Dreiecksbasierte spektrale Mehrgitterverfahren für elliptische Probleme

Diplomarbeit

von

Peter Norek

September 2014

Universität Duisburg-Essen, Campus Essen
Fachbereich Mathematik

Gutachter:
Prof. Dr. Wilhelm Heinrichs
Prof. Dr. Gerhard Starke

Danksagung

Zu Beginn möchte ich mich bei allen Menschen bedanken, die mir bei der Erstellung dieser Arbeit hilfreich zu Seite gestanden haben.

Ausdrücklich danke ich Herrn Prof. Dr. Heinrichs für die Bereitstellung dieses interessanten Themas und die wissenschaftliche Betreuung während der gesamten Arbeit.

Mein besonderer Dank gilt meiner Freundin und meiner Familie, die mein Studium ermöglicht haben und an mich geglaubt haben.

Ich danke ebenfalls Pascal Marquardt, Agnes Schwegmann und Johannes Schwegmann für das interessierte Korrekturlesen.

Zu guter Letzt möchte ich mich bei meinen Freunden, allen voran den ehemaligen und aktiven Mitgliedern der Arbeitsgruppe Ingenieurmathematik, für viele unterhaltsame Stunden in unserem Büro bedanken.

Inhaltsverzeichnis

1 Einleitung

Die vorliegende Diplomarbeit beschäftigt sich mit der Anwendung spektraler Mehrgitterverfahren auf zweidimensionale elliptische Randwertprobleme.

Die Randwertaufgaben werden in eine universellere Variationsformulierung übertragen. Zur Diskretisierung verwenden wir die dreiecksbasierte (engl. *triangular*) Spektrale-Elemente-Methode, kurz TSEM. Diese erreicht bei analytischen Lösungen spektrale Genauigkeit und bietet durch Referenzgebiete in Dreiecksgestalt eine hohe Flexibilität bzgl. des Gebiets, auf dem das Randwertproblem definiert ist. Auch ist eine adaptive Gitterwahl problemlos möglich.

Im Gegensatz zu viereckbasierten Verfahren, werden für TSEM zwei unterschiedliche Knotenmengen benötigt, um die spektrale Genauigkeit zu erhalten [14]. Es sind zum einen Interpolationspunkte, wie beispielsweise die Fekete-Punkte, und zum anderen Quadraturpunkte.

Das aus der Diskretisierung mit TSEM erhaltene lineare Gleichungssystem

$$A\mathbf{u} = \mathbf{b}$$

besteht aus einer dünnbesetzten Matrix A, welche eine Kondition von $\mathcal{O}(N^4)$ bzgl. des Polynomgrades N aufweist [14]. Um dieses System numerisch zu lösen, werden p-Mehrgitterverfahren eingesetzt [5, 13]. Durch Ausnutzung der Glättungseigenschaften klassischer Iterationsverfahren, wie etwa Gauß-Seidel-Iteration, bieten sie gute Konvergenzraten.

Alternativ kommen Schur-Komplement-Verfahren, welche aus der statischen Kondensation hevorgehen, zum Einsatz [15, 21]. Häufig müssen Vorkonditionierer eingesetzt werden.

Wir werden in dieser Arbeit beide Verfahren miteinander verbinden. So wird auf das lineare Gleichungssystem, das aus der Diskretisierung der Variationsaufgabe hervorgeht, die statische Kondensation angewandt. Anstatt jedoch nur das Schur-Komplement zu

betrachten, verwenden wir das umgeformte Gleichungssystem als Ganzes im Kontext der p-Mehrgitterverfahren.

Die Arbeit ist wie folgt aufgebaut:

In **Kapitel 2** wird die Idee der Spektralverfahren erläutert. Außerdem werden die elliptischen Randwertprobleme vorgestellt und in eine Variationsformulierung übertragen.

Kapitel 3 befasst sich mit den auf Dreiecksgebiete zugeschnittenen Komponenten, die für die hohe Konvergenz der dreiecksbasierten Spektrale-Elemente-Methode entscheidend sind.

In **Kapitel 4** beschreiben wir das Galerkin-Verfahren und die Berechnung und Assemblierung der Elementmatrizen basierend auf den Fekete-Punkten.

In **Kapitel 5** präsentieren wir die wichtigsten Aspekte der statischen Kondensation, welche auf das obige Gleichungssystem angewendet werden soll.

Kapitel 6 erörtert das p-Mehrgitterverfahren in Zusammenhang mit der dreiecksbasierten Spektrale-Elemente-Methode.

In **Kapitel 7** wird zunächst das Mehrgitterverfahren bestmöglich auf die statische Kondensation angepasst und anschließend mit den Mehrgitterverfahren ohne statische Kondensation verglichen.

Abschließend folgt im **Kapitel 8** eine Zusammenfassung der Ergebnisse und ein kurzer Ausblick auf mögliche Erweiterungen des Verfahrens.

2 Grundlagen

In vielen Bereichen der Physik werden Probleme und Phänomene, wie etwa Wärme- und Wellenausbreitung, Strömungen und Gravitationspotentiale, durch gewöhnliche und partielle Differentialgleichungen der Form

$$\mathcal{L}u = f \quad \text{in} \quad \Omega \subseteq \mathbb{R}^n \tag{2.1}$$

mit $\mathcal{L}$ als allgemeinen linearen Differentialoperator dargestellt. Speziell für partielle Differentialgleichungen existiert häufig keine analytische Lösung oder sie ist nicht berechenbar.

Die Finite-Differenzen- und Finite-Elemente-Methoden sind zwei etablierte Klassen von numerischen Verfahren, mit denen eine approximierte Lösung der Differentialgleichung ermittelt werden kann. Die Behandlung unregelmässiger Gebiete und die lokale Verfeinerung des Gitters ist ein Vorzug der Finiten-Elemente-Methode, während dies bei Verwendung der Finiten-Differenzen-Methode zu Schwierigkeiten führt. Dagegen ist die Konvergenzgeschwindigkeit der Finiten-Elemente-Methode für niedrige Polynomgrade gering, so dass für eine akzeptable Genauigkeit der Lösung sehr große lineare Gleichungssysteme gelöst werden müssen.

Die in den letzten Jahrzehnten entwickelte Spekralmethode erreichet dagegen *spektrale Konvergenz*, wenn die Lösung analytisch ist. Allerdings ist hierfür die Lösung eines linearen Gleichungssystems mit einer vollbesetzten Matrix nötig. Des Weiteren können unregelmäßige Gebiete nicht betrachtet werden.

Die *Spektrale-Elemente-Methode* vereinigt die Vorzüge der Spektralmethode und der Finiten-Elemente-Methode ohne die oben genannten Nachteile, indem das betrachtete Gebiet in mehrere Elemente zerlegt wird und auf jedem Element ein spektraler Ansatz verfolgt wird.

2.1 Spektrale Verfahren

Ausgehend von Canuto et al. in [2] lassen sich Spektrale Verfahren als *Methoden der gewichteten Residuen* auffassen, um die Differentialgleichung aus (2.1) mit passenden Randbedingungen zu lösen.

Es sei die Bilinearform $(u, v) = \int_\Omega u(x)v(x)w(x)\,dx$ mit einer Gewichtsfunktion w auf $X \times Y$ gegeben, so dass für das Residuum $r(u) = f - \mathcal{L}u$ gelten muss:

$$\text{Finde } u \in X \text{ mit} \quad (r(u), \varphi) = 0 \quad \text{für alle } \varphi \in Y. \tag{2.2}$$

In dieser Arbeit werden polynomiale Ansätze behandelt, indem der Ansatzraum X auf einen endlichdimensionalen Raum X_N beschränkt wird, der von den polynomialen Ansatzfunktionen $\psi_i,\ i = 1, \ldots, I$, aufgespannt wird. Somit ist die diskretisierte Lösung u^N eine Linearkombination der Ansatzfunktionen:

$$u(x) \approx u^N(x) = \sum_{i=1}^{I} \hat{u}_i \psi_i(x)$$

Abhängig von der Wahl des Testraum Y entstehen unterschiedliche Ansätze.

2.1.1 Kollokations-Ansatz

Wir erhalten das Kollokation-Verfahren bzw. Pseudo-Spektrale Verfahren, wenn als Testfunktionen die Dirac-Impulse δ, die jeweils um vorher festgelegte Kollokationspunkte $x_i,\ i = 1, \ldots, I$, verschoben sind, und die konstante Gewichtsfunktion $w \equiv 1$ gewählt werden:

$$\varphi(x) = \delta(x - x_i), \quad 1 \le i \le I$$

Dies bewirkt eine exakte Auswertung der Differentialgleichung an den Stützstellen

$$\mathcal{L}u^N(x_i) = f(x_i) \quad \text{für} \ \ 1 \le i \le I$$

und liefert so die zu lösenden Kollokationsgleichungen:

$$\sum_{j=1}^{I} \hat{u}_j \mathcal{L}\psi_j(x_i) = f(x_i), \quad 1 \le i \le I \tag{2.3}$$

Üblicherweise sind die Ansatzfunktionen die Lagrange-Interpolierenden zu den Kollokationspunkten, d. h. $\psi_i(x_j) = \delta_{ij}$, was dazu führt, dass der j-te Koeffiziente $\hat{u}_j$ der

physikalische Wert der Funktion u^N in den j-ten Kollokationspunkt ist: $\hat{u}_j = u(x_j)$. Die sich ergebenden Gleichungssysteme sind häufig unsymmetrisch.

2.1.2 Galerkin-Ansatz

Wird der Testraum $Y_N \subset Y$ derart gewählt, dass er mit dem diskretisierten Ansatzraum X_N übereinstimmt und dieser die wesentlichen Randbedingungen erfüllt, so erhält man den Galerkin-Ansatz:

$$(\mathcal{L}u^N, v) = (f, v) \quad \text{für alle } v \in X_N \tag{2.4}$$

Da die Ansatzfunktionen $\psi_i, i = 1, \ldots, I$, eine Basis von X_N bilden, ist (2.4) äquivalent zu

$$(\mathcal{L}u^N, \psi_i) = (f, \psi_i) \quad \text{für } 1 \leq i \leq I.$$

Wegen der Linearität von $\mathcal{L}$ folgt damit das zu lösende Gleichungssystem:

$$\sum_{j=1}^{I} \hat{u}_j(\mathcal{L}\psi_j, \psi_i) = (f, \psi_i), \quad 1 \leq i \leq I \tag{2.5}$$

Dieser Ansatz liefert eine *schwache Lösung* der ursprünglichen partiellen Differentialgleichung. Hat der lineare Differentialoperator $\mathcal{L}$ eine gewisse Struktur, ist es möglich die Terme $(\mathcal{L}\psi_j, \psi_i)$ analytisch umzuformen, um ein symmetrisches Gleichungssystem zu erhalten. Dies wird im nächsten Abschnitt vorgestellt.

2.2 Elliptische Randwertprobleme

Als grundlegende Problemstellung in dieser Arbeit werden Randwertaufgaben auf elliptischen partiellen Differentialgleichungen behandelt. Mit dieser Art von partiellen Differentialgleichungen werden typischerweise stationäre physikalische Probleme modelliert, wie sie beispielsweise bei der Berechnung von zeitlich unabhängigen Temperaturgefällen, elektrostatischen Ladungsverteilungen oder Gravitationspotentialen auftreten.

Wir betrachten das folgende elliptische Modell-Randwertproblem auf einem beschränk-

ten Gebiet $\Omega \subset \mathbb{R}^2$ mit Rand $\partial\Omega = \Gamma_{\mathcal{D}} \cup \Gamma_{\mathcal{N}}$:

$$-\Delta u + \lambda u = f \qquad \text{in } \Omega, \tag{2.6a}$$

$$u = g_{\mathcal{D}} \qquad \text{auf } \Gamma_{\mathcal{D}}, \tag{2.6b}$$

$$\frac{\partial u}{\partial \nu} = g_{\mathcal{N}} \qquad \text{auf } \Gamma_{\mathcal{N}}. \tag{2.6c}$$

Für $\lambda = 0$ entspricht (2.6a) der *Poisson-Gleichung* und andernfalls der *Helmholtz-Gleichung*. Die Angabe der Funktionswerte auf dem Randstück $\Gamma_{\mathcal{D}}$ in (2.6b) repräsentiert die *Dirichlet-Randbedingung*, während durch die Ableitung von u nach der äußeren Einheitsnormalen ν auf Randstück $\Gamma_{\mathcal{N}}$ in (2.6c) die *Neumann-Randbedingung* dargestellt wird.

Um die Galerkin-Methode anzuwenden, wird das Problem (2.6) in den Sobolevraum

$$H^1(\Omega) = \left\{ u \in L^2(\Omega) \mid \partial^\alpha u \text{ existiert für alle } \alpha \text{ mit } |\alpha| \leq 1 \text{ und liegt in } L^2(\Omega) \right\}$$

übertragen[1] und mittels der ersten Greenschen Formel verallgemeinert in die dazugehörige Variationsformulierung:

Finde $u \in V_{g_{\mathcal{D}}} = \{v \in H^1(\Omega) \mid v = g_{\mathcal{D}} \text{ auf } \Gamma_{\mathcal{D}}\}$, so dass

$$a(u,v) = F(v) \qquad \text{für alle } \; v \in V_0 = \left\{ v \in H^1(\Omega) \mid v = 0 \text{ auf } \Gamma_{\mathcal{D}} \right\} \tag{2.7a}$$

$$\text{mit } \; a(u,v) = \int_\Omega \nabla u \cdot \nabla v + \lambda uv \, dx \tag{2.7b}$$

$$\text{und } F(v) = \int_\Omega fv \, dx + \int_{\Gamma_{\mathcal{N}}} v g_{\mathcal{N}} \, ds. \tag{2.7c}$$

In diesem Kontext definieren wir die Bilinearform a und die Linearform F bzgl. eines Teilgebiets $M \subseteq \Omega$ durch

$$a_M(u,v) = \int_M \nabla u \cdot \nabla v + \lambda uv \, dx, \tag{2.8}$$

$$F_M(v) = \int_M fv \, dx + \int_{\Gamma_{\mathcal{N}} \cap M} v g_{\mathcal{N}} \, ds. \tag{2.9}$$

Bei inhomogenen Dirichlet-Randbedingung sind der Ansatzraum $V_{g_{\mathcal{D}}}$ und Testraum V_0 verschieden.

Um dies für die theoretische und praktische Betrachtung zu umgehen, wird die Lösung u in einen homogenen Anteil $u^{\mathcal{H}} \in V_0$ und einen inhomogenen Anteil $u^{\mathcal{D}} \in H^1(\Omega)$,

[1] $\partial^\alpha u$ bezeichnet die α-te schwache Ableitung von u

welcher die Dirichlet-Randbedingung (2.6b) erfüllt, aufgespalten, so dass $u = u^{\mathcal{H}} + u^{\mathcal{D}}$ gilt. Dieses Vorgehen wird auch *Anheben der Lösung* (engl. *lifting*) genannt.

Dadurch ist die Variationsformulierung (2.7) äquivalent zu:

$$\text{Finde } u^{\mathcal{H}} \in V_0 \text{ mit} \quad a(u^{\mathcal{H}}, v) = \hat{F}(v) := F(v) - a(u^{\mathcal{D}}, v) \quad \text{für alle } v \in V_0. \tag{2.10}$$

Auf diese Weise erhalten wir die *schwache Lösung* $u = u^{\mathcal{D}} + u^{\mathcal{H}} \in V_{g_{\mathcal{D}}}$.

Notwendig für die Lösbarkeit des Problems ist die Existenz einer Funktion $u^{\mathcal{D}} \in H^1(\Omega)$ mit $u^{\mathcal{D}}(\Gamma_{\mathcal{D}}) = g_{\mathcal{D}}$.

In den folgenden Kapiteln wird im Zuge der Implementierung der Spektralen-Elemente-Methode das Gebiet Ω in dreieckige Teilgebiete T_k zerlegt. Die jeweiligen Anteile der Variationsformulierung a_{T_k} bzw. F_{T_k} auf T_k werden durch Aufspaltung der Linearformen a bzw. F aus (2.10) berechnet und in ein Gesamtgleichungssystem zusammengestellt.

3 Besonderheiten der dreiecksbasierten Spektrale-Elemente-Methode

Die Effizienz der Spektrale-Elemente-Methode hängt wesentlich von der Wahl der Stützstellen auf dem Referenzgebiet ab.

In den klassischen viereckbasierten (engl. *quadrangle based*) Spektrale-Elemente-Methoden, kurz QSEM, können die Tensorprodukte von Gauß-Lobatto-Legendre-Punkten als Interpolations- und Integrationsknoten verwendet werden. Dadurch profitiert die Methode sowohl von den hervorragenden Interpolations- als auch Integrationseigenschaften.

Eine derartige Punktemenge auf Dreiecksgebieten ist nicht bekannt. Daher werden in diesem Kapitel zwei Arten von Punkten vorgestellt und anschließend in TSEM verwendet.

Einerseits die *Fekete*-Punkte, welche fast optimale Interpolationseigenschaften [19] aufweisen, und andererseits zwei Unterarten von Quadraturpunkten, die für die Approximation der Integrale benötigt werden.

Auch bei der Wahl einer orthogonalen Polynomialbasis ist eine einfache Verallgemeinerung durch Anwendung des Tensorprodukts auf eindimensionale Basisfunktionen wie in QSEM nicht möglich. Proriol[1], Koornwinder[2] und Dubiner[3] haben unabhängig voneinander eine passende Basis gefunden, welche im nächsten Abschnitt vorgestellt wird.

Die hier eingeführten Komponenten werden in Kapitel 4 für die Diskretisierung, ebenso wie im Abschnitt 6.3 für die Übergangsoperatoren benötigt.

Wir definieren zunächst das Standarddreieck

$$T_0 = \left\{(r,s) \in \mathbb{R}^2 \mid -1 \leq r,s;\ r+s \leq 0\right\}, \tag{3.1}$$

[1] [Proriol, Joseph. *„Sur une famille de polynomes á deux variables orthogonaux dans un triangle.“* Comptes Rendus Acad. Sci. Paris 245 (1957): 2459-2461.]

[2] [Koornwinder, Tom. *„Two-variable analogues of the classical orthogonal polynomials.“* In *Theory and applications of special functions* (ed. R. Askey). Academic Press, San Diego (1975): 235-495.]

[3] vgl. [6]

welches im Folgenden als Referenzdreieck genutzt wird. Die Dubiner-Polynome und die vorgestellten Punktemengen werden ebenfalls auf T_0 beschrieben.

3.1 Dubiner-Polynome

Um eine *gute* Funktionen-Basis mit günstigen Orthogonalitätseigenschaften und geringer Matrix-Kondition zu erhalten, werden in der viereckbasierten Spektralmethode üblicherweise Tensorprodukte von eindimensionalen Basisfunktionen eingesetzt. Diese Methode ist bei Dreiecksgebieten, also Nicht-Tensor-Gebieten, nicht umsetzbar.

Eine vielversprechende Klasse von orthogonalen Polynomialbasen auf Dreiecksgebieten sind die *Dubiner-Polynome*. Zunächst definieren wir eine Abbildung zwischen dem Standarddreieck T_0 und dem Standardquadrat $Q_0 = \{(r', s') \mid -1 \leq r', s' \leq 1\}$:

$$\Phi : Q_0 \longrightarrow T_0 : \begin{pmatrix} r' \\ s' \end{pmatrix} \longmapsto \begin{pmatrix} r \\ s \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(1+r')(1-s') - 1 \\ s' \end{pmatrix} \tag{3.2}$$

mit der inversen Zuordnung[4]:

$$\Phi^{-1} : T_0 \longrightarrow Q_0 : \begin{pmatrix} r \\ s \end{pmatrix} \longmapsto \begin{pmatrix} r' \\ s' \end{pmatrix} = \begin{pmatrix} 2\frac{1+r}{1-s} - 1 \\ s \end{pmatrix} \tag{3.3}$$

Die Funktion Φ hat die Jacobi-Determinante $\det\left(D_\Phi(r', s')\right) = \frac{1-s'}{2}$ und verursacht ein Kollabieren der oberen Kante von Q_0 auf den singulären Eckpunkt $(-1, 1) \in T_0$, wie sich der Abbildung 3.1 entnehmen lässt.

Definition 3.1. *Die* Dubiner-Polynome *bzw.* Proriol-Koornwinder-Dubiner-Polynome *sind gegeben durch*

$$g_{mn}(r, s) = P_m\left(2\frac{1+r}{1-s} - 1\right)\left(\frac{1-s}{2}\right)^m P_n^{(2m+1,0)}(s), \qquad m, n \in \mathbb{N}_0, \tag{3.4}$$

wobei $P_n^{(\alpha,\beta)}(x)$ das n-te Jacobi-Polynom bezüglich der Gewichtsfunktion $(1-x)^\alpha(1+x)^\beta$ und $P_m(x) = P_m^{(0,0)}(x)$ das m-te Legendre-Polynom bezeichnet.

Durch den Faktor $(\frac{1-s}{2})^m$ in (3.4) wird der Nenner des aus den Legendre-Polynoms resultierenden Ausdrucks beseitigt, welcher durch die Transformation (3.3) entstanden ist. Die Indizes m und n geben den maximalen Polynomgrad in r und s an. Infolgedessen ist g_{mn} ein Polynom in r, s vom Totalgrad $m + n$.

[4]Die Zuordnung bildet $T_0 \setminus \{(-1, 1)\}$ bijektiv auf $Q_0 \setminus \{[-1, 1] \times \{1\}\}$ ab.

