Hierarchische Matrizen bei Finite-Differenzen-Verfahren

Dissertation
zur Erlangung des Doktorgrades
der Fakultät für Mathematik, Informatik
und Naturwissenschaften
der Universität Hamburg

vorgelegt
im Fachbereich Mathematik
von

Dominik Enseleit
aus Hamburg

Hamburg
2013

Als Dissertation angenommen vom Fachbereich
Mathematik der Universität Hamburg

Auf Grund der Gutachten von Prof. Dr. Jens Struckmeier
und Prof. Dr. Sabine Le Borne

Hamburg, den 25. September 2013

Prof. Dr. Ulf Kühn
Leiter des Fachbereichs Mathematik

Inhaltsverzeichnis

1 Einleitung

Bei der Diskretisierung partieller Differentialgleichungen entstehen große lineare Gleichungssysteme, zu deren numerischer Lösung eine Vielzahl direkter und iterativer Verfahren zur Verfügung stehen. Durch die Weiterentwicklung bekannter und die Einführung neuer Methoden nimmt die Anzahl der verfügbaren Lösungsalgorithmen stetig zu. Bei der Auswahl eines geeigneten Verfahrens werden jedoch in der Regel nur die bekanntesten und „etabliertesten“ Vertreter berücksichtigt, obwohl möglicherweise ein weniger verbreiteter aber effizienterer Algorithmus zur Verfügung stünde.

Diese Problematik bildet den Ausgangspunkt der vorliegenden Arbeit. Das zu lösende Gleichungssystem stammt aus dem Strömungsmodell METRAS (Mesoskaliges Transport- und Strömungsmodell), welches am Fachbereich Meteorologie der Universität Hamburg entwickelt wurde (vgl. [SBL$^+$96]). Ein wesentlicher Anteil der Simulationszeit wird in diesem Modell zur Lösung linearer Gleichungssysteme benötigt, da in jedem Zeitschritt eines gelöst werden muss. In diesem Fall ist der Einsatz eines effizienten Lösungsalgorithmus von besonderer Bedeutung, weil bereits eine kleine Effizienzsteigerung bei der Lösung eines Gleichungssystems zu einer starken Beschleunigung der vollständigen Simulation führen kann. Die konkrete Fragestellung, die sich für diesen speziellen Anwendungsfall ergibt, lautet demnach, ob sich zur Beschleunigung der Simulation ein neues Verfahren zur Lösung der resultierenden linearen Gleichungssysteme einsetzen lässt.

Das Gleichungssystem im Modell METRAS stammt aus einer speziellen Finite-Differenzen-Diskretisierung, die zu einer dünnbesetzten Diskretisierungsmatrix führt. Derzeit werden zur Lösung iterative Verfahren eingesetzt, implementiert sind das präkonditionierte BiCGSTAB- und das IGCG-Verfahren (Idealized Generalized Conjugate Gradient). In [Sch07] wird darüber hinaus der Einsatz eines Mehrgitter-Verfahrens beschrieben.

Ein unter diesen Gesichtspunkten vielversprechender neuer Ansatz zur Lösung linearer Gleichungssysteme wurde Ende der 1990er Jahre von Wolfgang Hackbusch entwickelt: die Technik der Hierarchischen Matrizen ($\mathcal{H}$-Matrizen). Deren grundlegende Idee beruht auf der Einführung einer hierarchischen Blockpartitionierung der Matrix und anschließender Approximation geeigneter Matrixblöcke durch Niedrigrangmatrizen. Dies ermöglicht Operationen wie die Matrix-Vektor- und Matrix-Matrix-Multiplikation sowie die approximative Berechnung der Inversen und der LU-Zerlegung einer Matrix in der sogenannten $\mathcal{H}$-Arithmetik in fast linearer Komplexität durchzuführen. Die Berechnung einer approximativen $\mathcal{H}$-Inversen oder einer approximativen $\mathcal{H}$-LU-Zerlegung in fast linearer Komplexität ist im Hinblick auf die Lösung des Gleichungssystems im Modell METRAS von besonderem Interesse, da in jedem Zeitschritt das gleiche Gleichungssystem für unterschiedliche rechte Seiten gelöst werden muss. Sollte die $\mathcal{H}$-Matrix-Technik zur Lösung der resultierenden linearen Gleichungssysteme geeignet sein, dann könnte die

$\mathcal{H}$-Inverse bzw. $\mathcal{H}$-LU-Zerlegung entweder zur direkten Lösung des Gleichungssystems oder als Präkonditionierer zur Beschleunigung iterativer Verfahren genutzt werden.

Grundsätzlich kann die $\mathcal{H}$-Matrix-Technik zur Lösung einer Vielzahl linearer Gleichungssysteme eingesetzt werden. Entscheidend ist, dass die Genauigkeit der Approximation und die Effizienz des Verfahrens wesentlich von den Eigenschaften der entsprechenden Matrix abhängen. So ist beispielsweise bei der $\mathcal{H}$-Invertierung einer Matrix unklar, ob überhaupt eine (gute) $\mathcal{H}$-Matrix Approximation der Inversen existiert. Ist dies nicht der Fall, kann zwar mit Hilfe der $\mathcal{H}$-Arithmetik möglicherweise approximativ eine $\mathcal{H}$-Inverse berechnet werden, diese wird jedoch keine gute Approximation an die exakte Inverse darstellen. Aus diesem Grund ist vor dem Einsatz der $\mathcal{H}$-Matrix-Technik im Modell METRAS zu prüfen, ob die resultierenden Matrizen die erforderlichen Eigenschaften besitzen.

Das Ziel dieser Arbeit ist, einen Beitrag zur Beantwortung dieser grundlegenden Fragestellung zu leisten. Im Mittelpunkt der Überlegungen steht die theoretische Untersuchung der Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen von Finite-Differenzen-Matrizen. Dieses Resultat ist von wesentlicher Bedeutung, da darauf aufbauend weitere theoretische Untersuchungen – wie der Nachweis der Existenz einer $\mathcal{H}$-Matrix Approximation der Faktoren der LU-Zerlegung – erfolgen können. Die Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen konnte unter anderem bereits für Matrizen gezeigt werden, die aus der Diskretisierung elliptischer Randwertprobleme mittels Finiter-Elemente stammen ([BH03]). Auf Grundlage dieses Resultats wurde dessen Existenz ebenfalls für die Faktoren der LU-Zerlegung von Finite-Element-Matrizen in [Beb07] bewiesen. Die dort erzielten Ergebnisse gelten gleichzeitig für eine Finite-Differenzen-Matrix, wenn diese mit einer Diskretisierungsmatrix aus einer entsprechenden Finite-Element-Diskretisierung übereinstimmt. Wegen der speziellen Form der zugrunde liegenden Problemstellung im Modell METRAS ist jedoch nicht ersichtlich, ob sich die dort auftretenden Matrizen im Allgemeinen als Resultat einer Finite-Element-Diskretisierung interpretieren lassen. Daher können diese Ergebnisse nicht genutzt werden.

Da keine Untersuchungen bzw. Veröffentlichungen zur Thematik der $\mathcal{H}$-Matrizen für den Spezialfall der Finite-Differenzen-Matrizen bekannt sind, müssen zunächst einige grundlegende Überlegungen zur Untersuchung dieses Anwendungsfalls angestellt werden. Dazu erfolgt die Einführung geeigneter Begrifflichkeiten und Strategien, die zur Anpassung der $\mathcal{H}$-Matrix-Technik an das Umfeld der Finite-Differenzen-Verfahren erforderlich sind. Außerdem ist die Entwicklung eines neuen methodischen Ansatzes im Finite-Differenzen-Kontext erforderlich, der es ermöglicht, die grundlegende theoretische Fragestellung nach der Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen von Finite-Differenzen-Matrizen zu untersuchen. Da keine Resultate dieser Art bekannt sind, erfolgt die Untersuchung nicht direkt auf Grundlage des komplexen Gleichungssystems im Modell METRAS, sondern anhand eines Modellproblems. Bei diesem werden vereinfachende Annahmen getroffen, welche die Entwicklung eines methodischen Ansatzes ermöglichen. Die aus dem Modellproblem resultierende Finite-Differenzen-Matrix dient als Grundlage für die theoretischen Überlegungen. Die Auswahl des Modellproblems erfolgt dabei in Anlehnung an die charakteristischen Eigenschaften der Problemstellung im Modell METRAS, um anschließend eine möglichst einfache Übertragung der Ergebnisse

für diesen speziellen Fall zu ermöglichen.

Die theoretischen Untersuchungen werden durch numerische Tests zur Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen von Finite-Differenzen-Matrizen ergänzt, die hilfreiche Hinweise zur numerischen Umsetzung der $\mathcal{H}$-Matrix-Technik im Finite-Differenzen-Kontext liefern. Abschließende numerische Tests zum Einsatz der $\mathcal{H}$-Matrix-Technik im Modell METRAS unter Verwendung aller verfügbaren Konzepte, die bei der praktischen Umsetzung zur Effizienzsteigerung eingesetzt werden könnten, erfolgen in dieser Arbeit nicht.

Als Einführung in die beiden Themenkomplexe der Finite-Differenzen-Diskretisierungen und der Hierarchischen Matrizen werden in Kapitel 2 und 3 deren wesentliche Grundlagen zusammengefasst. Im zweiten Kapitel wird insbesondere das Modellproblem eingeführt und es werden diskrete Varianten der Poincaré- und der Cacciopoli-Ungleichung für Gitterfunktionen bewiesen, die als wesentliche Hilfsmittel im weiteren Verlauf der Arbeit benötigt werden. Im Anschluss an eine allgemeine Einführung der $\mathcal{H}$-Matrix-Technik in Kapitel 3 wird der spezielle Anwendungsfall der Finite-Differenzen-Matrizen thematisiert. Aufbauend auf diesen Grundlagen und Ergebnissen kann im folgenden Kapitel ein Resultat zur Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen von Finite-Differenzen-Matrizen für das Modellproblem im zweidimensionalen Fall erzielt werden. Eine Erweiterung auf den dreidimensionalen Fall durch Anpassung der Ergebnisse aus Kapitel 2 und 4 folgt in Kapitel 5. Das sechste Kapitel dient der numerischen Überprüfung der theoretischen Resultate aus den vorangegangenen Abschnitten anhand mehrerer Testprobleme, die insbesondere im Hinblick auf das lineare Gleichungssystem im Modell METRAS aufgestellt werden. Die numerischen Ergebnisse zu diesen Testproblemen weisen auf Schwächen der gewöhnlichen Vorgehensweise bei bestimmten Konstellationen der Koeffizienten im Modellproblem hin. Für diese kann jedoch eine Modifikation zur Verbesserung der Ergebnisse eingeführt werden. Darüber hinaus werden numerische Tests zur Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen von Finite-Differenzen-Matrizen durchgeführt, denen andere Problemstellungen als das Modellproblem zugrunde liegen. Im letzten Kapitel folgt eine Zusammenfassung der Ergebnisse.

Danksagung

Auf diesem Wege möchte ich Herrn Prof. Dr. Jens Struckmeier für die vertrauensvolle Zusammenarbeit sowie seine konstruktiven Ratschläge und Anregungen danken. Insbesondere für seine grundlegende Idee, die zu der spannenden Fragestellung dieser Arbeit führte, bin ich ihm sehr dankbar.

Frau Prof. Dr. Sabine Le Borne danke ich für ihr Interesse an meiner Arbeit und für ihre hilfreichen inhaltlichen Anmerkungen.

Meiner Familie gilt mein Dank für ihre große Unterstützung während der vergangenen Jahre. Judith – danke für Alles!

2 Grundlagen

In diesem Kapitel werden grundlegende Bezeichnungen und Konzepte im Zusammenhang mit Finite-Differenzen-Verfahren eingeführt, da das lineare Gleichungssystem im Modell METRAS aus einer Finite-Differenzen-Diskretisierung hervorgeht. Zusätzlich wird das Modellproblem angegeben, auf dessen Grundlage die weiteren theoretischen Untersuchungen erfolgen.

Zum Beweis der Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen der Diskretisierungsmatrix des Modellproblems in Kapitel 4 werden diskrete Varianten der Poincaré- und der Cacciopoli-Ungleichung benötigt. Da diese in der erforderlichen Form für Gitterfunktionen nicht bekannt sind, werden die entsprechenden Resultate in den Abschnitten 2.2 und 2.3 bewiesen.

2.1 Finite-Differenzen-Verfahren

Bei der Anwendung Finiter-Differenzen-Verfahren ist es erforderlich, ein Gitter einzuführen. In dieser Arbeit werden Gitter mit konstanter Schrittweite $h > 0$ betrachtet. Das d-dimensionale Gitter unendlicher Ausdehnung wird mit $h\mathbb{Z}^d$ bezeichnet und es wird die kurze Schreibweise $x_i := ih, i \in \mathbb{Z}$ zur Kennzeichnung der Koordinaten eines Gitterpunktes verwendet.

Im weiteren Verlauf beschränken sich die Untersuchungen auf den zweidimensionalen Fall von Rechteckgittern. Dabei bezieht sich der Ausdruck „Rechteckgitter“ nicht auf die einzelnen Gitterzellen, die bei konstanter Schrittweite quadratisch sind, sondern auf die Struktur des vollständigen Gitters. Ein Rechteckgitter

$$\Omega_h := \{(x_i, y_j) \in h\mathbb{Z}^2 : 1 \leq i \leq n, 1 \leq j \leq m\}$$

mit $n \in \mathbb{N}$ bzw. $m \in \mathbb{N}$ Gitterpunkten in x- bzw. y-Richtung und konstanter Gitterweite h wird als $n \times m$ Rechteckgitter bezeichnet. Außerdem wird

$$\overline{\Omega}_h := \{(x_i, y_j) \in h\mathbb{Z}^2 : 0 \leq i \leq n+1, 0 \leq j \leq m+1\}$$

definiert und die Menge der Randpunkte des Rechteckgitters ist durch $\partial\Omega_h := \overline{\Omega}_h \backslash \Omega_h$ gegeben. Zusätzlich wird die Menge der randnahen Punkte $\Gamma(\Omega_h)$ eingeführt, die durch die zu $\partial\Omega_h$ benachbarten Punkte aus Ω_h gegeben ist (vgl. Abbildung 2.1). Zwei Gitterpunkte $x, y \in \Omega_h$ heißen benachbart, wenn $\|x - y\|_2 = h$ gilt. Ein Gitter Ω_h heißt zusammenhängend, wenn sich je zwei Gitterpunkte $x, y \in \Omega_h$ durch eine Folge benachbarter Gitterpunkte aus Ω_h verbinden lassen.

