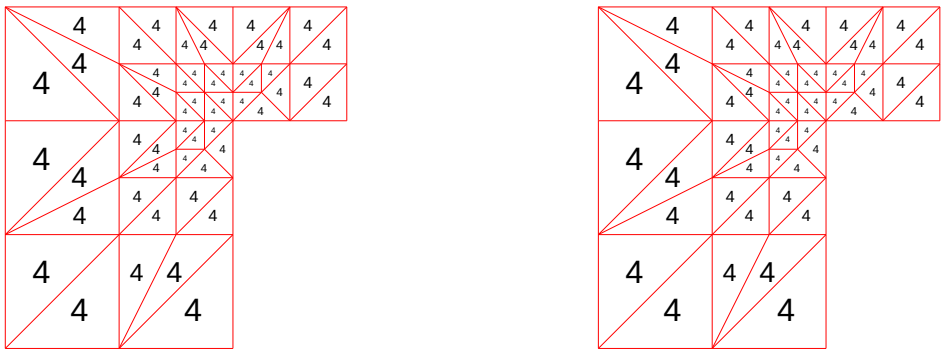


Abbildung 5.21: Strategie III - Quadrat - KS-Formfunktionen - Iterationslevel 0, 10, und 16

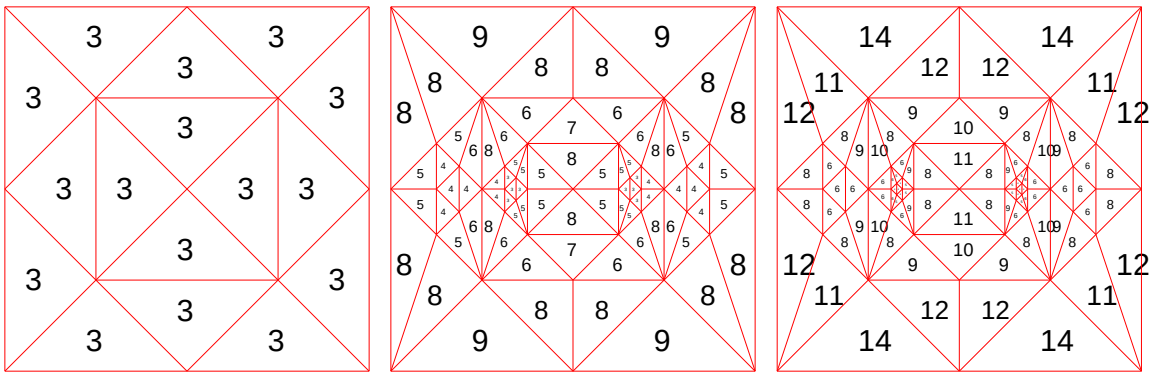


Abbildung 5.22: Strategie III - L-Gebiet - Lag-Formfunktionen - Iterationslevel 0, 15, und 25

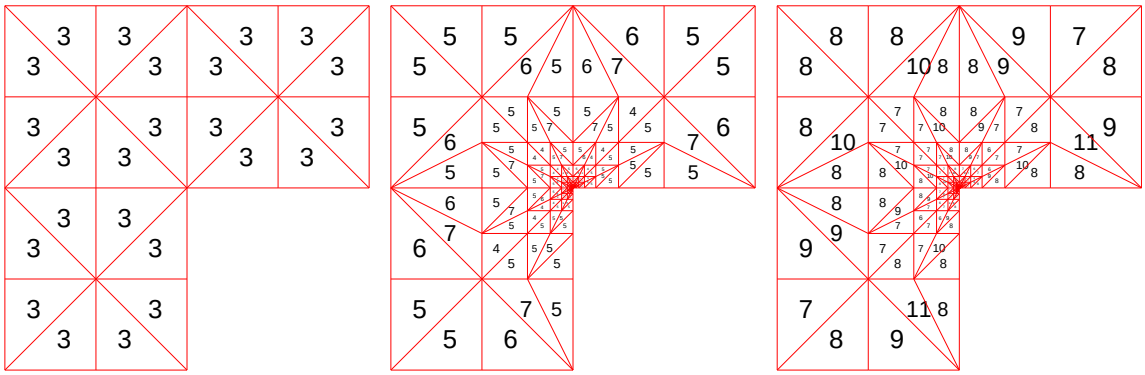


Abbildung 5.23: Zoom zur einspringenden Ecke: Level 15 (Vergrößerungsfaktor 2^{12}) - Level 25 (Vergrößerungsfaktor 2^{23})

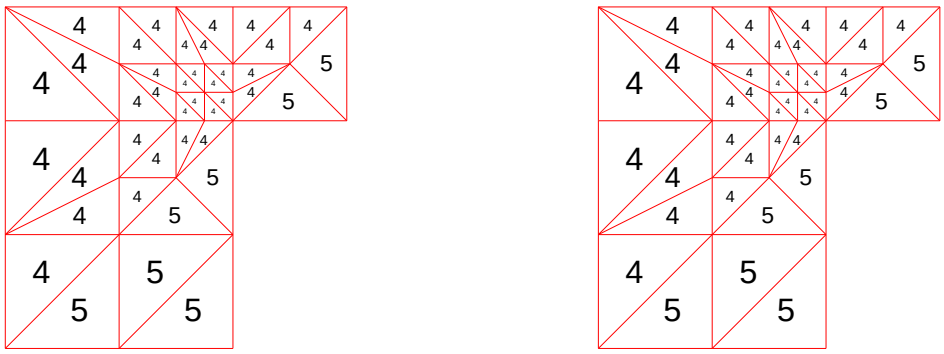


Abbildung 5.24: Strategie III - Quadrat - Lag-Formfunktionen - Iterationslevel 0, 10, und 16