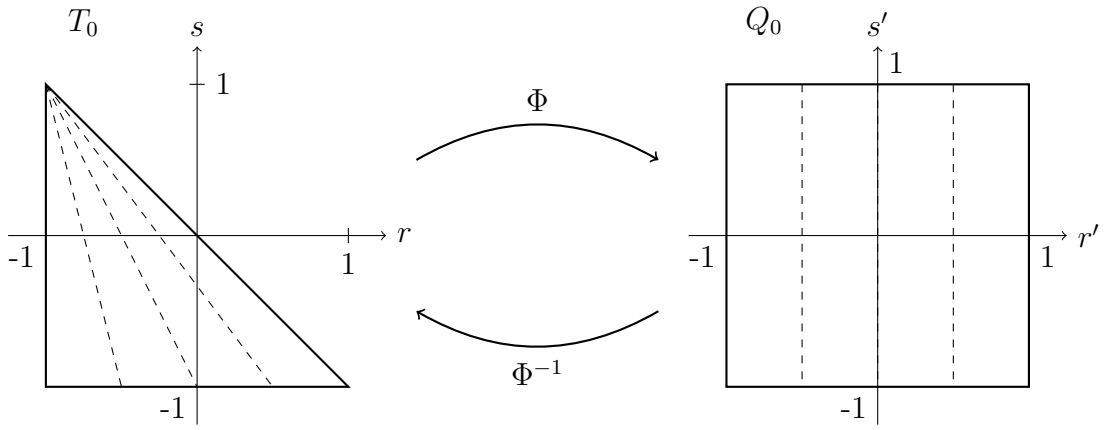


Abbildung 3.1: Darstellung der Funktion $\Phi : T_0 \longrightarrow Q_0$

Die Orthogonalität der Dubiner-Polynome bezüglich des L^2-Skalarproduktes auf T_0 folgt unter Ausnutzung des Transformationssatzes und der Orthogonalität der Jacobi- und Legendrepolynome aus:

$$(g_{mn}\,, g_{kl}) = \int_{T_0} g_{mn}(r,s) g_{kl}(r,s)\, dr\, ds \tag{3.5a}$$

$$= \int_{-1}^{1} L_m(r') L_k(r')\, dr' \cdot$$

$$\int_{-1}^{1} \left(\frac{1-s'}{2}\right)^{m+k} P_n^{(2m+1,0)}(s') P_l^{(2k+1,0)}(s') \frac{1-s'}{2}\, ds' \tag{3.5b}$$

$$= \frac{2}{(2m+1)\,(n+m+1)}\, \delta_{mk}\delta_{nl} \tag{3.5c}$$

Mit Hilfe dieser Polynome sind wir in der Lage, für einen festen Polynomgrad $N \in \mathbb{N}_0$ eine orthogonale Basis von

$$\mathcal{P}_N := \mathcal{P}_N(T_0) = \{p : T_0 \to \mathbb{R} \;\mid\; p \text{ ist Polynom, } \deg(p) \leq N\}\,, \tag{3.6}$$

also den Raum aller Polynome vom Totalgrad $\leq N$ zu bilden. $\mathcal{P}_N$ hat die Dimension $d := d_N = \frac{1}{2}(N+1)(N+2)$ und wird von der *Dubiner-Basis*

$$\mathcal{B}_N = \{g_{mn} \mid 0 \leq m, n;\, m+n \leq N\} \subset \mathcal{P}_N \tag{3.7}$$

aufgespannt.

Hinsichtlich einer einfacheren Indizierung der Basisfunktionen g_{mn} in $\mathcal{B}_N$ führen wir für festes $N \in \mathbb{N}_0$ folgende Bijektion ein:

$$\iota : \{(m,n) \mid 0 \le m,n \le N\} \longrightarrow \{1 \le k \le d_N\} \tag{3.8a}$$

$$(m,n) \longmapsto k = \frac{1}{2}(m+n+1)(m+n+2) - n \tag{3.8b}$$

Mit der Notation $g_k(r,s) = g_{mn}(r,s)$, wobei $k = \iota(m,n)$, erhalten wir schließlich die orthogonale Polynomialbasis bestehend aus den Dubiner-Polynomen vom Grad $\le N$:

$$\mathcal{B}_N = \{g_k \mid 0 \le k \le d_N\} \subset \mathcal{P}_N \tag{3.9}$$

Ein Vorteil der Indizierung der Dubiner-Polynome durch (3.8) ist die Sortierung der Basisfunktionen nach dem Polynomgrad, d. h.

$$\deg g_k \le \deg g_l \quad \text{für} \quad k < l.$$

Somit erhalten wir eine nach Polynomgrad geordnete hierarchische Basis.

3.2 Fekete-Punkte

Im weiteren Verlauf unseres Verfahrens benötigen wir eine Punktemenge, welche gute Interpolationseigenschaften bei Verwendung einer Polynomialbasis aufweist. Dazu betrachten wir das Interpolationsproblem auf dem Standard-Gebiet T_0 aus (3.1).

Sei $N \in \mathbb{N}$ und $\mathcal{P}_N$ der Raum aller Polynome vom Totalgrad $\le N$ aus (3.6) mit Dimension $d := d_N = \frac{1}{2}(N+1)(N+2)$. Ferner sei $\{\varphi_1, \ldots, \varphi_d\}$ eine Basis von $\mathcal{P}_N$. Wir suchen nun eine Punktemenge

$$\Pi = \left\{\xi^i : \ i = 1, \ldots, n\right\} \subseteq T_0,$$

für die der Interpolationsoperator

$$\mathcal{I}_N : C^0(T_0) \to \mathcal{P}_N, \quad \mathcal{I}_N f(\xi^i) = f(\xi^i), \quad i = 1, \ldots, d$$

wohl-definiert ist, d. h. die Menge Π *unisolvent* ist, und gute Näherungen liefert. Das

bedeutet insbesondere, dass die sogenannte *Lebesgue-Konstante*

$$\Lambda_N = \|\mathcal{I}_N\|_\infty = \max_{\|f\|_\infty=1} \|\mathcal{I}_N f\|_\infty$$

möglichst klein ist. Denn es lässt sich zeigen [11, § 3.3.1], dass für alle $f \in C^0(T_0)$ gilt:

$$\|f - \mathcal{I}_N f\|_\infty \leq (1 + \Lambda_N) \inf_{p \in \mathcal{P}_N} \|f - p\|_\infty$$

Somit ist die Lebesgue-Konstante ein Maß für die Abweichung der Interpolierenden $\mathcal{I}_N f$ zur bestmöglichen Approximation in der Supremumsnorm. Bei Verwendung von äquidistanten Punkten für Π wächst die Lebesgue-Konstante exponentiell [10], was sich durch das *Runge-Phänomen* beobachten lässt.

Dahingegen scheint die Bestimmung von sogenannten *Lebesgue-Punkten*, die durch eine minimale Lebesgue-Konstante charakterisiert sind, kaum durchführbar.

Als Alternative bieten sich die *Fekete-Punkte* an, die einfacher berechenbar sind.

Zunächst definieren wir die *(verallgemeinerte) Vandermonde-Matrix*

$$V(\xi^1, \ldots, \xi^d) = \left(\varphi_j(\xi^i)\right)_{i,j} \in \mathbb{R}^{(d,d)} \tag{3.10}$$

zu der Punktemenge $\left\{\xi^1, \ldots, \xi^d\right\}$ und einer vorgegebenen Basis $\{\varphi_1, \ldots, \varphi_d\}$ von $\mathcal{P}_N$.

Definition 3.2. *Die Punktemenge* $\Pi = \left\{\xi^1, \ldots, \xi^d\right\}$ *heißt* Fekete-Punkte *auf* T_0 *zum Polynomgrad* N*, falls durch sie für eine feste Basis* $\{\varphi_i\}_{i=1,\ldots,d}$ *von* $\mathcal{P}_N$ *die Determinante der verallgemeinerten Vandermonde-Matrix maximiert wird, d. h.*

$$\left|V(\xi^1, \ldots, \xi^d)\right| = \max_{\zeta^i \in T_0} \left|V(\zeta^1, \ldots, \zeta^d)\right|.$$

Diese Punkte sind unabhängig von der gewählten Basis von $\mathcal{P}_N$, da eine Wechsel der Basis die Determinante nur um eine von den Punkten unabhängige Konstante verändern wird [11, § 3.3.4].

Seien die Lagrangeschen Interpolierenden der Fekete-Punkte gegeben durch $L_j \in \mathcal{P}_N,\ L_j(\xi^i) = \delta_{ij}$ für $1 \leq i, j \leq d$, so lässt sich nachrechnen [19]

$$L_j(\xi) = \frac{V(\xi^1, \ldots, \xi^{j-1}, \xi, \xi^{j+1}, \ldots, \xi^d)}{V(\xi^1, \ldots, \xi^d)}, \tag{3.11}$$

$$\Lambda_N = \max_{\xi \in T_0} \sum_{j=1}^{d} |L_j(\xi)|. \tag{3.12}$$

N	6	9	12	15	18
Λ_N	4,17	6,80	9,60	9,97	13,5

Tabelle 3.1: Verlauf der Lebesgue-Konstante Λ_N der von Taylor et al. [19] berechneten Fekete-Punkte in Abhängigkeit vom Grad N

Wegen Gleichung (3.11) und der Definition 3.2 gilt $|L_j(\xi)| \leq 1$ für alle $\xi \in T_0$ und so erhalten wir mit (3.12) eine akzeptable Abschätzung für die Lebesgue-Konstante zu den Fekete-Punkten:

$$\Lambda_N = \max_{\xi \in T_0} \sum_{j=1}^{d} |L_j(\xi)| \leq N.$$

Die tatsächlichen Lebesgue-Konstanten der in dieser Arbeit verwendeten Fekete-Punkte sind in Tabelle 3.1 dargestellt.

Da die Fekete-Punkte sowohl auf dem Intervall, als auch auf Tensorprodukt-Gebieten (wie beispielsweise Quadrate oder Würfel) die (Tensorprodukte der) Gauß-Lobatto-Punkte sind[5], stellen die Fekete-Punkte eine mögliche Verallgemeinerung der Gauß-Lobatto-Punkte dar.

Demnach können die Fekete-Punkte, wie in [19] nahegelegt, als Stützstellen für die numerische Quadratur einer Funktion $f \in C^0(T_0)$ auf T_0 aufgefasst werden:

$$\int_{T_0} f(\xi) d\xi \approx \sum_{i=1}^{d} f(\xi^i) w_i \tag{3.13}$$

Die Gewichte w_i lassen sich aus dem folgenden, in w_i linearen, Gleichungssystem berechnen, das sich direkt aus der Orthogonalität der Dubiner-Polynome (3.5) und bei Verwendung der Notation (3.9) ergibt:

$$\sum_{j=1}^{d} w_j g_k(\xi^j) \stackrel{!}{=} \int_{T_0} g_k(\xi)\, d\xi = \int_{T_0} g_k(\xi) g_1(\xi)\, d\xi = 2\delta_{1k}, \quad 1 \leq k \leq d \tag{3.14}$$

In [14] wurde jedoch gezeigt, dass die Quadraturformel (3.13) lediglich exakt für $f \in \mathcal{P}_N$ ist und keine spektrale Konvergenz der TSEM liefert.

Bei der von uns verwendeten dreiecksbasierten SEM greifen wir daher lediglich zur Interpolation auf die Fekete-Punkte zurück, welche Taylor et al. [19] näherungsweise durch lokale Maximierung der verallgemeinerten Vandermonde-Matrix mittels Gradien-

[5] [L. Bos, M. A. Taylor und B. A. Wingate. *„Tensor product Gauss-Lobatto points are the Fekete points for the cube'* Math. Comp. **70** (2001), S. 1543-1547]

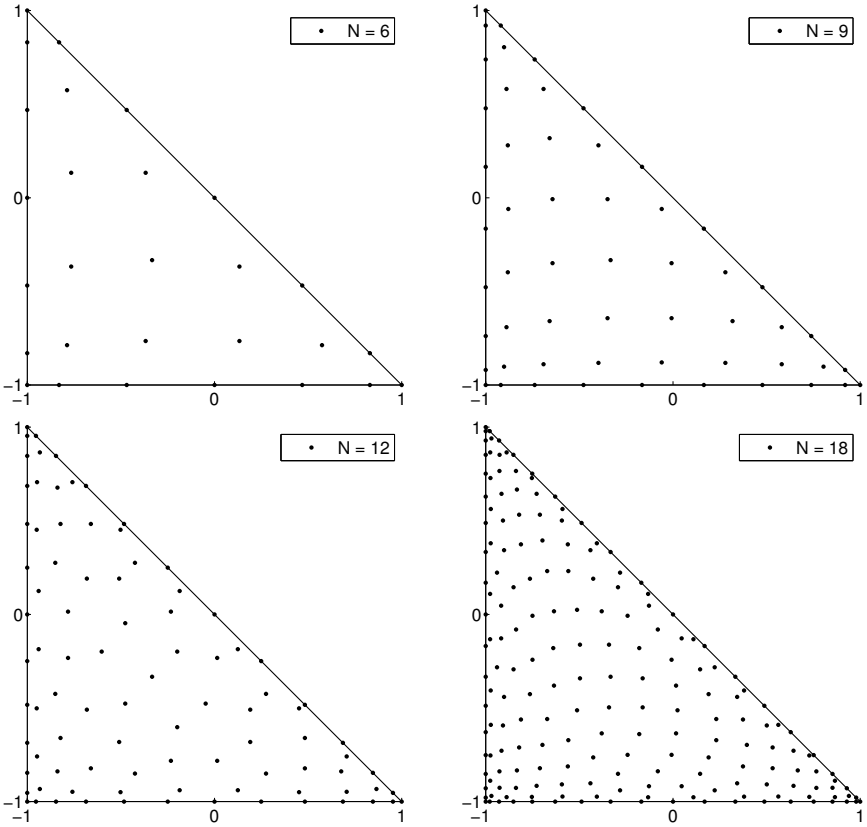


Abbildung 3.2: Die Fekete-Punkte auf T_0 für die Polynomgrade $N = 6, 9, 12$ und 18

tenverfahren bestimmt haben.

In Abbildung 3.2 sind die so bestimmten Fekete-Punkte auf dem Standard-Dreieck T_0 beispielhaft für $N = 6, 9, 12$ und 18 dargestellt.

Entlang jeder Kante liegen genau $N + 1$ Punkte nach der Legendre-Gauß-Lobatto-Verteilung. Im Inneren von T_0 befinden sich die übrigen $\tilde{d} = \frac{1}{2}(N-1)(N-2)$ Punkte.

3.3 Quadraturformeln auf Dreiecken

In der Spektralmethode wie auch in der Finite-Elemente-Methode ist es üblicherweise nötig, mehrdimensionale Integrale, die bei der Erstellung des linearen Gleichungssystem auftreten, analytisch oder numerisch auszuwerten. Insbesondere in den verwendeten Galerkin-Verfahren treten zweidimensionale Integrale auf dreieckigen Teilgebieten T_k des Grundgebiets Ω auf.

Da die Integranden häufig keine trivialen Ausdrücke darstellen, ist eine analytische Auswertung schwierig. Stattdessen werden als praktikable Alternativen die folgenden

numerischen Quadraturformeln auf Dreiecken vorgestellt.

3.3.1 Kollabierte Gauß-Quadratur

Ausgehend von der eindimensionalen Gauß-Quadraturformel

$$\int_{-1}^{1} \omega^{\alpha}(x) u(x)\, dx \approx \sum_{i=1}^{M} w_i^{\alpha} u(x_i^{\alpha}) \tag{3.15}$$

mit Gewichtsfunktion $\omega^{\alpha}(x) = (1-x)^{\alpha}$ für ein $M \in \mathbb{N}$ und $\alpha > -1$, welche exakt ist für alle Polynome u vom Grad $\leq 2M-1$, ist es notwendig das Integrationsgebiet vom Standard-Dreieck T_0 auf das Standard-Quadrat Q_0 zu transformieren. Durch die Abbildung Φ aus (3.2) mit der Jacobi-Determinante $\det\left(D_{\Phi}(r',s')\right) = \frac{1-s'}{2}$ und unter Einbeziehung des Transformationssatzes entsteht die Quadraturformel:

$$\int_{T_0} u(r,s)\, dr\, ds = \int_{-1}^{1}\int_{-1}^{1} u(r',s')\frac{1-s'}{2}\, ds'\, dr' \approx \sum_{i=1}^{M}\sum_{j=1}^{M} w_{ij}\, u(\eta_{ij}) \tag{3.16a}$$

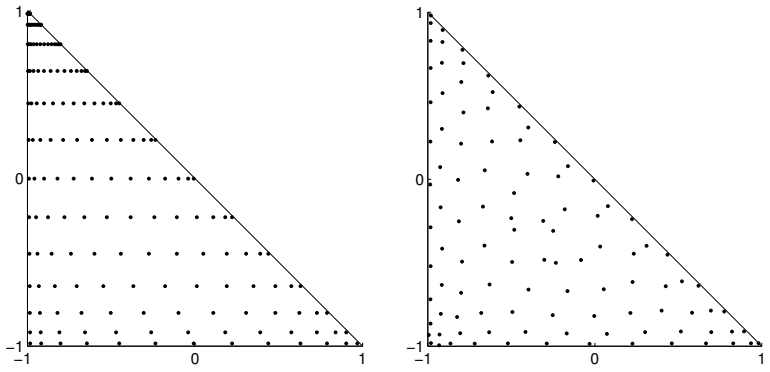
$$\text{mit} \quad \eta_{ij} = (x_i^0, x_j^1) \in Q_0 \quad \text{und} \quad w_{ij} = \frac{1}{2} w_i^0 w_j^1, \quad 1 \leq i,j \leq M \tag{3.16b}$$

Die Quadraturgewichte w_{ij} und Quadraturpunkte η_{ij} ergeben sich durch ein Tensorprodukt aus der Formel (3.15) für $\alpha = 0$ bei Integration nach r' und $\alpha = 1$ bei Integration nach s'. Somit ist die Formel (3.16) in jeder Varibale exakt für alle Polynome vom Grad $\leq 2M-1$ und insbesondere für alle Polynome aus $\mathcal{P}_{2M-1}$.

Ein Nachteil dieser Formel liegt in der unnötigen Häufung von Quadraturpunkten nahe des singulären Eckpunkts $(-1,1)$, was sich etwa in Abbildung 3.3a beobachten lässt.

Außerdem hat der zu dieser Quadraturformel passende Galerkin-Ansatzraum $\mathcal{P}_M(T_0)$ auf T_0 die Dimension $\dim \mathcal{P}_M(T_0) = \frac{1}{2}(M+1)(M+2)$, so dass er durch etwa $\frac{1}{2}M^2$ unisolvente Punkte beschrieben wird. Allerdings werden etwa die doppelte Anzahl an Quadraturpunkten benötigt.

Die Verteilung der kollabierten Gauß-Punkte ist beispielsweise für $M = 12$, wodurch Polynome vom Grad ≤ 23 exakt integriert werden, in Abbildung 3.3 dargestellt.



(a) Kollabierte Gauß-Punkte (b) Taylor-Wingate-Bos-Punkte

Abbildung 3.3: Verteilung der Quadraturpunkte zum exakten Integrieren in $\mathcal{P}_{23}$

3.3.2 Quadratur nach Taylor-Wingate-Bos

Die oben genannten Nachteile der "kollabierten Gauss-Quadratur" haben früh dazu geführt nach besseren Quadraturpunkten zu suchen. Bereits 1971 hat Stroud[6] eine Übersicht unterschiedlicher Quadraturformeln auf diversen Gebieten zusammengetragen, welche von Cools in [3, 4] fortgeführt wurde.

In dieser Diplomarbeit wird von den 2007 in [20] ermittelten Werten Gebrauch gemacht. Taylor et al. verwenden einen sogenannten *Kardinal-Funktion-Algorithmus*, der effizient Quadraturpunkte und -gewichte auf beliebigen Gebieten berechnet. Dieser liefert besonders für höhere Polynomgrade Formeln mit den wenigsten Knotenpunkten unter den bis dahin bekannten Formeln auf dem Dreieck und wird daher kurz vorgestellt.

Für $N \in \mathbb{N}$ und einer orthogonalen Basis $\mathcal{B}_N$ des Polynomialraums $\mathcal{P}_N = \mathcal{P}_N(T_0)$, beispielsweise der Dubiner-Basis aus (3.9), wird mit einer unisolventen Menge $\boldsymbol{z} = \{z_1, \ldots, z_d\}$ bestehend aus $d = \dim \mathcal{P}_N = \frac{(N+1)(N+2)}{2}$ Punkten in T_0 gestartet. Das Ziel ist die Exaktheit der Quadraturformel in $\mathcal{P}_{N+E}$ für ein möglichst hohes $E \in \mathbb{N}$:

$$F_m(\boldsymbol{z}) = \sum_{i=1}^{d} w_i g_m(z_i) \stackrel{!}{=} \int_{\mathcal{T}_0} g_m \, d\xi = 2\delta_{1m} \quad \text{für} \quad g_m \in \mathcal{B}_{N+E} \subset \mathcal{P}_{N+E} \tag{3.17}$$

Entsprechend des Gleichungssystems (3.14) werden die von $\boldsymbol{z}$ abhängigen Gewichte $w(\boldsymbol{z}) = \{w_1, \ldots, w_d\}$ stets derart berechnet, dass die Formel exakt für alle $f \in \mathcal{P}_N$ ist. Somit muss statt (3.17) nur die Nullstelle der Funktion $F(\boldsymbol{z}) = (F_m(\boldsymbol{z}) \;:\; N <$

[6] [A. H. Stroud. *„Approximate calculation of multiple integrals."* Englewood Chiffs, New Jersey: Prentice Hall (1971)]

$\deg g_m \leq N + E$) berechnet werden. Dies geschieht iterativ durch das Newton-Verfahren

$$\boldsymbol{z}^{(k+1)} = \boldsymbol{z}^{(k)} - \left(D_F\left(\boldsymbol{z}^{(k)}\right)\right)^{-1} F(\boldsymbol{z}^{(k)}),$$

wobei die Inverse der Jacobi-Matrix $D_F\left(\boldsymbol{z}^{(k)}\right)$ durch die Pseudoinverse berechnet wird. Um eine Unterbestimmtheit des Systems zu erreichen, wird der Zusatzgrad E derart gewählt, dass $\dim \mathcal{P}_{N+E} \leq 2 \dim \mathcal{P}_N$ gilt.

Wesentlich für die Effizienz des Algorithmus ist die Auswertung der Kardinalfunktionen $L(\xi; \boldsymbol{z})$, d. h. der Lagrange-Interpolierenden zu $\boldsymbol{z}$. Durch Aufstellen der Identitäten

$$\frac{\partial L_i}{\partial z_j}(\xi; \boldsymbol{z}) = -L_j(\xi; \boldsymbol{z}) \frac{\partial L_i}{\partial \xi}(z_j; \boldsymbol{z})$$
$$\text{und} \qquad \frac{\partial w_i}{\partial z_j} = -w_j \frac{\partial L_i}{\partial \xi}(z_j; \boldsymbol{z})$$

wird die Berechnung der auftretenden partiellen Ableitungen im Newton-Verfahren verbessert.