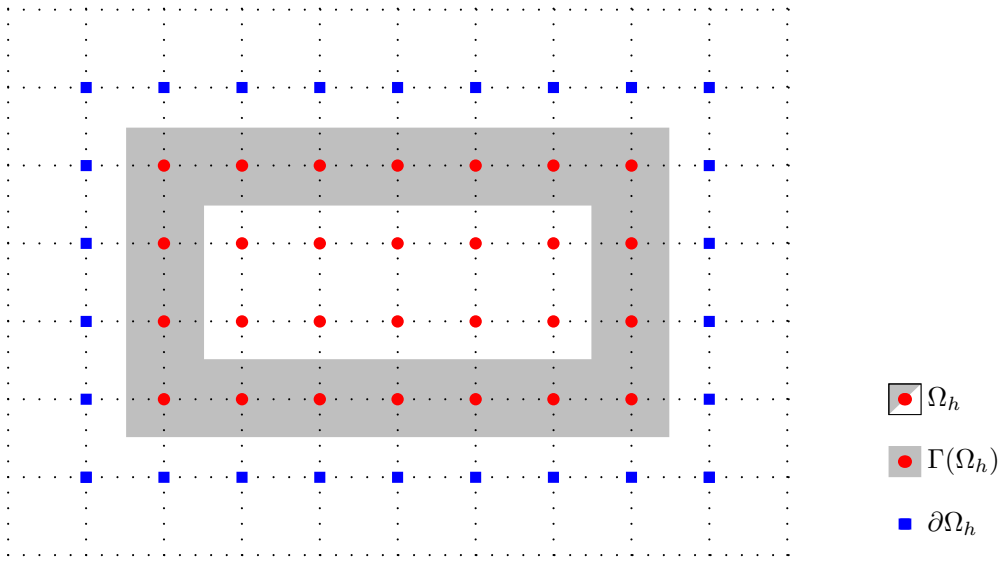


Abbildung 2.1: Bezeichnungen am Rechteckgitter

Für Gitter $\Omega_h, \tilde{\Omega}_h \subset h\mathbb{Z}^2$ lassen sich der Durchmesser $\operatorname{diam}(\Omega_h)$ und der Abstand $\operatorname{dist}(\Omega_h, \tilde{\Omega}_h)$ durch

$$\operatorname{diam}(\Omega_h) := \max\left\{\left\|x' - x''\right\|_2 : x', x'' \in \Omega_h\right\}$$
$$\operatorname{dist}_\infty(\Omega_h, \tilde{\Omega}_h) := \min\left\{\left\|x - y\right\|_\infty : x \in \Omega_h, y \in \tilde{\Omega}_h\right\}$$

einführen, wobei die Definitionen für die in dieser Arbeit verwendeten Normen – bei der Berechnung des Durchmessers die Euklidische Norm und zur Bestimmung der Distanz die Maximumsnorm – angegeben sind.

Der Raum der (reellwertigen) Gitterfunktionen auf dem Gitter Ω_h wird mit $\mathcal{D}_h(\Omega_h)$ bezeichnet und für Gitterfunktionen $u \in \mathcal{D}_h(\overline{\Omega}_h)$ mit $u|_{\partial\Omega_h} = 0$ wird die Bezeichnung $\mathcal{D}_{h,0}(\overline{\Omega}_h)$ verwendet. Für eine Gitterfunktion $u \in \mathcal{D}_h(\Omega_h)$ werden folgende Differenzenoperatoren eingeführt:

$$u_x(x_i) := \frac{1}{h}\left(u(x_{i+1}) - u(x_i)\right), \quad x_i, x_{i+1} \in \Omega_h \qquad \text{(Vorwärtsdifferenz)}$$
$$u_{\bar{x}}(x_i) := \frac{1}{h}\left(u(x_i) - u(x_{i-1})\right), \quad x_i, x_{i-1} \in \Omega_h \qquad \text{(Rückwärtsdifferenz).}$$

Die Vorwärts- oder Rückwärtsdifferenz zwischen einem Punkt aus Ω_h und einem Punkt, der nicht zu Ω_h gehört, wird zu null gesetzt. Analog zum Gradienten wird der Vektor, dessen Komponenten durch die Vorwärtsdifferenzen der Gitterfunktion $u \in \mathcal{D}_h(\Omega_h)$ bezüglich aller Koordinatenrichtungen gegeben sind, mit $\nabla_h u$ bezeichnet.

Analog zur Produktregel lässt sich eine diskrete Produktregel für Gitterfunktionen $u, v \in \mathcal{D}_h(\Omega_h)$ einführen:

$$(vu)_x(x_i) = v_x(x_i)u(x_{i+1}) + v(x_i)u_x(x_i), \quad x_i, x_{i+1} \in \Omega_h. \tag{2.1}$$

Außerdem kann wie bei der partiellen Integration die partielle Summation durchgeführt werden, die für den eindimensionalen Fall eines Gitters $\Omega_h = \{ih : 1 \leq i \leq n\}$ mit $n \in \mathbb{N}$ in der folgenden Form für $u, v \in \mathcal{D}_h\left(\overline{\Omega}_h\right)$ gegeben ist:

$$h \sum_{i=0}^{n} v_x u = -h \sum_{i=0}^{n} v(x_{i+1}) u_x + v(x_{n+1}) u(x_{n+1}) - v(x_0) u(x_0).$$

Hierbei wurde die Kurzschreibweise $v_x = v_x(x_i)$ (analog für u und u_x) verwendet. Später wird insbesondere die Variante

$$h \sum_{i=0}^{n} v_x u = -h \sum_{i=1}^{n} v u_{\bar{x}} + v(x_{n+1}) u(x_n) - v(x_0) u(x_0)$$

eingesetzt. Die Erweiterungen der diskreten Produktregel und der partiellen Summation auf den höherdimensionalen Fall können direkt auf Grundlage der eindimensionalen Darstellungen erfolgen.

Analog zur L^2-Norm wird die L_h^2-Norm für Gitterfunktionen $u \in \mathcal{D}_h\left(\Omega_h\right)$ durch

$$\|u\|_{L_h^2(\Omega_h)} := \left(h^d \sum_{x \in \Omega_h} |u(x)|^2 \right)^{\frac{1}{2}}$$

definiert.

Bezeichnet man mit $|\Omega_h|$ die Anzahl der Gitterpunkte aus Ω_h, so ist der Mittelwert einer Gitterfunktion $u \in \mathcal{D}_h\left(\Omega_h\right)$ durch

$$\overline{u} := \frac{1}{|\Omega_h|} \sum_{x \in \Omega_h} u(x) \tag{2.2}$$

gegeben.

2.1.1 Das Modellproblem

Die theoretischen Untersuchungen zur Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen von Finite-Differenzen-Matrizen in Kapitel 4 werden nicht direkt für die Diskretisierungsmatrix aus dem Modell METRAS durchgeführt, sondern für ein vereinfachtes Modellproblem. Dieses Vorgehen dient der Herleitung eines methodischen Ansatzes, ohne alle Details der ursprünglichen Problemstellung zu berücksichtigen. Gleichzeitig sollte die Wahl des Modellproblems so erfolgen, dass im Anschluss eine Übertragung des entwickelten Ansatzes auf die Problemstellung im Modell METRAS leicht möglich ist.

Um ein in diesem Sinne geeignetes Modellproblem aufzustellen, sind die speziellen Gegebenheiten im Modell METRAS zu beachten. Deshalb werden in diesem Abschnitt die wesentlichen Grundlagen des Modells beschrieben. Dabei ist von besonderem Interesse, in welchem Zusammenhang die Lösung des linearen Gleichungssystems erforderlich ist und auf welche Weise dieses gebildet wird. Eine ausführliche Beschreibung des

vollständigen Modells findet man in [SBL+96]. Die folgenden Ausführungen basieren auf der kompakten Darstellung in [Sch07].

Das Modell METRAS wurde zur Simulation kleinräumiger atmosphärischer Prozesse der Mesoskala entwickelt und basiert auf den Reynolds-gemittelten dreidimensionalen Navier-Stokes-Gleichungen. Zur Kennzeichnung der gemittelten Größen wird die Bezeichnung $\bar{\cdot}^R$ verwendet, um eine Verwechslung mit dem Mittelwert einer Gitterfunktion $\bar{\cdot}$ nach (2.2) zu vermeiden. Die im Modell verwendeten Approximationen beschränken sich auf die Boussinesq- und die anelastische Approximation.

Der untere Rand des Modellgebiets wird durch die Topographiehöhe $z_s(x, y)$ vorgegeben und es können nichtäquidistante Schrittweiten für das Modellgitter in allen Koordinatenrichtungen gewählt werden. Dies ermöglicht die Verfeinerung des Gitters in vorgegebenen Bereichen, ohne dass die Gesamtzahl der Gitterpunkte zu stark zunimmt.

Die Modellgleichungen werden nicht in kartesischen Koordinaten (x, y, z) gelöst, sondern es wird eine Koordinatentransformation des Modellgitters auf ein äquidistantes Gitter mit den Koordinaten $(\hat{x}, \hat{y}, \hat{z})$ vorgenommen:

$$\begin{aligned} \hat{x} &= \hat{x}(x) \\ \hat{y} &= \hat{y}(y) \\ \hat{z} &= \hat{z}(\eta), \end{aligned}$$

wobei die vertikale Koordinate durch

$$\eta = z_t \frac{z - z_s}{z_t - z_s}$$

gegeben ist. Dabei beschreibt die Konstante z_t die Höhe des oberen Modellrandes, so dass man dort $\eta = z_t$ und am unteren Modellrand $(z = z_s)$ $\eta = 0$ erhält. In Abbildung 2.2 sind exemplarisch Koordinatenebenen für konstante η-Koordinate angegeben.

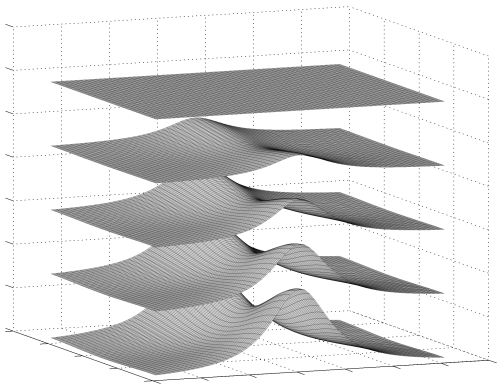


Abbildung 2.2: Koordinatenebenen für konstante η-Koordinate

Die kontravarianten Komponenten $(\hat{u}, \hat{v}, \hat{w})$ des Geschwindigkeitsvektors im neuen Koordinatensystem lassen sich in Abhängigkeit des Geschwindigkeitsvektors in kartesischen Koordinaten (u, v, w) durch

$$\begin{aligned}\hat{u} &= u\frac{\partial \hat{x}}{\partial x}\\ \hat{v} &= v\frac{\partial \hat{y}}{\partial y}\\ \hat{w} &= u\frac{\partial \hat{z}}{\partial x} + v\frac{\partial \hat{z}}{\partial y} + w\frac{\partial \hat{z}}{\partial z}\end{aligned}$$

angeben.

Im Modell wird die Dichte ρ zerlegt in

$$\rho = \rho_0(z) + \tilde{\rho}(x, y, z)$$

mit einem Grundzustand ρ_0 und einem mesoskaligen Anteil $\tilde{\rho}$. Im Zuge der Boussinesq-Approximation wird in allen Termen, bis auf den Auftriebsterm, die Dichte durch den Grundzustand ρ_0 ersetzt.

Außerdem wird der Druck analog zerlegt in

$$p = p_0(z) + p_g(x, y, z) + p_1(x, y, z) + p_2(x, y, z).$$

Die Anteile p_0 und p_1 ergeben sich aus dem hydrostatischen Gleichgewicht mit ρ_0 und $\tilde{\rho}$:

$$\begin{aligned}\frac{\partial p_0}{\partial \hat{z}} &= -g\rho_0\frac{\partial z}{\partial \hat{z}}\\ \frac{\partial p_1}{\partial \hat{z}} &= -g\tilde{\rho}\frac{\partial z}{\partial \hat{z}}.\end{aligned}$$

Für den geostrophischen Druck p_g wird der Zusammenhang zum geostrophischen Wind genutzt:

$$\begin{aligned}u_g &= -\frac{1}{\rho_o f}\left\{\frac{\partial \hat{y}}{\partial y}\frac{\partial p_g}{\partial \hat{y}} + \frac{\partial \hat{z}}{\partial y}\frac{\partial p_g}{\partial \hat{z}}\right\}\\ v_g &= \frac{1}{\rho_o f}\left\{\frac{\partial \hat{x}}{\partial x}\frac{\partial p_g}{\partial \hat{x}} + \frac{\partial \hat{z}}{\partial x}\frac{\partial p_g}{\partial \hat{z}}\right\}\end{aligned}$$

mit dem Coriolisparameter f. Der „dynamische“ Druck p_2 wird aus numerischen Gründen eingeführt. Da in diesem Zusammenhang die Lösung des linearen Gleichungssystems erforderlich ist, folgen dazu später weitere Erläuterungen.