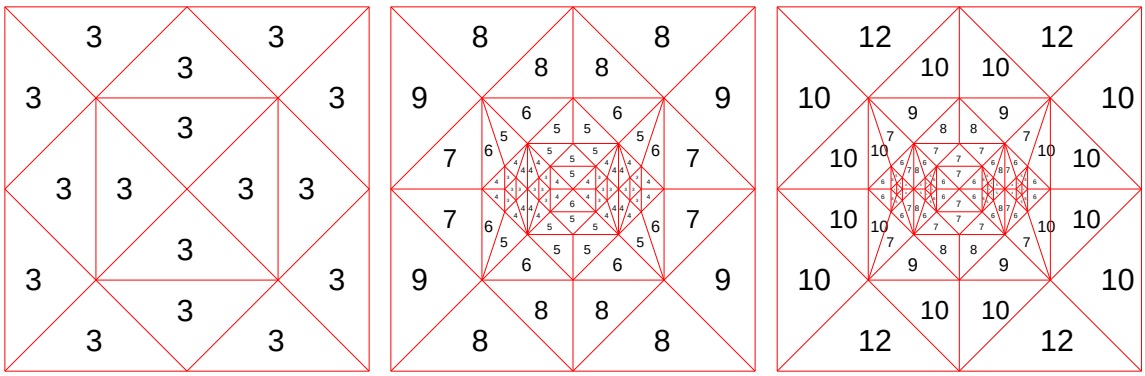


Abbildung 5.25: Strategie I - KS-Formfunktionen

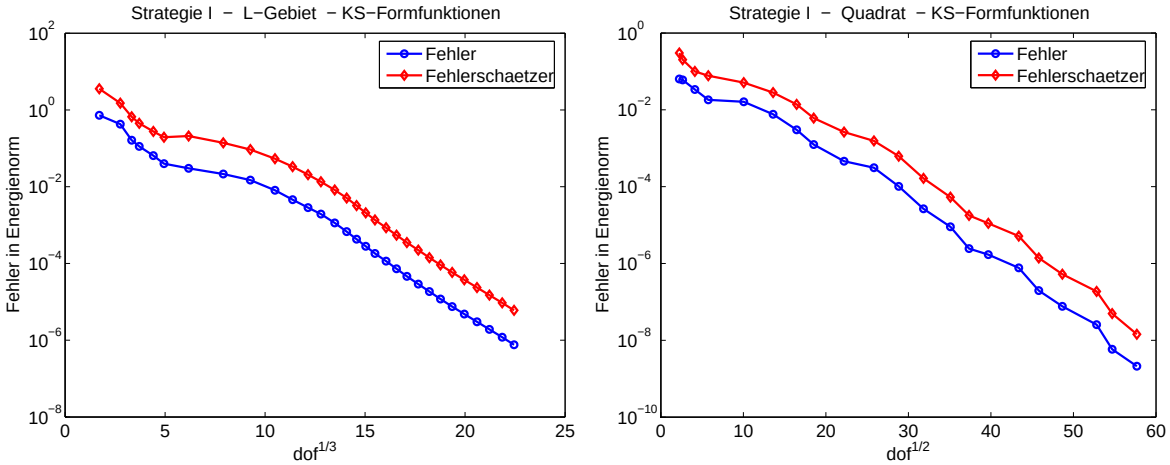


Abbildung 5.26: Strategie II - KS-Formfunktionen

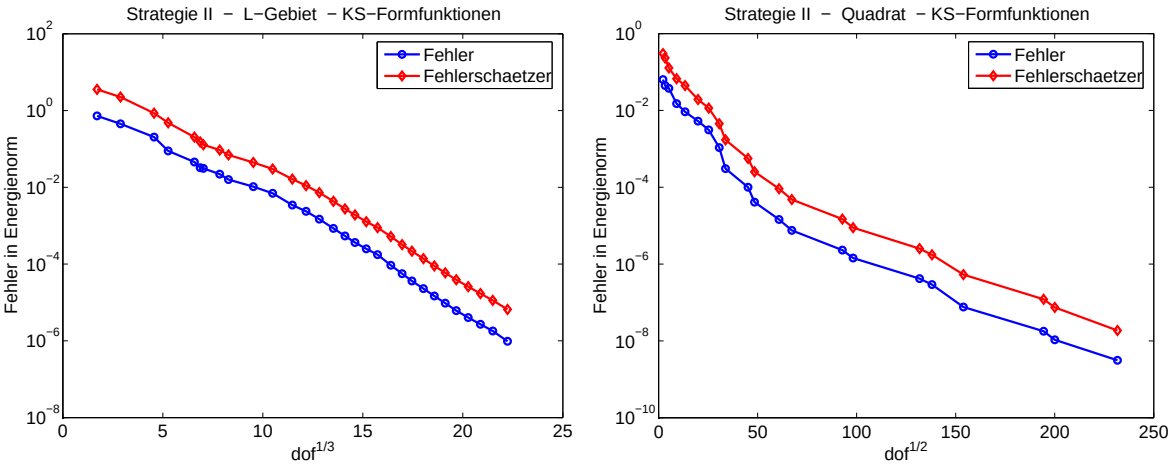


Abbildung 5.27: Strategie III - KS-Formfunktionen

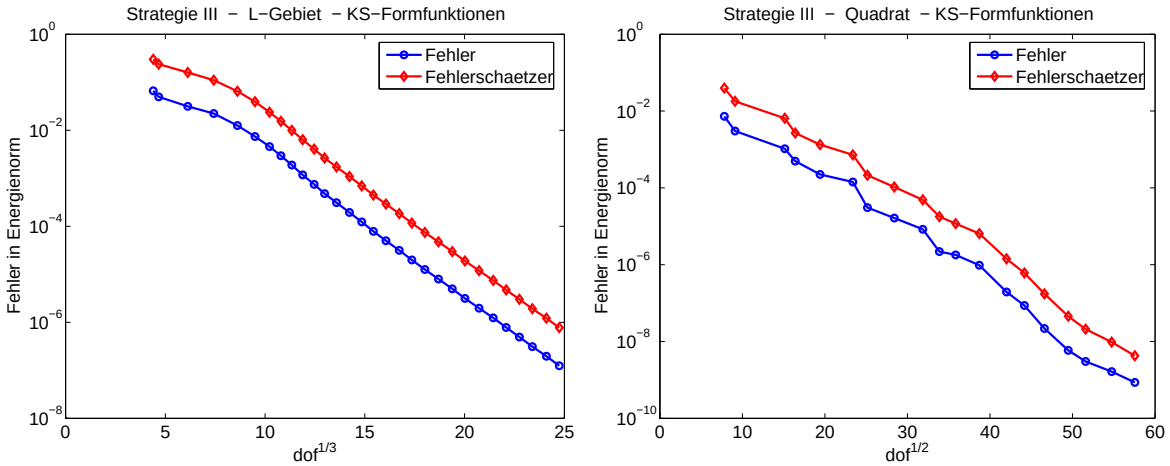


Abbildung 5.28: Strategie I - Lag-Formfunktionen

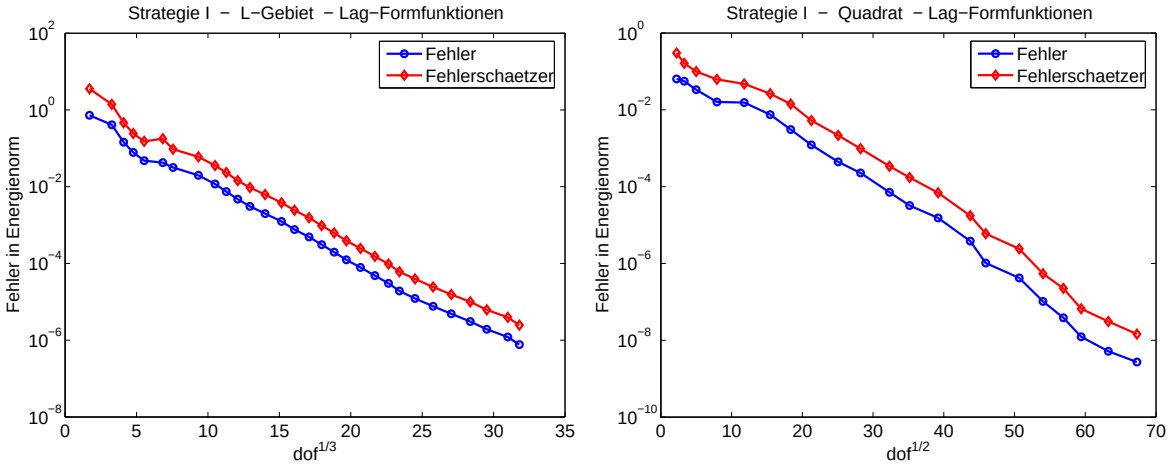


Abbildung 5.29: Strategie II - Lag-Formfunktionen

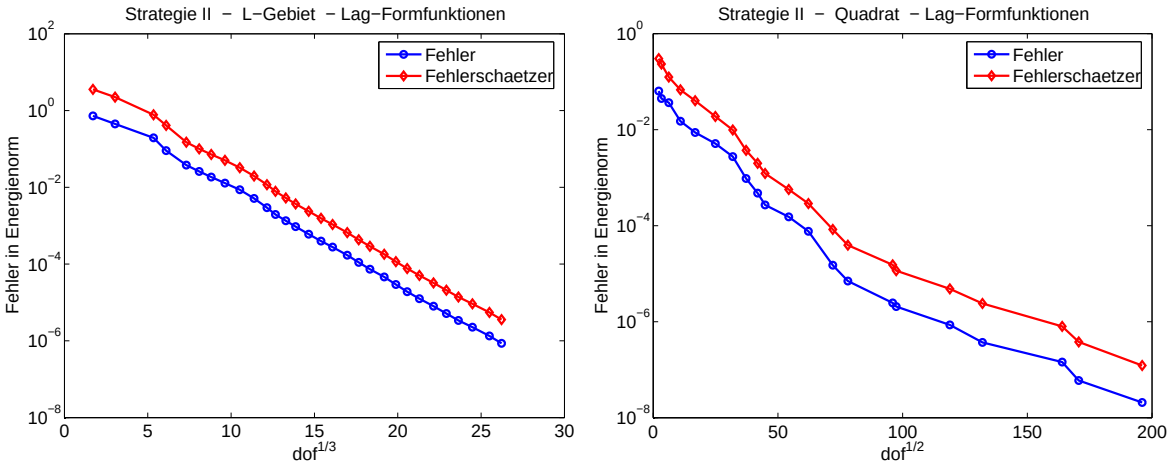


Abbildung 5.30: Strategie III - Lag-Formfunktionen

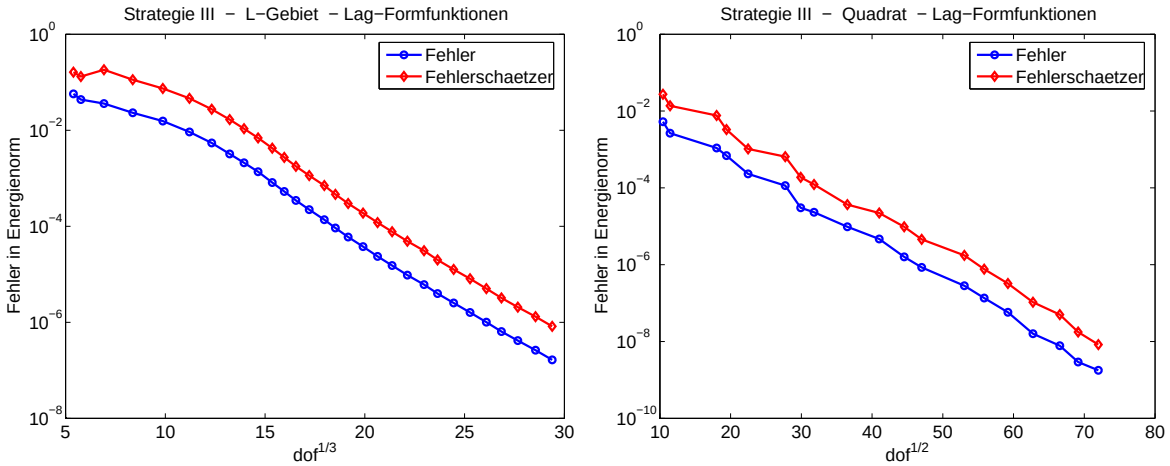


Abbildung 5.31: Strategie I

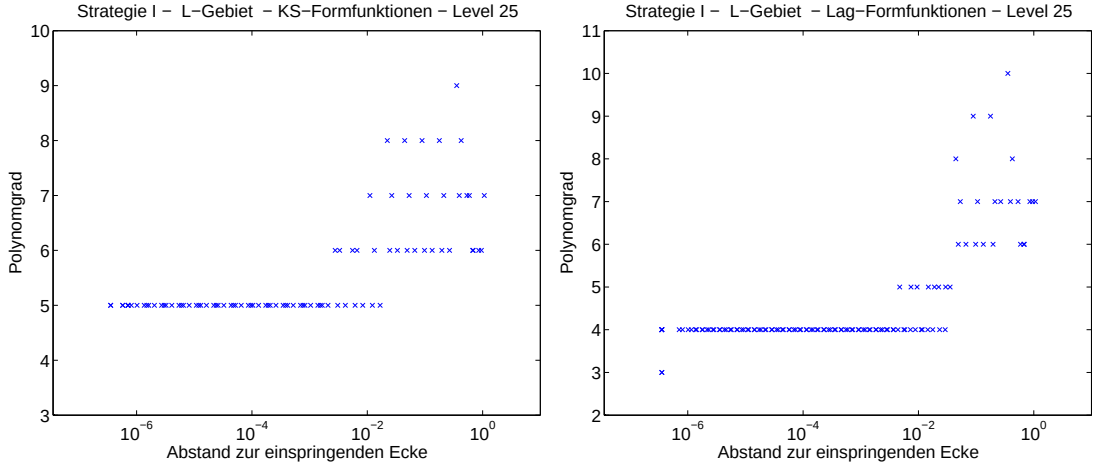


Abbildung 5.32: Strategie II

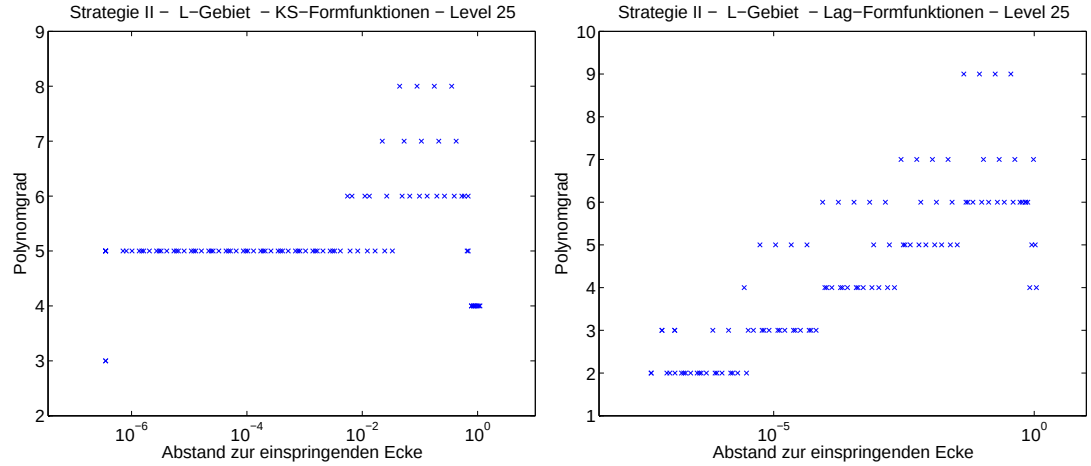


Abbildung 5.33: Strategie III

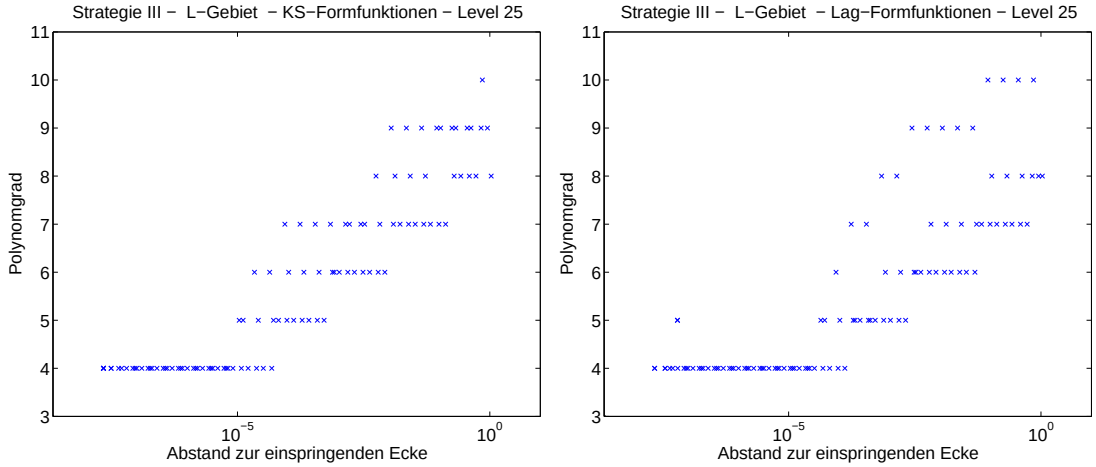


Abbildung 5.34: L-Gebiet - KS-Formfunktionen