Auf diese Weise wurden für $N = 1, \ldots, 14$ die Quadraturformeln der Form

$$\int_{\mathcal{T}_0} u(r, s)\, dr\, ds \approx \sum_{i=1}^{m} w_i\, u(z_i) \tag{3.18}$$

mit $d = \dim \mathcal{P}_N$ berechnet. Die erhaltenen Quadraturpunkte z_i liegen alle im Gebiet T_0 und alle Gewichte w_i sind positiv.

Die Verteilung der Punkte ist beispielsweise für $N = 13$ und $E = 10$ in Abbildung 3.3 dargestellt. Im Vergleich zur “kollabierten Gauß-Quadratur“ wird die verringerte Knotenanzahl ab Exaktheitsgrad 7 in Tabelle 3.2 ersichtlich.

Exaktheitsgrad	kollabierte Gauß-Formel	Taylor-Wingate-Bos-Formel	
	Anzahl Punkte M^2	Zusatzgrad E	Anzahl Punkte $\frac{1}{2}(N+1)(N+2)$
3	4	-	-
4	-	2	6
5	9	2	10
7	16	3	15
9	25	4	21
11	36	5	28
13	49	6	36
17	81	-	-
18	-	8	66
21	121	9	91
23	144	10	105
25	169	11	120
29	225	-	-
35	324	-	-

Tabelle 3.2: Gegenüberstellung der Punktanzahl in den vorgestellten Quadraturformeln in Abhängigkeit vom Polynomgrad der exakten Integration, hier kurz „Exaktheitsgrad“ (da ist $2M-1$ bzw. $N+E$) genannt

4 Diskretisierung

In diesem Kapitel wird eine Diskretisierung der Variationsformulierung (2.10), die aus den elliptischen Randwertproblem (2.6) hevorgegangen ist, durch TSEM vorgestellt. Wir erhalten ein lineares Gleichungssystem $A\mathbf{u} = \mathbf{b}$.

Es wird ein *nodaler* Ansatz verfolgt, d. h. die Koeffizienten des Lösungsvektors $\mathbf{u}$ entsprechen den physikalischen Wert der berechneten Funktion in den Knotenpunkt (x_i, y_i): $u_i = u(x_i, y_i)$.

Ausgehend von einer Triangulierung des Gebiets Ω betrachten wir die Transformation des Referenzgebiets T_0 auf eine beliebiges Dreieck, anschließend die Berechnung der Steifigkeits- und Massenelementmatrizen, sowie die Assemblierung zum endgültigen linearen Gleichungssystem.

Wir orientieren uns an der in [11] verwendeten Notation. Im Folgenden gehen wir von folgenden Voraussetzungen aus:

T_0	Das Standard-Gebiet $\{(r, s) \mid -1 \leq r, s\ ; r + s \leq 1\}$;
N	Polynomgrad auf einem Dreiecksgebiet;
$d = d_N$	Anzahl der Fekete-Punkte auf einem Gebiet,
	Dimension der Dubiner-Basis $\mathcal{B}_N$;
M	Anzahl der Dreiecks-Elemente im Gebiet Ω.

4.1 Triangulierung und Nummerierung der Knoten

Um das elliptische Randwertproblem (2.6) bzw. die dazugehörige Variationsformulierung (2.10) numerisch behandeln zu können, wird das beschränkte, polygonale[1] Gebiet Ω in abgeschlossene Dreiecke $T_k, 1 \leq k \leq M$, zerlegt. Wir arbeiten ausschließlich mit *konformen Triangulierungen* $\mathcal{T}_h = \{T_1, \ldots, T_M\}$, das bedeutet:

- $\bar{\Omega} = \bigcup_{T_k \in \mathcal{T}_h} T_k$

[1] Der Rand von Ω wird durch ein Polygon gebildet.

- Für alle $T_k, T_l \in \mathcal{T}_h$ mit $k \neq l$ gilt:
  $T_k \cap T_l \neq \emptyset$, dann $T_k \cap T_l$ entweder ein Eckpunkt beider Elemente oder eine gemeinsame Kante.

Da im späteren Verlauf dieser Arbeit das p-Mehrgitterverfahren auf den im Allgemeinen unstrukturierten Gittern, die durch Abbildung der Fekete-Punkte auf die Elemente T_k entstanden sind, operiert, muss eine unkomplizierte *globale Nummerierung* gefunden werden, die für verschiedene Polynomgrade jeden unterschiedlichen Knotenpunkt einen eindeutigen Index zuweist.

Ausgehend von obiger Triangulierung seien M_V bzw. M_E die Anzahl aller verschiedenen Eckpunkte bzw. verschiedenen Elementkanten[2], wovon $M_{V_\mathcal{D}}$ und $M_{E_\mathcal{D}}$ die Anzahl der Eckpunkte und Kanten angibt, die auf dem Randstück $\Gamma_\mathcal{D}$ mit Dirichlet-Randbedingung (2.6b) liegen. Die Eckpunkte und Kanten werden jeweils eindeutig und unabhängig vom Polynomgrad N durchnummeriert, so dass jeweils stets diejenigen aus dem Dirichlet-Randstück zuletzt $\Gamma_\mathcal{D}$ einsortiert wurden.

Die Menge aller verschiedenen Knotenpunkte, die sich abhängig von Polynomgrad N durch Transformation der Fekete-Punkte auf die Elemente T_k für $1 \leq k \leq M$ ergeben, wird mit $\Omega_N \subset \Omega$ bezeichnet. Dabei werden Punkte, die zwar aus verschiedenen Elementen stammen, aber physikalisch an gleicher Stelle liegen, zu einem Punkt zusammengefasst, anders als es beispielsweise bei der *Diskontinuierlichen Galerkin-Methode* der Fall ist.

Es folgt sukzessive die globale Nummerierung aller Knotenpunkte aus Ω_N:

i) Erst die $\frac{1}{2}(N-1)(N-2)$ Fekete-Punkte im inneren jedes Elements T_k, $k = 1, \ldots, M$, entsprechend ihrer lokalen Nummerierung;

ii) die $M_V - M_{V_\mathcal{D}}$ Eckpunkte im Inneren von Ω oder auf $\Gamma_\mathcal{N}$ entsprechend ihrer Nummerierung;

iii) nacheinander die $N-1$ Punkte (ohne Eckpunkte) auf Kante Nr. i, welche im Inneren von Ω oder auf $\Gamma_\mathcal{N}$ liegt, für $i = 1, \ldots, M_E - M_{E_\mathcal{D}}$, angefangen vom Eckpunkt mit dem niedrigeren Index zum größeren hin;

iv) die übrigen $M_{V_\mathcal{D}}$ Eckpunkte (auf $\Gamma_\mathcal{D}$) entsprechend ihrer Nummerierung;

v) zuletzt die Punkte auf den übrigen $M_{E_\mathcal{D}}$ Kanten (auf $\Gamma_\mathcal{D}$) entsprechend Eintrag ii) sortiert;

[2]Hierbei steht V für engl. *vertex* und E für engl. *edge*.

i	1	2	3	4	5	6	7	8	9	10
$\mathrm{map}_1^3(i)$	1	6	5	8	11	12	4	3	10	9
$\mathrm{map}_2^3(i)$	2	8	7	6	14	13	3	4	16	15

Tabelle 4.1: Die Mapping-Vektoren der beiden Elemente T_1 und T_2 zur Triangulierung aus Abbildung 4.1.

Auf diese Weise stehen die inneren Elementpunkte blockweise vor den sonstigen Knotenpunkten auf den Elementrändern, wobei zuletzt alle Punkte auf dem Dirichlet-Randstück $\Gamma_{\mathcal{D}}$ folgen. Insgesamt beträgt die *Anzahl aller Freiheitsgrade*

$$n = |\,\Omega_N| = M_V + (N-1)M_E + \frac{(N-1)(N-2)}{2}M. \tag{4.1}$$

Die Anzahl aller Punkte auf $\Gamma_{\mathcal{D}}$ lautet $n_{\mathcal{D}} = M_{V_{\mathcal{D}}} + (N-1)M_{E_{\mathcal{D}}}$, so dass die *Anzahl der inneren Freiheitsgrade* $\tilde{n} = n - n_{\mathcal{D}}$ entspricht. Damit ist die Menge der Knotenpunkte derart durchnummeriert, dass

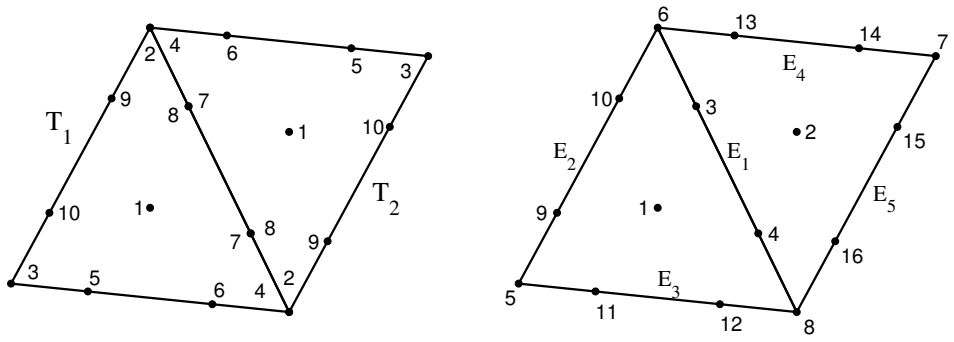
$$\Omega^N = \{\underbrace{(x_1, y_1), \ldots, (x_{\tilde{n}}, y_{\tilde{n}})}_{\in \Omega \setminus \Gamma_{\mathcal{D}}}, \underbrace{(x_{\tilde{n}+1}, y_{\tilde{n}+1}), \ldots, (x_n, y_n)}_{\in \Gamma_{\mathcal{D}}}\} \tag{4.2}$$

gilt.

Die Fekete-Punkte auf dem Standard-Dreieck T_0 sind ähnlich durchnummeriert: Zunächst die Punkte im Inneren, anschließend folgen die Eckpunkte gegen den Uhrzeigersinn, und zuletzt die Kantenpunkte gegen den Uhrzeigersinn, angefangen auf der gegenüberliegenden Kante des ersten Eckpunktes $\bar{P}_1$.

In der Praxis wird für jedes Element $T_k \in \mathcal{T}_h$ und jeden verwendeten Polynomgrad $N \in \mathbb{N}$ ein Mapping-Vektor $\mathrm{map}_k^N \in \mathbb{N}^d$ mitgespeichert, der die globalen Nummern der der lokalen Knoten enthält.

Beispielhaft illustriert Abbildung 4.1 und Tabelle 4.1 die obige lokale und globale Nummerierung den Vektor map_k^N für eine Triangulierung bestehend aus zwei Elementen und Polynomgrad $N = 3$.



(a) lokale Nummerierung der Elemente T_k

(b) globale Nummerierung und Angabe der Kanten E_i

Abbildung 4.1: Darstellung der der (a) lokalen und (b) globalen Nummerierungen für Fekete-Polynomgrad $N = 3$ auf einer Gebiets-Triangulierung bestehend aus den Dreiecken T_1 und T_2.

4.2 Bestimmung der Elementmatrizen auf dem Standard-Element T_0

Ausgehend von einem festen Polynomgrad $N \in \mathbb{N}$ sowie $d = \dim \mathcal{P}_N = \frac{1}{2}(N+1)(N+2)$ definieren wir in Analogie zu Abschnitt 3.2 die quadratische bzw. rechteckige Matrix

$$V_\xi = V(\xi_1, \ldots, \xi_d) = (g_j(\xi_i))_{ij} \in \mathbb{R}^{d \times d} \tag{4.3}$$

$$V_\eta = V(\eta_1, \ldots, \eta_m) = (g_j(\eta_i))_{ij} \in \mathbb{R}^{m \times d} \tag{4.4}$$

als die *verallgemeinerte Vandermonde-Matrix* auf dem Standard-Dreieck T_0 zu den Fekete-Punkten $\xi_i,\ 1 \le i \le d$, aus Abschnitt 3.2 bzw. zu den Quadratur-Punkten $\eta_j,\ 1 \le j \le m$, aus Abschnitt 3.3, welche passend zum Polynomgrad gewählt wurden. Ferner seien

$$L_i \in \mathcal{P}_N(T_0), \quad L_i(\xi_j) = \delta_{ij} \quad \text{für} \ \ 1 \le i, j \le d, \tag{4.5}$$

die Lagrangeschen Interpolationspolynome vom Grad $\le N$ zu den Fekete-Punkten.

Sei $v \in \mathcal{P}_N(T_0)$ die Projektion einer Funktion in den Polynomialraum $\mathcal{P}_N$. v lässt sich sowohl als Linearkombination der Lagrange-Polynome L_i wie auch der Dubiner-Basis $\mathcal{B}_N$ (vgl. (3.9)) darstellen:

$$v(r,s) = \sum_{i=1}^{d} v_i L_i(r,s) = \sum_{k=1}^{d} \hat{v}_k g_k(r,s) \tag{4.6}$$

Für die Koeffizienten von v bzgl. der beiden Polynomial-Basen gilt[3]:

$$\mathbf{v} = \mathbf{v}^{\xi} = (v(\xi_i))_i \quad \text{und} \quad \hat{\mathbf{v}} = V_{\xi}^{-1}\mathbf{v} \tag{4.7}$$

Zur Bestimmung der Elementmatrizen mittels der Quadraturformeln aus Abschnitt 3.3 muss die Funktion in den Quadratur-Punkten η_j ausgewertet werden. Formel (4.7) zusammen mit der analogen Beziehung $V_{\eta}\hat{\mathbf{v}} = \mathbf{v}^{\eta}$ liefern hierfür zunächst die lineare Transformation

$$\mathbf{v}^{\eta} = (v(\eta_i))_i = V_{\eta}V_{\xi}^{-1}\mathbf{v}^{\xi} \ \in \mathbb{R}^m. \tag{4.8}$$

Nun sind wir in der Lage, das L^2-Skalarprodukt auf T_0 zu diskretisieren und in Abhängigkeit von den Funktionswerten in den Fekete-Punkten darzustellen. Bei Verwendung der Quadraturformeln (3.16) oder (3.18) derart, dass sie Polynome bis zum Grad $2N$ exakt integrieren, gilt mit (4.8) für alle $u, v \in \mathcal{P}_N$

$$(u, v)_{T_0} = \sum_{i=1}^{m} u(z_i)v(z_i)w_i = (\mathbf{v}^{\eta})^T W \mathbf{u}^{\eta} = (\mathbf{v}^{\xi})^T M \mathbf{u}^{\xi} \tag{4.9}$$

mit symmetrischer *Massenelementmatrix* $M = V_{\xi}^{-T}V_{\eta}^T W V_{\eta}V_{\xi}^{-1} \in \mathbb{R}^{d\times d}$ und Gewichtsdiagonalmatrix $W = \operatorname{diag}(w_1, \ldots, w_m) \in \mathbb{R}^{m\times m}$, deren Diagonaleinträge die Gewichte aus der verwendeten Quadraturformel sind. Bei Exaktheit der verwendeten Quadraturformeln bis Grad $2N-1$ wird das Skalarprodukt zwar nicht mehr exakt integriert, aber die Näherung bleibt akzeptabel und ist vom Exaktheitsgrad identisch mit den üblichen Gauss-Tensor-Produkt-Formeln in QSEM.

Um das Funktional $F(v)$ aus (2.7c) approximieren zu können, wird (4.9) für $v \in \mathcal{P}_N(T_0)$ und $f \in H^1(T_0)$ abgewandelt zu

$$F_{T_0}(v) = (f, v)_{T_0} \approx (\mathbf{v}^{\xi})^T \tilde{F}\, \mathbf{f}^{\eta} \tag{4.10}$$

mit $\tilde{F} = V_{\xi}^{-T}V_{\eta}^T W \in \mathbb{R}^{d\times m}$. Im Allgemeinen kann das Funktional nicht exakt berechnet werden, da nur $f \in H^1(T_0)$ gilt.

Die Wahl der Lagrange-Polynome L_i für die beliebige Funktion v in (4.7) liefert die

[3]Hier und im Folgenden wird die *fett*-Schreibweise symbolisch für Vektoren verwendet, d. h. $\mathbf{v} = (v_1, \ldots, v_d)^T$, $\hat{\mathbf{v}} = (\hat{v}_1, \ldots, \hat{v}_d)^T$, $\mathbf{v}^{\xi} = (v(\xi_1), \ldots, v(\xi_d))^T$, $\mathbf{v}^{\eta} = (v(\eta_1), \ldots, v(\eta_m))^T$ u.s.w..

Basen-Transformationsformel:

$$\begin{bmatrix} L_1(r,s) \\ \vdots \\ L_n(r,s) \end{bmatrix} = V_\xi^{-T} \begin{bmatrix} g_1(r,s) \\ \vdots \\ g_n(r,s) \end{bmatrix} \tag{4.11}$$

Bei Berechnung der partiellen Ableitung nach r bzw. s einer Funktion $v \in \mathcal{P}_N(T_0)$ wird die Ableitung erst auf die Lagrangeschen Basisfunktionen in (4.6) und anschließend mittels Gleichung (4.11) auf die Dubiner-Basis übertragen. Dies führt zu:

$$\frac{\partial \mathbf{v}^\eta}{\partial \alpha} = \left(\frac{\partial v}{\partial \alpha}(\eta_i) \right)_i = D_\alpha \mathbf{v} \quad \text{für} \quad \alpha \in \{r,s\} \quad \text{mit} \tag{4.12a}$$

$$D_\alpha = A_\alpha V_F^{-1} \quad \text{und} \quad A_\alpha = \left(\frac{\partial}{\partial \alpha} g_j(\eta_i) \right)_{ij} \in \mathbb{R}^{m \times d}. \tag{4.12b}$$

Die praktische Berechnung der auftretenden Terme, insbesondere der Ableitung des Dubiner-Polynoms g_j nach r bzw. s, ist unter Zuhilfenahme von einfachen Rekursionsformeln für die Jacobi-Polynome effizient möglich.

Analog zu (4.9) gestaltet sich die Berechnung der partiellen Ableitungen von u und v in den L^2-Skalarprodukt unter Verwendung von (4.12):

$$\left(\frac{\partial u}{\partial \alpha}, \frac{\partial v}{\partial \beta} \right)_{T_0} = (\mathbf{v}^\xi)^T S_{\alpha\beta} \mathbf{u}^\xi \quad \text{für} \quad \alpha, \beta \in \{r,s\} \tag{4.13}$$

mit unsymmetrischer Elementmatrix $S_{\alpha\beta} = V_\xi^{-T} A_\alpha^T W A_\beta V_\xi^{-1} \in \mathbb{R}^{d \times d}$ und Gewichtsmatrix W, wie oben.

4.3 Elementmatrizen auf beliebigen Element T_k

Die Berechnung der Elementmatrizen auf einem beliebigen Element $T_k \in \mathcal{T}_h$ mit den Eckpunkten $P_i = (x_i, y_i) \in \Omega$ für $i = 1,2,3$ soll dadurch vereinfacht werden, dass die Berechnung auf das Standard-Element T_0 zurückgeführt wird. Hierfür benutzen wir eine lineare Bijektion von T_0 auf ein beliebiges Dreieck T_k:

$$\chi : T_0 \longrightarrow T_k : \begin{bmatrix} r \\ s \end{bmatrix} \longmapsto \begin{bmatrix} x(r,s) \\ y(r,s) \end{bmatrix} = \frac{1}{2} \begin{bmatrix} x_2 + x_3 + (x_2 - x_1)r + (x_3 - x_1)s \\ y_2 + y_3 + (y_2 - y_1)r + (y_3 - y_1)s \end{bmatrix} \tag{4.14}$$

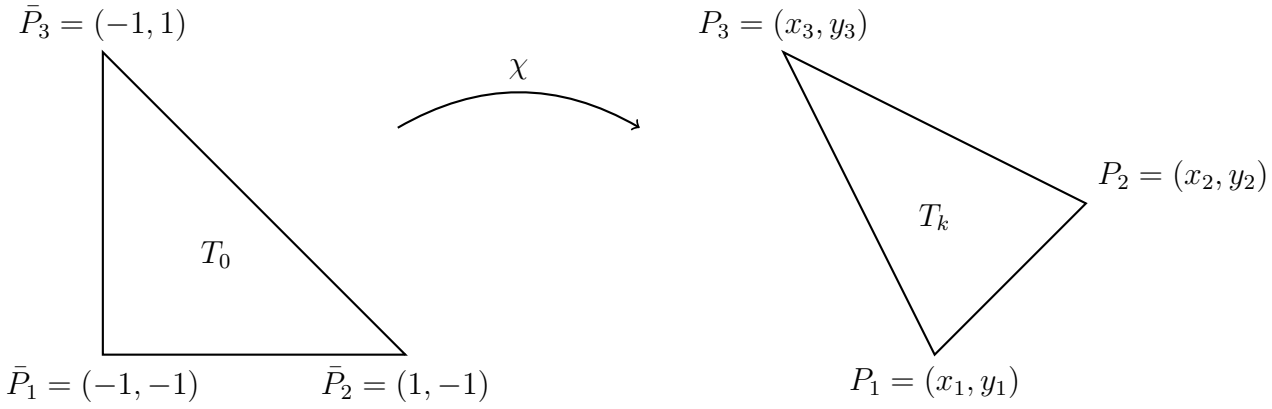


Abbildung 4.2: Die *Mapping*-Funktion χ bildet T_0 auf ein T_k ab.

Wie in Abbildung 4.2 veranschaulicht, bildet χ jeweils die Eckpunkte von T_0 auf die Eckpunkte von T_k ab: $\chi(\bar{P}_i) = P_i$ für $i = 1, 2, 3$.

Wichtig ist die Nummerierung der Eckpunkte von T_k gegen den Uhrzeigersinn, so dass die Jacobideterminate die halbe Größe des Flächeninhalts von T_k hat, d. i.

$$J_k := \det D_\chi \equiv \frac{1}{4}\left[(x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1)\right] = \frac{1}{2}\lambda(T_k) > 0.$$

Das L^2-Skalarprodukt auf T_k von $u, v \in \mathcal{P}_N(T_k)$ wird aus (4.9) mit dem Transformationssatz berechnet zu:[4]

$$(u, v)_{T_k} = J_k (u, v)_{T_0} = (\mathbf{v}^\xi)^T M_k \mathbf{u}^\xi. \tag{4.15}$$

mit Massenelementmatrix $M_k = J_k M \in \mathbb{R}^{d \times d}$ zu Element T_k.