Unter Berücksichtigung dieser Gegebenheiten und Einführung der Bezeichnungen

$$\alpha^* = \frac{\partial x}{\partial \hat{x}}\frac{\partial y}{\partial \hat{y}}\frac{\partial z}{\partial \hat{z}}$$

und

$$m = \begin{pmatrix} \rho_0 \overline{u}^R & \rho_0 \overline{v}^R & \rho_0 \overline{w}^R \end{pmatrix}^T$$

erhält man die folgenden Gleichungen, welche die Grundlage des Modells METRAS bilden:

$$\frac{\partial \alpha^* m}{\partial t} = -\mathcal{A}\left(\alpha^* m\right) - \mathcal{P}\left(p_1 + p_2\right) - \mathcal{B}\left(\tilde{\rho}\right) + \mathcal{C}\left(\overline{u}^R - u_g, \overline{v}^R - v_g, \overline{w}^R\right) - F_m \tag{2.3}$$

$$\nabla \cdot \left(\rho_0 \overline{\mathbf{u}}^R\right) = \frac{\partial}{\partial \hat{x}}\left(\overline{u}^R \frac{\partial \hat{x}}{\partial x} \rho_0 \alpha^*\right) + \frac{\partial}{\partial \hat{y}}\left(\overline{v}^R \frac{\partial \hat{y}}{\partial y} \rho_0 \alpha^*\right) + \frac{\partial}{\partial \hat{z}}\left(\overline{\hat{w}}^R \rho_0 \alpha^*\right) = 0. \tag{2.4}$$

Dabei wurde die Kurzschreibweise aus [Sch07] verwendet, nach der die entsprechenden Terme folgendermaßen gegeben sind:

$$\begin{aligned}
\text{Advektion: } \mathcal{A}\left(\chi\right) &= \frac{\partial}{\partial \hat{x}}\left(\overline{\hat{u}}^R \chi\right) + \frac{\partial}{\partial \hat{y}}\left(\overline{\hat{v}}^R \chi\right) + \frac{\partial}{\partial \hat{z}}\left(\overline{\hat{w}}^R \chi\right) \\
\text{Druckgradient: } \mathcal{P}\left(p\right) &= \begin{pmatrix} \alpha^* \frac{\partial \hat{x}}{\partial x}\frac{\partial p}{\partial \hat{x}} + \alpha^* \frac{\partial \hat{z}}{\partial x}\frac{\partial p}{\partial \hat{z}} \\ \alpha^* \frac{\partial \hat{y}}{\partial y}\frac{\partial p}{\partial \hat{y}} + \alpha^* \frac{\partial \hat{z}}{\partial y}\frac{\partial p}{\partial \hat{z}} \\ \alpha^* \frac{\partial \hat{z}}{\partial z}\frac{\partial p}{\partial \hat{z}} \end{pmatrix} \\
\text{Auftrieb: } \mathcal{B}\left(\rho\right) &= \begin{pmatrix} 0 & 0 & \rho \alpha^* g \end{pmatrix}^T \\
\text{Corioliskraft: } \mathcal{C}\left(\overline{u}^R, \overline{v}^R, \overline{w}^R\right) &= \begin{pmatrix} f \rho_0 \alpha^* \overline{v}^R - f' \rho_0 \alpha^* \overline{w}^R \\ -f \rho_0 \alpha^* \overline{u}^R + f' \rho_0 \alpha^* \overline{w}^R \\ \rho_0 \alpha^* f' \left(\overline{u}^R - \overline{v}^R d\right) \end{pmatrix} \\
\text{Turbulente Diffusion: } F_m &= \begin{pmatrix} F_{\overline{u}^R} & F_{\overline{v}^R} & F_{\overline{w}^R} \end{pmatrix}^T.
\end{aligned} \tag{2.5}$$

Dabei bezeichnen f und f' die Coriolisparameter und g die Schwerebeschleunigung. Für die dritte Komponente des Druckgradienten wird im Modell

$$\mathcal{P}_3 = \alpha^* \left[\left(\frac{\partial \hat{z}}{\partial x}\right)^2 + \left(\frac{\partial \hat{z}}{\partial y}\right)^2 + \left(\frac{\partial \hat{z}}{\partial z}\right)^2\right] \frac{\partial p}{\partial \hat{z}} + \alpha^* \left[\frac{\partial \hat{x}}{\partial x}\frac{\partial \hat{z}}{\partial x}\frac{\partial p}{\partial \hat{x}} + \frac{\partial \hat{y}}{\partial y}\frac{\partial \hat{z}}{\partial y}\frac{\partial p}{\partial \hat{y}}\right] \tag{2.6}$$

verwendet.

Die numerische Lösung der Gleichungen (2.3) und (2.4) erfolgt in mehreren Schritten. Dazu wird zunächst als Lösung von (2.3) eine vorläufige Geschwindigkeit $\mathbf{u}^*$, unter Verwendung des Drucks p_2^{n-1} zum vorherigen Zeitschritt (gekennzeichnet durch $n-1$), berechnet. Diese erfüllt die Gleichung (2.4) im Allgemeinen nicht, so dass im Anschluss eine Korrektur von $\mathbf{u}^*$ berechnet werden muss. Aus diesem Grund wird die Druckkorrektur $\hat{p}_2 = p_2^n - p_2^{n-1}$ eingeführt. Zu deren Bestimmung lässt sich unter Verwendung von (2.4) die Gleichung

$$\nabla \cdot \left(\mathcal{P}\left(\hat{p}_2\right)\right) = \frac{1}{\Delta t} \nabla \cdot \left(\rho_0 \mathbf{u}^*\right) \tag{2.7}$$

herleiten. Durch diesen Ansatz kann mit Hilfe der berechneten Druckkorrektur $\hat{p}_2$ der Druck zum nächsten Zeitschritt und aus der vorläufigen Geschwindigkeit $\mathbf{u}^*$ die Geschwindigkeit zum nächsten Zeitschritt, welche die Bedingung (2.4) erfüllt, bestimmt werden.

Die Gleichung (2.7) wird im Modell durch homogene Neumann-Randwerte ergänzt und das resultierende Randwertproblem mittels Finiter-Differenzen diskretisiert. Daraus ergibt sich das lineare Gleichungssystem, das in jedem Zeitschritt zur Bestimmung der Druckkorrektur $\hat{p}_2$ gelöst werden muss.

Bei der Diskretisierung der Gleichung (2.7) ist Folgendes zu berücksichtigen: Damit die Gleichung (2.4) im Modell numerisch erfüllt ist, muss die Diskretisierung von (2.7) durch die Hintereinanderausführung der im Modell verwendeten Diskretisierungen der Gleichungen (2.4) und (2.5) erfolgen. Um zu verdeutlichen, dass sich zur Diskretisierung der Ableitungen in den beiden Gleichungen unterschiedliche Differenzenapproximationen verwenden lassen, werden zur Darstellung die allgemeinen Bezeichnungen δ bzw. $\tilde{\delta}$ für eine beliebige Finite-Differenzen-Approximation der ersten Ableitung in (2.4) bzw. (2.5) eingeführt. Beispielsweise könnten in (2.4) Vorwärts- und in (2.5) Rückwärtsdifferenzen zur Diskretisierung verwendet werden, so dass sich $\delta_x u = u_x$ und $\tilde{\delta}_x u = u_{\bar{x}}$ (analog für die anderen Koordinatenrichtungen) ergeben würden. Unter Verwendung dieser Bezeichnungen ergibt sich eine allgemeine Darstellung der Diskretisierung von (2.4) durch

$$\nabla_h \cdot (\rho_0 \mathbf{u}) = \delta_{\hat{x}} (\rho_0 \gamma_1 u) + \delta_{\hat{y}} (\rho_0 \gamma_2 v) + \delta_{\hat{z}} (\rho_0 \gamma_3 \hat{w}) \tag{2.8}$$

und von (2.5) unter Berücksichtigung von (2.6) durch

$$\mathcal{P}_h(p) = \begin{pmatrix} c_1 \tilde{\delta}_{\hat{x}} p + c_2 \tilde{\delta}_{\hat{z}} p \\ c_3 \tilde{\delta}_{\hat{y}} p + c_4 \tilde{\delta}_{\hat{z}} p \\ c_5 \tilde{\delta}_{\hat{z}} p + c_6 \tilde{\delta}_{\hat{x}} p + c_7 \tilde{\delta}_{\hat{y}} p \end{pmatrix}. \tag{2.9}$$

Die aus der Koordinatentransformation resultierenden Größen aus (2.4) und (2.5) sind in den (im Allgemeinen ortsabhängigen) Koeffizienten γ_i, $i \in \{1, 2, 3\}$ und c_j, $j \in \{1, \ldots, 7\}$ zusammengefasst. Das zu lösende Gleichungssystem ergibt sich nach Gleichung (2.7) demnach durch das Einsetzen der Komponenten von $\mathcal{P}_h(\hat{p}_2)$ nach (2.9) anstelle von $\rho_0 \mathbf{u}$ in (2.8) und der anschließenden Hintereinanderausführung der jeweiligen im Modell verwendeten Differenzenapproximationen $\tilde{\delta}$ und δ. Aus diesem Ansatz resultiert die spezielle Struktur des linearen Gleichungssystems im Modell METRAS.

Die auftretenden Koeffizienten c_j, $j \in \{1, \ldots, 7\}$ und γ_i, $i \in \{1, 2, 3\}$ hängen von der Topographie und den vorgegebenen Gitterweiten im Modellgebiet ab. Im allgemeinen Fall einer dreidimensionalen Problemstellung mit nicht verschwindender Topographie und einem nicht äquidistanten Gitter ergibt sich ein 15-Punkte-Differenzenstern. Für den Spezialfall mit verschwindender Topographie und äquidistanten Schrittweiten vereinfacht sich dieser auf den Standard 7-Punkte-Differenzenstern aus der direkten Diskretisierung des Laplace-Operators. Die exakte Darstellung der 15 Einträge des Differenzensterns ist in [SBL+96] oder [Sch07] zu finden.

Aufgrund dieser speziellen Struktur der Problemstellung im Modell METRAS wird als Modellproblem das diskrete Randwertproblem

$$\begin{aligned} -Lu &= f \quad \text{in } \Omega_h \\ u &= 0 \quad \text{auf } \partial\Omega_h \end{aligned} \tag{2.10}$$

verwendet, wobei Ω_h durch ein $n \times m$ Rechteckgitter mit konstanter Schrittweite h, $f \in \mathcal{D}_h(\Omega_h)$ und der Differenzenoperator durch

$$Lu = (au_x)_{\bar{x}} + (du_y)_{\bar{y}} \tag{2.11}$$

mit ortsabhängigen Koeffizienten $a, d \in \mathcal{D}_h\left(\overline{\Omega}_h\right)$, $a, d > 0$ gegeben sind.

Im Vergleich zur Problemstellung im Modell METRAS wurden im Modellproblem mehrere Vereinfachungen vorgenommen. Die Beschränkung auf den zweidimensionalen Fall erfolgte, um eine übersichtliche Darstellung der wesentlichen Überlegungen zu ermöglichen. Eine Verallgemeinerung zum dreidimensionalen Fall ist meist aufwendig, aber unproblematisch (vgl. Kapitel 5). Die Beschränkung auf Rechteckgitter mit konstanter Schrittweite h wurde auf Grundlage der Gegebenheiten des numerischen Gitters im Modell METRAS vorgenommen, das durch ein Gitter mit konstanter Schrittweite gegeben ist. Die Struktur des Differenzenoperators ist in Anlehnung an (2.8) und (2.9) so gewählt, dass sich die Diskretisierungsmatrix aus der Hintereinanderausführung zweier Diskretisierungen ergibt, wobei die allgemeinen Differenzenoperatoren δ und $\tilde{\delta}$ im Modellproblem durch Rückwärts- bzw. Vorwärtsdifferenzen konkretisiert wurden. Die allgemeinen ortsabhängigen Koeffizienten $a, d \in \mathcal{D}_h\left(\overline{\Omega}_h\right)$ wurden in Anlehnung an die Koeffizienten in (2.8) gewählt. Wie im Modell METRAS ergibt sich bei Verwendung der konstanten Koeffizienten $a = d = 1$ die Standard-Diskretisierung des Laplace-Operators. Um die Besonderheiten im Zusammenhang mit Neumann-Randbedingungen zu umgehen (vgl. Abschnitt 6.2.2), wurde die Randbedingung im Modellproblem durch homogene Dirichlet-Randwerte ersetzt.

Der Differenzenstern für das Modellproblem am Gitterpunkt $(x_i, y_j) \in \Omega_h$ ergibt sich zu

$$\frac{1}{h^2}\begin{bmatrix} & -d(x_i, y_j) & \\ -a(x_{i-1}, y_j) & c(x_i, y_j) & -a(x_i, y_j) \\ & -d(x_i, y_{j-1}) & \end{bmatrix}$$

mit $c(x_i, y_j) = [a(x_i, y_j) + a(x_{i-1}, y_j) + d(x_i, y_j) + d(x_i, y_{j-1})]$. Nach Elimination der Randpunkte erhält man ein lineares Gleichungssystem der Form

$$L_h u = f_h \tag{2.12}$$

mit $L_h \in \mathbb{R}^{I \times I}, f_h \in \mathbb{R}^I$, $I = \{1, \ldots, nm\}$. Die Struktur der Matrix L_h hängt von der Nummerierung der Gitterpunkte bzw. der Anordnung der Indexmenge ab. Sie ist jedoch stets dünnbesetzt und symmetrisch. Außerdem weist sie weitere Besonderheiten auf, die im Zusammenhang mit der Diskretisierung elliptischer Differentialgleichungen mittels Finiter-Differenzen-Verfahren von Bedeutung sind. Dazu wird das Konzept der M-Matrix eingeführt:

Definition 2.1.1 *Eine Matrix $A = (a_{ij})_{i,j \in I} \in \mathbb{R}^{I \times I}$ heißt* M*-Matrix, wenn*

$$a_{ii} > 0 \qquad \text{für alle } i \in I, \tag{2.13}$$

$$a_{ij} \leq 0 \qquad \text{für alle } i \neq j, \tag{2.14}$$

$$A \text{ nichtsingulär und } A^{-1} \geq 0 \tag{2.15}$$

gilt, wobei die letzte Ungleichung elementweise zu verstehen ist.

Die Bedingungen (2.13) und (2.14) sind für die Diskretisierungsmatrix L_h erfüllt. Zum Nachweis der M-Matrix Eigenschaft der Diskretisierungsmatrix kann ein Kriterium aus [Hac86] verwendet werden.

Definition 2.1.2 *1. Eine Matrix A heißt irreduzibel, wenn keine Permutationsmatrix P existiert, so dass die resultierende Matrix die Gestalt*

$$PAP^T = \begin{pmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & \tilde{A}_{22} \end{pmatrix}$$

besitzt.

2. *Eine Matrix $A = (a_{ij})_{i,j \in I}$ heißt diagonaldominant, falls*

$$\sum_{j \in I, j \neq i} |a_{ij}| < |a_{ii}| \tag{2.16}$$

für alle $i \in I$ gilt.

3. *Eine Matrix $A = (a_{ij})_{i,j \in I}$ heißt schwach diagonaldominant, falls*

$$\sum_{j \in I, j \neq i} |a_{ij}| \leq |a_{ii}|$$

für alle $i \in I$ gilt.

4. *Eine Matrix A heißt irreduzibel diagonaldominant, falls A irreduzibel und schwach diagonaldominant ist und für mindestens einen Index $i \in I$ die Ungleichung* (2.16) *erfüllt ist.*

Kriterium 2.1.3 ([Hac86, Kriterium 4.3.10]) *Ist eine Matrix A mit den Eigenschaften* (2.13) *und* (2.14) *diagonaldominant oder irreduzibel diagonaldominant, so ist A eine* M*-Matrix.*

Durch die Vorgabe $a, d > 0$ ist die Diskretisierungsmatrix des Modellproblems stets irreduzibel diagonaldominant, so dass sie unabhängig von der Wahl der Koeffizienten nach Kriterium 2.1.3 eine M-Matrix ist. Im Hinblick auf die Invertierung in der $\mathcal{H}$-Arithmetik ist folgendes Resultat hilfreich.

Kriterium 2.1.4 ([Hac86, Kriterium 4.3.24]) *Ist eine symmetrische Matrix mit positiven Diagonalelementen diagonaldominant oder irreduzibel diagonaldominant, so ist sie positiv definit.*

Demnach ist die Diskretisierungsmatrix ebenfalls positiv definit, so dass alle Hauptuntermatrizen invertierbar sind. Dies ermöglicht die praktische Durchführung der $\mathcal{H}$-Invertierung für das Modellproblem (vgl. Abschnitt 3.1.3).