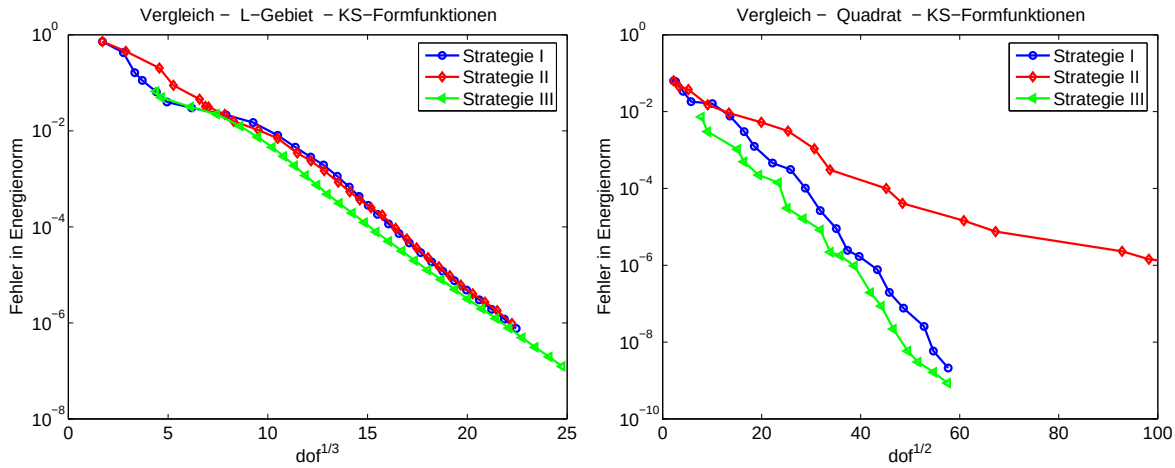


Abbildung 5.35: L-Gebiet - Lag-Formfunktionen

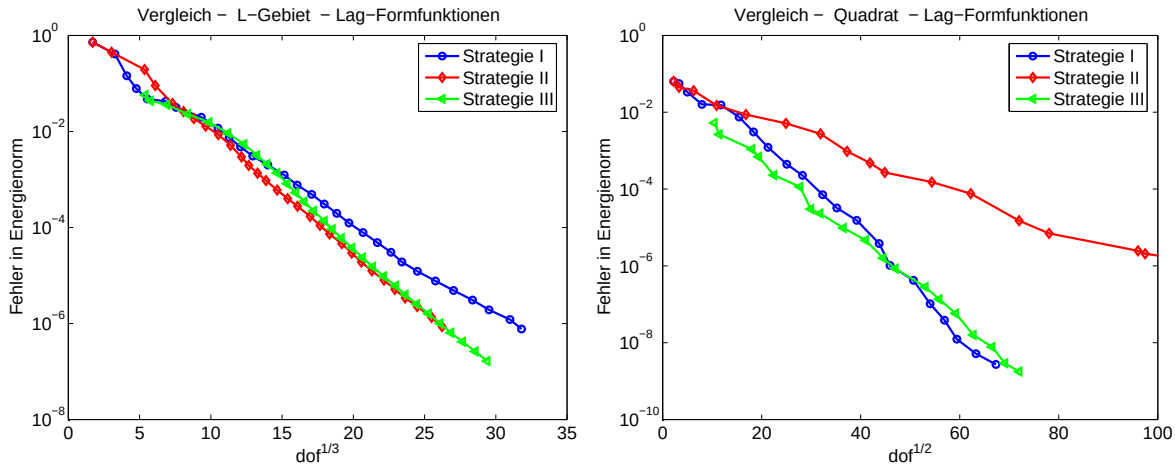


Tabelle 5.2: Strategie I - L-Gebiet - KS-Formfunktionen

Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}	Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}
0	24	1	0	14	13	214	5	14	10
1	24	2	0	14	14	238	5	14	2
2	24	3	0	8	15	262	5	14	2
3	24	3	0	14	16	286	6	14	6
4	24	4	0	10	17	310	6	14	26
5	24	5	6	2	18	334	6	14	24
6	34	5	14	0	19	358	6	14	28
7	58	5	14	2	20	382	7	14	40
8	82	5	26	4	21	406	8	14	42
9	118	5	14	6	22	430	8	14	48
10	142	5	14	8	23	454	8	14	58
11	166	5	14	2	24	478	8	14	66
12	190	5	14	16	25	502	9	14	78

Tabelle 5.3: Strategie II - L-Gebiet - KS-Formfunktionen

Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}	Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}
0	24	1	16	4	13	284	5	12	36
1	60	2	4	52	14	308	5	12	26