Entsprechend (4.10) gelangen wir zu einer Approximation für das Funktional F_{T_k} auf Element T_k für $v \in \mathcal{P}_N(T_k)$ und $f \in H^1(T_k)$ durch

$$F_{T_k}(v) = (f, v)_{T_k} \approx (\mathbf{v}^\xi)^T \tilde{F}_k \, \mathbf{f}^\eta \tag{4.16}$$

mit $\tilde{F}_k = J_k \tilde{F} \in \mathbb{R}^{d \times m}$ und $\tilde{F}$ aus (4.10).

Zur Bestimmung der Steifigkeitselementmatrix muss zunächst der Differentialoperator bzgl. x und y auf r und s übertragen werden. Durch die Inverse der Jacobi-Matrix von

[4] Auf die Angabe der Abbildungsfunktion χ durch $u \circ \chi$ u.s.w. wird hier und im Folgenden der besseren Übersichtlichkeit halber verzichtet, wenn sich die Transformation aus dem Kontext ergibt.

χ und mittels zwei-dimensionaler Kettenregel erhalten wir

$$\nabla = \begin{bmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{bmatrix} = \frac{1}{2J_k} \begin{bmatrix} y_3 - y_1 & y_1 - y_2 \\ x_1 - x_3 & x_2 - x_1 \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial r} \\ \frac{\partial}{\partial s} \end{bmatrix}. \tag{4.17}$$

Angewandt auf das L^2-Skalarprodukt der Gradienten von $u, v \in \mathcal{P}_N(T_k)$

$$\begin{aligned} (\nabla u, \nabla v)_{T_k} &= a_k \left(\frac{\partial u}{\partial r}, \frac{\partial v}{\partial r} \right)_{T_0} + b_k \left[\left(\frac{\partial u}{\partial r}, \frac{\partial v}{\partial s} \right)_{T_0} + \left(\frac{\partial u}{\partial s}, \frac{\partial v}{\partial r} \right)_{T_0} \right] + c_k \left(\frac{\partial u}{\partial s}, \frac{\partial v}{\partial s} \right)_{T_0} \\ \text{mit} \quad a_k &= \frac{1}{4J_k} \left[(x_3 - x_1)^2 + (y_3 - y_1)^2 \right], \\ b_k &= -\frac{1}{4J_k} \left[(x_2 - x_1)(x_3 - x_1) + (y_2 - y_1)(y_3 - y_1) \right], \\ c_k &= \frac{1}{4J_k} \left[(x_2 - x_1)^2 + (y_2 - y_1)^2 \right]. \end{aligned}$$

Die Skalarprodukte auf der rechten Seite werden durch (4.13) ausgewertet und so ist

$$(\nabla u, \nabla v)_{T_k} = (\mathbf{v}^\xi)^T S_k \mathbf{u}^\xi, \quad \text{wobei} \tag{4.18}$$

$$S_k = a_k S_{rr} + b_k (S_{rs} + S_{sr}) + c_k S_{ss} \in \mathbb{R}^{d \times d} \tag{4.19}$$

die symmetrische Steifigkeitselementmatrix zu T_k ist.

Der Anteil des k-ten Elements in der Bilinearform a, wie sie in (2.7a) definiert ist, kann bestimmt werden zu

$$a_{T_k}(u, v) = (\nabla u, \nabla v)_{\mathcal{T}_k} + \lambda (u, v)_{\mathcal{T}_k} = (\mathbf{v}^\xi)^T \left[S_k + \lambda M_k \right] \mathbf{u}^\xi. \tag{4.20}$$

4.4 Galerkin-Verfahren und Assemblierung zu einem linearen Gleichungssystem

Im Folgenden betrachten wir für $\lambda \geq 0$ die Randwertaufgabe

$$-\Delta u + \lambda u = f \qquad \text{in } \Omega, \tag{4.21a}$$

$$u = g_{\mathcal{D}} \qquad \text{auf } \Gamma_{\mathcal{D}}, \tag{4.21b}$$

$$\frac{\partial u}{\partial n} = 0 \qquad \text{auf } \Gamma_{\mathcal{N}}, \tag{4.21c}$$

auf dem beschränkten, polygonalen Gebiet $\Omega \in \mathbb{R}^2$ mit Rand $\partial\Omega = \Gamma_D \cup \Gamma_N$. Die schwache Formulierung bzw. Variationsformulierung ist, wie in Abschnitt 2.2 gezeigt, gegeben durch (2.10):

$$\text{Finde } u^H \in V_0 \text{ mit} \quad a(u^H, v) = \hat{F}(v) = F(v) - a(u^{\mathcal{D}}, v) \quad \text{für alle } v \in V_0. \tag{4.22}$$

Dann lautet die schwache Lösung: $u^N = u^H + u^{\mathcal{D}} \in H^1(\Omega)$.

Notwendig ist die Existenz einer Funktion $u^{\mathcal{D}} \in H^1(\Omega)$ mit $u^{\mathcal{D}}(\Gamma_{\mathcal{D}}) = g_{\mathcal{D}}$. Während für homogene Dirichlet-Randbedingung, d. h. $g_{\mathcal{D}} = 0$ in (4.21b), $u^{\mathcal{D}} \equiv 0$ gewählt wird, ist der allgemeine inhomogene Fall etwas aufwändiger und wird im Folgenden betrachtet.

4.4.1 Galerkin-Verfahren

Wir wollen nun die Variationsformulierung (4.22) diskretisieren und eine passende Näherung für $u^{\mathcal{D}}$ angeben.

Sei ein fester Polynomgrad $N \in \mathbb{N}$ gegeben. Ferner sei eine konforme Triangulierung $\mathcal{T}_h$ des Gebiets Ω mit passender Nummerierung der Fekete-Knotenpunkte aller Elemente $\Omega_N = \{(x_i, y_i) \mid 1 \leq i \leq n\}$ wie in Abschnitt 4.1 gegeben.

Als Diskretisierung des Raumes $H^1(\Omega)$ definieren wir den endlich-dimensionalen Raum

$$V^N = \left\{ v \in C^0(\Omega) \mid v_{|T_k} \in \mathcal{P}_N(T_k) \text{ für alle } T_k \in \mathcal{T}_h \right\} \subset H^1(\Omega). \tag{4.23}$$

mit Dimension $\dim(V^N) = n$.

V^N wird entsprechend der Nummerierung aller Knotenpunkte von den *nodalen Basisfunktionen*

$$\psi_i \in V^N \quad \text{mit} \quad \psi_i(x_j, y_j) = \delta_{ij}, \ 1 \leq i, j \leq n, \tag{4.24}$$

aufgespannt.

Für $i \in \{1, \ldots, n\}$ sei $\hat{\mathcal{T}}^i = \{T_k \in \mathcal{T}_h \mid (x_i, y_i) \in T_k\}$ die Menge der Elemente, in denen der i-te Knotenpunkt liegt. Dann ist die i-te Basisfunktion auf dem k-ten Element

$$\psi_i(x, y) = \begin{cases} L^k_{m_i}(x, y), & \text{wenn } T_k \in \tilde{\mathcal{T}}^i \\ 0, & \text{wenn } T_k \notin \tilde{\mathcal{T}}^i, \end{cases} \quad \text{für alle } (x, y) \in T_k, \quad 1 \leq k \leq n, \tag{4.25}$$

wobei $m_i = (\mathrm{map}^N_k)^{-1}(i)$ die lokale Nummer des transformierten Fekete-Punktes (x_i, y_i) im k-ten Element repräsentiert und $L^k_{m_i}$ das m_i-te Lagrangesche Polynom zu den Fekete-Punkten auf T_k ist. Diese zu (4.24) äquivalente Darstellung ist wohl-definiert und stetig

insbesondere auf den Elementrändern.

Analog zur Definition (4.23) wird der obige Lösungsraum V_0 der Variationsformulierung (4.22) diskretisiert zu

$$V_0^N = \left\{ v \in V^N \;\middle|\; v = 0 \text{ auf } \Gamma_{\mathcal{D}} \right\} \subset V_0 \subset H_0^1(\Omega). \tag{4.26}$$

V_0^N wird von den nodalen Basisfunktionen $\psi_i(x, y),\ i = 1, \ldots, \tilde{n}$ aufgespannt, welche durch die Knotenpunkte im Inneren von Ω und auf $\Gamma_{\mathcal{N}}$ charakterisiert sind. Zu beachten ist, dass hier im Gegensatz zu V^N nur die ersten $\tilde{n}$ der insgesamt n nodalen Basisfunktionen verwendet werden, da die übrigen auf $\Gamma_{\mathcal{D}}$ nicht verschwinden und nicht in V_0^N liegen.

Damit sind wir in der Lage, die Variationsformulierung (4.22) zu diskretisieren und gelangen zum *Galerkin-Verfahren*:

$$\text{Finde } u^H \in V_0^N \text{ mit} \quad a(u^H, v) = \hat{F}(v) = F(v) - a(u^{\mathcal{D}}, v) \quad \text{für alle } v \in V_0^N, \tag{4.27}$$

wobei zuvor $u^{\mathcal{D}}$ entsprechend der Randbedingung (4.21b) gewählt wird. Die *Galerkin-Lösung* lautet dann:

$$u^N = u^H + u^{\mathcal{D}} \in H^1(\Omega), \tag{4.28}$$

wobei $u^H \in V_0^N$ die Formulierung (4.27) erfüllt. Zudem entspricht $u^{\mathcal{D}} \in H^1(\Omega)$ der Funktion $g_{\mathcal{D}}$ auf $\Gamma_{\mathcal{D}}$.

Die homogene Lösung u^H lässt sich als Linearkombination der Basisfunktionen $\psi_i \in V_0^N,\ 1 \leq i \leq \tilde{n}$, darstellen:

$$u^H(x, y) = \sum_{i=1}^{\tilde{n}} u_i^H \psi_i(x, y) \quad \text{ und } \ u_i^H = u^H(x_i, y_i)$$

Aufgrund der Linearität von a und $\hat{F}$ ist (4.27) äquivalent zur Lösung des linearen Gleichungssystems:

$$\sum_{j=1}^{\tilde{n}} u_j^H a(\psi_j, \psi_i) = \hat{F}(\psi_i) = F(\psi_i) - a(u^{\mathcal{D}}, \psi_i), \quad 1 \leq i \leq \tilde{n} \tag{4.29}$$

Des Weiteren ist eine Funktion $u^{\mathcal{D}} \in H^1(\Omega)$ zu wählen, die die Dirichlet-Randbedingung in (4.21b) erfüllt.

Ist diese homogen, d. h. $g_{\mathcal{D}} = 0$ auf $\Gamma_{\mathcal{D}}$, wird $u^{\mathcal{D}} \equiv 0$ in Ω gewählt, so dass in (4.29) der Term $a(u^{\mathcal{D}}, \psi_i)$ für alle i verschwindet.

Bei allgemeiner inhomogener Dirichlet-Randbedingung (4.21b), gestaltet sich die Angabe einer passenden Funktion schwieriger. Wir wählen daher

$$u^{\mathcal{D}}(x,y) = \sum_{j=\tilde{n}+1}^{n} \psi_j(x,y) g_{\mathcal{D}}(x_j,y_j) \in V^N \subset H^1(\Omega).$$

Diese Funktion erfüllt die Bedingung (4.21b) zwar nicht auf ganz $\Gamma_{\mathcal{D}}$. Sie liefert jedoch eine gute Approximation der Randfunktion, da die Punkte auf den jeweiligen Elementrändern entsprechend der Gauß-Lobatto-Legendre-Verteilung liegen, und sie exakt für alle Knotenpunkte auf dem Rand $\Gamma_{\mathcal{D}}$ ist. Bei dieser Wahl für $u^{\mathcal{D}}$ verschwindet die Funktion in den übrigen Knotenpunkten:

$$u^{\mathcal{D}}(x_i,y_i) = 0, \quad \text{falls } (x_i,y_i) \in \Omega_N \setminus \Gamma_{\mathcal{D}}, \text{ d. i. für alle } i = 1,\dots,\tilde{n}$$

Für die Galerkin-Lösung (4.28) folgt demnach:

$$u^N(x_i,y_i) = \begin{cases} u^H(x_i,y_i), & \text{falls } (x_i,y_i) \in \Omega_N \setminus \Gamma_{\mathcal{D}}, \\ g_{\mathcal{D}}(x_i,y_i), & \text{falls } (x_i,y_i) \in \Gamma_{\mathcal{D}}. \end{cases} \tag{4.30}$$

Entsprechend (4.30) definieren wir den Vektor $\mathbf{u^N} = (u_i^N)_i = (u^N(x_i,y_i))_i \in \mathbb{R}^n$.

Eingesetzt in (4.29) gelangen wir zusammen mit Definition (2.7c) zu dem Gleichungssystem

$$\sum_{j=1}^{\tilde{n}} u_j^N a(\psi_j,\psi_i) = (f,\psi_i) - \sum_{l=\tilde{n}+1}^{n} g_{\mathcal{D}}(x_l,y_l) a(\psi_l,\psi_i) \quad \text{für } i = 1,\dots,\tilde{n}. \tag{4.31}$$

4.4.2 Assemblierung

Wir werden (4.30) und das Gleichungssystem (4.31) in das lineare Gleichungssystem

$$L\mathbf{u^N} = \mathbf{b}, \tag{4.32}$$

mit $L \in \mathbb{R}^{n\times n}$ und $\mathbf{b} \in \mathbb{R}^n$ überführen, wobei der Lösungsvektor $\mathbf{u^N}$ den Funktionswerten der approximierten Lösung in den Knotenpunkten entspricht. Für $i = 1,\dots,\tilde{n}$ repräsentiert die i-te Zeile von (4.32) die i-te Bedingung der Galerkin-Formulierung (4.31), während die übrigen Zeilen für $i = \tilde{n}+1,\dots,n$ die Dirichlet-Randbedingung zum i-ten Knotenpunkt (x_i,y_i), wie in (4.30) dargestellt, angeben.

In Anlehnung an die Ausführung[5] von Knabner und Angermann in [12] soll eine kurze Darstellung der elementweisen Assemblierung der Matrix L und der rechten Seite $\mathbf{b}$ vorgestellt werden, die auf den Elementmatrizen und -vektoren aus Abschnitt 4.3 basiert und die Randbedingungen (4.21b),(4.21c) erfüllt.

Dafür werden wir sukzessive für jedes Element $T_k,\ k = 1, \ldots, M$, die Anteile $L^{(k)}$ bzw. $\mathbf{b^{(k)}}$ zum Gesamtsystem (4.32), die sich aus (4.31) und (4.30) ergeben, bestimmen und in die Matrix L bzw. Vektor $\mathbf{b}$ aufsummieren (engl. *assemble*).

Zunächst zerlegen wir die in (4.31) auftretenden Integrale auf die jeweiligen Elemente $T_k \in \mathcal{T}_h$. So wird beispielsweise

$$A_{ij} := a_\Omega(\psi_j, \psi_i) = \sum_{k=1}^{M} A_{ij}^{(k)} \quad \text{mit } A_{ij}^{(k)} = a_{T_k}(\psi_j, \psi_i)$$

aufsummiert. Wie in [12, § 2.2] beschrieben, kann $A_{ij}^{(k)}$ nur dann ungleich 0 sein, wenn der i-te und der j-te Knotenpunkt in T_k liegen, da ansonsten der Träger[6] einer der Basisfunkionen ψ_i oder ψ_j keine Schnittmenge positiven Maßes mit T_k hat. Somit müssen die Terme $A_{ij}^{(k)}$ nur für die Werte von i und j betrachtet werden, wenn sowohl i, als auch j der globalen Nummer eines Knotenpunktes in Element T_k entsprechen, da für die sonstigen Werte für i und j der Term $A_{ij}^{(k)}$ verschwindet.

Es wird von der lokalen Nummerierung der Knotenpunkte in T_k ausgegangen, so dass jeweils der Knotenpunkt mit der lokalen Nummer $\imath \in \{1, \ldots, d\}$ die globale Nummer $m_\imath := \mathrm{map}_k^N(\imath)$ trägt.[7] Auf diese Weise bauen wir

$$\left(A_{m_\imath m_\jmath}^{(k)}\right)_{\imath,\jmath=1,\ldots,n} \text{ als } \left(\tilde{A}_{\imath\jmath}^{(k)}\right)_{\imath,\jmath=1,\ldots,n} \tag{4.33}$$

auf. Bei Verwendung der Steifigkeitselementmatix S_k aus (4.20) und der Darstellung der Basisfunktionen in (4.25) gilt

$$\begin{aligned} \tilde{A}_{\imath\jmath}^{(k)} &= A_{m_\imath m_\jmath}^{(k)} = a_{T_k}(\psi_{m_\jmath}, \psi_{m_\imath}) \\ &= a_{T_k}(L_\jmath^k, L_\imath^k) = (\mathbf{L}_\imath^\mathbf{k})^T S_k \mathbf{L}_\jmath^\mathbf{k} \\ &= (S_k)_{\imath\jmath}\,, \end{aligned}$$

wobei $L_\imath^k \in \mathcal{P}_N(T_k)$ das Langrangesche Polynom zu den Fekete-Punkten auf T_k ist,

[5] In § 2.4 aus [12] wird die Assemblierung allerdings nur für lineare FEM vorgestellt.

[6] Der Träger einer Funktion $f : D \longrightarrow \mathbb{R}$ ist der Abschluss der Nicht-Nullstellen-Menge von f in D, d. i. $\mathrm{supp}(f) = \overline{\{x \in D \mid f(x) \neq 0\}}$

[7] map_k^N bezeichnet den Mapping-Vektor zu T_k aus Abschnitt 4.1

welches in den $\imath$-ten lokalen Knotenpunkt 1 wird. Derart gelangen wir zu

$$\tilde{A}^{(k)} = S_k. \tag{4.34}$$

Mit gleicher Überlegung lassen sich die Ausdrücke (f, ψ_i) der rechten Seite von (4.31) auf die Elemente T_k zerlegen.

Analog zur Darstellung (4.33) wollen wir die Anteile $L^{(k)}$ bzw. $\mathbf{b^{(k)}}$ aus $\tilde{L}^{(k)}$ bzw. $\tilde{\mathbf{b}}^{(\mathbf{k})}$ berechnen.

Nach obiger Überlegung geschieht dies durch folgenden Algorithmus:

1. Setze zuerst:
$$\tilde{L}^{(k)} = S_k \quad \text{und} \quad \tilde{\mathbf{b}}^{(\mathbf{k})} = \tilde{F}_k \, \mathbf{f}^{\eta}, \tag{4.35}$$
wobei S_k Matrix aus (4.20), $\tilde{F}_k$ Rechteckmatrix aus (4.16) und $\mathbf{f}^{\eta} \in \mathbb{R}^m$ Vektor der Funktionswerte von f in den auf T_k transformierten Quadraturpunkten η.

2. Für alle $\jmath = 1, \ldots, d$ mit $m_\jmath > \tilde{n}$, d. h. der Knotenpunkt $(x_{m_\jmath}, y_{m_\jmath})$ gehört zu $\Gamma_{\mathcal{D}}$:
   - Modifiziere den $\imath$-ten Eintrag $(\tilde{\mathbf{b}}^{(\mathbf{k})})_\imath$ für $\imath \neq \jmath$ zu $\tilde{\mathbf{b}}^{(\mathbf{k})} - (\tilde{L}^{(k)})_{\imath\jmath} g_{\mathcal{D}}(x_{m_\jmath}, y_{m_\jmath})$,
   - ersetze die $\jmath$-te Spalte und $\jmath$-te Zeile von $\tilde{L}^{(k)}$ durch den $\jmath$-ten Einheitsvektor,
   - setze den $\jmath$-ten Eintrag von $\tilde{\mathbf{b}}^{(\mathbf{k})}$ auf $g_{\mathcal{D}}(x_{m_\jmath}, y_{m_\jmath})$.

Nachdem die Anteile bestimmt sind, werden sie in die Matrix L bzw. rechte Seite $\mathbf{b}$ von (4.32) über eine Schleife assembliert:
Für $\imath, \jmath = 1, \ldots, d$ setze

$$L_{m_\imath m_\jmath} := L_{m_\imath m_\jmath} + \tilde{L}^{(k)}_{\imath\jmath}, \tag{4.36a}$$
$$\mathbf{b}_{m_\imath} := \mathbf{b}_{m_\imath} + \tilde{\mathbf{b}}^{(k)}_{\imath}. \tag{4.36b}$$

Schließlich erlangen wir auf diese Weise das lineare Gleichungssystem (4.32)

$$L\mathbf{u}^{\mathbf{N}} = \mathbf{b}$$

mit symmetrischer Matrix $L \in \mathbb{R}^{n \times n}$. Die Gestalt des Systems sieht dabei wie folgt aus:

$$\left[\begin{array}{c|c} L_{\mathcal{H}} & \mathbb{O} \\ \hline \mathbb{O}^T & \mathbb{I}_{n_{\mathcal{D}}} \end{array}\right] \cdot \left[\begin{array}{c} \mathbf{u^N}_{\mathcal{H}} \\ \hline \mathbf{u^N}_{\mathcal{D}} \end{array}\right] = \left[\begin{array}{c} \mathbf{b}_{\mathcal{H}} \\ \hline \mathbf{b}_{\mathcal{D}} \end{array}\right] \tag{4.37}$$

mit $L_{\mathcal{H}} \in \mathbb{R}^{\tilde{n} \times \tilde{n}}$, einer $n_{\mathcal{D}} \times n_{\mathcal{D}}$-dimensionalen Einheitsmatrix $\mathbb{I}_{n_{\mathcal{D}}}$, sowie der Nullmatrix $\mathbb{O} \in \mathbb{R}^{\tilde{n} \times n_{\mathcal{D}}}$. Charakteristisch ist die Aufteilung des Gleichungssystems in den oberen Teil $L_{\mathcal{H}}\mathbf{u^N}_{\mathcal{H}} = \mathbf{b}_{\mathcal{H}}$, der äquivalent zur Variationsformulierung (4.31) ist, und den unteren Teil $\mathbb{I}_{n_{\mathcal{D}}}\mathbf{u^N}_{\mathcal{D}} = \mathbf{b}_{\mathcal{D}}$, der die Dirichlet-Randbedingung in Form von (4.30) darstellt. Es ist eine Folge der Nummerierung aus Abschnitt 4.1 und insbesondere der Sortierung in (4.2).