2.1.2 Diskrete Greensche Funktion

Analog zum Konzept der Greenschen Funktion kann zu Finite-Differenzen-Diskretisierungen auf dem Gitter Ω_h mit Differenzenoperator L die diskrete Greensche Funktion $g_h(x,\xi) \in \mathcal{D}_h(\Omega_h \times \Omega_h)$ eingeführt werden. Diese ist durch $h^{-2}L_h^{-1}$ gegeben und besitzt ähnliche Eigenschaften wie im kontinuierlichen Fall.

Der Definitionsbereich lässt sich auf $\overline{\Omega}_h \times \overline{\Omega}_h$ erweitern, indem

$$\tilde{g}_h(x,\xi) = \begin{cases} g_h(x,\xi) & x,\xi \in \Omega_h \\ 0 & x \in \partial\Omega_h \text{ oder } \xi \in \partial\Omega_h \end{cases}$$

gesetzt wird. Analog zur (kontinuierlichen) Greenschen Funktion ist die diskrete Greensche Funktion für festes $\xi \in \Omega_h$ Lösung des diskreten Randwertproblems

$$\begin{aligned} \left(-L\tilde{g}_h\right)(x,\xi) &= \delta_\xi(x) && x \in \Omega_h \\ \tilde{g}_h(x,\xi) &= 0 && x \in \partial\Omega_h \end{aligned}$$

mit

$$\delta_\xi(x) := \begin{cases} h^{-2} & \text{falls } x = \xi \\ 0 & \text{falls } x \neq \xi. \end{cases}$$

Der Differenzenoperator bezieht sich in diesem Fall auf die Variable x. Die Lösung u des diskreten Dirichlet-Problems (2.10) lässt sich demnach darstellen als

$$u(x) = h^2 \sum_{\xi \in \Omega_h} \tilde{g}_h(\xi, x) f(\xi), \quad x \in \overline{\Omega}_h.$$

Der Faktor h^{-2} in δ_ξ wurde eingeführt, damit die Summation aus der diskreten Darstellung der Lösung die Integration der kontinuierlichen Darstellung mit der Greenschen Funktion $G(x,\xi)$

$$u(x) = \int_\Omega G(\xi, x) f(\xi)\, d\xi$$

approximiert.

Für festes $\xi \in \Omega_h$ ist $\tilde{g}_h(\cdot,\xi)$ eine Gitterfunktion auf $\overline{\Omega}_h$ und es gilt

$$(L\tilde{g}_h)(x,\xi) = 0 \text{ für alle } x \in \Omega_h \backslash \{\xi\}.$$

Die diskrete Greensche Funktion erfüllt demnach für alle $x \in \Omega_h \backslash \{\xi\}$ die homogene Differenzengleichung mit dem Differenzenoperator L, woraus die folgende Definition motiviert wird:

Definition 2.1.5 (Diskret *L*-harmonische Funktionen)
Eine Gitterfunktion $u \in \mathcal{D}_h\left(\overline{\Omega}_h\right)$ heißt diskret L-harmonisch im Gitterpunkt $x \in \Omega_h$, wenn die Differenzengleichung

$$(Lu)(x) = 0$$

erfüllt ist.

Gilt für $K_h \subset \Omega_h$

$$(Lu)(x) = 0 \quad \textit{für alle } x \in K_h,$$

so heißt u *diskret harmonisch auf* K_h*. Der Raum der diskret* L*-harmonischen Gitterfunktionen auf einem Gitter* K_h *wird mit* $Z_h^L(K_h)$ *bezeichnet.*

Ist demnach ein Teilgitter $X_h \subset \Omega_h$ gegeben, dann gilt für die diskrete Greensche Funktion zum Modellproblem für jedes $\xi \in \Omega_h \backslash X_h$

$$g_h(\cdot, \xi) \in Z_h^L(X_h).$$

Zum Beweis einer speziellen Form der diskreten Cacciopoli-Ungleichung werden Gitterfunktionen $u \in \mathcal{D}_{h,0}\left(\overline{\Omega}_h\right)$ betrachtet, die diskret L-harmonisch auf einem Teilgitter $K_h \cap \Omega_h$ sind, wobei das Gitter $K_h \subset h\mathbb{Z}^2$ auch Gitterpunkte enthalten kann, die nicht zu Ω_h gehören. Dazu wird die Bezeichnung $Z_{h,0}^L(K_h; \Omega_h)$ eingeführt.

2.2 Diskrete Poincaré-Ungleichung

Als wesentliches Hilfsmittel zum Beweis von Lemma 4.1.1 in Abschnitt 4.1.1 wird eine diskrete Variante der Poincaré-Ungleichung für Gitterfunktionen mit verschwindendem Mittelwert benötigt. Ein ähnliches Resultat ist beispielsweise in [Sül91] oder [Tem77] angegeben. Dieses liefert eine Ungleichung für den speziellen Fall von Gitterfunktionen, die homogene Randwerte besitzen. In Abschnitt 4.1.1 wird jedoch ein alternatives Ergebnis für Gitterfunktionen mit verschwindendem Mittelwert benötigt. Außerdem ist es erforderlich, die Konstante in der Ungleichung explizit angeben zu können. Ein solches Resultat ist nicht bekannt, so dass die Ungleichung im folgenden Abschnitt in der benötigten Form bewiesen wird.

Im kontinuierlichen Fall für konvexe Gebiete Ω kann die Ungleichung mit der Konstanten

$$C_p = \frac{\operatorname{diam}(\Omega)}{\pi}$$

gezeigt werden (vgl. [Beb03]). Eine Übertragung des Beweises für Gitterfunktionen ist nicht möglich. Es lassen sich jedoch aus dem Beweis der Ungleichung für Gitterfunktionen mit homogenen Randwerten grundlegende Ideen (Verwendung von Teleskopsummen und Abschätzung mittels der Ungleichung von Cauchy-Schwarz) übernehmen. Der Beweis des eindimensionalen Resultats kann daher auf ähnliche Weise erfolgen, nur die Bestimmung der Konstanten fällt durch das Auftreten von Mehrfachsummen komplexer aus.

Die Erweiterung auf den höherdimensionalen Fall ist jedoch deutlich aufwendiger als für Gitterfunktionen mit homogenen Randwerten und es muss eine neue Beweisstrategie eingeführt werden. Bei dieser wird der höherdimensionale auf den eindimensionalen Fall zurückgeführt. Der darauf aufbauende Beweis für Gitterfunktionen mit verschwindendem Mittelwert im höherdimensionalen Fall wird jedoch durch das Auftreten von Mehrfachsummen deutlich komplexer. Dadurch fallen die anschließenden Abschätzungen und insbesondere die Abschätzung der Konstanten in der Ungleichung deutlich aufwendiger aus als für Gitterfunktionen mit homogenen Randwerten. Die Gültigkeit der diskreten

Ungleichung lässt sich auf diese Weise sowohl im eindimensionalen Fall als auch für Rechteckgitter Ω_h im zweidimensionalen Fall mit der Konstanten

$$C_h = \frac{\operatorname{diam}(\Omega_h)}{\sqrt{2}}$$

zeigen. In Abschnitt 5.1 erhält man auch für den dreidimensionalen Fall von Quadergittern die Gültigkeit der Ungleichung mit der gleichen Konstanten.

2.2.1 Eindimensionale Ungleichung

Zum Beweis der eindimensionalen diskreten Poincaré-Ungleichung wird zunächst ein Lemma gezeigt:

Lemma 2.2.1 *Sei $\Omega_h = \{x_i \in h\mathbb{Z}^1 : 1 \leq i \leq n\}$ ein Gitter mit $n \in \mathbb{N}$ Gitterpunkten und $u \in \mathcal{D}_h(\Omega_h)$ eine Gitterfunktion. Dann gilt*

$$u(x_i) - \overline{u} = \frac{h}{n} \sum_{\kappa=1}^{n-1} c(\kappa, i) u_x(\kappa h)$$

mit

$$c(\kappa, i) = \begin{cases} \kappa & 1 \leq \kappa \leq i-1 \\ -(n-\kappa) & i \leq \kappa \leq n-1 \end{cases}$$

für alle $x_i \in \Omega_h$.

Beweis Da Ω_h zusammenhängend ist, erhält man für $u \in \mathcal{D}_h(\Omega_h)$ und $x_i, x_j \in \Omega_h$ im Fall $x_i < x_j$

$$u(x_i) - u(x_j) = \sum_{t=i}^{j-1} (u(x_t) - u(x_{t+1})) = -h \sum_{t=i}^{j-1} u_x(x_t)$$

und für $x_j < x_i$ analog

$$u(x_i) - u(x_j) = h \sum_{t=j}^{i-1} u_x(x_t).$$

Summiert man über j und teilt durch n, erhält man

$$\begin{aligned} u(x_i) - \overline{u} &= \frac{1}{n} \sum_{j=1}^{i-1} (u(x_i) - u(x_j)) + \frac{1}{n} \sum_{j=i+1}^{n} (u(x_i) - u(x_j)) \\ &= \frac{h}{n} \sum_{j=1}^{i-1} \sum_{t=j}^{i-1} u_x(x_t) - \frac{h}{n} \sum_{j=i+1}^{n} \sum_{t=i}^{j-1} u_x(x_t) \\ &= \frac{h}{n} \sum_{t=1}^{i-1} t u_x(x_t) - \frac{h}{n} \sum_{t=i}^{n-1} (n-t) u_x(x_t), \end{aligned}$$

woraus die Behauptung folgt. ∎

Unter Verwendung von Lemma 2.2.1 lässt sich direkt die diskrete Poincaré-Ungleichung für Gitterfunktionen mit verschwindendem Mittelwert im eindimensionalen Fall beweisen:

Satz 2.2.2 *Sei $\Omega_h \subset h\mathbb{Z}^1$ ein zusammenhängendes Gitter mit der konstanten Schrittweite h. Dann gilt*

$$\|u - \bar{u}\|_{L_h^2(\Omega_h)} \leq C_h \|u_x\|_{L_h^2(\Omega_h)} \quad \textit{für alle } u \in \mathcal{D}_h(\Omega_h) \tag{2.17}$$

mit der Konstanten

$$C_h = \frac{\operatorname{diam}(\Omega_h)}{\sqrt{2}}.$$

Beweis Das Gitter sei ohne Einschränkung durch $\Omega_h = \{x_i \in h\mathbb{Z}^1 : 1 \leq i \leq n\}$ mit $n \in \mathbb{N}$ Gitterpunkten gegeben. Nach Lemma 2.2.1 gilt dann für $u \in \mathcal{D}_h(\Omega_h)$

$$u(x_i) - \bar{u} = \frac{h}{n} \sum_{\kappa=1}^{n-1} c(\kappa, i) u_x(\kappa h)$$

mit

$$c(\kappa, i) = \begin{cases} \kappa & 1 \leq \kappa \leq i-1 \\ -(n-\kappa) & i \leq \kappa \leq n-1 \end{cases}$$

für alle $x_i \in \Omega_h$. Daher kann folgendermaßen summiert und abgeschätzt werden:

$$\begin{aligned}
h \sum_{i=1}^{n} |u(x_i) - \bar{u}|^2 &= h \sum_{i=1}^{n} \left| \frac{h}{n} \sum_{\kappa=1}^{n-1} c(\kappa, i) u_x(\kappa h) \right|^2 \\
&\leq h \frac{h^2}{n^2} \sum_{i=1}^{n} \left(\sum_{\kappa=1}^{n-1} c(\kappa, i)^2 \right) \left(\sum_{\kappa=1}^{n-1} |u_x(\kappa h)|^2 \right) \\
&= \frac{h^2}{n^2} \sum_{i=1}^{n} \left(\sum_{k=1}^{i-1} k^2 + \sum_{k=i}^{n-1} (n-k)^2 \right) \left(h \sum_{\kappa=1}^{n-1} |u_x(\kappa h)|^2 \right) \\
&= \frac{h^2}{n^2} \frac{1}{6} (n^4 - n^2) \left(h \sum_{\kappa=1}^{n-1} |u_x(\kappa h)|^2 \right) \\
&= \frac{h^2 (n^2-1)}{6} \left(h \sum_{\kappa=1}^{n-1} |u_x(\kappa h)|^2 \right).
\end{aligned}$$

Demnach ist die Ungleichung (2.17) für den Fall $n = 1$ mit $C_h = 0$ erfüllt, so dass die Abschätzung der Konstanten in der Ungleichung für $n > 1$ erfolgen kann. Unter dieser Voraussetzung erhält man

$$n^2 - 1 = (n-1)^2 \left(1 + \frac{2}{(n-1)} \right) \leq 3(n-1)^2,$$

so dass

$$\frac{h^2(n^2-1)}{6} \leq \frac{h^2(n-1)^2}{2} = \frac{\operatorname{diam}(\Omega_h)^2}{2}$$

gilt, woraus die Gültigkeit der Ungleichung (2.17) mit $C_h = \frac{\operatorname{diam}(\Omega_h)}{\sqrt{2}}$ folgt. ∎

2.2.2 Zweidimensionale Ungleichung

Zum Beweis der zweidimensionalen diskreten Poincaré-Ungleichung werden die Ergebnisse aus dem eindimensionalen Fall genutzt. Zunächst wird erneut ein Lemma (analog zu Lemma 2.2.1) gezeigt:

Lemma 2.2.3 *Sei $\Omega_h = \{(x_i, y_j) \in h\mathbb{Z}^2 : 1 \leq i \leq n, 1 \leq j \leq m\}, n, m \in \mathbb{N}$ ein $n \times m$ Rechteckgitter mit konstanter Schrittweite h und $u \in \mathcal{D}_h(\Omega_h)$ eine Gitterfunktion auf Ω_h mit dem Mittelwert $\bar{u}$. Dann gilt*

$$u(x_i, y_l) - \bar{u} = \frac{h}{nm} \sum_{j=1}^{n} \sum_{\kappa=1}^{n+m-2} \tilde{c}(\kappa, i, l) \partial_\kappa^{j,l} u$$

für alle $(x_i, y_l) \in \Omega_h$ mit

$$\tilde{c}(\kappa, i, l) = \begin{cases} \frac{m}{n} c_1(\kappa, i) & 1 \leq \kappa \leq n-1 \\ c_2(\kappa - (n-1), l) & n \leq \kappa \leq n+m-2 \end{cases}$$

$$\partial_\kappa^{j,l} u = \begin{cases} u_x(\kappa h, y_l) & 1 \leq \kappa \leq n-1 \\ u_y(x_j, (\kappa - (n-1))h) & n \leq \kappa \leq n+m-2, \end{cases}$$

$$c_1(\kappa, i) = \begin{cases} \kappa & 1 \leq \kappa \leq i-1 \\ -(n-\kappa) & i \leq \kappa \leq n-1 \end{cases}$$

und

$$c_2(\kappa, l) = \begin{cases} \kappa & 1 \leq \kappa \leq l-1 \\ -(m-\kappa) & l \leq \kappa \leq m-1. \end{cases}$$