2	64	3	12	18	15	332	5	12	20
3	92	4	26	34	16	356	6	12	36
4	126	4	8	4	17	380	6	12	42
5	134	4	6	4	18	404	6	12	68
6	140	5	10	18	19	428	7	12	56
7	158	5	6	12	20	452	7	12	44
8	164	5	12	28	21	476	7	12	60
9	188	5	12	30	22	500	8	12	62
10	212	5	12	48	23	524	8	12	66
11	236	5	12	16	24	548	8	12	72
12	260	5	12	24	25	572	8	12	84

Tabelle 5.4: Strategie III - L-Gebiet - KS-Formfunktionen

Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}	Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}
0	24	3	0	6	13	306	7	14	42
1	24	4	10	0	14	330	7	14	44
2	42	4	14	0	15	354	7	14	56
3	66	4	14	14	16	378	8	14	68
4	90	4	14	8	17	402	8	14	68
5	114	5	14	6	18	426	8	14	78
6	138	5	14	0	19	450	8	14	88
7	162	5	14	4	20	474	9	14	96
8	186	5	14	8	21	498	9	14	96
9	210	6	14	16	22	522	9	14	112
10	234	6	14	16	23	546	10	14	120
11	258	6	14	30	24	570	10	14	116
12	282	6	14	42	25	594	10	14	124