5 Statische Kondensation

Die in diesem Kapitel vorgestellte Methode soll das lineare Gleichungssystem (4.32), welches in Kapitel 4 erarbeitet wurde, in eine für das Mehrgitterverfahren günstigere Form überführen. Die statische Kondensation wird lediglich zur Vorkonditionierung des Systems (4.32) eingesetzt, ohne dabei die Größe des System auf das sogenannte Schur-Komplement zu verkleinern.

Erst im Anschluss wird das Mehrgitterverfahren auf dem kompletten Gitter angewendet. Auf diese Weise bleibt das System zwar groß, die günstigen Interpolationseigenschaften der Fekete-Punkte werden hingegen durch das Streichen der inneren Elementpunkte nicht beeinträchtigt.

5.1 Struktur des assemblierten linearen Gleichungssystems

Seien $N \in \mathbb{N}$ der Polynomgrad und $M \in \mathbb{N}$ die Anzahl der Dreiecke in der Triangulierung $\mathcal{T}_h$. Ausgehend von der im Kapitel 4 beschriebenen Diskretisierung der Variationsaufgabe, erhalten wir das symmetrische Gleichungssystem

$$A\mathbf{u} = \mathbf{b}. \tag{5.1}$$

Während die Reihenfolge der Knotenpunkte aus Ω_N für die Assemblierung von (5.1) außer der eindeutigen Zuordnung keinen tatsächlichen Mehrwert hat, kann die in diesem Kapitel präsentierte Methode davon profitieren.

Entsprechend der Nummerierung aus Abschnitt 4.1 lassen sich die Knotenpunkte Ω_N aufteilen in

$$\Omega_i = \{(x_1, y_1), \ldots, (x_{n_i}, y_{n_i})\} = \bigcup_{T_k \in \mathcal{T}_h} \operatorname{int}(T_k) \cap \Omega_N, \tag{5.2a}$$

$$\Omega_r = \{(x_{n_i+1}, y_{n_i+1}), \ldots, (x_n, y_n)\} = \bigcup_{T_k \in \mathcal{T}_h} \partial T_k \cap \Omega_N. \tag{5.2b}$$

mit $n_i = |\Omega_i| = \tilde{d}M = \frac{1}{2}(N-1)(N-2)M$ und $n_r = |\Omega_r| = M_V + (N-1)M_E$. Das bedeutet, dass die ersten n_i Knoten die Fekete-Punkte im Inneren jedes Gebiets darstellen, gefolgt von den Punkten auf allen Elementrändern. Diese Aufteilung ist nicht zu verwechseln mit (4.2), wo nur zwischen dem Inneren und dem Rand des Gesamtgebiets Ω unterschieden wurde.

Die Segmentierung der Knoten in (5.2) überträgt sich auf das Gleichungssystem (5.1) zu

$$\begin{bmatrix} A_i & A_c \\ A_c^T & A_r \end{bmatrix} \begin{bmatrix} \mathbf{u}_i \\ \mathbf{u}_r \end{bmatrix} = \begin{bmatrix} \mathbf{b}_i \\ \mathbf{b}_r \end{bmatrix}. \tag{5.3}$$

Der Anteil $\mathbf{u}_i \in \mathbb{R}^{n_i}$ bzw. $\mathbf{u}_r \in \mathbb{R}^{n_r}$ (analog $\mathbf{b}_i$ und $\mathbf{b}_r$) stellt die innere bzw. die Randkomponente von $\mathbf{u}$ dar. Die Teilmatrix A_i bildet die Wechselbeziehung von den inneren zu den inneren Funktionen über alle Elemente ab. Entsprechend geben A_r die Rand-zu-Rand- und A_c die Innere-zu-Rand-Interaktion wieder.

Die Interaktion erfolgt hierbei über die nodalen Basisfunktionen aus (4.24). Aus der Formulierung (4.25) wird ersichtlich, dass die Basisfunktionen von zwei inneren Knotenpunkten aus verschiedenen Elementen nicht miteinander interagieren, da beide Funktionen auf dem Elementrand verschwinden und keinen gemeinsamen Träger besitzen.

Dies hat zur Folge, dass die Matrix A_i Blockdiagonalgestalt annimmt:

$$A_i = \begin{bmatrix} A_i^1 & & & \\ & A_i^2 & & \\ & & \ddots & \\ & & & A_i^M \end{bmatrix} \tag{5.4}$$

Die Blöcke $A_i^k \in \mathbb{R}^{\tilde{d}\times\tilde{d}}$ mit $\tilde{d} = \frac{1}{2}(N-1)(N-2)$ überschneiden sich nicht. Somit ist das Invertieren bzw. Auflösen der Matrix A_i günstig.

5.2 Idee der statischen Kondensation

Die Motivation der Methode liegt darin, die Abhängigkeit der Elementrandknoten aus Ω_r von den Elementinnenknoten Ω_i zu trennen.

Das Gleichungssystem (5.3) wird hierzu mit der nichtsingulären Matrix

$$\begin{bmatrix} A_i^{-1} & 0 \\ -A_c^T M_i^{-1} & I \end{bmatrix}$$

multipliziert, wobei $0 \in \mathbb{R}^{n_i \times n_r}$ die Nullmatrix und I die Einheitsmatrix der Dimension $n_r \times n_r$ darstellt. Wir erhalten das zu (5.1) äquivalente Gleichungssystem

$$\begin{bmatrix} I & A_i^{-1}A_c \\ 0 & A_r - A_c^T A_i^{-1} A_c \end{bmatrix} \begin{bmatrix} \mathbf{u}_i \\ \mathbf{u}_r \end{bmatrix} = \begin{bmatrix} A_i^{-1}\mathbf{b}_i \\ \mathbf{b}_r - A_c^T A_i^{-1}\mathbf{b}_i \end{bmatrix}. \tag{5.5}$$

Dieses Vorgehen wird *statische Kondensation* genannt.

Um das System (5.1) numerisch effizient in (5.5) zu überführen, wird die Struktur der Matrix A_i ausgenutzt. Aus (5.4) wird ersichtlich:

$$A_i^{-1} = \begin{bmatrix} (A_i^1)^{-1} & & & \\ & (A_i^2)^{-1} & & \\ & & \ddots & \\ & & & \left(A_i^M\right)^{-1} \end{bmatrix} \tag{5.6}$$

Also genügt es zur Ermittlung der Inversen A_i^{-1}, die einzelnen, nichtüberschneidenden Blöcke A_i^k zu invertieren. Damit verringert sich der Aufwand erheblich.

Durch die statische Kondensation bleibt die Matrix in gleicher Weise dünnbesetzt.

Zur Lösung des linearen Gleichungssystems (5.5) wird üblicherweise zuerst das verkleinerte Teilsystem

$$\left(A_r - A_c^T A_i^{-1} A_c\right) \mathbf{u}_r = \mathbf{b}_r - A_c^T A_i^{-1}\mathbf{b}_i \tag{5.7}$$

gelöst, vergleiche zum Beispiel [15, 21]. Die Matrix $A_r - A_c^T A_i^{-1} A_c$ der Dimension $n_r \times n_r$ wird als das *Schur-Komplement* zu A bezeichnet.

Nachdem die Funktionswerte auf den Elementrändern mittels (5.7) berechnet wurden, lassen sich die Funktionswerte $\mathbf{u}_i$ der Elementinnenpunkte durch

$$\mathbf{u}_i = A_i^{-1}\mathbf{b}_i - A_i^{-1} A_c \mathbf{u}_r$$

ermitteln.

Wir betrachten hingegen im nächsten Kapitel das komplette System (5.5), um es mit dem Mehrgitterverfahren zu lösen. Dadurch werden die günstigen Interpolationseigenschaften der Fekete-Punkte während der Gitterübergänge (Prolongation und Restriktion) nicht beeinträchtigt.

6 Spektrale p-Mehrgitterverfahren

In den vorherigen Kapiteln wurde das ursprüngliche Randwertproblem (2.6) in eine Variationsformulierung überführt und anschließend mittels des Galerkin-Verfahrens in Form eines linearen Gleichungssystems

$$A\mathbf{u} = \mathbf{b} \tag{6.1}$$

mit einer großen, dünn besetzten Matrix $A \in \mathbb{R}^{n \times n}$ diskretisiert.

Während für kleinere Probleme direkte Löser, wie etwa die Gauß-Elimination, das Cholesky- oder das QR-Verfahren, angewendet werden können, ist dies für große Probleme kaum möglich, da der Rechenaufwand im Allgemeinen mit $O(n^3)$ zunimmt. Aber auch klassische Iterationsverfahren, beispielsweise das Jacobi- oder Gauss-Seidel-Verfahren, sind ohne zusätzliche Hilfsmittel ungeeignet, da die Konvergenzrate für wachsendes n immer schlechter wird.

Mehrgitterverfahren bieten sich als effiziente Alternative an und nutzen die Glättungseigenschaften klassischer Iterationsverfahren, beispielsweise des SOR-Verfahrens.

Im Folgenden wird ein p-Mehrgitterverfahren für die im Kapitel 4 erarbeitete TSEM vorgestellt. Wir halten uns in diesem Kapitel an die Ausführung in [13] von Pasquetti und Rappetti, sowie an die Darstellungen aus [12].

6.1 SOR-Iterationsverfahren

Die klassischen Iterationsverfahren, wozu auch das SOR-Verfahren gehört, weisen eine sogenannte Glättungseigenschaft auf. So werden die kurzwelligen Fehleranteile sehr schnell reduziert bzw. *geglättet*, während langwellige Fehleranteile nur sehr langsam gesenkt werden.

Wird beispielsweise zur Veranschaulichung das Gauß-Seidel-Verfahren auf das lineare Gleichungssystem eines elliptischen Randwertproblems (zum Beispiel der Poisson-Aufgabe, siehe Abschnitt 7.3) mit einem zufälligen Startvektor angewandt, so lässt sich

eine rasante Glättung des Fehlers nach nur wenigen Schritten beobachten. Dagegen wird der niederfrequente Fehleranteil nur sehr langsam gedämpft, was lediglich einer Fehlerreduktion der Ordnung $1 - \mathcal{O}(n^{-2})$ entspricht.[1]

Um das Iterationsverfahren einzuführen, wird die Matrix A aus (6.1) aufgespalten zu

$$A = L + D + R,$$

in die strikte untere Dreiecksmatrix L, die Diagonalmatrix D und die strikte obere Dreiecksmatrix R von A. Das *SOR-Verfahren* (engl. *successive overrelaxation*) entspricht der Iteration

$$\mathbf{u}^{(k+1)} = N_\omega^{-1} M_\omega \mathbf{u}^{(k)} + N_\omega^{-1}\mathbf{b} \tag{6.2a}$$

$$\text{mit} \quad M_\omega = -\frac{1}{\omega}\left((\omega - 1)D + \omega R\right) \tag{6.2b}$$

$$N_\omega = \frac{1}{\omega}(D + \omega L). \tag{6.2c}$$

für $k = 0, 1, 2, \ldots$ und einem Startvektor $\mathbf{u}^{(0)} \in \mathbb{R}^n$. Die Lösung des Gleichungssystems in jedem Schritt ist wegen der Dreiecksgestalt der Matrix N_ω effizient möglich. Das SOR-Verfahren eignet sich besonders für dünnbesetzte Matrizen, wie sie beispielsweise bei der Spektrale-Elemente-Methode auftreten.

Für den Wert $\omega = 1$ erhalten wir das *Gauß-Seidel-Verfahren*:

$$\mathbf{u}^{(k+1)} = -(D + L)^{-1} R \mathbf{u}^{(k)} + (D + L)^{-1}\mathbf{b} \tag{6.3a}$$

$$\text{bzw.} \quad (D + L)(\mathbf{u}^{(k+1)} - \mathbf{u}^{(k)}) = \mathbf{b} - A\mathbf{u}^{(k)} \tag{6.3b}$$

Basierend auf einer affin-linearen Fixpunktiterationsfunktion ist das Verfahren (6.3) genau dann konvergent, wenn der Spektralradius der Iterationsmatrix echt kleiner 1 ist, d. h. $\varrho((D + L)^{-1} R) < 1$.

Ein hinreichendes Kriterium für das allgemeine SOR-Verfahren liefert

Satz 6.1. *Ist A symmetrisch und positiv definit und gilt $0 < \omega < 2$, so konvergiert das SOR-Verfahren für alle Startvektoren $\mathbf{u}^{(0)} \in \mathbb{R}^n$.*

Beweis: siehe z.B. [8, S. 97].

[1] Vergleiche dazu [12, § 5.1]

6.2 Idee des p-Mehrgitterverfahrens

Das Konzept der Mehrgitterverfahren besteht darin, die klassischen Iterationsverfahren auf unterschiedlich feinen Gittern einzusetzen, um alle Fehleranteile rasch zu senken.

Ausgehend von einem beliebigen Anfangsvektor ist nach nur wenigen Schritten des SOR-Verfahrens der Fehler geglättet und die Oszillationen beseitigt, so dass der Fehler auf einem gröberen Gitter darstellbar ist. Die zuvor niederfrequenten Anteile sind nun hochfrequent und können wieder geglättet werden. Dies wird wiederholt, bis man eine gute Approximation auf dem feinsten Gitter erhält.

Grundsätzlich konzentrieren wir uns auf das p-Mehrgitterverfahren, bei dem der Polynomgrad der Ansatzräume V^N variiert wird, während die Triangulierung, im Gegensatz zum h-Mehrgitterverfahren, unverändert bleibt.

Es wird daher von einer festen Triangulierung $\mathcal{T}_h$ des Gebiets Ω und einer Folge von Polynomgraden

$$N_1 < N_2 < \ldots < N_l$$

ausgegangen. Auf jeder Stufe j betrachten wir neben dem Gitter $\Omega_j := \Omega_{N_j}$, welches in Abschnitt 4.1 definiert wurde, den zugehörigen Ansatzraum $V^j := V^{N_j}$ der Dimension $n_j := n_{N_j}$ (siehe Abschnitt 4.4). Dann sind die Ansatzräume geschachtelt:

$$V^1 \subset V^2 \subset \ldots \subset V^l \subset H^1(\Omega)$$

Wir werden zunächst die Übergangs-Operatoren und die auf jedem Gitter benötigten Hilfsmatrizen A_j vorstellen und anschließend einen Mehrgitter-Algorithmus präsentieren, der die obige Beschreibung konkretisiert.

6.3 Übergänge zwischen den Gittern

Zu den grundlegenden Bestandteilen der Mehrgitterverfahren gehören die Übergangs-Operatoren, die den Fehler bzw. das Residuum von einem Gitter auf das nächste projizieren.

Wie zuvor erwähnt, ändern sich beim p-Mehrgitterverfahren nur die Polynomgrade auf jeder Stufe, während die Triangulierung gleich bleibt. Folglich ist es ausreichend, die Übergänge auf nur einem beliebigem Element $T_k \in \mathcal{T}_h$ zu betrachten und die Operatoren daraufhin elementweise auf den Gesamtvektor anzuwenden, ähnlich der Assemblierung aus Abschnitt 4.4.

Der Einfachheit halber wird der Übergang zwischen einem groben und einem feinem Gitter, jeweils gekennzeichnet[2] durch c und f, beschrieben.

Seien also $T_k \in \mathcal{T}_h$ und die Polynomgrade $N_c, N_f \in \mathbb{N}$ des groben bzw. feinen Gitters derart fest vorgegeben, dass $N_c < N_f$. Auf dem groben Gitter benennen wir die auf T_k transformierten Fekete-Punkte[3] $\mathbf{x}^c = \left\{x_1^c, \ldots, x_{n_c}^c\right\} \subset T_k$ mit den zugehörigen Lagrangeschen Basisfunktionen $\left\{\varphi_1^c, \ldots, \varphi_{n_c}^c\right\}$ aus den Polynomialraum $\mathcal{P}_{N_c}(T_k)$ und die Dimension $n_c = \dim \mathcal{P}_{N_c}(T_k) = \frac{1}{2}(N_c+1)(N_c+2)$. Dementsprechend folgen die Bezeichnung $\boldsymbol{\xi}^f = \left\{\xi_1^f, \ldots, \xi_{n_f}^f\right\}$, $\mathbf{x}^f = \left\{x_1^f, \ldots, x_{n_f}^f\right\}$, $\left\{\varphi_1^f, \ldots, \varphi_{n_f}^f\right\}$ und n_f auf dem feinen Gitter.

6.3.1 Prolongation

Der Übergang vom groben zum feinen Gitter wird *Prolongation* genannt. Um die Vorteile des hohen Polynomgrades der *TSEM* für den Prolongations-Operator auszunutzen, betrachten wir zunächst die Interpolationsfunktion auf dem groben Gitter

$$I_c : \mathbb{R}^{n_c} \longrightarrow \mathcal{P}_{N_c}(T_k) : \mathbf{u} \longmapsto I_c\mathbf{u}(x) = \sum_{i=1}^{n_c} u_i \varphi_i^c(x), \tag{6.4}$$

die das Interpolationspolynom zum Vektor der Funktionswerte bestimmt.

Die Prolongation eines Vektors $\mathbf{u}^c \in \mathbb{R}^{n_c}$ wird auf natürliche Weise durch Auswertung der Interpolationsfunktion $I_c\mathbf{u}^c$ in den feinen Gitter $\mathbf{x}^f$ realisiert:
Für $i = 1, \ldots, n_f$ setze $u_i^f := I_c\mathbf{u}^c(x_i^f)$ und folglich gilt

$$\mathbf{u}^f = P\mathbf{u}^c, \quad \text{mit} \quad P = \left[\varphi_j^c(x_i^f)\right]_{ij} \in \mathbb{R}^{n_f \times n_c}. \tag{6.5}$$

Eine solche Wahl entspricht der *kanonischen Prolongation* auf einem Element, die von Hackbusch [9, § 3.6] eingeführt wurde.

Obwohl die Matrix P scheinbar von dem Element T_k abhängt, wird P nur einmal berechnet und kann für alle Elemente verwendet werden, wie folgende Berechnung zeigt:

$$P_{ij} = \varphi_j^c(\chi(\xi_i^f)) = L_j^c(\xi_i^f), \quad 1 \le i \le n_f,\ 1 \le j \le n_c,$$

wobei L_j^c das j-te Lagrangesche Polynom aus (4.5) ist. Mit der verallgemeinerten Van-

[2] c für engl. *coarse* und f für engl. *fine*.
[3] das ist $x_i^c = \chi(\xi_i^c)$, $1 \le i \le n_c$ mit Abbildung χ aus (4.14) und den Fekete-Punkten $\boldsymbol{\xi}^c = \left\{\xi_1^c, \ldots, \xi_{n_c}^c\right\} \subset T_0$ für Polynomgrad N_c.

dermonde-Matrix V_{ξ^c} für $\boldsymbol{\xi}^c = \boldsymbol{\xi}$ aus (4.3) und den Dubiner-Polynomen g_k aus Abschnitt 3.1 gelangen wir mittels (4.11) zu

$$P = \tilde{V} V_{\xi^c}^{-1} \quad \text{und} \quad \tilde{V} = \left[g_j(\xi_i^f) \right]_{ij} \in \mathbb{R}^{n_f \times n_c}. \tag{6.6}$$

Somit ist die Prolongationsmatrix einfach berechenbar. Sie ist auf jedem Element gleich und kann elementweise analog zur Assemblierung (4.36) angewendet werden.

6.3.2 Restriktion

Die *Restriktion* stellt als Gegenstück zur Prolongation den Übergang vom groben zum feinen Gitter dar. Im Gegensatz zur Prolongation gibt es hierbei keine naheliegende Wahl. Wir betrachten daher vier Restriktionsstrategien.

Die ersten drei Varianten basieren auf der Überlegung, zu den aus den Funktionswerten bestehenden Vektor $\mathbf{u}^f \in \mathbb{R}^{n_f}$ die Interpolierende[4] $I_f \mathbf{u}^f \in \mathcal{P}_{N_f}(T_k)$ zu bestimmen und anschließend die Koeffizienten bzgl. der auf T_k transformierten Dubiner-Basis $\left\{ \tilde{g}_1, \ldots, \tilde{g}_{n_f} \right\}$, wobei $\tilde{g}_i(x) = g_i(\chi^{-1}(x))$ für $x \in T_k$, zu beeinflussen. Der Koeffizientenvektor $\hat{\mathbf{u}}^f$ resultiert aus (4.7) zu

$$\hat{\mathbf{u}}^f = V_{\xi^f}^{-1} \mathbf{u}^f$$

mit der verallgemeinerten Vandermonde-Matrix V_{ξ^f} zu $\boldsymbol{\xi}^f$.

Nun wird der Koeffizientenvektor durch Multiplikation mit einer Diagonalmatrix $Q \in \mathbb{R}^{n_f \times n_f}$, welche abhängig von der Restriktion gewählt ist, angepasst. Der Vorteil liegt hier in der gezielten Dämpfung hoch- (oder niederfrequenter) Anteile, die sich als Koeffizienten der hierarchischen Dubiner-Basisfunktionen äußern.

Schließlich wird die neue Funktion, die eine Linearkombination aus der transformierten Dubiner-Basis und dem veränderten Koeffizientenvektor ist, in den groben Gitter $\mathbf{x}^c$ ausgewertet, so dass wir zu dem restringierten Vektor $\mathbf{u}^c$ gelangen. Zusammenfassend gilt die Darstellung für die ersten drei Restriktionen:

$$\mathbf{u}^c = R \mathbf{u}^f \quad \text{und} \quad R = \breve{V} Q V_{\xi^f}^{-1} \in \mathbb{R}^{n_c \times n_f} \tag{6.7}$$

mit der verallgemeinerten Vandermonde-Matrix $V_{\xi^f} \in \mathbb{R}^{n_f \times n_f}$, der Matrix $\breve{V} = [g_j(\xi_i^c)]_{ij} \in \mathbb{R}^{n_c \times n_f}$ und einer Diagonalmatrix $Q = \operatorname{diag}(q_1, \ldots, q_{n_f}) \in \mathbb{R}^{n_f \times n_f}$.

[4] in Analogie zu (6.4)

Restriktion durch Interpolation

Wird die Einheitsmatrix für Q gewählt, bedeutet das keine Veränderung des Koeffizientenvektors, so dass letztlich die Interpolationsfunktion $I_f \mathbf{u}^f$ auf den groben Gitter ausgewertet wird. So gilt hier analog zur Prolongation (6.5): $R = \left[\varphi_j^f(x_i^c)\right]_{ij}$.

Allerdings ist zu erwarten, dass die Auflösung der hochfrequenten Anteile auf einem groben Gitter, insbesondere durch Interpolation, nicht ausreichend ist.