Außerdem lässt sich abschätzen:

$$|u(x_i, y_l) - \bar{u}|^2 \leq \frac{h^2}{(nm)^2} n \left(\sum_{\kappa=1}^{n+m-2} \tilde{c}(\kappa, i, l)^2 \right) \sum_{j=1}^{n} \sum_{\kappa=1}^{n+m-2} \left| \partial_\kappa^{j,l} u \right|^2.$$

Beweis In einem Rechteckgitter Ω_h existiert zu je zwei Punkten $(x_i, y_l), (x_j, y_k) \in \Omega_h$ immer ein rechtwinkliger Weg über benachbarte Gitterpunkte aus Ω_h, der über den Punkt $(x_j, y_l) \in \Omega_h$ verläuft. Demnach kann in der Darstellung

$$u(x_i, y_l) - u(x_j, y_k) = (u(x_i, y_l) - u(x_j, y_l)) + (u(x_j, y_l) - u(x_j, y_k))$$

für beide Summanden auf der rechten Seite Lemma 2.2.1 aus dem eindimensionalen Fall verwendet werden, so dass man

$$\frac{1}{n}\sum_{j=1}^{n}\left(u(x_i,y_l)-u(x_j,y_l)\right)=\frac{h}{n}\sum_{\kappa=1}^{n-1}c_1(\kappa,i)u_x(\kappa h,y_l)$$

und

$$\frac{1}{m}\sum_{k=1}^{m}\left(u(x_j,y_l)-u(x_j,y_k)\right)=\frac{h}{m}\sum_{\kappa=1}^{m-1}c_2(\kappa,l)u_y(x_j,\kappa h)$$

für alle $(x_i,y_l),(x_j,y_k)\in\Omega_h$ mit

$$c_1(\kappa,i)=\begin{cases}\kappa & 1\le\kappa\le i-1\\ -(n-\kappa) & i\le\kappa\le n-1\end{cases}$$

und

$$c_2(\kappa,l)=\begin{cases}\kappa & 1\le\kappa\le l-1\\ -(m-\kappa) & l\le\kappa\le m-1\end{cases}$$

erhält. Demnach ergibt sich

$$\begin{aligned}u(x_i,y_l)-\bar{u}&=\frac{h}{nm}\left(\sum_{k=1}^{m}\sum_{\kappa=1}^{n-1}c_1(\kappa,i)u_x(\kappa h,y_l)+\sum_{j=1}^{n}\sum_{\kappa=1}^{m-1}c_2(\kappa,l)u_y(x_j,\kappa h)\right)\\&=\frac{h}{nm}\left(\sum_{j=1}^{n}\frac{m}{n}\sum_{\kappa=1}^{n-1}c_1(\kappa,i)u_x(\kappa h,y_l)+\sum_{j=1}^{n}\sum_{\kappa=1}^{m-1}c_2(\kappa,l)u_y(x_j,\kappa h)\right)\\&=\frac{h}{nm}\sum_{j=1}^{n}\left(\sum_{\kappa=1}^{n-1}\frac{m}{n}c_1(\kappa,i)u_x(\kappa h,y_l)+\sum_{\kappa=1}^{m-1}c_2(\kappa,l)u_y(x_j,\kappa h)\right).\end{aligned}$$

Daraus folgt der erste Teil der Behauptung.

Durch zweimalige Anwendung der Ungleichung von Cauchy-Schwarz erhält man den zweiten Teil der Behauptung:

$$\begin{aligned}|u(x_i,y_l)-\bar{u}|^2&\le\frac{h^2}{(nm)^2}n\sum_{j=1}^{n}\left|\sum_{\kappa=1}^{n+m-2}\tilde{c}(\kappa,i,l)\partial_\kappa^{j,l}u\right|^2\\&\le\frac{h^2}{(nm)^2}n\left(\sum_{\kappa=1}^{n+m-2}\tilde{c}^2(\kappa,i,l)\right)\sum_{j=1}^{n}\sum_{\kappa=1}^{n+m-2}\left|\partial_\kappa^{j,l}u\right|^2.\end{aligned}$$

■

Die folgenden Ergebnisse werden ebenfalls zum Beweis der zweidimensionalen diskreten Poincaré-Ungleichung verwendet:

Lemma 2.2.4 *Sei $\tilde{c}$ wie in Lemma 2.2.3 mit $n, m \in \mathbb{N}$ gegeben, dann gilt*

$$\sum_{i=1}^{n}\sum_{l=1}^{m}\sum_{\kappa=1}^{n+m-2} \tilde{c}^2(\kappa,i,l) = \frac{m^2}{6}\left(m(n^2-1)+n(m^2-1)\right).$$

Beweis Mittels der Definition von $\tilde{c}$ lässt sich die Summe direkt berechnen durch

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{l=1}^{m}\sum_{\kappa=1}^{n+m-2} \tilde{c}^2(\kappa,i,l) &= \sum_{i=1}^{n}\sum_{l=1}^{m}\left(\sum_{\kappa=1}^{n-1}\frac{m^2}{n^2}c_1^2(\kappa,i)+\sum_{\kappa=1}^{m-1}c_2^2(\kappa,l)\right)\\
&= \sum_{l=1}^{m}\left(\frac{m^2}{n^2}\sum_{i=1}^{n}\sum_{\kappa=1}^{n-1}c_1^2(\kappa,i)\right)+\left(\sum_{i=1}^{n}\sum_{l=1}^{m}\sum_{\kappa=1}^{m-1}c_2^2(\kappa,l)\right)\\
&= \frac{m^3}{n^2}\frac{1}{6}(n^4-n^2)+\frac{1}{6}n(m^4-m^2)\\
&= \frac{m^2}{6}\left(m(n^2-1)+n(m^2-1)\right).
\end{aligned}$$

∎

Lemma 2.2.5 *Sei $\tilde{c}$ wie in Lemma 2.2.3 mit $n, m \in \mathbb{N}$ gegeben, dann gilt*

$$\max_{l\in[1,m]}\sum_{i=1}^{n}\sum_{\kappa=1}^{n+m-2} \tilde{c}^2(\kappa,i,l) = \frac{1}{6}m\left(m(n^2-1)+n(2m^2-3m+1)\right).$$

Beweis Nach Definition von $\tilde{c}$ erhält man

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{\kappa=1}^{n+m-2} \tilde{c}(\kappa,i,l)^2 &= \sum_{i=1}^{n}\sum_{\kappa=1}^{n-1}\frac{m^2}{n^2}c_1^2(\kappa,i)+\sum_{i=1}^{n}\sum_{\kappa=1}^{m-1}c_2^2(\kappa,l)\\
&= \frac{m^2}{n^2}\sum_{i=1}^{n}\sum_{\kappa=1}^{n-1}c_1^2(\kappa,i)+\sum_{i=1}^{n}\sum_{\kappa=1}^{m-1}c_2^2(\kappa,l)\\
&= \frac{m^2}{n^2}\frac{1}{6}(n^4-n^2)+n\left(\sum_{\kappa=1}^{l-1}\kappa^2+\sum_{\kappa=l}^{m-1}(m-\kappa)^2\right).
\end{aligned}$$

Für die Summation ergibt sich

$$\sum_{\kappa=1}^{l-1}\kappa^2+\sum_{\kappa=l}^{m-1}(m-\kappa)^2 = l(lm-m^2-m)+\frac{1}{6}\left(2m^3+3m^2+m\right) =: \gamma(l).$$

Das Maximum von $\gamma(l)$ wird für $l \in [1,m]$ am Rand angenommen und dort erhält man

$$\gamma(1) = \gamma(m) = \frac{1}{6}\left(2m^3-3m^2+m\right).$$

Zusammengefasst ergibt dies

$$\max_{l\in[1,m]}\sum_{i=1}^{n}\sum_{\kappa=1}^{n+m-2} \tilde{c}(\kappa,i,l)^2 = \frac{1}{6}m\left(m(n^2-1)+n(2m^2-3m+1)\right),$$

womit die Behauptung gezeigt ist. ∎

Lemma 2.2.6 *Es gelten die Ungleichungen*

$$\frac{(n^2-1)+\frac{n}{m}(2m^2-3m+1)}{(n-1)^2+(m-1)^2} \leq 3$$

und

$$\frac{\frac{m}{n}(n^2-1)+(m^2-1)}{(n-1)^2+(m-1)^2} \leq 3$$

für alle $n, m \in \mathbb{N}$ mit $n, m \geq 2$.

Beweis Für $n, m \in \mathbb{N}$ mit $n, m \geq 2$ gilt

$$\begin{aligned}(n^2-1)+\frac{n}{m}(2m^2-3m+1) &\leq n^2-1+2nm-\frac{5}{2}n \\ &\leq 2n^2-\frac{5}{2}n+m^2-1.\end{aligned}$$

Mittels vollständiger Induktion kann gezeigt werden, dass $2n^2-\frac{5}{2}n \leq 3(n-1)^2$ und $m^2-1 \leq 3(m-1)^2$ für $n, m \in \mathbb{N}$ mit $n, m \geq 2$ gelten, woraus der erste Teil der Behauptung folgt.

Für den zweiten Teil lässt sich die Abschätzung

$$\frac{m}{n}(n^2-1)+(m^2-1) \leq \frac{1}{2}m^2+\frac{1}{2}n^2-\frac{2}{n}+m^2-1$$

nutzen. Mittels vollständiger Induktion kann gezeigt werden, dass $\frac{3}{2}m^2-1 \leq 3(m-1)^2$ und $\frac{1}{2}n^2-\frac{2}{n} \leq 3(n-1)^2$ für $n, m \in \mathbb{N}$ mit $n, m \geq 2$ gilt, woraus der zweite Teil der Behauptung folgt. ∎

Satz 2.2.7 *Sei $\Omega_h \subset h\mathbb{Z}^2$ ein Rechteckgitter mit konstanter Schrittweite h. Dann gilt*

$$\|u-\bar{u}\|_{L_h^2(\Omega_h)} \leq C_h \|\nabla_h u\|_{L_h^2(\Omega_h)} \quad \text{für alle } u \in \mathcal{D}_h(\Omega_h) \tag{2.18}$$

mit der Konstanten

$$C_h = \frac{\operatorname{diam}(\Omega_h)}{\sqrt{2}}.$$

Beweis Das Rechteckgitter sei ohne Einschränkung durch ein $n \times m$ Rechteckgitter $\Omega_h = \{(x_i, y_j) \in h\mathbb{Z}^2 : 1 \leq i \leq n, 1 \leq j \leq m\}$ mit $n, m \in \mathbb{N}$ gegeben. Für $n = 1$ oder $m = 1$ folgt die Behauptung aus dem eindimensionalen Resultat in Satz 2.2.2. Daher kann im Folgenden $n, m \geq 2$ angenommen werden. Nach Lemma 2.2.3 gilt für Gitterfunktionen $u \in \mathcal{D}_h(\Omega_h)$ die Abschätzung

$$|u(x_i, y_l)-\bar{u}|^2 \leq \frac{h^2}{(nm)^2} n \left(\sum_{\kappa=1}^{n+m-2} \tilde{c}(\kappa, i, l)^2\right) \sum_{j=1}^{n} \sum_{\kappa=1}^{n+m-2} \left|\partial_\kappa^{j,l} u\right|^2 \tag{2.19}$$

für alle $(x_i, y_l) \in \Omega_h$.

Summiert man (2.19) über i und l und multipliziert mit h^2, erhält man

$$\begin{aligned}
\|u-\bar{u}\|^2_{L^2_h(\Omega_h)} &\leq h^2 \sum_{i=1}^{n}\sum_{l=1}^{m} \frac{h^2}{(nm)^2} n \left(\sum_{\kappa=1}^{n+m-2} \tilde{c}(\kappa,i,l)^2 \right) \sum_{j=1}^{n} \sum_{\kappa=1}^{n+m-2} \left| \partial_\kappa^{j,l} u \right|^2 \\
&= h^2 \sum_{i=1}^{n}\sum_{l=1}^{m} \frac{h^2}{(nm)^2} n \left(\sum_{\kappa=1}^{n+m-2} \tilde{c}(\kappa,i,l)^2 \right) \sum_{j=1}^{n} \left(\sum_{\kappa=1}^{n-1} |u_x(\kappa h, y_l)|^2 \right. \\
&\quad \left. + \sum_{\kappa=1}^{m-1} |u_y(x_j,(\kappa-1)h)|^2 \right) \\
&= h^2 \sum_{i=1}^{n}\sum_{l=1}^{m} \frac{h^2}{(nm)^2} n \left(\sum_{\kappa=1}^{n+m-2} \tilde{c}(\kappa,i,l)^2 \right) \sum_{j=1}^{n} \sum_{\kappa=1}^{n-1} |u_x(\kappa h, y_l)|^2 \\
&\quad + h^2 \sum_{i=1}^{n}\sum_{l=1}^{m} \frac{h^2}{(nm)^2} n \left(\sum_{\kappa=1}^{n+m-2} \tilde{c}(\kappa,i,l)^2 \right) \sum_{j=1}^{n} \sum_{\kappa=1}^{m-1} |u_y(x_j, \kappa h)|^2 \\
&= h^2 \sum_{i=1}^{n}\sum_{l=1}^{m} \frac{h^2}{(nm)^2} n^2 \left(\sum_{\kappa=1}^{n+m-2} \tilde{c}(\kappa,i,l)^2 \right) \sum_{\kappa=1}^{n-1} |u_x(\kappa h, y_l)|^2 \\
&\quad + \sum_{i=1}^{n}\sum_{l=1}^{m} \frac{h^2}{(nm)^2} n \left(\sum_{\kappa=1}^{n+m-2} \tilde{c}(\kappa,i,l)^2 \right) h^2 \sum_{j=1}^{n} \sum_{\kappa=1}^{m-1} |u_y(x_j, \kappa h)|^2 \\
&\leq \left(h^2 \sum_{l=1}^{m} \sum_{\kappa=1}^{n-1} |u_x(\kappa h, y_l)|^2 \right) \cdot \frac{h^2}{m^2} \max_{l\in[1,m]} \left[\sum_{i=1}^{n} \left(\sum_{\kappa=1}^{n+m-2} \tilde{c}(\kappa,i,l)^2 \right) \right] \\
&\quad + \left(h^2 \sum_{j=1}^{n} \sum_{\kappa=1}^{m-1} |u_y(x_j, \kappa h)|^2 \right) \cdot \frac{h^2}{m^2 n} \sum_{i=1}^{n}\sum_{l=1}^{m} \left(\sum_{\kappa=1}^{n+m-2} \tilde{c}(\kappa,i,l)^2 \right).
\end{aligned}$$