Tabelle 5.5: Strategie I - L-Gebiet - Lag-Formfunktionen

Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}	Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}
0	24	1	0	14	13	274	5	28	14
1	24	2	0	14	14	322	5	34	20
2	24	3	0	10	15	378	6	20	40
3	24	3	0	12	16	418	6	28	36
4	24	4	8	2	17	466	6	20	52
5	36	4	12	0	18	506	7	32	52
6	48	4	24	2	19	560	7	42	52
7	84	4	18	4	20	622	8	46	56
8	114	4	12	0	21	690	9	20	74
9	138	4	12	8	22	730	10	56	58
10	162	4	16	2	23	810	10	66	82
11	194	4	20	12	24	900	10	66	84
12	234	4	20	34	25	994	10	82	88

Tabelle 5.6: Strategie II - L-Gebiet - Lag-Formfunktionen

Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}	Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}
0	24	1	16	4	13	318	7	16	52
1	60	2	4	52	14	342	7	16	54
2	64	3	12	14	15	366	7	16	54
3	92	4	6	44	16	390	7	16	84
4	102	4	16	2	17	414	7	16	66
5	126	4	16	4	18	438	7	16	68
6	150	4	16	16	19	462	7	16	94
7	174	4	16	26	20	486	8	16	82
8	198	5	16	38	21	510	8	16	84
9	222	5	16	34	22	534	8	16	96
10	246	5	16	18	23	558	9	16	114
11	270	5	16	30	24	582	9	16	110
12	294	6	16	30	25	606	10	16	112

Tabelle 5.7: Strategie III - L-Gebiet - Lag-Formfunktionen

Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}	Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}
0	24	3	0	6	13	292	7	16	32
1	24	4	6	2	14	316	7	16	18
2	34	5	12	0	15	340	7	16	30
3	52	5	16	2	16	364	8	16	46
4	76	5	16	8	17	388	9	16	48
5	100	5	16	10	18	412	10	16	54
6	124	5	16	6	19	436	10	16	64
7	148	5	16	0	20	460	10	16	76
8	172	5	16	4	21	484	10	16	66
9	196	5	16	10	22	508	10	16	86
10	220	6	16	6	23	532	10	16	94
11	244	6	16	8	24	556	11	16	100
12	268	6	16	20	25	580	11	16	98

Tabelle 5.8: Strategie I - Quadrat - KS-Formfunktionen

Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}	Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}
0	16	1	0	4	11	116	8	0	50
1	16	2	0	8	12	116	9	0	28
2	16	3	0	8	13	116	10	0	28
3	16	4	8	4	14	116	11	0	58
4	32	4	20	4	15	116	11	0	42
5	68	4	12	18	16	116	11	0	38
6	80	4	0	28	17	116	12	0	64
7	80	5	8	24	18	116	13	0	30
8	92	6	12	26	19	116	13	0	40
9	116	6	0	46	20	116	14	0	48
10	116	7	0	44					

Tabelle 5.9: Strategie II - Quadrat - KS-Formfunktionen

Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}	Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}
0	16	1	8	0	11	704	8	84	118
1	28	1	12	12	12	808	8	360	158
2	40	2	12	24	13	1488	8	144	96
3	64	3	48	8	14	1652	8	724	496
4	148	3	36	90	15	3244	8	248	144
5	200	4	60	60	16	3544	8	24	1542
6	276	4	40	86	17	3568	8	912	120
7	328	4	12	66	18	5584	8	80	304
8	340	5	72	132	19	5664	8	308	2124
9	452	6	32	58	20	6164	8	60	316
10	496	7	128	132					

Tabelle 5.10: Strategie III - Quadrat - KS-Formfunktionen

Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}	Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}
0	16	3	0	8	11	88	10	0	48
1	16	4	12	4	12	88	10	0	28
2	40	5	0	12	13	88	11	0	34
3	40	6	8	12	14	88	12	0	36
4	52	7	8	28	15	88	12	0	24
5	64	7	0	22	16	88	13	0	38
6	64	7	8	24	17	88	14	0	36
7	76	8	0	44	18	88	14	0	44
8	76	9	0	24	19	88	15	0	28
9	76	9	0	24	20	88	15	0	24
10	76	9	8	24					

Tabelle 5.11: Strategie I - Quadrat - Lag-Formfunktionen

Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}	Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}
0	16	1	0	4	11	68	8	0	34
1	16	2	0	6	12	68	9	0	38
2	16	3	0	12	13	68	9	0	16
3	16	3	8	0	14	68	9	0	40
4	32	3	20	0	15	68	9	0	28
5	68	3	0	26	16	68	10	0	22
6	68	4	0	20	17	68	11	0	26
7	68	5	0	28	18	68	11	0	30
8	68	6	0	26	19	68	12	0	30
9	68	6	0	36	20	68	12	0	30
10	68	7	0	18					

Tabelle 5.12: Strategie II - Quadrat - Lag-Formfunktionen

Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}	Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}
0	16	1	8	0	11	520	7	84	74
1	28	1	12	12	12	624	7	16	82
2	40	2	12	20	13	644	7	100	208
3	64	3	48	12	14	808	7	16	20
4	148	3	36	80	15	824	7	260	164
5	200	3	48	64	16	1292	8	112	124
6	252	4	40	52	17	1440	8	400	428
7	304	4	28	52	18	2104	8	112	52
8	340	4	16	52	19	2236	8	172	604
9	356	5	52	152	20	2496	8	48	136
10	428	6	84	92					

Tabelle 5.13: Strategie III - Quadrat - Lag-Formfunktionen

Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}	Level	$\#\mathcal{T}$	p_{max}	h_{ref}	p_{ref}
0	16	3	0	4	11	80	9	16	33
1	16	4	8	4	12	104	9	0	28
2	32	4	0	8	13	104	10	0	34
3	32	5	0	16	14	104	10	0	42
4	32	6	12	4	15	104	11	0	34
5	56	6	0	16	16	104	11	0	26
6	56	6	0	12	17	104	12	0	24
7	56	7	0	34	18	104	13	0	41
8	56	8	16	9	19	104	13	0	52
9	80	8	0	28	20	104	14	0	43
10	80	9	0	24					

Literaturverzeichnis

[Ada75] R. A. Adams. *Sobolev Spaces.* Academic Press, 1975.

[AF03] Robert A. Adams and John J.F. Fournier. *Sobolev Spaces.* Academic Press, second edition, 2003.

[AS97] M. Ainsworth and B. Senior. Aspects of an adaptive *hp*-finite element method: adaptive strategy, conforming approximation and efficient solvers. *Comput. Meth. Appl. Mech. Engrg.*, 150:65–87, 1997.

[AS98] M. Ainsworth and B. Senior. An adaptive refinement strategy for *hp*-finite-element computations. *Appl. Numer. Math.*, 26:165—178, 1998.

[AS99] M. Ainsworth and B. Senior. *hp*-finite element procedures on non-uniform geometric meshes. In M. Bern, J. Flaherty, and M. Luskin, editors, *Grid generation and adaptive algorithms*, pages 1–27. Springer Verlag, 1999. IMA Vol. Math. Appl. 113.

[BD81] I. Babuška and M. R. Dorr. Error estimates for the combined h and p version of the finite element method. *Numer. Math.*, 37:257–277, 1981.

[BG86a] I. Babuška and W. Gui. The h, p and hp versions of the finite element method in one dimension, part I, the error analysis of the p version. *Numer. Math.*, 49:577–612, 1986.

[BG86b] I. Babuška and W. Gui. The h, p and hp versions of the finite element method in one dimension, part II, the error analysis of the h and hp versions. *Numer. Math.*, 49:613–657, 1986.

[BG86c] I. Babuška and B.Q. Guo. The $h-p$ version of the finite element method. Part 1: The basic approximation results. *Computational Mechanics*, 1:21–41, 1986.