Restriktion durch Projektion

Als alternative Idee zur Interpolations-Restriktion werden in dieser Variante alle Anteile, die nicht auf dem groben Gitter dargestellt werden können, abgeschnitten. Dies wird realisiert durch die Festlegung der Diagonaleinträge von Q:

$$q_i = \begin{cases} 1, & \text{für } \ 1 \leq i \leq n_c, \\ 0, & \text{für } \ n_c + 1 \leq i \leq n_f. \end{cases}$$

Folglich gehen nur die Anteile in den restrigierten Vektor mit ein, die zu einem Dubiner-Polynom gehören, dessen Polynomgrad $\leq n_c$ ist.

Restriktion durch (Cosinus-)Filterung

Die dritte Variante präsentiert einen Mittelweg der beiden ersten Methoden, indem eine Gewichtung vorgestellt wird, die die Fehleranteile entsprechend ihrer Frequenz berücksichtigt:

$$q_i = \frac{1}{2}\left(\cos\left(\frac{N(i)}{N_f}\pi\right) + 1\right) \quad \text{mit} \ \ N(j) = \deg(g_i) \ \ 1 \leq i \leq n_f.$$

Die Diagonaleinträge sind eine monoton fallende, endliche Folge $1 = q_0 \geq q_1 \geq \ldots \geq q_{n_f} = 0$. Dies bewirkt bei Anwendung der Restriktion eine kontinuierlich ansteigende Dämpfung der höher werdenden Frequenzen, die zur Funktion des Vektors $\mathbf{u}^f$ gehören.

Restriktion durch Transposition

Die vierte Variante der Restriktion wird definiert als Transposition der Prolongationsmatrix

$$R = P^T = \left[\varphi_i^c\left(x_j^f\right)\right]_{ij} \in \mathbb{R}^{n_f \times n_c}.$$

Diese von Hackbusch [9] als *kanonische Restriktion* bezeichnete Matrix wird häufig, beispielsweise in [1, § 5.1], als einfache Wahl der Restriktion gewählt.

Die i-te grobe Basisfunktion φ_i^c lässt sich darstellen als Linearkombination der Fein-Gitter-Basis in Form von $\varphi_i^c = \sum_{j=1}^{n_f} \varphi_i^c\left(x_j^f\right)\varphi_j^f$. So gilt in Zusammenhang mit der Variationsformulierung

$$\left(I_f \mathbf{u}^f, \varphi_i^c\right)_{T_k} = \sum_{j=1}^{n_f} R_{ij}\left(I_f \mathbf{u}^f, \varphi_j^f\right)_{T_k},$$

was die Wahl der Restriktion durch Transposition untermauert.

6.4 Aufstellen der Grobgitter-Matrizen

Im Verlauf des Mehrgitterverfahrens wird der Defekt d der berechneten Lösung auf ein gröberes Gitter Ω_j mit $1 \leq j < l$ restringiert. Dort wird, wie in Abschnitt 6.5 genauer dargelegt, für $j > 1$ wieder das Mehrgitterverfahren angewandt oder im Fall $j = 1$ auf dem gröbsten Gitter direkt gelöst.

In beiden Fällen benötigen wir eine Matrix $A_j \in \mathbb{R}^{n_j \times n_j}$, die konsistent zu der Matrix A_{j+1} auf der nächst-feineren Stufe ist. Pasquetti und Rapetti [13] präsentieren zwei Varianten, die auch in [9, § 3.7] vorgeschlagen wurden.

6.4.1 Direktes Aufstellen

Die Grobgittermatrizen $A_j \in \mathbb{R}^{n_j \times n_j}$, $1 \leq j < l$, werden ebenso aufgestellt wie die Matrix A_l des zu lösenden Gleichungssystems $A_l \mathbf{u}_l = \mathbf{b}_l$ auf dem feinsten Gitter.

Somit wird für jeden Polynomgrad N_j des Mehrgitterverfahrens die Matrix A_j entsprechend der Galerkinformulierung (4.31) zu einer Matrix der Form (4.37) assembliert. Die rechte Seite der Gleichungssysteme wird hingegen nur auf dem feinsten Gitter benötigt, da auf den tieferen Stufen die rechte Seite im Laufe der Mehrgitteriteration aus der Restriktion des Defekts gebildet wird.

Ein Vorteil dieser Variante ist die Unabhängigkeit des Aufbaus der Matrizen A_j von den anderen Stufen $k \neq j$. Allerdings muss in jeder Stufe j eine komplette Assemblierung der Hilfsmatrix A_j durchgeführt werden.

Das direkte Aufstellen wurde u.A. in [17] für das spektrale p-Mehrgitterverfahren eingesetzt.

6.4.2 Aggregationsmethode

Ein alternatives Aufbauen der Grobgitter-Matrizen wird durch *Aggregation* erreicht.

Im Verlauf der Assemblierung der Matrix A_l auf dem feinsten Gitter, werden hierbei die Hilfsmatrizen A_j auf den gröberen Gittern für $j = d-1, d-2, \ldots, 1$ sukzessive aufgestellt. So wird während der Assemblierung der zu Dreieck T_k gehörenden Steifigkeitselementmatrix $S_k^l = S_k$ in (4.35) die Elementmatrizen S_k^j auf den tieferen Stufen berechnet. Dies geschieht durch

$$S_k^j = R_{j+1} S_k^{j+1} P_j, \quad \text{für } j = l-1, l-2, \ldots, 1$$

wobei die Prolongationsmatrix P_j und die Restriktionsmatrix R_{j+1} der Stufe j entsprechend und nur auf einem Element operierend gewählt wurden.

Erst danach werden die Randbedingungen auf die Steifigkeitselementmatrix S_k und auf die sogenannten Galerkin-Produkte $R_{j+1} S_k^j P_j$ angewandt und in die Gesamtmatrizen A_l bzw. A_j assembliert.

Ein Nachteil dieser Methode ist die Abhängigkeit von der Feingittermatrix A_l.

Wie in [13, S. 672] hervorgehoben, ist die Wahl der Galerkin-Produkte verknüpft mit der Restriktion durch Transposition.

6.5 Mehrgitteralgorithmus

Aufbauend auf Abschnitt 6.2 wird im Folgenden der Ablauf eines l-stufigen Mehrgitterverfahrens kurz vorgestellt.

Zu den Polynomgraden $N_1 < \ldots < N_l$ betrachten wir die Gitter $\Omega_j = \Omega_{N_j}$ mit $n_j = n_{N_j}$ Freiheitsgraden. Um das Verfahren auf das lineare Gleichungssystem

$$A_l \mathbf{u}_l = \mathbf{b}_l$$

des feinsten Gitters anzuwenden, werden die Hilfsmatrizen $A_j \in \mathbb{R}^{n_j \times n_j}$ für $j = l-1, l-2, \ldots, 1$ auf den gröberen Gittern nach Abschnitt 6.4 gewählt.

Ebenso werden die Prolongationsmatrix $P_j \in \mathbb{R}^{n_{j+1} \times n_j}$ und die Restriktionsmatrix $R_j \in \mathbb{R}^{n_{j-1} \times n_j}$ entsprechend Abschnitt 6.3 bestimmt.

Wir bezeichnen die ν-fache Anwendung des SOR-Glätters auf dem Gitter Ω_j mit $\mathcal{S}_j^\nu$.

Wir betrachten zunächst einen Zyklus der Mehrgitteriteration.

Mehrgitteriteration MG$(l, \mathbf{u}_l^k, \mathbf{b})$

$(k+1)$-ter Zyklus auf der Ebene l mit Startwert $\mathbf{u}_l^k$:

1. *Vorglättung:*

   Führe ν Glättungsschritte durch:

   $$\mathbf{u}_l^{k,1} = \mathcal{S}_l^\nu \mathbf{u}_l^k$$

2. *Grobgitterkorrektur:*

   Berechne den Defekt $\mathbf{d}_l = \mathbf{b} - A_l \mathbf{u}_l^{k,1}$ und die Restriktion $\mathbf{b}_{l-1} = R_l \mathbf{d}_l$.
   Bestimme zu

   $$A_{l-1}\tilde{\mathbf{v}}_{l-1} = \mathbf{b}_{l-1}$$

   i) bei $l = 2$ die exakte Lösung und setze $\mathbf{v}_{l-1} = \tilde{\mathbf{v}}_{l-1}$,

   ii) bei $l > 2$ eine Näherung der Lösung durch γ Schritte der Mehrgitteriteration MG$(l-1, \mathbf{v}_{l-1}^i, \mathbf{b}_{l-1})$ auf Stufe $l-1$ mit Startwert $\mathbf{v}_{l-1}^0 = 0 \in \mathbb{R}^{n_{l-1}}$.
   Setze $\mathbf{v}_{l-1} = \mathbf{v}_{l-1}^\gamma$.

   Setze daraufhin

   $$\mathbf{u}_l^{k,2} = \mathbf{u}_l^{k,1} + P_{l-1}\mathbf{v}_{l-1}.$$

3. *Nachglättung:*

   Führe ν Glättungsschritte durch:

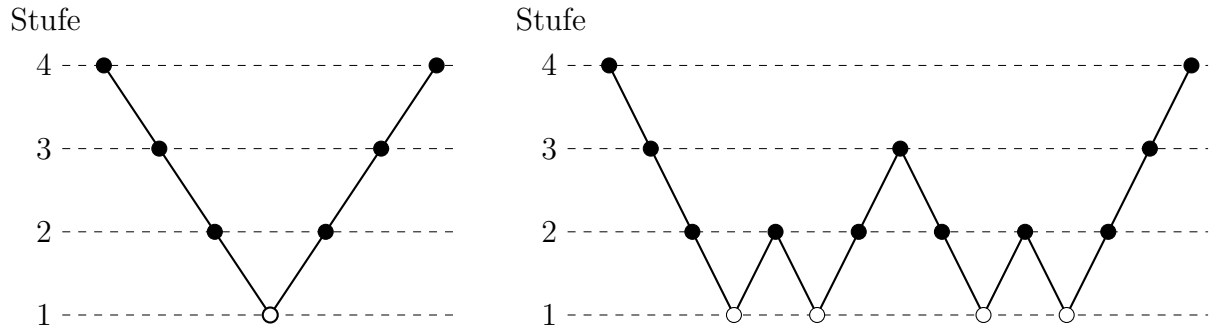
   $$\mathbf{u}_l^{k+1} = \mathcal{S}_l^\nu \mathbf{u}_l^{k,2}$$

In jeder Iteration wird auf der feinsten Stufe mit dem SOR-Verfahren jeweils ν-mal vor- und nachgeglättet.

Abhängig von γ hat der obige Algorithmus verschiedene Abläufe. Für $\gamma = 1$ ergibt sich der *V-Zyklus* und für $\gamma = 2$ der *W-Zyklus*. Die nach der Form der Ablaufdiagramme benannten Zyklen sind in Abbildung 6.1 veranschaulicht.

Um das Mehrgitterverfahren anzuwenden, wird der Startvektor $\mathbf{u}_l^0 = 0 \in \mathbb{R}^{n_l}$ auf dem feinsten Gitter gewählt und die Iteration MG$(l, \mathbf{u}_l^k, \mathbf{b}_l)$ für $k = 0, 1, 2, \ldots$ so lange ausgeführt, bis das Residuum eine vorgegebene Schranke ε unterschreitet, d. h.

$$\left\| r^k \right\| = \left\| A_l \mathbf{u}_l^k - \mathbf{b}_l \right\| < \varepsilon.$$



(a) Der V-Zyklus ($\gamma = 1$)

(b) Der W-Zyklus ($\gamma = 2$)

Abbildung 6.1: Ablaufdiagramme einer Mehrgitteriteration auf $l = 4$ Gitterstufen für a) den V-Zyklus und b) den W-Zyklus. In den schwarzen Kreisen wird geglättet und in den weißen direkt gelöst.

Wir werden in Kapitel 7 als Abbruchkriterium $\left\|r^k\right\|_\infty < \varepsilon = 1.0e-7$ wählen.

7 Numerische Ergebnisse

In diesem Kapitel werden die zuvor beschriebenen Verfahren in der Praxis getestet.

Zunächst wird die Friedrichs-Keller-Triangulierung vorgestellt, die häufig als Muster-triangulierung quadratischer Grundgebiete eingesetzt wird. Es folgt eine Analyse des von uns vorgestellten Mehrgitterverfahrens mit statischer Kondensation, um günstige Parameter im Mehrgitterverfahren zu erhalten. Das so optimierte Verfahren wird auf unterschiedliche Randwertaufgaben angewendet und mit den Mehrgitterverfahren aus [13] verglichen.

Die Berechnung der Matrixkonditionen erfolgt in MATLAB durch Aufruf der Funktion *condest* für dünnbesetzte Matrizen.

Für Polynomgrad $N \in \mathbb{N}$ benutzen wir wie bei QSEM Quadraturformeln die zumindest Polynome vom Grad $2N-1$ exakt integrieren. Dabei bevorzugen wir die Quadraturpunkte aus Abschnitt 3.3.2. Nur für die Polynomgrade $N = 15$ und $N = 18$ werden die Quadraturpunkte aus Abschnitt 3.3.1 verwendet, da in [20] für derart hohe Polynomgrade keine Formeln berechnet werden konnten.

Das Mehrgitterverfahren wird abgebrochen, wenn für das Residuum gilt:

$$\|\text{res}\| < 1.0e-7$$

7.1 Friedrichs-Keller-Triangulierung

Die *Friedrichs-Keller-Triangulierung* bzw. *Quadratgittertriangulierung* stellt eine gleichmäßige, strukturierte Triangulierung des quadratischen Grundgebiets
$\Omega = [-1,1] \times [-1,1]$ dar.

Sei $Y \in \mathbb{N}$ gegeben. Die Triangulierung entsteht durch Zerlegung von Ω in Y^2 Quadrate der Seitenlänge $h = 2/Y$, die in zwei Dreiecke geteilt werden. Wir erhalten $M = 2Y^2$ kongruente, rechtwinklige Dreiecke. Die Zerlegung besitzt $M_V = (Y+1)^2$ Eckpunkte, sowie $M_E = 3Y^2 + 2Y$ Kanten. Ist $N \in \mathbb{N}$ der Polynomgrad der Fekete-Punkte auf jedem Element, so gibt es insgesamt $n = (NY+1)^2$ Punkte.

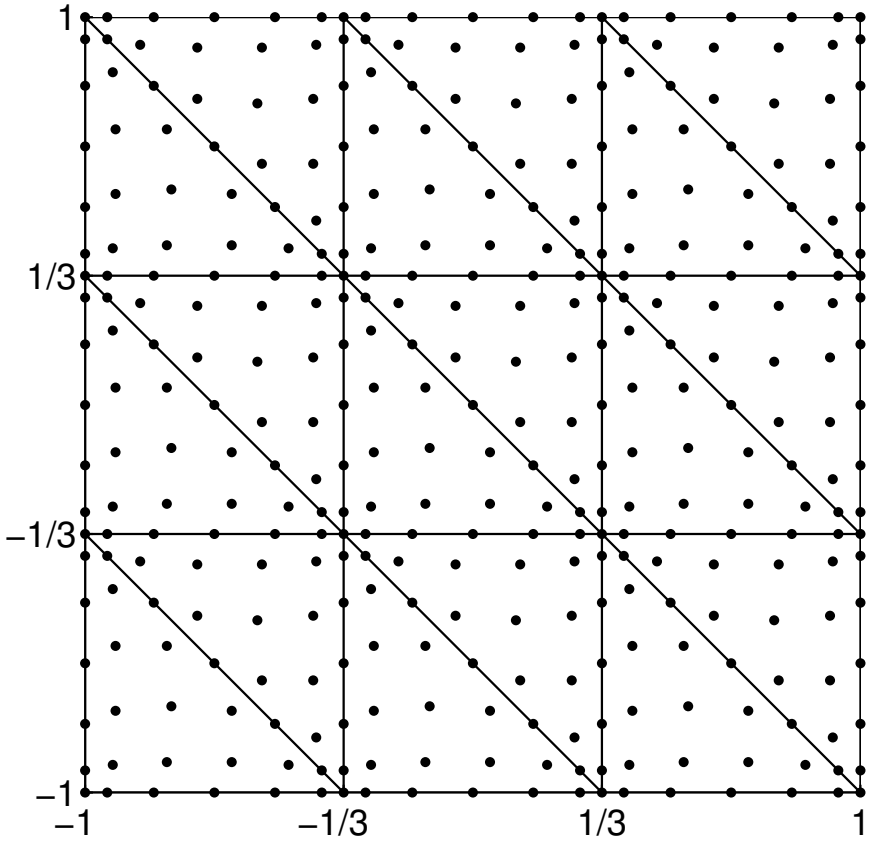


Abbildung 7.1: Triangulierung FKT_3 mit globalen Fekete-Knoten-Punkten für $N = 6$

Wir werden diese Gebietszerlegung mit FKT_Y für ein $Y \in \mathbb{N}$ bezeichnen.

In Abbildung 7.1 ist beispielhaft die Zerlegung FKT_3 mit den Fekete-Punkten für Polynomgrad $N = 6$ illustriert.

7.2 Konvergenzanalyse des Mehrgitterverfahrens mit statischer Kondensation

In diesem Abschnitt soll die Konvergenz des TSEM-Mehrgitterverfahrens untersucht werden, welches angewendet wird auf ein lineares Gleichungssystem, dass zuvor, wie im Kapitel 5 erörtert, durch die statische Kondensation verändert wurde.

Bei der Beschreibung der Mehrgitterverfahren haben wir verschiedene Restriktionsvarianten vorgestellt. Ebenfalls gibt es zwei Möglichkeiten die Grobgitter-Matrizen aufzustellen. Des Weiteren muss neben der Wahl des Glätters (Parameter ω im SOR-Verfahren) die optimale Anzahl der Glättungsschritte, die jeweils vor und nach der

Grobgitter-Korrektur durchgeführt werden, für diese Variation des Mehrgitterverfahrens ermittelt werden.

Das Mehrgitterverfahren ohne statische Kondensation wurde von Dolean et al. [5], sowie von Pasquetti and Rapetti [13] bereits untersucht. Wir konzentrieren uns daher auf die Variante mit statischer Kondensation und vergleichen die Werte miteinander.

7.2.1 Wahl der Restriktion und der Grobgittermatrix

Wir beginnen damit, eine passende Restriktionsstrategie zu finden.

Als Beispielproblem wird die folgende Helmholtzaufgabe mit homogener Dirichlet-Randbedingung behandelt:

$$-\Delta u + u = f \qquad \text{in } \Omega = (-1,1)^2 \tag{7.1}$$

$$u = 0 \qquad \text{auf } \partial\Omega \tag{7.2}$$

Die rechte Seite sei mit $f(x,y) = (2\pi^2 + 1)u(x,y)$ passend zur exakten Lösung

$$u_e(x,y) = \sin(\pi x)\sin(\pi y)$$

gewählt. Das Gebiet Ω sei durch die Friedrichs-Keller-Triangulierung FKT_2 mit insgesamt 8 Elementen zerlegt.

Wir betrachten die vier Restriktionstrategien aus Abschnitt 6.3.2 nacheinander. Die Grobgittermatrix wird jeweils durch das direkte Aufstellen, wie in 6.4.1 beschrieben, ermittelt.

Die Aggregationsmethode zur Bestimmung der Grobgittermatrix wird nur in Verbindung mit der Transposition als Restriktionsstrategie verwendet. Somit werden diese Kombinationen untersucht:

I-D Restriktion durch Interpolation; direktes Aufstellen

P-D Restriktion durch Projektion; direktes Aufstellen

F-D Restriktion durch Filterung; direktes Aufstellen

T-D Restriktion durch Transposition; direktes Aufstellen

T-A Restriktion durch Transposition; Aggregationsmethode

Bei dieser Versuchsreihe werden V-Zyklen und jeweils 4 Gauß-Seidel-Iterationen zur Vor- und Nachglättung benutzt.

Die Entwicklungen der Residuen in Abhängigkeit von der Anzahl der Glättungsiterationen sind in den Abbildungen 7.2 dargestellt. Hier und im Folgendem bezeichnen wir beispielsweise das Mehrgitterverfahren auf den Gittern mit den Polynomgraden $N_1 = 3, N_2 = 6, N_3 = 12$ mit Mehrgitter(3,6,12) bzw. MG(3,6,12).

Auffällig ist die sehr gute Konvergenzrate der Kombinationen T-D und T-A. Bei der Wahl der Restriktion durch Transposition spielt es so gut wie keine Rolle, mit welchen Verfahren die Grobgittermatrix aufgestellt wurde. Die Aggregationsmethode scheint leicht schlechter zu konvergieren bei Vorkonditionierung mit statischer Kondensation. Dies liegt wahrscheinlich an der fehlenden Symmetrie.

Daraus folgt unmittelbar, dass wir nur noch die Restriktionsstrategie durch Transposition verwenden werden. Auch beschränken wir uns auf die Variante des direkten Aufstellens der Grobgittermatrizen.

7.2.2 Wahl des Glätters

Nachdem die Wahl der Restriktion und der Grobgittermatrix feststeht, soll in diesem Abschnitt der Relaxationsparameter ω des SOR-Verfahrens und die Anzahl der Glättungsschritte vor und nach einem Gitterübergang möglichst optimal gewählt werden.

Hierzu folgen wir der Methode aus [5, § 3.2]. Demnach wird das Residuum $\mathbf{r}^n$ in einem V-Zyklus eines Zweigitterverfahrens mit jeweils ν Vor- und Nachglättungen verkleinert zu

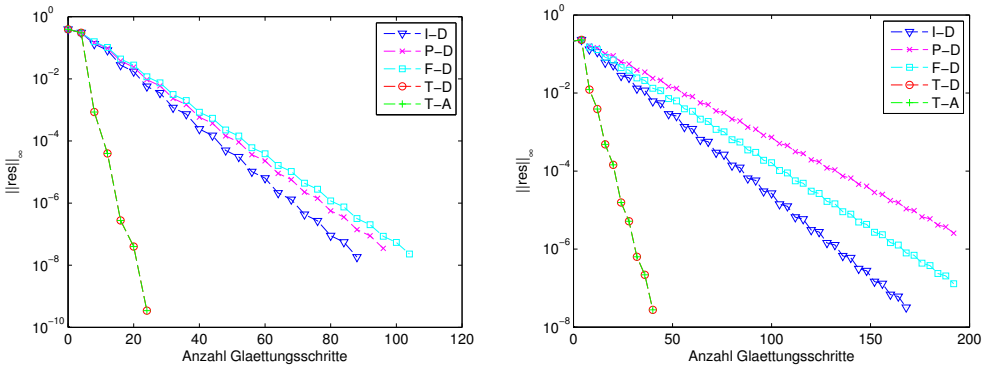
$$\begin{aligned} \mathbf{r}^{n+2\nu+1} &= T\mathbf{r}^n \\ \text{mit} \quad T &= A_f(N_\omega^{-1}M_\omega)^\nu A_f^{-1}(I - A_f P A_c^{-1} R)A_f(N_\omega^{-1}M_\omega)^\nu A_f^{-1}, \end{aligned}$$

wobei $A_f = N_\omega - M_\omega$ die Zerlegung der Feingittermatrix aus (6.2b) und (6.2c) ist, P und R die Prolongations- und Restriktionsmatrizen sind und A_c die Grobgittermatrix darstellt.