Mit Hilfe der Ergebnisse aus Lemma 2.2.4 und Lemma 2.2.5, kann folgende Abschätzung angegeben werden:

$$\begin{aligned}
\|u-\bar{u}\|^2_{L^2_h(\Omega_h)} &\leq \frac{h^2}{m}\frac{1}{6} \left(m(n^2-1) + n(2m^2-3m+1) \right) h^2 \sum_{l=1}^{m}\sum_{\kappa=1}^{n-1} |u_x(\kappa h, y_l)|^2 \\
&\quad + \frac{h^2}{6n} \left(m(n^2-1) + n(m^2-1) \right) h^2 \sum_{j=1}^{n}\sum_{\kappa=1}^{m-1} |u_y(x_j, \kappa h)|^2 \\
&= \frac{\operatorname{diam}(\Omega_h)^2}{6} \left[\frac{(n^2-1) + \frac{n}{m}(2m^2-3m+1)}{(n-1)^2+(m-1)^2} h^2 \sum_{l=1}^{m}\sum_{\kappa=1}^{n-1} |u_x(\kappa h, y_l)|^2 \right. \\
&\quad \left. + \frac{\frac{m}{n}(n^2-1) + (m^2-1)}{(n-1)^2+(m-1)^2} h^2 \sum_{j=1}^{n}\sum_{\kappa=1}^{m-1} |u_y(x_j, \kappa h)|^2 \right].
\end{aligned}$$

Mit Lemma 2.2.6 erhält man daraus

$$\|u-\bar{u}\|^2_{L^2_h(\Omega_h)} \leq \frac{\operatorname{diam}(\Omega_h)^2}{6} 3 \left(h^2 \sum_{l=1}^{m} \sum_{\kappa=1}^{n-1} |u_x(\kappa h, y_l)|^2 + h^2 \sum_{j=1}^{n} \sum_{\kappa=1}^{m-1} |u_y(x_j, \kappa h)|^2 \right),$$

woraus die Behauptung folgt. ∎

2.3 Diskrete Cacciopoli-Ungleichung

Neben der diskreten Poincaré-Ungleichung wird im weiteren Verlauf der Arbeit eine diskrete Variante der Cacciopoli-Ungleichung für diskret L-harmonische Funktionen auf Rechteckgittern für den Differenzenoperator aus (2.11) benötigt. Wie im Fall der diskreten Poincaré-Ungleichung ist ein Resultat dieser Form für Gitterfunktionen nicht bekannt.

Ein ähnliches Resultat ist jedoch in [CFL28] für diskret harmonische Funktionen (dies entspricht dem Modellproblem mit konstanten Koeffizienten $a = d = 1$) auf Quadratgittern angegeben. Eine Erweiterung des Resultats für Rechteckgitter ist unproblematisch. Außerdem lassen sich die Überlegungen auf diskret L-harmonische Funktionen übertragen, wenn man sich auf den Fall eines Differenzenoperators vom Typ (2.11) mit konstanten Koeffizienten $a(x_i, y_j) = a \in \mathbb{R}^+$ und $d(x_i, y_j) = d \in \mathbb{R}^+$ beschränkt. Durch diese Vorgehensweise kann die diskrete Cacciopoli-Ungleichung für diskret L-harmonische Funktionen (mit konstanten Koeffizienten $a, d \in \mathbb{R}^+$) in Abschnitt 2.3.1 gezeigt werden, wobei die Konstante in der Ungleichung durch $C_{caccio} = 1$ gegeben ist.

Eine Erweiterung dieser Vorgehensweise auf den Fall des allgemeinen Differenzenoperators vom Typ (2.11) mit variablen Koeffizienten $a, d \in \mathcal{D}_h\left(\overline{\Omega}_h\right)$ ist nicht möglich. Daher muss zur Untersuchung des allgemeinen Modellproblems ein anderer Ansatz gewählt werden. Ein weiteres Resultat für Gitterfunktionen ist nicht bekannt, so dass ein neuer Ansatz entwickelt werden muss. Dazu werden Ideen zum Beweis der (kontinuierlichen) Cacciopoli-Ungleichung aus [Hac09] auf den Fall von Gitterfunktionen übertragen. In diesem Zusammenhang ist die Verwendung der diskreten Produktregel (2.1) erforderlich. Diese unterscheidet sich von der kontinuierlichen Variante dadurch, dass die beteiligten Gitterfunktionen an unterschiedlichen Gitterpunkten ausgewertet werden. Dies führt dazu, dass der Beweis der diskreten Variante der Cacciopoli-Ungleichung aufwendiger ausfällt als der in [Hac09]. Das entsprechende Resultat wird in Abschnitt 2.3.2 beschrieben und führt zu einer Ungleichung mit der Konstanten $C_{caccio} = 3\sqrt{2}$. Da diese Konstante deutlich größer ausfällt, als bei der Beschränkung auf konstante Koeffizienten, werden beide Resultate angegeben.

Im Zuge der Untersuchungen werden die Bezeichnungen

$$\lambda_{max} := \max_{x \in \overline{\Omega}_h} \{a(x), d(x)\} \text{ und } \lambda_{min} := \min_{x \in \overline{\Omega}_h} \{a(x), d(x)\}$$

verwendet. Da $a, d > 0$ vorausgesetzt wird, gilt $0 < \lambda_{min} \leq \lambda_{max}$. Außerdem wird eine spezielle Schachtelung von Gittern benötigt, die in der folgenden Konstruktion angegeben ist.

Konstruktion 2.3.1 (Gitterschachtelung)
Ausgehend von einem Rechteckgitter $K_h \subset h\mathbb{Z}^2$ werden zur Konstruktion einer Schachtelung von $l+1$ Gittern

$$K_h =: K_h^0 \subset K_h^1 \subset \cdots \subset K_h^l$$

mit $K_h^i \subset h\mathbb{Z}^2, i = 1, \ldots, l$ angefangen bei K_h^0 die Gitter sukzessive um die Randpunkte erweitert, so dass die Konstruktion der Gitterschachtelung durch

$$K_h^{i+1} = K_h^i \cup \partial K_h^i, \quad i = 0, \ldots, l-1$$

erfolgt.

Für den Beweis der diskreten Cacciopoli-Ungleichung in Satz 2.3.2 für konstante Koeffizienten werden zur Kennzeichnung spezieller Teilmengen von Gitterpunkten eines $n \times m$ Rechteckgitters $K_h = \{(x_i, y_j) \in h\mathbb{Z}^2 : 1 \le i \le n, 1 \le j \le m\} \subset h\mathbb{Z}^2$ die folgenden Bezeichnungen eingeführt:

$$\begin{aligned} \Gamma^{x+}(K_h) &:= \{(x_n, y_j) \in h\mathbb{Z}^2 : 1 \le j \le m\} \subset \Gamma(K_h) \\ \Gamma^{x-}(K_h) &:= \{(x_1, y_j) \in h\mathbb{Z}^2 : 1 \le j \le m\} \subset \Gamma(K_h) \\ \partial^{x+} K_h &:= \{(x_{n+1}, y_j) \in h\mathbb{Z}^2 : 1 \le j \le m\} \subset \partial K_h \\ \partial^{x-} K_h &:= \{(x_0, y_j) \in h\mathbb{Z}^2 : 1 \le j \le m\} \subset \partial K_h. \end{aligned} \tag{2.20}$$

Für die y-Richtung lassen sich analoge Bezeichnungen verwenden. In Abbildung 2.3 sind exemplarisch die Teilmengen $\Gamma^{x+}(K_h), \Gamma^{y-}(K_h), \partial^{x+} K_h$ und $\partial^{y-} K_h$ veranschaulicht.

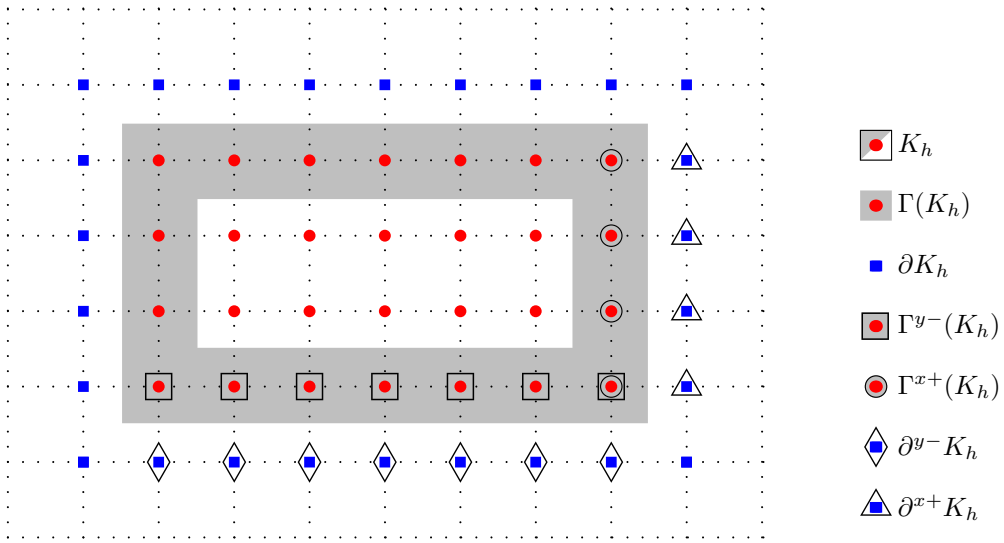


Abbildung 2.3: Bezeichnungen am Rand des Rechteckgitters

2.3.1 Konstante Koeffizienten

Unter Verwendung der eingeführten Bezeichnungen lässt sich das folgende Resultat auf Grundlage der Ausführungen in [CFL28] für den Fall konstanter Koeffizienten $a, d \in \mathbb{R}^+$ beweisen:

Satz 2.3.2 *Sei $K_h \subset h\mathbb{Z}^2$ ein Rechteckgitter mit konstanter Schrittweite h und der Differenzenoperator L von der Form (2.11) mit konstanten Koeffizienten $a, d \in \mathbb{R}^+$. Aus $K_h^0 := K_h$ wird, wie in Konstruktion 2.3.1 beschrieben, eine Schachtelung bis zum Gitter $K_h^l, l \geq 1$ gebildet. Dann gilt*

$$\|\nabla_h u\|_{L_h^2(K_h)} \leq C_{caccio} \frac{\sqrt{\kappa}}{\operatorname{dist}_\infty(K_h, \Gamma(K_h^l))} \|u\|_{L_h^2(K_h^l)}$$

mit $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$ und $C_{caccio} = 1$ für alle $u \in Z_h^L(K_h^l)$.

Beweis Aus $a, d \in \mathbb{R}^+$ folgt $\lambda_{max} \geq \lambda_{min} > 0$, so dass folgende Abschätzung vorgenommen werden kann:

$$\begin{aligned}
\|\nabla_h u\|_{L_h^2(K_h)}^2 &= h^2 \sum_{K_h \setminus \Gamma^{x+}(K_h)} u_x^2 + h^2 \sum_{K_h \setminus \Gamma^{y+}(K_h)} u_y^2 \\
&\leq \frac{1}{\lambda_{min}} \frac{1}{2} \left(h^2 \sum_{\partial^{x-} K_h \cup K_h \setminus \Gamma^{x+}(K_h)} a u_x^2 + h^2 \sum_{K_h} a u_x^2 \right. \qquad (2.21) \\
&\quad \left. + h^2 \sum_{\partial^{y-} K_h \cup K_h \setminus \Gamma^{y+}(K_h)} d u_y^2 + h^2 \sum_{K_h} d u_y^2 \right). \qquad (2.22)
\end{aligned}$$

Für die weiteren Überlegungen werden die Bezeichnungen $u^{x+}(x_i, y_j) = u(x_{i+1}, y_j)$, $u^{x-}(x_i, y_j) = u(x_{i-1}, y_j)$ und analoge Bezeichnungen für die y-Richtung eingeführt. Damit lässt sich die erste Teilsumme aus (2.21) unter Verwendung von partieller Summation darstellen als

$$\begin{aligned}
h^2 \sum_{\partial^{x-} K_h \cup K_h \setminus \Gamma^{x+}(K_h)} a u_x^2 &= h^2 \sum_{\partial^{x-} K_h \cup K_h \setminus \Gamma^{x+}(K_h)} a u_x \frac{1}{h} u^{x+} - h^2 \sum_{\partial^{x-} K_h \cup K_h \setminus \Gamma^{x+}(K_h)} a u_x \frac{1}{h} u \\
&= h^2 \sum_{K_h} a (u_x)^{x-} \frac{1}{h} u - h^2 \sum_{K_h} a u_x \frac{1}{h} u \\
&\quad - h^2 \sum_{\partial^{x-} K_h} a u_x \frac{1}{h} u + h^2 \sum_{\Gamma^{x+}(K_h)} a u_x \frac{1}{h} u \\
&= -h^2 \sum_{K_h} (a u_x)_{\bar{x}} u \\
&\quad - h \sum_{\partial^{x-} K_h} a u_x u + h \sum_{\Gamma^{x+}(K_h)} a u_x u. \qquad (2.23)
\end{aligned}$$

Analog kann für die zweite Teilsumme aus (2.21) vorgegangen werden, so dass sich

$$\begin{aligned} h^2 \sum_{K_h} a u_x^2 &= h^2 \sum_{K_h} a u_x \frac{1}{h} u^{x+} - h^2 \sum_{K_h} a u_x \frac{1}{h} u \\ &= h^2 \sum_{\partial^{x+} K_h \cup K_h \setminus \Gamma^{x-}(K_h)} a (u_x)^{x-} \frac{1}{h} u - h^2 \sum_{K_h} a u_x \frac{1}{h} u \\ &= -h^2 \sum_{K_h} (a u_x)_{\bar{x}} u + h \sum_{\Gamma^{x+}(K_h)} a u_x u^{x+} - h \sum_{\partial^{x-} K_h} a u_x u^{x+} \end{aligned} \tag{2.24}$$

ergibt. Verwendet man die gleichen Überlegungen für die Teilsummen aus (2.22), ergeben sich die Darstellungen

$$h^2 \sum_{\partial^{y-} K_h \cup K_h \setminus \Gamma^{y+}(K_h)} d u_y^2 = -h^2 \sum_{K_h} (d u_y)_{\bar{y}} u - h \sum_{\partial^{y-} K_h} d u_y u + h \sum_{\Gamma^{y+}(K_h)} d u_y u \tag{2.25}$$

und

$$h^2 \sum_{K_h} d u_y^2 = -h^2 \sum_{K_h} (d u_y)_{\bar{y}} u + h \sum_{\Gamma^{y+}(K_h)} d u_y u^{y+} - h \sum_{\partial^{y-} K_h} d u_y u^{y+}. \tag{2.26}$$