[BG86d] I. Babuška and B.Q. Guo. The $h-p$ version of the finite element method. Part 2: General results and applications. *Computational Mechanics*, 1:203–220, 1986.

[BG88] I. Babuška and B. Guo. The *hp* version of the finite element method for domains with curved boundaries. *SIAM J. Numer. Anal.*, 25:837–861, 1988.

[BG00] Ivo Babuska and Benqi Guo. Optimal estimates for lower and upper bounds of approximation errors in the p-version of the finite element method in two dimensions. *Numer. Math.*, 85(2):219–255, 2000.

[BL76] J. Bergh and J. Löfström. *Interpolation Spaces*. Springer Verlag, 1976.

[BM92] C. Bernardi and Y. Maday. *Approximations spectrales de problèmes aux limites elliptiques*. Mathématiques & Applications. Springer Verlag, 1992.

[BM97] C. Bernardi and Y. Maday. Spectral methods. In P.G. Ciarlet and J.L. Lions, editors, *Handbook of Numerical Analysis, Vol. 5*. North Holland, 1997.

[BOV01] C. Bernardi, R.G. Owens, and J. Valenciano. An error indicator for mortar element solutions to the Stokes problem. *IMA J. Numer. Anal.*, 21:857–886, 2001.

[Bra97] D. Braess. *Finite Elemente - Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer Lehrbuch, 1997.

[BS87] I. Babuška and M. Suri. The optimal convergence rate of the p-version of the finite element method. *SIAM J. Numer. Anal.*, 24:750–776, 1987.

[BS92] U. Banerjee and M. Suri. The effect of numerical quadrature in the p-version of the finite element method. *Math. Comput.*, 59(199):1–20, 1992.

[BS94] I. Babuška and M. Suri. The p and h-p versions of the finite element method, basic principles and properties. *SIAM review*, 36(4):578–632, 1994.

[BSK81] I. Babuška, B. Szabó, and I. N. Katz. The p-version of the finite element method. *SIAM J. Numer. Anal.*, 18:515–545, 1981.

[Cia76] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland Publishing Company, 1976.

[CMM91] Zhiqiang Cai, Jan Mandel, and Steve McCormick. The finite volume element method for diffusion equations on general triangulations. *SIAM J. Numer. Anal.*, 28(2):392–402, 1991.

[DL93] R.A. DeVore and G.G. Lorentz. *Constructive Approximation*. Springer Verlag, 1993.

[EGH00] R. Eymard, T. Gallouët, and R. Herbin. *Finite volume methods.* Ciarlet, P. G. (ed.) et al., Handbook of numerical analysis. Vol. 7, Amsterdam: North-Holland/ Elsevier., 2000.

[EM04] T. Eibner and J.M. Melenk. An adaptive strategy for hp-FEM based on testing for analyticity. Technical Report 04-10, SFB 393, TU Chemnitz, `http://www.tu-chemnitz.de/sfb393`, 2004.

[EM05a] T. Eibner and J.M. Melenk. Fast algorithms for setting up the stiffness matrix in hp-fem: a comparison. In *HERCMA 2005*, `http://www.tu-chemnitz.de/sfb393`, 2005.

[EM05b] T. Eibner and J.M. Melenk. Multilevel preconditioner for boundary concentrated hp-FEM. Technical Report 05-13, SFB 393, TU Chemnitz, `http://www.tu-chemnitz.de/sfb393`, 2005.

[EM06a] T. Eibner and J.M. Melenk. Local error analysis of the boundary concentrated FEM. *IMA Journal of Numerical Analysis*, erscheint 2006.

[EM06b] T. Eibner and J.M. Melenk. Quadrature analysis for *hp*-FEM on triangular and tetrahedral meshes. in prep. 2006.

[GR80] I.S. Gradshteyn and I.M. Ryzhik. *Table of Integrals, Series, and Products, corrected and enlarged edition*. Academic Press, New York, 1980.

[GR92] Ch. Großmann and H.-G. Roos. *Numerik partieller Differentialgleichungen*. Teubner Studienbücher, 1992.

[Gri85] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Pitman, 1985.

[GRT93] H. Goering, H.G. Roos, and L. Tobiska. *Finite-Element-Methode*. Akademie Verlag, 1993.

[Hac94] W. Hackbusch. *Iterative Solution of Large Sparse Systems of Equations*, volume 95 of *Applied Mathematical Sciences*. Springer, 1994.

[Hac95] W. Hackbusch. *Integral Equations. Theory and Numerical Treatment*. Birkhäuser, 1995.

[Hac96] W. Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen*. Teubner, 1996.

[HMS01] N. Heuer, M.E. Mellado, and E.P. Stephan. *hp*-adaptive two-level methods for boundary integral equations on curves. *Computing*, 67(4):305–334, 2001.

[Hör90] L. Hörmander. *An Introduction Complex Analysis in Several Variables*. North Holland, 1990.

[HS05] P. Houston and E. Süli. A note on the design of *hp*-adaptive finite element methods for elliptic partial differential equations. *Comput. Meth. Appl. Mech. Engrg.*, 194:229–243, 2005.

[HSS03] P. Houston, B. Senior, and E. Süli. Sobolev regularity estimation for hp-adaptive finite element methods. In F. Brezzi, A. Buffa, S. Corsaro, and A. Murli, editors, *Numerical Mathematics and Advanced Applications, ENUMATH 2001*, pages 631–656. Springer-Verlag, 2003.

[JL01] M. Jung and U. Langer. *Methode der Finiten Elemente für Ingenieure: eine Einführung in die numerischen Grundlagen und Computersimulation*. Teubner, 2001.

[KM02] B.N. Khoromskij and J.M. Melenk. An efficient direct solver for the boundary concentrated FEM in 2D. *Computing*, 69:91–117, 2002.

[KM03] B.N. Khoromskij and J.M. Melenk. Boundary concentrated finite element methods. *SIAM J. Numer. Anal.*, 41(1):1–36, 2003.

[KS99] G.E. Karniadakis and S.J. Sherwin. *Spectral/hp Element Methods for CFD*. Oxford University Press, 1999.

[Mav94] Catherine Mavriplis. Adaptive mesh strategies for the spectral element method. *Comput. Methods Appl. Mech. Engrg.*, 116(1-4):77–86, 1994. ICOSAHOM'92 (Montpellier, 1992).

[Mel01] J.M. Melenk. On condition numbers in *hp*-FEM with Gauss-Lobatto based shape functions. *J. Comput. Appl. Math.*, 139:21–48, 2001.

[Mel02] J.M. Melenk. *hp finite element methods for singular perturbations*, volume 1796 of *Lecture Notes in Mathematics*. Springer Verlag, 2002.

[MGS01] J.M. Melenk, K. Gerdes, and C. Schwab. Fully discrete *hp*-FEM: fast quadrature. *Comput. Meth. Appl. Mech. Engrg.*, 190:4339–4364, 2001.

[Mn90] Y. Maday and E.M. Rønquist. Optimal error analysis of spectral methods with emphasis on non-constant coefficients and deformed geometries. *Comput. Meth. Appl. Mech. Engrg.*, 80:91–115, 1990.

[Mor66] C.B. Morrey. *Multiple Integrals in the Calculus of Variations*. Springer Verlag, 1966.

[MW01] J.M. Melenk and B. Wohlmuth. On residual-based a posteriori error estimation in *hp*-FEM. *Advances in Comp. Math.*, 15:311–331, 2001.