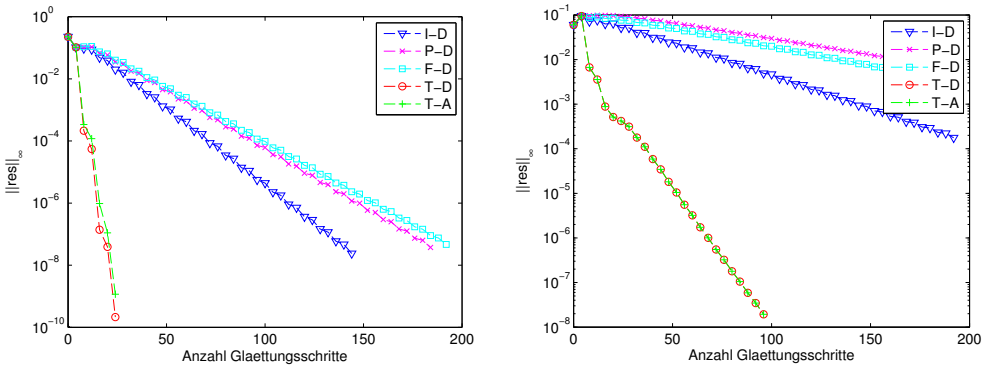
Daraus folgt die Abschätzung

$$\begin{aligned} \left\|\mathbf{r}^{2\nu+1+n}\right\| &\leq \|T\| \, \|\mathbf{r}^n\| \equiv \rho^{2\nu+1} \, \|\mathbf{b}^n\| \,, \\ \rho(\omega,\nu) &= \|T\|^{1/(2\nu+1)} \,. \end{aligned}$$

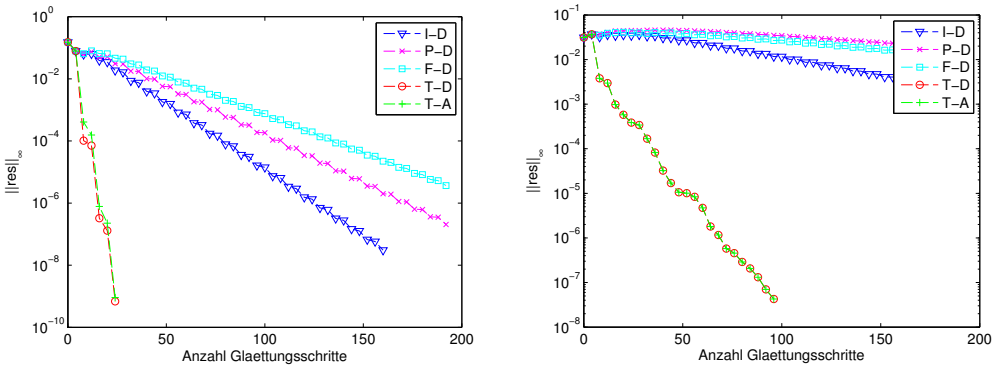
Der Wert $\rho(\omega, \nu)$ ist eine Näherung des Spektralradius, wobei die Arbeit durch die Glät-



(a) Mehrgitter(3,6)



(b) Mehrgitter(3,6,12)



(c) Mehrgitter(3,6,12,18)

Abbildung 7.2: Residuenplots zu den Restriktionsstrategien auf unterschiedlichen Gittern; links: mit statischer Kondensation, rechts: ohne statische Kondensation

ter nivelliert wurde.

Die Konturlinien des Mehrgitterverfahren mit statischer Kondensation auf der Triangulierung FKT_2 sind für unterschiedliche Gitter in Abbildung 7.3 festgehalten. Es zeigt sich, dass das Gauß-Seidel-Verfahren, d. h. das SOR-Verfahren für $\omega = 1$, die besten Werte liefert. Es empfiehlt sich, die Anzahl der Glättungsiterationen auf $\nu = 4$ je Vor- bzw. Nachglättung zu belassen.

Ähnliche Ergebnisse liefern die Darstellungen der Konturlinien für das Mehrgitterverfahren ohne Kondensation, welche in [5] abgebildet sind.

7.3 Dirichlet-Problem auf dem Einheitsquadrat

Gegeben sei die Poissongleichung

$$-\Delta u = f \qquad \text{in } \Omega = (-1,1)^2, \tag{7.3a}$$

$$u = g \qquad \text{auf } \partial\Omega. \tag{7.3b}$$

mit exakter Lösung

$$u_e = \sin(2x+y)\sin(x-1)\sin(1+y)$$

und passender rechter Seite f und g. Der Plot der Lösung ist in Abbildung 7.4 illustriert.

Wir betrachten zunächst die Friedrichs-Keller-Triangulierung FKT_{10} bestehend aus 200 Dreiecken. Nach Abschnitt 4.4 erhalten wir das lineare Gleichungssystem

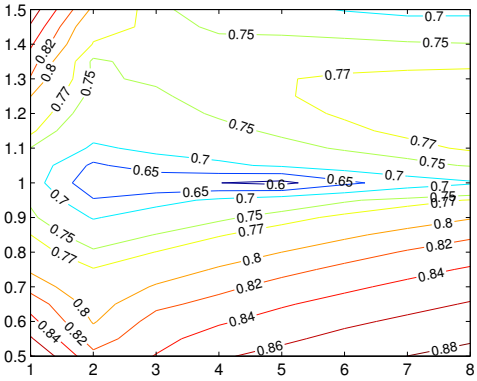
$$A_G \mathbf{u} = \mathbf{b}_G. \tag{7.4}$$

Der Index G soll andeuten, dass das System (7.4) direkt aus dem Galerkin-Verfahren entstammt. Im Gegensatz dazu bezeichnen wir das lineare Gleichungssystem, das sich anschließend durch die Anwendung der statischen Kondensation ergibt mit

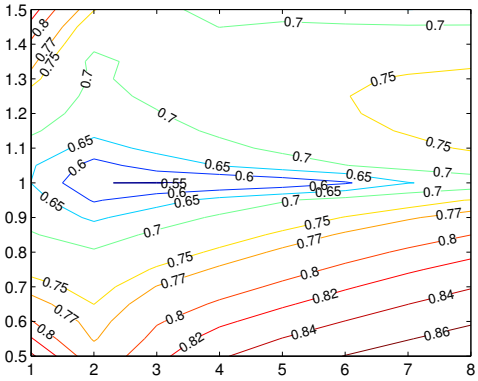
$$A_{SK} \mathbf{u} = \mathbf{b}_{SK}. \tag{7.5}$$

Die Konditionen der Matrizen sind in Abbildung 7.5 festgehalten. Hierbei fällt auf, dass die Kondition durch die statische Kondensation deutlich gesenkt werden konnte.

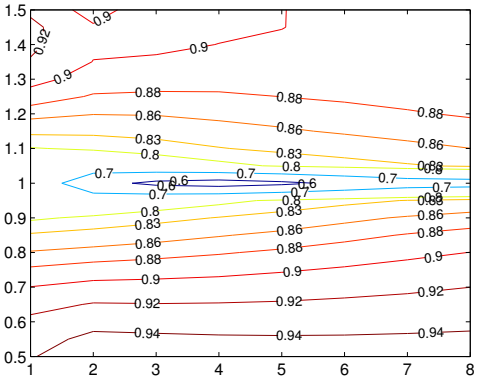
Wir wenden nun das Mehrgitterverfahren bestehend aus V-Zyklen und jeweils 4 Gauß-Seidel-Iterationen zur Vor- und Nachglättung auf die beiden Systeme an. Mit MG_G bzw. MG_{SK} bezeichnen wir die Anwendung des Mehrgitterverfahrens auf (7.4) bzw. (7.5).



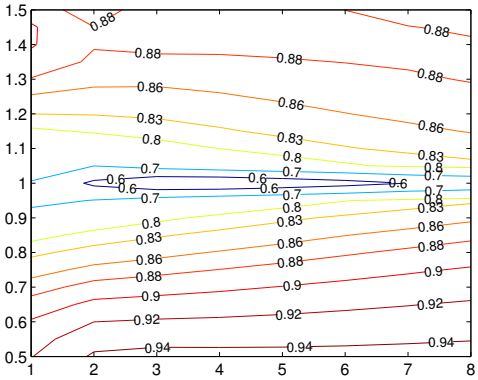
(a) Gittergrade: $N_c = 3$, $N_f = 6$



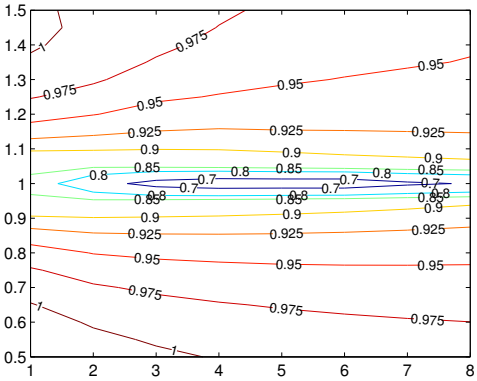
(b) $N_c = 3$, $N_f = 6$



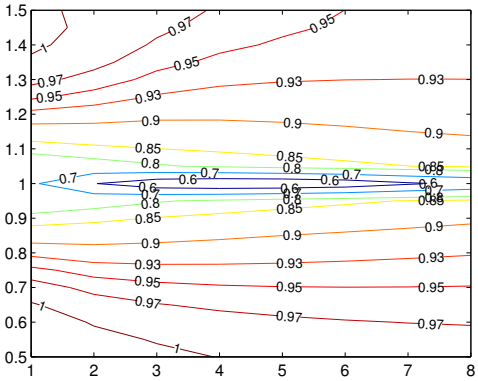
(c) $N_c = 6$, $N_f = 12$



(d) $N_c = 6$, $N_f = 12$



(e) $N_c = 9$, $N_f = 18$



(f) $N_c = 9$, $N_f = 18$

Abbildung 7.3: Konturlinien zu $\rho(w, m)$ auf FKT_2 bzgl. der Supremumsnorm $\|\cdot\|_\infty$ (links) und der Euklidischen Norm $\|\cdot\|_2$ (rechts)

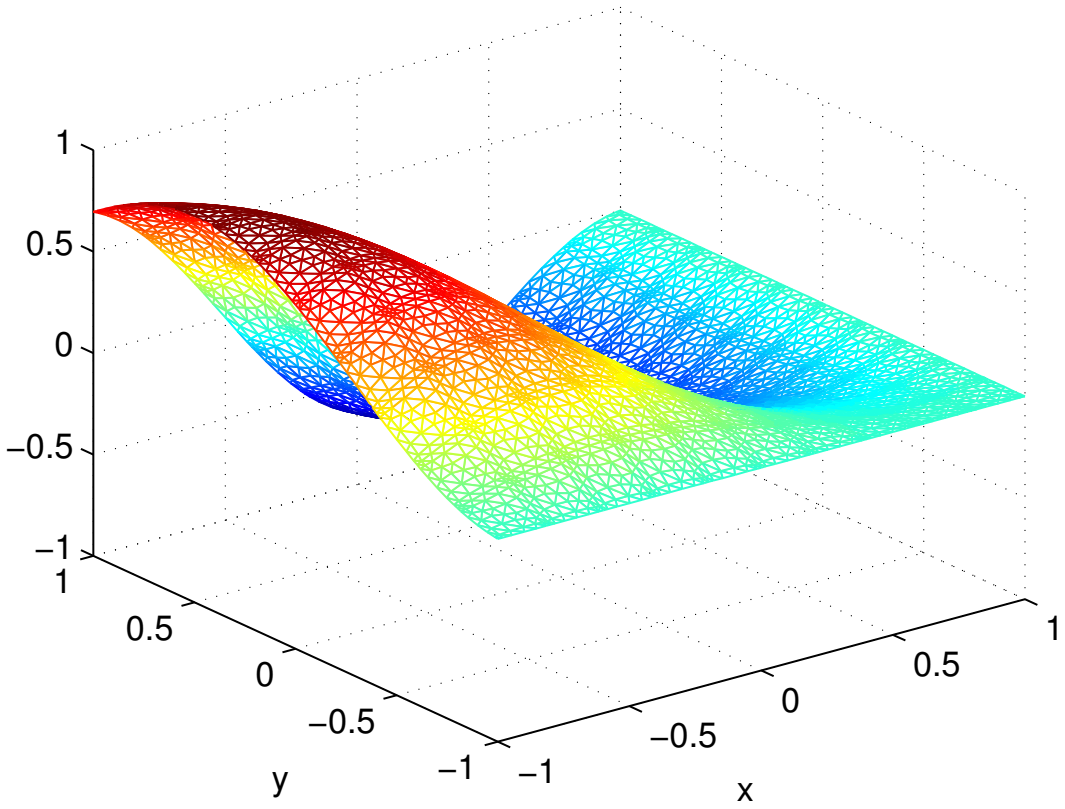


Abbildung 7.4: Plot von u_e auf Triangulierung FKT_{10} für $N = 6$

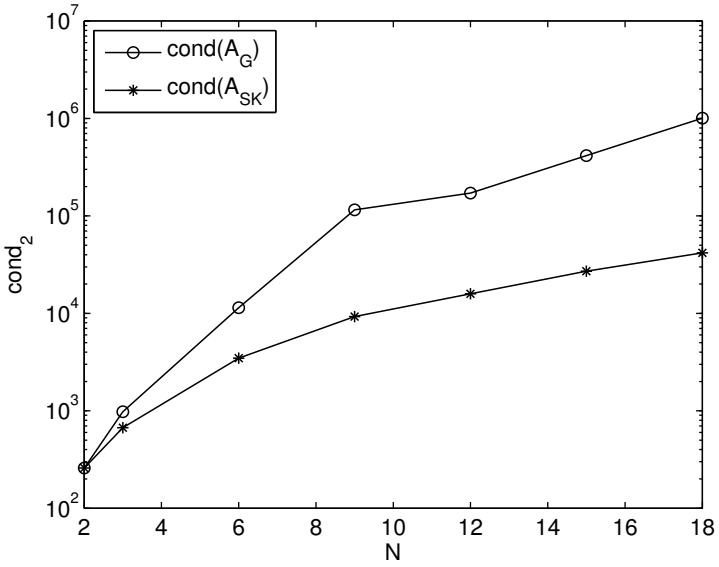


Abbildung 7.5: Kondition von A_{SK} und A_G für FKT_{10}

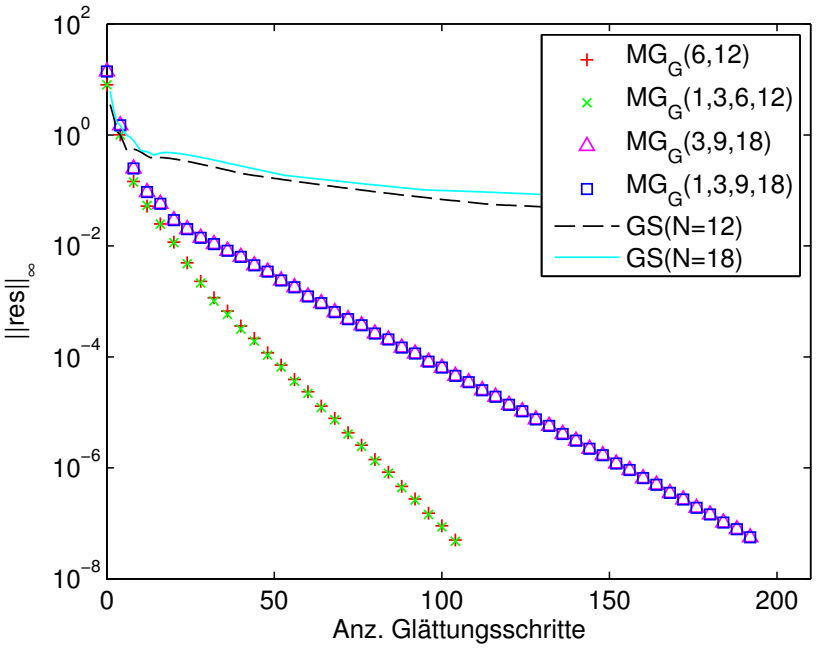
Abbildung 7.6 zeigt die Entwicklung der Residuen auf unterschiedlichen Gittern für beide Mehrgittervarianten. Der Vergleich zum Gauß-Seidel-Glätter demonstriert den Wirkungsgrad der Mehrgitterverfahren. Zu beachten ist die unterschiedliche Skalierung der Iterationsachsen. In allen Fällen ist das Verfahren mit statischer Kondensation dem nicht vorkonditionierten Mehrgitterverfahren deutlich überlegen in der Konvergenzrate.

Außerdem ist ersichtlich, dass die Verwendung mehrerer Gitter bis zur Stufe $N = 1$, kaum Einfluss auf die Konvergenzrate hat. Lediglich das MG_{SK} zeigt leichte Einbußen, die allerdings wegen der hohen Konvergenz innerhalb des nächsten V-Zyklus beseitigt sind.

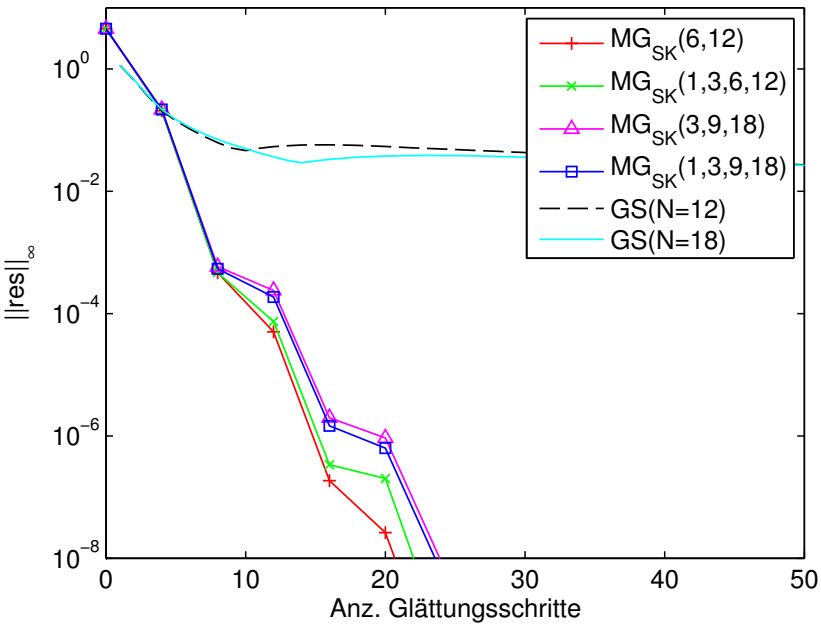
Offensichtlich stellt sich das Mehrgitterverfahren MG_{SK} in der Anzahl an Glätteriterationen wesentlich besser dar als MG_G. Zuvor muss allerdings A_{SK} aus A_G berechnet werden. Um die Matrix in (5.5) zu ermitteln, ist es insbesondere erforderlich, A_i aus (5.5) zu invertieren. Dies sind rechenintensive Operationen.

Daher werden in Tabelle 7.1 die Zeiten angegeben, die für das Mehrgitterverfahren und - im Falle von A_{SK} - für die statische Kondensation benötigt werden. Die Assemblierungszeit wird außer Acht gelassen, da sie in beiden Verfahren identisch abläuft. Wir stellen fest, dass MG_{SK} weniger Rechenzeit benötigt als MG_G. Der Großteil der CPU-Zeit wird für die Berechnung der statischen Kondensation benötigt.

Da die Residuen in Abbildung 7.6 gleichmäßig bzgl. der Logarithmusskala abnehmen, kann die Konvergenzrate ρ gut approximiert werden. Bei der Berechnung der Konver-



(a) Das Mehrgitterverfahren MG_G ohne statische Kondensation



(b) Das Mehrgitterverfahren MG_{SK} mit statischer Kondensation

Abbildung 7.6: Residuenentwicklung der Mehrgitterverfahren und der Gauß-Seidel-Glätter

Methode	MG(6,12)	MG(1,3,6,12)	MG(3,9,18)	MG(1,3,9,18)
MG_G	2.061	2.037	9.326	9.837
MG_{SK}+SK	0.651	0.626	2.304	2.414
-davon nur SK	0.280	0.293	1.801	1.790

Tabelle 7.1: Gegenüberstellung der benötigten CPU-Zeiten (in Sekunden) für das Berechnen der Mehrgitterverfahren und der statischen Kondensation

Methode	MG(6,12)	MG(1,3,6,12)	MG(3,9,18)	MG(1,3,9,18)
MG_G	0.8338	0.8335	0.9042	0.9042
MG_{SK}	0.3581	0.3861	0.4337	0.4269

Tabelle 7.2: Approximierte Konvergenzrate ρ zu den Ergebnissen aus Abbildung 7.6

genzrate gehen wir davon aus, dass sich das Residuum $\mathbf{r}_i$ nach i Glättungsiterationen durch

$$\|\mathbf{r}_i\|_\infty \approx \rho^i \|\mathbf{r}_0\|_\infty$$

annähern lässt. Wir erhalten

$$\rho \approx \frac{\|\mathbf{r}_i\|_\infty}{\|\mathbf{r}_0\|_\infty}^{1/i}.$$

Die so ermittelten Konvergenzraten sind in Tabelle 7.2 aufgeführt.

7.4 Randwertproblem auf unstrukturiertem Gitter

In diesem Abschnitt werden wir die Robustheit des Mehrgitterverfahrens bzgl. des Gitters und der Randbedingungen überprüfen. Wir betrachten ein polygonales Gebiet Ω, welches in Abbildung 7.7 dargestellt ist. Die abgebildete Triangulierung umfasst 235 Elemente. Das innere Randstück werde mit $\Gamma_\mathcal{D}$ und das äußere Randstück mit $\Gamma_\mathcal{N}$ bezeichnet.