Die Addition der vier Summen (2.23) - (2.26) ergibt zusammengefasst die Darstellung

$$\begin{aligned} &h^2 \sum_{\partial^{x-} K_h \cup K_h \setminus \Gamma^{x+}(K_h)} a u_x^2 + h^2 \sum_{K_h} a u_x^2 + h^2 \sum_{\partial^{y-} K_h \cup K_h \setminus \Gamma^{y+}(K_h)} d u_y^2 + h^2 \sum_{K_h} d u_y^2 \\ &\quad = -2h^2 \sum_{K_h} \left[(a u_x)_{\bar{x}} + (d u_y)_{\bar{y}} \right] u \\ &\qquad - h \sum_{\partial^{x-} K_h} a u_x u + h \sum_{\Gamma^{x+}(K_h)} a u_x u + h \sum_{\Gamma^{x+}(K_h)} a u_x u^{x+} - h \sum_{\partial^{x-} K_h} a u_x u^{x+} \qquad (2.27) \\ &\qquad - h \sum_{\partial^{y-} K_h} d u_y u + h \sum_{\Gamma^{y+}(K_h)} d u_y u + h \sum_{\Gamma^{y+}(K_h)} d u_y u^{y+} - h \sum_{\partial^{y-} K_h} d u_y u^{y+}. \qquad (2.28) \end{aligned}$$

Die Summen aus (2.27) lassen sich kombinieren zu

$$\begin{aligned} &- h \sum_{\partial^{x-} K_h} a u_x u + h \sum_{\Gamma^{x+}(K_h)} a u_x u + h \sum_{\Gamma^{x+}(K_h)} a u_x u^{x+} - h \sum_{\partial^{x-} K_h} a u_x u^{x+} \\ &\qquad = \sum_{\partial^{x-} K_h} a \left(u^2 - \left(u^{x+} \right)^2 \right) + \sum_{\Gamma^{x+}(K_h)} a \left(\left(u^{x+} \right)^2 - u^2 \right) \end{aligned}$$

und die aus (2.28) zu

$$\begin{aligned} &- h \sum_{\partial^{y-} K_h} d u_y u + h \sum_{\Gamma^{y+}(K_h)} d u_y u + h \sum_{\Gamma^{y+}(K_h)} d u_y u^{y+} - h \sum_{\partial^{y-} K_h} d u_y u^{y+} \\ &\qquad = \sum_{\partial^{y-} K_h} d \left(u^2 - \left(u^{y+} \right)^2 \right) + \sum_{\Gamma^{y+}(K_h)} d \left(\left(u^{y+} \right)^2 - u^2 \right). \end{aligned}$$

Beachtet man, dass

$$\sum_{\partial^{x-}K_h} a\left(u^{x+}\right)^2 = \sum_{\Gamma^{x-}(K_h)} au^2 \text{ und } \sum_{\Gamma^{x+}(K_h)} a\left(u^{x+}\right)^2 = \sum_{\partial^{x+}K_h} au^2$$

(analog für die y-Richtung) gilt, kann die Summe der Ausdrücke (2.27) und (2.28) demnach abgeschätzt werden durch

$$\begin{aligned}
&-h\sum_{\partial^{x-}K_h} au_x u + h\sum_{\Gamma^{x+}(K_h)} au_x u + h\sum_{\Gamma^{x+}(K_h)} au_x u^{x+} - h\sum_{\partial^{x-}K_h} au_x u^{x+} \\
&\qquad - h\sum_{\partial^{y-}K_h} du_y u + h\sum_{\Gamma^{y+}(K_h)} du_y u + h\sum_{\Gamma^{y+}(K_h)} du_y u^{y+} - h\sum_{\partial^{y-}K_h} du_y u^{y+} \\
&\qquad \leq \left[\sum_{\partial K_h} c_{ad}u^2 - \sum_{\Gamma(K_h)} c_{ad}u^2\right]
\end{aligned}$$

mit

$$c_{ad}(x_i, y_j) := \begin{cases} a & \text{für } (x_i, y_j) \in \partial^{x+}K_h \cup \partial^{x-}K_h \cup \Gamma^{x+}(K_h) \cup \Gamma^{x-}(K_h) \\ d & \text{für } (x_i, y_j) \in \partial^{y+}K_h \cup \partial^{y-}K_h \cup \Gamma^{y+}(K_h) \cup \Gamma^{y-}(K_h). \end{cases}$$

Da für $u \in Z_h^L(K_h^l)$

$$\sum_{K_h^l} \left[(au_x)_{\bar{x}} + (du_y)_{\bar{y}}\right] u = 0$$

erfüllt ist, ergibt sich insgesamt die Abschätzung

$$\|\nabla_h u\|^2_{L_h^2(K_h)} \leq \frac{1}{\lambda_{min}} \frac{1}{2} \left[\sum_{\Gamma(K_h^1)} c_{ad}u^2 - \sum_{\Gamma(K_h)} c_{ad}u^2\right]$$

für alle $u \in Z_h^L(K_h)$. Die Argumente lassen sich für $u \in Z_h^L(K_h^l)$ auf beliebige Schachtelungen K_h^k, K_h^{k+1} mit $k \leq l$ übertragen, so dass man zusammen mit

$$\|\nabla_h u\|^2_{L_h^2(K_h)} \leq \|\nabla_h u\|^2_{L_h^2(K_h^k)}$$

die κ Ungleichungen

$$\|\nabla_h u\|^2_{L_h^2(K_h)} \leq \frac{1}{\lambda_{min}} \frac{1}{2} \left[\sum_{\Gamma(K_h^{k+1})} c_{ad}u^2 - \sum_{\Gamma(K_h^k)} c_{ad}u^2\right], \quad l \geq \kappa > k \geq 0$$

erhält. Durch Addition der Ungleichungen ergibt sich

$$\kappa \|\nabla_h u\|^2_{L_h^2(K_h)} \leq \frac{1}{2}\frac{1}{\lambda_{min}} \left[\sum_{\Gamma(K_h^\kappa)} c_{ad}u^2 - \sum_{\Gamma(K_h^0)} c_{ad}u^2\right] \leq \frac{1}{2}\frac{1}{\lambda_{min}} \sum_{\Gamma(K_h^\kappa)} c_{ad}u^2.$$

Erneute Summation von $\kappa = 1$ bis $\kappa = l$ führt zu dem Resultat

$$\frac{1}{2} l^2 \left\| \nabla_h u \right\|^2_{L^2_h(K_h)} \leq \frac{1}{2} \frac{1}{\lambda_{min}} \sum_{K^l_h} c_{ad} u^2 \leq \frac{1}{2} \frac{\lambda_{max}}{\lambda_{min}} \sum_{K^l_h} u^2.$$

Mit $\operatorname{dist}_\infty(K_h, \Gamma(K^l_h)) = lh$ $(l \geq 1)$ ergibt sich die Behauptung. ∎

2.3.2 Variable Koeffizienten

Da der Differenzenoperator im Modellproblem mit variablen Koeffizienten $a, d \in \mathcal{D}_h\left(\overline{\Omega}_h\right)$ angegeben ist, lässt sich das Resultat aus Satz 2.3.2 für diesen allgemeinen Fall nicht nutzen. Darüber hinaus kann der Beweis nicht dementsprechend angepasst werden, so dass ein alternativer Ansatz gewählt werden muss. Um ein Resultat für variable Koeffizienten zu erhalten, wird die Beweisidee zur Cacciopoli-Ungleichung aus [Hac09, Lemma 11.3.3] auf den Fall von Gitterfunktionen übertragen.

Dazu wird eine diskrete Abschneidefunktion $\eta \in \mathcal{D}_h\left(K^l_h\right)$ eingeführt, die durch

$$\eta(x_i, y_j) := \begin{cases} 1 & (x_i, y_j) \in K_h \\ 1 - \frac{k}{l} & (x_i, y_j) \in \Gamma(K^k_h), k = 1, \ldots, l \end{cases}$$

definiert ist. Wegen des linearen Verlaufs zwischen $\Gamma(K_h)$ und $\Gamma(K^l_h)$ erhält man insbesondere $\eta(x_i, y_j) = 0$ für $(x_i, y_j) \in \Gamma(K^l_h)$ und $|\eta_x|, |\eta_y| \leq \frac{1}{hl}$.

Außerdem wird die diskrete Produktregel verwendet, mit der man für die Abschneidefunktion η und eine Gitterfunktion $u \in \mathcal{D}_h\left(K^l_h\right)$ folgende Identitäten erhält:

$$\begin{aligned} \eta(x_{i+1})^2 u_x &= (\eta^2 u)_x - \eta_x u \left(\eta(x_{i+1}) + \eta\right) \\ \eta^2 u_x &= (\eta^2 u)_x - \eta_x u(x_{i+1}) \left(\eta(x_{i+1}) + \eta\right) \\ \eta(x_{i+1}) \eta u_x &= (\eta^2 u)_x - \eta_x \left(u(x_{i+1}) \eta(x_{i+1}) + u\eta\right). \end{aligned} \tag{2.29}$$

Mit Hilfe dieser Bezeichnungen kann das folgende Resultat gezeigt werden:

Satz 2.3.3 *Sei $K_h \subset h\mathbb{Z}^2$ ein Rechteckgitter mit konstanter Schrittweite h und der Differenzenoperator L von der Form (2.11) mit ortsabhängigen Koeffizienten $a, d \in \mathcal{D}_h\left(\overline{\Omega}_h\right)$, $a, d > 0$ gegeben. Aus $K^0_h := K_h$ wird, wie in Konstruktion 2.3.1 beschrieben, eine Schachtelung bis zum Gitter $K^l_h, l \geq 1$ gebildet. Dann gilt*

$$\left\| \nabla_h u \right\|_{L^2_h(K_h)} \leq C_{caccio} \frac{\sqrt{\kappa}}{\operatorname{dist}_\infty(K_h, \Gamma(K^l_h))} \left\| u \right\|_{L^2_h(K^l_h)}$$

mit $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$ und $C_{caccio} = 3\sqrt{2}$ für alle $u \in Z^L_h(K^l_h)$.

Beweis Das Rechteckgitter sei ohne Einschränkung durch ein $n \times m$ Rechteckgitter

$$K_h = \{(x_i, y_j) \in h\mathbb{Z}^2 : 1 \leq i \leq n, 1 \leq j \leq m\}$$

gegeben. Die Erweiterung des Rechteckgitters K_h nach Konstruktion 2.3.1 ergibt demnach das Gitter

$$K_h^l = \{(x_i, y_j) \in h\mathbb{Z}^2 : -l+1 \leq i \leq n+l, -l+1 \leq j \leq m+l\}.$$

Aus $a, d > 0$ folgt $\lambda_{max} \geq \lambda_{min} > 0$, so dass die folgende Abschätzung vorgenommen werden kann:

$$\begin{aligned}
\|\nabla_h u\|^2_{L_h^2(K_h)} &= h^2 \sum_{i=1}^{n-1}\sum_{j=1}^{m} u_x^2 + h^2 \sum_{i=1}^{n}\sum_{j=1}^{m-1} u_y^2 \\
&= \frac{1}{4}\left(h^2 \sum_{i=1}^{n-1}\sum_{j=1}^{m} \left(\eta(x_{i+1}, y_j) + \eta\right)^2 u_x^2 + h^2 \sum_{i=1}^{n}\sum_{j=1}^{m-1} \left(\eta(x_i, y_{j+1}) + \eta\right)^2 u_y^2 \right) \\
&\leq \frac{1}{4}\frac{1}{\lambda_{min}} h^2 \sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1} \left\| \begin{pmatrix} a^{\frac{1}{2}} \left(\eta(x_{i+1}, y_j) + \eta\right) u_x \\ d^{\frac{1}{2}} \left(\eta(x_i, y_{j+1}) + \eta\right) u_y \end{pmatrix} \right\|^2 .
\end{aligned}$$

Verwendet man die diskrete Produktregel nach (2.29), dann kann die folgende Darstellung angegeben werden:

$$\begin{aligned}
&h^2 \sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1} \left(\eta(x_{i+1}, y_j) + \eta\right)^2 u_x a u_x \\
&\quad = h^2 \sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1} \Big\{ 4(\eta^2 u)_x a u_x - a u_x \eta_x \Big[u\left(\eta(x_{i+1}, y_j) + \eta\right) \\
&\qquad + u(x_{i+1}, y_j)\left(\eta(x_{i+1}, y_j) + \eta\right) + 2\left(u(x_{i+1}, y_j)\eta(x_{i+1}, y_j) + u\eta\right) \Big] \Big\} \\
&\quad = h^2 \sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1} \Big\{ 4(\eta^2 u)_x a u_x - a u_x \eta_x \Big[u\left(\eta(x_{i+1}, y_j) + 3\eta\right) \\
&\qquad + u(x_{i+1}, y_j)\left(3\eta(x_{i+1}, y_j) + \eta\right) \Big] \Big\} .
\end{aligned}$$

Die diskrete Produktregel in (2.29) gilt ebenfalls für die Finite-Differenzen-Diskretisierung in y-Richtung, so dass man analog die folgende Darstellung erhält:

$$\begin{aligned}
&h^2 \sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1} \left(\eta(x_i, y_{j+1}) + \eta\right)^2 u_y d u_y \\
&\quad = h^2 \sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1} \Big\{ 4(\eta^2 u)_y d u_y - d u_y \eta_y \Big[u\left(\eta(x_i, y_{j+1}) + 3\eta\right) \\
&\qquad + u(x_i, y_{j+1})\left(3\eta(x_i, y_{j+1}) + \eta\right) \Big] \Big\} .
\end{aligned}$$

Mittels partieller Summation ergibt sich (unter Berücksichtigung von $\eta = 0$ auf $\Gamma(K_h^l)$)

$$\begin{aligned} h^2 \sum_{i=-l+1}^{n+l-1} \sum_{j=-l+1}^{m+l-1} (\eta^2 u)_x a u_x &= -h^2 \sum_{i=-l+2}^{n+l-1} \sum_{j=-l+1}^{m+l-1} (\eta^2 u)\, (a u_x)_{\bar{x}} \\ &\quad + h \sum_{j=-l+1}^{m+l-1} \eta^2(x_{n+l}, y_j) u(x_{n+l}, y_j) a(x_{n+l-1}, y_j) u_x(x_{n+l-1}, y_j) \\ &\quad - h \sum_{j=-l+1}^{m+l-1} \eta^2(x_{-l+1}, y_j) u(x_{-l+1}, y_j) a(x_{-l+1}, y_j) u_x(x_{-l+1}, y_j) \\ &= -h^2 \sum_{i=-l+1}^{n+l-1} \sum_{j=-l+1}^{m+l-1} (\eta^2 u)\, (a u_x)_{\bar{x}} \end{aligned}$$

und

$$h^2 \sum_{i=-l+1}^{n+l-1} \sum_{j=-l+1}^{m+l-1} (\eta^2 u)_y d u_y = -h^2 \sum_{i=-l+1}^{n+l-1} \sum_{j=-l+1}^{m+l-1} (\eta^2 u)\, (d u_y)_{\bar{y}}\,.$$