[Neč64] J. Nečas. Sur la coercivité des formes sesquilinéaires elliptiques. *Rev. Roumaine de Math. Pures et App.*, 9(1):47–69, 1964.

[Neč67] J. Nečas. *Les méthodes directes en théorie des équations elliptiques*. Masson, 1967.

[Nep86] S. V. Nepomnyaschikh. *Domain Decomposition and Schwarz Methods in a Subspace for the Approximate Solution of Elliptic Boundary Value Problems*. PhD thesis, Computing Center of the Siberian Branch of the USSR Academy of Sciences, Novosibirsk, 1986.

[Ors80] S.A. Orszag. spectral methods for problems in complex geometries. *J. Comput. Phys.*, 37:70–92, 1980.

[Osw94] P. Oswald. *Multilevel Finite Element Approximation*. Teubner Skripten zur Numerik. Teubner, 1994.

[QV97] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*. Springer, 1997.

[Sam84] A.A Samarskii. *Theorie der Differenzenverfahren*. Akademische Verlagsgesellschaft Geest & Portig K.-G., Leipzig, 1984.

[Sch98] C. Schwab. *p- and hp-Finite Element Methods*. Oxford University Press, 1998.

[SK95] S.J. Sherwin and G.E. Karniadakis. A new triangular and tetrahedral basis for high-order (*hp*) finite element methods. *Internat. J. Numer. Meths. Engrg.*, 38:3775–3802, 1995.

[SK96] S.J. Sherwin and G.E. Karniadakis. Tetrahedral *hp* finite elements: Algorithms and flow simulations. *J. Comput. Phys.*, 124:14–45, 1996.

[SMPZ05] Schöberl, J.M. Melenk, C. Pechstein, and S. Zaglmayr. Additive schwarz preconditioning for p-version triangular and tetrahedral finite elements. Technical Report 2005–11, RICAM, `http://www.ricam.oeaw.ac.at`, 2005.

[SS04] S. Sauter and C. Schwab. Teubner, 2004.

[Ste03] O. Steinbach. *Numerische Näherungsverfahren für elliptische Randwertprobleme: Finite Elemente und Randelemente*. Teubner, 2003.

[STW90] Albert H. Schatz, Vidar Thome, and Wolfgang L. Wendland. *Mathematical theory of finite and boundary element methods. Lecture notes of the seminar 'Mathematische Theorie der finiten Element- und Randelementmethoden' organized by the 'Deutsche Mathematiker-Vereinigung' and held in Duesseldorf, Germany, from June 7th to 14th, 1987.* DMV Seminar, 15. Basel etc.: Birkhuser Verlag. 276 p. sFr. 52.00; DM 58.00 , 1990.

[Sze75] B. Szegö. *Orthogonal Polynomials*. American Mathematical Society, fourth edition, 1975.

[Tri95] H. Triebel. *Interpolation Theory, Function Spaces, Differential Operators*. Johann Ambrosius Barth, 2 edition, 1995.

[TW05] A. Toselli and O. Widlund. *Domain Decomposition Methods — Algorithms and Theory*. Springer Verlag, 2005.

[Ver98] R. Verfürth. *Numerische Behandlung von Differentialgleichungen II - Vorlesungsskriptum*. http://www.ruhr-uni-bochum.de/num1/rv/publikationsliste.html, 1998.

[Yse99] H. Yserentant. Coarse grids spaces for domains with a complicated boundary. *Numerical Algorithms*, 21:387–392, 1999.

[Zha92] X. Zhang. Multilevel Schwarz methods. *Numer. Math.*, 63:521–539, 1992.

Thesen

zur Dissertation

Randkonzentrierte und adaptive hp-FEM

zur Erlangung des akademischen Grades eines „Dr. rer. nat.“

an der Technischen Universität Chemnitz
Fakultät für Mathematik

vorgelegt von **Dipl.-Math.-techn. Tino Eibner**

1. Betrachten wir das Verhältnis von Fehler gegenüber Freiheitsgraden, so ist die *hp*-Version der Finiten-Element-Methode bei geeignet gewählter Netzstruktur und Polynomgradverteilung sowohl der klassischen *h*-FEM als auch der *p*-FEM überlegen. Insbesondere kann für eine große Klasse von Aufgaben gezeigt werden, dass mit der *hp*-FEM exponentielle Konvergenz möglich ist (siehe [BG86c, BG86d, BG88]). Nichtsdestotrotz ist der Einsatz der *hp*-FEM aber auch mit einigen zu lösenden Schwierigkeiten verbunden. So ist zum Beispiel die Implementation einer *hp*-FEM wesentlich anspruchsvoller als die Implementation einer klassischen *h*- bzw. *p*-FEM und insbesondere gilt das Aufstellen der Steifigkeitsmatrix hierbei als sehr rechenintensiv und aufwändig. Ein weiterer wichtiger Punkt ist die Frage nach der für ein konkretes Problem bzw. für eine konkrete Problemklasse geeignetesten Netzstruktur und Polynomgradverteilung. Speziell für eine adaptive Steuerung dieser Größen ist hierbei entscheidend, ob ein zur Verfeinerung bestimmtes Element in mehrere kleine Elemente (*h*-Verfeinerung) geteilt wird oder ob stattdessen lieber die Approximationsordnung erhöht werden sollte (*p*-Verfeinerung). Die Wichtigkeit dieser Entscheidung wird durch die Tatsache belegt, dass nur die passende Netzstruktur und Polynomgradverteilung zu exponentieller Konvergenz führt.

   In der vorliegenden Arbeit werden all die eben als Schwierigkeiten herausgestellten Punkte näher untersucht und die theoretischen Ergebnisse stets durch Testrechnungen mit dem von uns implementierten 2D-*hp*-FEM-Programmpaket ADURAKON verifiziert.

2. Wir untersuchen verschiedene Algorithmen für das Assemblieren der Steifigkeitsmatrix beziehungsweise der „on the fly“ Matrix-Vektor-Multiplikation und betrachten dabei sowohl den Fall von variablen Koeffizienten als auch die Vereinfachungen, die sich bei konstanten Koeffizienten ergeben. Neben dem Vorstellen von bekannten Algorithmen (Standard-Algorithmus und Summenfaktorisierung aus [KS99]) verallgemeinern wir hierbei auch die in [MGS01] vorgestellte Spektral-Galerkin-Methode sowohl für Dreiecks- als auch Tetraederelemente. Wesentlich für diese Verallgemeinerung ist dabei die Definition an die Quadratur angepasster Formfunktionen.

3. Die von uns definierten und an die Quadratur angepassten Formfunktionen Φ^{Lag} enthalten, gegenüber der bei den Karniadakis & Sherwin-Formfunktionen Φ^{KS} ([KS99]) verwendeten minimal notwendigen Anzahl von Formfunktionen, auf dem Dreieck die doppelte und auf dem Tetraeder die 6fache Anzahl innerer Formfunktionen. Dadurch erhalten wir einerseits eine leicht verbesserte Approximationseigenschaft, andererseits

wird es aber auch schwieriger an die Rechenzeiten der Summenfaktorisierung bezüglich Φ^{KS}-Formfunktionen heranzukommen bzw. diese Rechenzeit zu unterbieten. Asymptotisch ist die Rechenzeit für das Aufstellen der Steifigkeitsmatrix mittels Spektral-Galerkin-Methode mit einem Aufwand von $O(p^{2d})$ der Summenfaktorisierung mit Aufwand $O(p^{2d+1})$ überlegen. Wegen $\#\Phi^{Lag} \geq \#\Phi^{KS}$ bedarf es aber konkreter Testrechnungen, um zu sehen, ob tatsächlich ein für die Praxis relevanter Bereich existiert, in dem es sich lohnt, die Spektral-Galerkin-Methode zu verwenden.