Wir betrachten die Helmholtzaufgabe

$$-\Delta u + u = f \qquad \text{in } \Omega, \tag{7.6a}$$

$$u = 0 \qquad \text{auf } \Gamma_\mathcal{D}, \tag{7.6b}$$

$$\frac{\partial u}{\partial \nu} = 0 \qquad \text{auf } \Gamma_\mathcal{N}, \tag{7.6c}$$

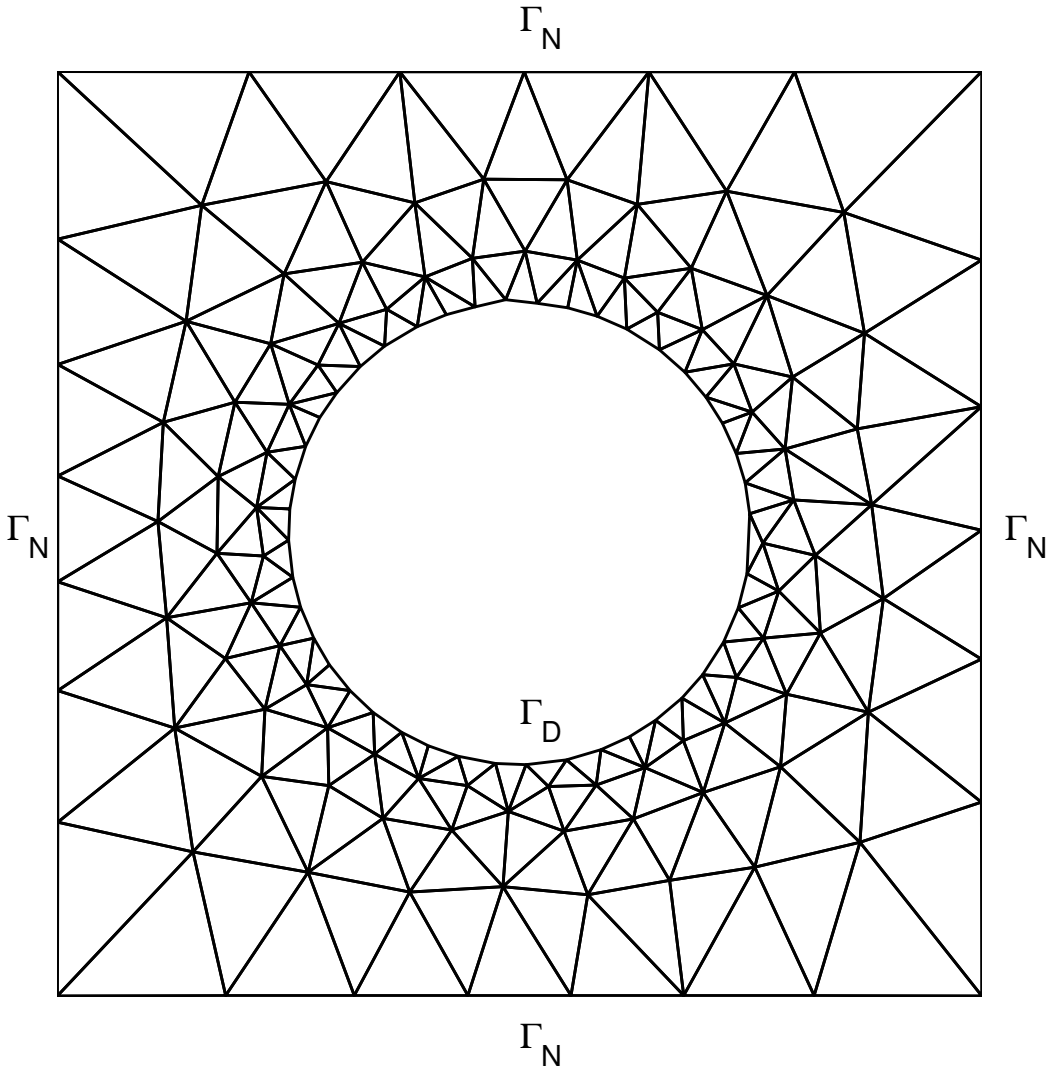


Abbildung 7.7: Triangulierung des Gebiets Ω

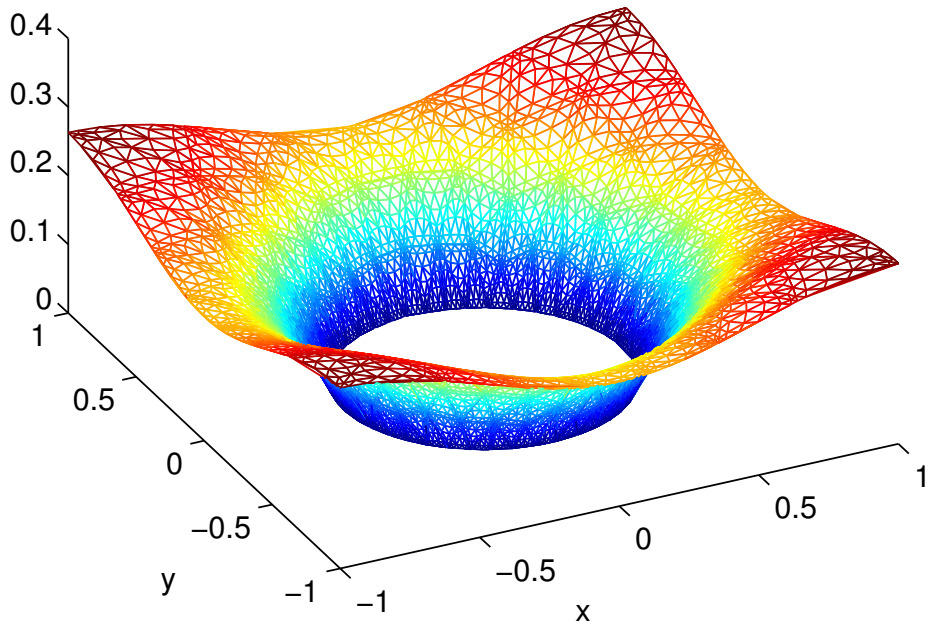


Abbildung 7.8: Plot der berechneten Lösung für $N = 6$

mit rechter Seite $f = 1$ in (7.6a) und mit homogenen Randbedingungen, sowohl auf dem Dirichlet-Randstück $\Gamma_{\mathcal{D}}$, als auch auf dem Neumann-Randstück $\Gamma_{\mathcal{N}}$. Eine Näherung der Lösung für den Polynomgrad $N = 6$ ist in Abbildung 7.8 dargestellt.

Die Randwertaufgabe (7.3) wird wie zuvor mit Hilfe der TSEM diskretisiert. Wie im Abschnitt 7.3 bezeichnet das lineare Gleichungssystem $A_G\mathbf{u} = \mathbf{b}_G$ die aus der Galerkin-Diskretisierung hervorgegangene Form (4.37). Nach Anwendung der statischen Kondensation erhalten wir das System $A_{SK}\mathbf{u} = \mathbf{b}_{SK}$.

Die Konditionen der beiden Matrizen sind in einer Log-Log-Skala der Abbildung 7.9 angegeben. Anhand der Hilfslinien lässt sich erkennen, dass die Kondition von A_G mit der Rate $\mathcal{O}(N^4)$ steigt, während für die vorkonditionierte Matrix A_{SK} nur $\mathcal{O}(N^2)$ gilt. Die genauen Werte sind in der Tabelle 7.3 aufgeführt.

In Abbildung 7.10 werden die Residuenentwicklungen beider Mehrgitterverfahren dargestellt. Auch hier sind die Iterationsachsen verschieden skaliert. Im direkten Vergleich zu den Historien aus Abbildung 7.10 scheint die Konvergenzrate von MG_G hier deutlich besser zu sein. Allerdings ist hier das Anfangsresiduum viel geringer.

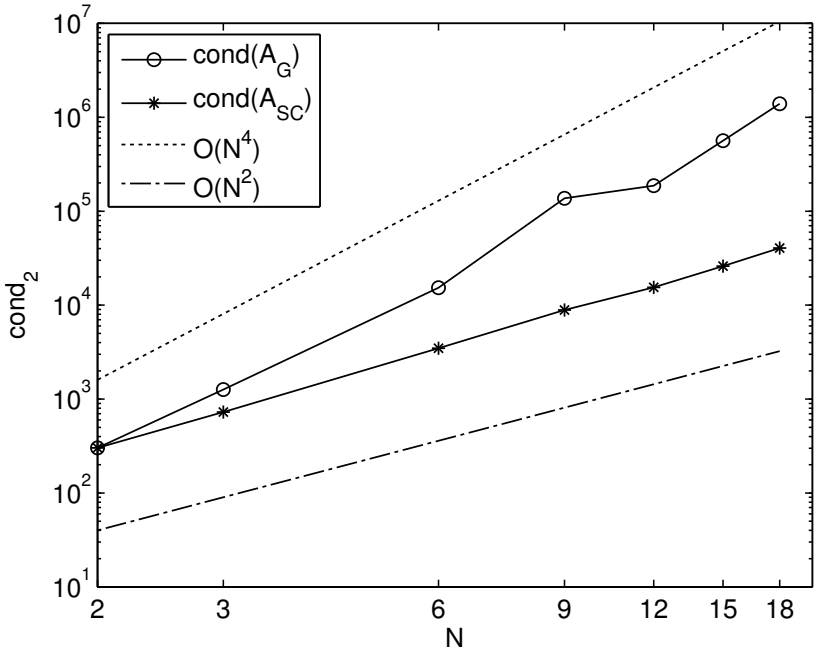


Abbildung 7.9: Verlauf der Kondition in Abghängigkeit von N

N	$\text{cond}(A_G)$	$\text{cond}(A_{SK})$
2	3.0246e+02	3.0246e+02
3	1.2613e+03	7.2946e+02
6	1.5356e+04	3.4747e+03
9	1.3728e+05	8.8909e+03
12	1.8690e+05	1.5511e+04
15	5.6351e+05	2.6006e+04
18	1.3958e+06	4.0666e+04

Tabelle 7.3: Konditionsverlauf bzgl. $\|\cdot\|_\infty$ in Abhängigkeit von N.

Methode	MG(6,12)	MG(1,3,6,12)	MG(3,9,18)	MG(1,3,9,18)
MG_G	0.8438	0.8442	0.8754	0.8756
MG_{SK}	0.4780	0.5552	0.5759	0.5843

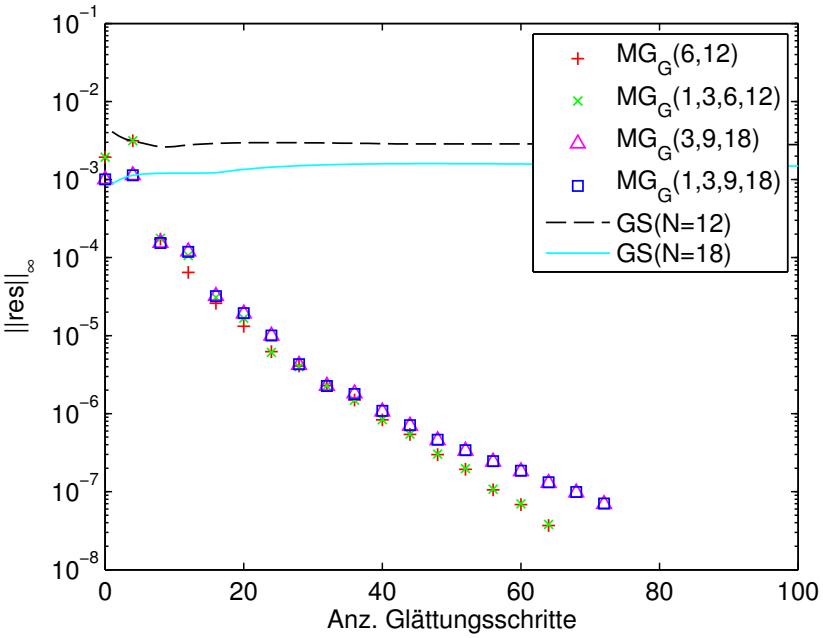
Tabelle 7.4: Approximierte Konvergenzraten ρ zu den Ergebnissen aus Abbildung 7.10

Methode	MG(6,12)	MG(1,3,6,12)	MG(3,9,18)	MG(1,3,9,18)
MG_G	1.386	1.268	3.983	4.137
MG_{SK}+SK	0.501	0.692	2.704	2.733
-davon nur SK	0.338	0.342	2.167	2.138

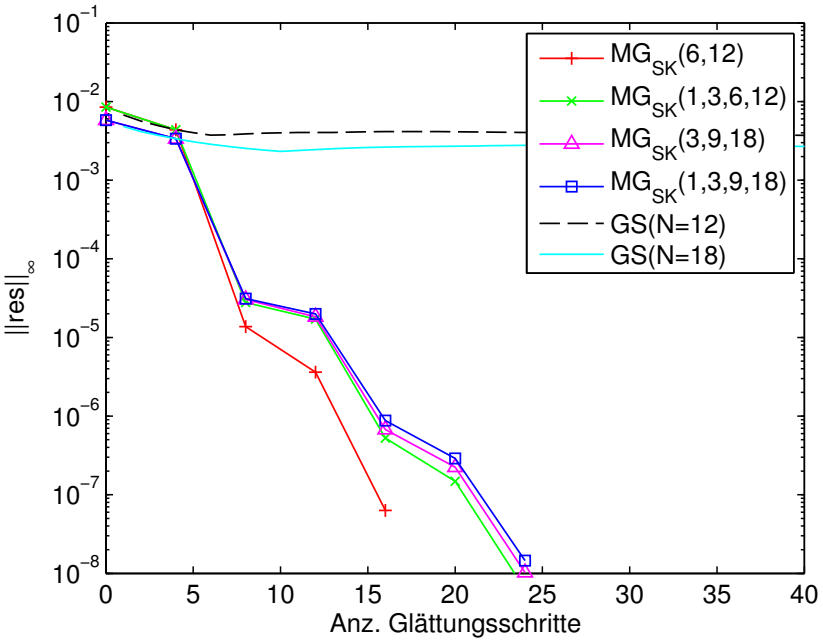
Tabelle 7.5: Die benötigten CPU-Zeiten (in Sekunden) für das Berechnen der Mehrgitterverfahren und der statischen Kondensation

Für eine bessere Vergleichbarkeit werden in Tabelle 7.4 die approximierten Konvergenzraten ρ zu den Daten aus Abbildung 7.10 vermerkt. Im Vergleich zu Tabelle 7.2 sind die Raten von MG_G insgesamt gleich geblieben und haben sich bzgl. der Polynomgrade ein wenig angenähert. Das Verfahren MG_{SK} zeigt hier Einbußen in der Konvergenzgeschwindigkeit. Dennoch bleibt es hinsichtlich der Konvergenz und der Iterationsschritte überlegen.

Abschließend folgen in Tabelle 7.5 die benötigten CPU-Zeiten für das Berechnen der Mehrgitterverfahren und der statischen Kondensation für die Beispielaufgabe (7.6). Analog zu den Residuenhistorien aus Abbildung 7.10 ist die verkürzte Vefahrensdauer von MG_G prägnant. Die statische Kondensation nimmt besonders für höhere Polynomgrade die meiste Zeit in Anspruch.



(a) Das Mehrgitterverfahren MG_G



(b) Das Mehrgitterverfahren MG_{SK}

Abbildung 7.10: Konvergenzentwicklungen der Mehrgitterverfahren und der Gauß-Seidel-Glätter

8 Fazit

Zum Abschluss der Arbeit wollen wir unsere Ergebnisse kurz zusammenfassen.

Nachdem die dreiecksbasierte Spektrale-Elemente-Methode auf Basis der Fekete-Punkte bestimmt wurde und entsprechende Elementmatrizen für die nodalen Basisfunktionen vorgestellt wurden, haben wir ein passendes p-Mehrgitterverfahren nach [13] angegeben.

Zur Vorkonditionierung des linearen Gleichungssystems haben wir die statische Kondensation vorgestellt, welche üblicherweise bei der Schur-Komplement-Methode zum Einsatz kommt. Die Restriktionsstrategie, die Grobgitterkorrektur und der Glätter des Mehrgitterverfahrens wurden auf das so vorkonditionierte lineare Gleichungssystem abgestimmt.

Ein Vergleich der numerischen Ergebnisse der hier präsentierten Methode mit den Verfahren aus [13] bestätigt den Vorzug dieser Methode. Dabei wurde auf die Robustheit bzgl. der Triangulierung und der Randbedingungen geprüft und die Rechenzeiten verglichen.

Ebenfalls wurde gezeigt, dass sich die benötigte Rechenzeit für unsere Methode auf das Berechnen der statischen Kondensation konzentriert, während das Mehrgitterverfahren innerhalb weniger V-Zyklen das Abbruchkriterium erreicht hatte. Die meisten Rechenoperationen werden für die Invertierung einer Blockdiagonalmatrix benötigt.

Diese Beobachtung macht das Verfahren für eine Parallelisierung empfehlenswert, bei der die Invertierung der einzelnen Blöcke problemlos auf mehrere Prozessoren verteilt werden kann.

Insgesamt ist die Anwendung dieser Methode auf komplexere Geometrien (wie zum Beispiel krummlinige oder dreidimensionale Gebiete) mit adaptiven Gittern für zukünftige Forschungen interessant.

Tabellenverzeichnis

Abbildungsverzeichnis

Literatur

[1] Dietrich Braess. *Finite Elemente*. Springer-Verlag Berlin Heidelberg, 2007.

[2] Claudio. Canuto u. a. *Spectral Methods - Fundamentals in Single Domains*. Scientific Computation, Berlin, Heidelberg : Springer Berlin Heidelberg, 2006.

[3] Ronald Cools. „Monomial cubature rules since “Stroud”: a compilation — part 2“. In: *Journal of Computational and Applied Mathematics* **112**.1–2 (1999), S. 21–27.

[4] Ronald Cools und Philip Rabinowitz. „Monomial cubature rules since “Stroud”: a compilation“. In: *Journal of Computational and Applied Mathematics* **48**.3 (1993), S. 309–326.

[5] Victorita Dolean, Richard Pasquetti und Francesca Rapetti. „p-Multigrid for Fekete Spectral Element Method“. In: *Domain Decomposition Methods in Science and Engineering XVII*. Hrsg. von Ulrich Langer u. a. Bd. 60. Lecture Notes in Computational Science and Engineering. Springer Berlin Heidelberg, 2008, S. 485–492.

[6] Moshe Dubiner. „Spectral Methods on Triangles and Other Domains“. In: *J. Sci. Comput.* **6**.4 (Dez. 1991), S. 345–390.

[7] L. Fejer. „Bestimmung derjenigen Abszissen eines Intervalles für welche die Quadratsumme der Grundfunktionen der Lagrangeschen Interpolation im Intervalle $[-1, 1]$ ein möglichst kleines Maximum besitzt“. In: *Ann. Scuola Norm. Sup. Pisa Sci. Fis. Mt. Ser. II* **12** (1932), S. 263–276.

[8] Wolfgang Hackbusch. *Iterative Lösung großer schwachbesetzter Gleichungssysteme*. Bd. 69. Leitfäden der angewandten Mathematik und Mechanik. Stuttgart: Teubner, 1991, S. 382.

[9] Wolfgang Hackbusch. *Multi-grid methods and applications*. Bd. 4. Springer series in computational mathematics. Berlin [u.a.]: Springer, 1985, S. XIV, 377.

[10] J. S. Hesthaven. „From Electrostatics to Almost Optimal Nodal Sets for Polynomial Interpolation in a Simplex“. In: *SIAM J. Numer. Anal.* **35**.2 (Apr. 1998), S. 655–676.

[11] G. Karniadakis und S. Sherwin. *Spectral/hp Element Methods for Computational Fluid Dynamics: Second Edition*. Numerical Mathematics and Scientific Computation. OUP Oxford, 2005.

[12] Peter Knabner und Lutz Angermann. *Numerik Partieller Differentialgleichungen/ Numerical Analysis of Partial Differential Equations: Eine Anwendungsorientierte Einführung/ an Application-oriented Introduction*. Springer-Lehrbuch Masterclass. Springer Berlin Heidelberg, 2000.

[13] Richard Pasquetti und Francesca Rapetti. „p-Multigrid Method for Fekete-Gauss Spectral Element Approximations of Elliptic Problems“. In: *Commun. Comput. Phys.* **5** (No. 2-4 Feb. 2009), S. 667–682.

[14] Richard Pasquetti und Francesca Rapetti. „Spectral Element Methods on Unstructured Meshes: Comparisons and Recent Advances“. In: *J. Sci. Comput.* **27**.1-3 (Juni 2006), S. 377–387.

[15] Richard Pasquetti u. a. „Neumann–Neumann–Schur complement methods for Fekete spectral elements“. English. In: *Journal of Engineering Mathematics* **56**.3 (2006), S. 323–335.

[16] C. Pozrikidis. *Introduction to Finite and Spectral Element Methods using MATLAB*. Taylor & Francis, 2005.

[17] Einar M. Rønquist und Anthony T. Patera. „Spectral element multigrid. I. - Formulation and numerical results“. In: *J. Sci. Comput.* **2**.4 (1987), S. 389–406.

[18] A. H. Stroud. *Approximate calculation of multiple integrals*. Englisch. Englewood Chiffs, New Jersey: Prentice-Hall, 1971.

[19] M. A. Taylor, B. A. Wingate und R. E. Vincent. „An Algorithm for Computing Fekete Points in the Triangle“. In: *SIAM J. Numer. Anal.* **38**.5 (Okt. 2000), S. 1707–1720.

[20] M. Taylor, B. Wingate und L. Bos. „A Cardinal Function Algorithm for Computing Multivariate Quadrature Points“. In: *SIAM Journal on Numerical Analysis* **45**.1 (2007), S. 193–205.

[21] T. Warburton, L. F. Pavarino und J. S. Hesthaven. „A Pseudo-spectral Scheme for the Incompressible Navier-Stokes Equations Using Unstructured Nodal Elements“. In: *J. Comput. Phys.* **164**.1 (Okt. 2000), S. 1–21.

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die Arbeit selbstständig erstellt und keine anderen als die angegebenen Hilfsmittel verwendet habe.

Ort, Datum, Unterschrift (Peter Norek)