4. Unsere numerischen Ergebnisse belegen, dass insbesondere in 2D das Assemblieren der Steifigkeitsmatrix mittels Spektral-Galerkin-Methode zu Rechenzeitersparnissen führt. Auch bei der, insbesondere bei Verwendung eines PCG-Lösers, möglichen Alternative, einer „on the fly“ Matrix-Vektor-Multiplikation, reduziert sich in 2D unter Ausnutzung der Spektral-Galerkin-Ideen die notwendige Rechenzeit deutlich gegenüber dem auf Summenfaktorisierung basierenden Algorithmus. Für den 3D Fall ist die benötigte Rechenzeit für das Aufstellen der Steifigkeitsmatrix mittels Spektral-Galerkin-Methode trotz 6facher Anzahl innerer Formfunktionen nahezu identisch zur Summenfaktorisierung in Verbindung mit den Karniadakis & Sherwin-Formfunktionen.

5. Ein in der *hp*-FEM verbreitetes Verfahren zur Dimensionsreduktion ist die statische Kondensation. Die statische Kondensation ist wegen der vergrößerten Anzahl innerer Formfunktionen für $\Phi^{(Lag)}$-basierte Steifigkeitsmatrizen langsamer als bei Verwendung von $\Phi^{(KS)}$. Für 2D sind die Rechenzeiten für das Aufstellen der kondensierten $\Phi^{(Lag)}$-basierten Steifigkeitsmatrix mittels Spektral-Galerkin-Algorithmus und die Rechenzeit für das Aufstellen der kondensierten $\Phi^{(KS)}$-basierten Steifigkeitsmatrix mittels Summenfaktorisierung bis zu einem Polynomgrad $p_k \leq 20$ nahezu identisch. In 3D gilt dies nur für Polynomgrade $p_k \leq 8$.

6. Eine spezielle Version der *hp*-FEM ist die in [KM03] erstmals vorgestellte randkonzentrierte Finite-Element-Methode. Mittels a priori vorgegebener Netzstruktur und Polynomgradverteilung stellt sie eine überaus effektive und leistungsfähige Methode dar, die insbesondere für elliptische Randwertaufgaben mit analytischen Koeffizienten geeignet ist, deren Lösungen jedoch durch Randeffekte von geringer globaler Regularität sind.

   In der vorliegenden Arbeit beweisen wir für die randkonzentrierte FEM eine gegenüber der globalen H^1-Konvergenz verbesserte lokale Konvergenzrate im Gebietsinneren und konstruieren zwei Vorkonditionierer.

7. Die Konstruktion der auf der Additiv-Schwarz-Methode (siehe [Osw94, TW05]) basierenden Vorkonditionierer für die randkonzentrierte *hp*-FEM ist von der Dimension des Gebiets Ω unabhängig, ebenso wie der Beweis der optimalen Konditionszahlen $O(1)$ für die vorkonditionierten Systeme. Numerische Beispiele in 2D und theoretische Komplexitätsabschätzungen demonstrieren zudem die Wirksamkeit und Effizienz der Vorkonditionierer. Speziell für Probleme mit gemischten- oder Neumann-Randbedingungen reduziert sich die für das Lösen des FE-Gleichungssystems notwendige Rechenzeit deutlich gegenüber einer reinen CG-Iteration. Konkrete Testrechnungen zeigen, dass wir bereits ab einer Problemgröße mit circa 2000 Unbekannten signifikante Rechenzeiteinsparungen erzielen können, dass die Rechenzeit für das Lösen des FE-Gleichungssystems mittels Vorkonditionierer nahezu optimal in der Problemgröße skaliert und dass diese

Rechenzeit in der gleichen Größenordnung wie die Rechenzeit für das Aufstellen der Steifigkeitsmatrix für konstante Koeffizienten liegt.

8. Wir beweisen, dass eine auf dem Referenztetraeder definierte L^2-Funktion genau dann analytisch auf einer Umgebung des Referenztetraeders ist, wenn ihre Zerlegungskoeffizienten bezüglich einer geeignet gewählten L^2-Orthogonalbasis exponentiell abklingen.

9. Wir nutzen das von uns bewiesene Kriterium über die Analytizität einer Funktion auf dem Tetraeder bzw. das 2D Analogon über die Analytizität einer Funktion auf dem Dreieck, um eine auf dem Abklingen der Legendre-Zerlegungskoeffizienten basierende adaptive *hp*-Strategie auf Dreiecks- und Tetraedervernetzungen zu verallgemeinern. Vergleichsrechnungen in 2D mit zwei weiteren adaptiven *hp*-Strategien demonstrieren die Konkurrenzfähigkeit des von uns verallgemeinerten Algorithmus.

Literaturverzeichnis zu den Thesen

[BG86c] I. Babuška and B.Q. Guo. The $h - p$ version of the finite element method. Part 1: The basic approximation results. *Computational Mechanics*, 1:21–41, 1986.

[BG86d] I. Babuška and B.Q. Guo. The $h - p$ version of the finite element method. Part 2: General results and applications. *Computational Mechanics*, 1:203–220, 1986.

[BG88] I. Babuška and B. Guo. The *hp* version of the finite element method for domains with curved boundaries. *SIAM J. Numer. Anal.*, 25:837–861, 1988.

[KM03] B.N. Khoromskij and J.M. Melenk. Boundary concentrated finite element methods. *SIAM J. Numer. Anal.*, 41(1):1–36, 2003.

[KS99] G.E. Karniadakis and S.J. Sherwin. *Spectral/hp Element Methods for CFD.* Oxford University Press, 1999.

[MGS01] J.M. Melenk, K. Gerdes, and C. Schwab. Fully discrete *hp*-FEM: fast quadrature. *Comput. Meth. Appl. Mech. Engrg.*, 190:4339–4364, 2001.

[Osw94] P. Oswald. *Multilevel Finite Element Approximation.* Teubner Skripten zur Numerik. Teubner, 1994.

[TW05] A. Toselli and O. Widlund. *Domain Decomposition Methods — Algorithms and Theory.* Springer Verlag, 2005.

Lebenslauf

Persönliche Daten

Name:	Tino Eibner
Geburtsdatum:	25.04.1978
Geburtsort:	Karl-Marx-Stadt
Familienstand:	ledig

Schulausbildung

09/1984 - 08/1990	Polytechnische Oberschule in Karl-Marx-Stadt
09/1990 - 08/1992	Spezialschule math.-naturw.-techn. Richtung in Chemnitz
09/1992 - 06/1996	Johannes-Kepler-Gymnasium in Chemnitz Abschluss: Abitur

Grundwehrdienst

11/1996 - 08/1997	Grundwehrdienst

Studium

10/1997 - 08/2002	Studium der Technomathematik an der TU Chemnitz Abschluss als Diplom-Technomathematiker

Wissenschaftlicher Werdegang

seit 09/2002	wissenschaftlicher Mitarbeiter im DFG-Sonderforschungsbereich 393 „Parallele Numerische Simulation für Physik und Kontinuumsmechanik“

Erklärung

Ich erkläre an Eides Statt, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Chemnitz, den 23. Januar 2006