Für $u \in Z_h^L(K_h^l)$ folgt demnach

$$\begin{aligned} h^2 \sum_{i=-l+1}^{n+l-1} \sum_{j=-l+1}^{m+l-1} &\left\{ (\eta^2 u)_x a u_x + (\eta^2 u)_y d u_y \right\} \\ &= -h^2 \sum_{i=-l+1}^{n+l-1} \sum_{j=-l+1}^{m+l-1} (\eta^2 u) \left[(a u_x)_{\bar{x}} + (d u_y)_{\bar{y}} \right] = 0, \end{aligned}$$

so dass man insgesamt

$$\begin{aligned} h^2 \sum_{i=-l+1}^{n+l-1} \sum_{j=-l+1}^{m+l-1} &\left\{ (\eta(x_{i+1}, y_j) + \eta)^2 u_x a u_x + (\eta(x_i, y_{j+1}) + \eta)^2 u_y d u_y \right\} \\ &= -h^2 \sum_{i=-l+1}^{n+l-1} \sum_{j=-l+1}^{m+l-1} \Big\{ a u_x \eta_x \Big[u\,(\eta(x_{i+1}, y_j) + 3\eta) \\ &\qquad + u(x_{i+1}, y_j)\,(3\eta(x_{i+1}, y_j) + \eta) \Big] + d u_y \eta_y \Big[u\,(\eta(x_i, y_{j+1}) + 3\eta) \\ &\qquad + u(x_i, y_{j+1})\,(3\eta(x_i, y_{j+1}) + \eta) \Big] \Big\} \end{aligned}$$

erhält. Außerdem können die betragsmäßigen Abschätzungen

$$\begin{aligned} \Big| -a u_x \eta_x \Big[u\,(\eta(x_{i+1}, y_j) + 3\eta) + u(x_{i+1}, y_j)\,(3\eta(x_{i+1}, y_j) + \eta) \Big] \Big| \\ \leq 3 \frac{\sqrt{\lambda_{max}}}{hl} \left| a^{\frac{1}{2}} u_x \right| |\eta(x_{i+1}, y_j) + \eta|\,(|u(x_{i+1}, y_j)| + |u|) \end{aligned}$$

und

$$\left|-du_y\eta_y\Big[u\left(\eta(x_i,y_{j+1})+3\eta\right)+u(x_i,y_{j+1})\left(3\eta(x_i,y_{j+1})+\eta\right)\Big]\right| \leq 3\frac{\sqrt{\lambda_{max}}}{hl}\left|d^{\frac{1}{2}}u_y\right|\left|\eta(x_i,y_{j+1})+\eta\right|\left(|u(x_i,y_{j+1})|+|u|\right)$$

vorgenommen werden. Somit erhält man

$$\begin{aligned}
&h^2\sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1}\left\|\begin{pmatrix}a^{\frac{1}{2}}\left(\eta(x_{i+1},y_j)+\eta\right)u_x\\ d^{\frac{1}{2}}\left(\eta(x_i,y_{j+1})+\eta\right)u_y\end{pmatrix}\right\|^2\\
&\quad=\left|h^2\sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1}\left\{\left(\eta(x_{i+1},y_j)+\eta\right)^2u_xau_x+\left(\eta(x_i,y_{j+1})+\eta\right)^2u_ydu_y\right\}\right|\\
&\quad\leq 3\frac{\sqrt{\lambda_{max}}}{hl}h^2\sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1}\begin{pmatrix}\left|a^{\frac{1}{2}}u_x\right|\left|\eta(x_{i+1},y_j)+\eta\right|\\ \left|d^{\frac{1}{2}}u_y\right|\left|\eta(x_i,y_{j+1})+\eta\right|\end{pmatrix}^T\begin{pmatrix}|u|\\|u|\end{pmatrix}\\
&\quad+3\frac{\sqrt{\lambda_{max}}}{hl}h^2\sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1}\begin{pmatrix}\left|a^{\frac{1}{2}}u_x\right|\left|\eta(x_{i+1},y_j)+\eta\right|\\ \left|d^{\frac{1}{2}}u_y\right|\left|\eta(x_i,y_{j+1})+\eta\right|\end{pmatrix}^T\begin{pmatrix}|u(x_{i+1},y_j)|\\|u(x_i,y_{j+1})|\end{pmatrix}\\
&\quad\leq 3\frac{\sqrt{\lambda_{max}}}{hl}h^2\sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1}\left\|\begin{pmatrix}\left|a^{\frac{1}{2}}u_x\right|\left|\eta(x_{i+1},y_j)+\eta\right|\\ \left|d^{\frac{1}{2}}u_y\right|\left|\eta(x_i,y_{j+1})+\eta\right|\end{pmatrix}\right\|\left\|\begin{pmatrix}|u|\\|u|\end{pmatrix}\right\|\\
&\quad+3\frac{\sqrt{\lambda_{max}}}{hl}h^2\sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1}\left\|\begin{pmatrix}\left|a^{\frac{1}{2}}u_x\right|\left|\eta(x_{i+1},y_j)+\eta\right|\\ \left|d^{\frac{1}{2}}u_y\right|\left|\eta(x_i,y_{j+1})+\eta\right|\end{pmatrix}\right\|\left\|\begin{pmatrix}|u(x_{i+1},y_j)|\\|u(x_i,y_{j+1})|\end{pmatrix}\right\|\\
&\quad\leq 6\sqrt{2}\frac{\sqrt{\lambda_{max}}}{hl}\left(h^2\sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1}\left\|\begin{pmatrix}\left|a^{\frac{1}{2}}u_x\right|\left|\eta(x_{i+1},y_j)+\eta\right|\\ \left|d^{\frac{1}{2}}u_y\right|\left|\eta(x_i,y_{j+1})+\eta\right|\end{pmatrix}\right\|^2\right)^{\frac{1}{2}}\|u\|_{L_h^2(K_h^l)}.
\end{aligned}$$

Fasst man die Ergebnisse zusammen, ergibt sich

$$\begin{aligned}
\|\nabla_h u\|_{L_h^2(K_h)}&\leq\frac{1}{2}\frac{1}{\sqrt{\lambda_{min}}}\left(h^2\sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1}\left\|\begin{pmatrix}a^{\frac{1}{2}}\left(\eta(x_{i+1},y_j)+\eta\right)u_x\\ d^{\frac{1}{2}}\left(\eta(x_i,y_{j+1})+\eta\right)u_y\end{pmatrix}\right\|^2\right)^{\frac{1}{2}}\\
&\leq\sqrt{\frac{\lambda_{max}}{\lambda_{min}}}\frac{3\sqrt{2}}{hl}\|u\|_{L_h^2(K_h^l)}
\end{aligned}$$

für alle $u\in Z_h^L(K_h^l)$, woraus mit $\operatorname{dist}_\infty(K_h,\Gamma(K_h^l))=hl$ die Behauptung folgt. ∎

Spezielle diskrete Cacciopoli-Ungleichung

Da die diskrete Cacciopoli-Ungleichung im Anwendungsfall in Kapitel 4 nicht für Gitterfunktionen $u\in Z_h^L(K_h^l)$, sondern für $u\in Z_{h,0}^L(K_h^l;\Omega_h)$ benötigt wird, muss für diesen Fall eine spezielle diskrete Cacciopoli-Ungleichung formuliert werden.

Satz 2.3.4 *Seien Ω_h und $K_h \subset \Omega_h$ Rechteckgitter mit der konstanten Schrittweite h und der Differenzenoperator L von der Form (2.11) mit ortsabhängigen Koeffizienten $a, d \in \mathcal{D}_h\left(\overline{\Omega}_h\right)$, $a, d > 0$ gegeben. Aus $K_h^0 := K_h$ wird, wie in Konstruktion 2.3.1 beschrieben, eine Schachtelung bis zum Gitter $K_h^l, l \geq 1$ mit $\Omega_h \backslash K_h^l \neq \emptyset$ gebildet. Dann gilt*

$$\|\nabla_h u\|_{L_h^2(K_h)} \leq C_{caccio} \frac{\sqrt{\kappa}}{\operatorname{dist}_\infty(K_h, \Gamma(K_h^l))} \|u\|_{L_h^2(K_h^l \cap \Omega_h)}$$

mit $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$ und $C_{caccio} = 3\sqrt{2}$ für alle $u \in Z_{h,0}^L(K_h^l; \Omega_h)$.

Beweis Aufgrund von $K_h \subset \Omega_h$ können die ersten Abschätzungen aus dem Beweis von Satz 2.3.3 direkt übernommen werden. Die Gitterfunktion u wird durch 0 auf $K_h^l \backslash \overline{\Omega}_h$ fortgesetzt. Zu beachten ist, dass nur $u \in Z_h^L(K_h^l \cap \Omega_h)$ gilt, so dass in der Summe

$$h^2 \sum_{i=-l+1}^{n+l-1} \sum_{j=-l+1}^{m+l-1} \left\{(\eta^2 u)_x a u_x + (\eta^2 u)_y d u_y\right\} = h^2 \sum_{i=-l+1}^{n+l-1} \sum_{j=-l+1}^{m+l-1} (\eta^2 u) \left[(a u_x)_{\bar{x}} + (d u_y)_{\bar{y}}\right]$$

für Gitterpunkte $(x_i, y_j) \notin K_h^l \cap \Omega_h$ nicht unbedingt $\left[(a u_x)_{\bar{x}} + (d u_y)_{\bar{y}}\right] = 0$ gelten muss. Durch die Nullfortsetzung gilt jedoch in diesen Punkten $u(x_i, y_j) = 0$, so dass die Summe insgesamt verschwindet.

Unter Berücksichtigung der Nullfortsetzung von u erhält man

$$\|u\|_{L_h^2(K_h^l)} = \|u\|_{L_h^2(K_h^l \cap \Omega_h)},$$

woraus die Behauptung folgt. ∎

3 Hierarchische Matrizen

In diesem Kapitel erfolgt die Einführung aller Grundlagen zur Thematik der $\mathcal{H}$-Matrizen, die im weiteren Verlauf benötigt werden. Im Zusammenhang mit der $\mathcal{H}$-Matrix-Technik sind zahlreiche Veröffentlichungen erschienen, von denen die Monographie [Hac09] besonders hervorzuheben ist. Neben einer umfangreichen Einführung in die Thematik sind auch mehrere Kapitel zu verschiedenen Anwendungsfällen enthalten. In der ebenfalls sehr umfangreichen Monographie [Beb08] wird insbesondere auf die Behandlung von Matrizen eingegangen, die aus der Diskretisierung elliptischer Randwertprobleme stammen. In den genannten Arbeiten sind in umfassenden Literaturverzeichnissen weitere Verweise zu Veröffentlichungen aus den verschiedensten Anwendungsgebieten aufgeführt.

Da in dieser Arbeit die Anwendbarkeit der $\mathcal{H}$-Matrix-Technik im Zusammenhang mit Finite-Differenzen-Matrizen untersucht wird, sind die weiteren Ausführungen auf diesen Anwendungsfall ausgerichtet. Die Einführung der Grundlagen beschränkt sich daher auf die wesentlichen Informationen, die im Zusammenhang mit der Untersuchung von Finite-Differenzen-Matrizen benötigt werden. Insbesondere wird nicht auf Einzelheiten zur $\mathcal{H}$-Arithmetik und deren numerischer Umsetzung eingegangen. Informationen zu diesen Aspekten findet man in den oben angegebenen Monographien oder beispielsweise in der Dissertation [Gra01], deren Anhang ausführliche Informationen zur Implementierung von $\mathcal{H}$-Matrizen in der Programmiersprache **C** enthält.

In Abschnitt 3.1 werden zunächst Grundlagen zur Theorie der $\mathcal{H}$-Matrizen eingeführt, welche eine allgemeine Definition der Menge der $\mathcal{H}$-Matrizen in Abschnitt 3.1.2 ermöglichen. Die Darstellung der allgemeinen Grundlagen und die Einführung von Definitionen und Notationen erfolgt in Anlehnung an die Ausführungen in den Monographien [Hac09] und [Beb08]. Abschnitt 3.2 enthält die Beschreibung der Anpassungen, die zur Verwendung der $\mathcal{H}$-Matrix-Technik für den bisher noch nicht behandelten speziellen Fall der Finite-Differenzen-Matrizen erforderlich sind. Dazu werden die Konzepte aus Abschnitt 3.1 für diesen Fall konkretisiert und durch grundlegende Überlegungen ergänzt, welche den Ausgangspunkt der Untersuchungen in Kapitel 4 bilden.

3.1 Grundlagen

Im Zusammenhang mit der Darstellung von Matrizen und Teilblöcken von Matrizen werden einige spezielle Schreibweisen verwendet. Sind endliche Indexmengen I (Zeilenindizes) und J (Spaltenindizes) gegeben, schreibt man

$$M \in \mathbb{R}^{I \times J}$$

für Matrizen $(M_{ij})_{i \in I, j \in J}$. Da der Fall unendlicher Indexmengen in dieser Arbeit keine Berücksichtigung findet, wird an Stelle von „endlicher Indexmenge“ kurz „Indexmenge“

geschrieben. Die Elemente einer Indexmenge sind im Allgemeinen nicht angeordnet und die Formulierung der $\mathcal{H}$-Matrix-Technik erfolgt unabhängig von einer Anordnung. Lediglich bei Anwendungsfällen wie der Berechnung der LU-Zerlegung ist eine Anordnung erforderlich. In diesem Fall wird von der sogenannten „internen“ Anordnung Gebrauch gemacht, die sich bei der Erstellung des Clusterbaums ergibt und in Abschnitt 3.2.1 eingeführt wird. Zur Durchführung der numerischen Tests in Kapitel 6 wird diese ebenfalls verwendet.

Zur Bezeichnung eines Teilblocks der Matrix $M \in \mathbb{R}^{I\times J}$ zu den Teilmengen $\tau \subset I$ und $\sigma \subset J$ wird

$$M|_{\tau\times\sigma} \in \mathbb{R}^{\tau\times\sigma}$$

für $(M_{ij})_{i\in\tau,j\in\sigma}$ verwendet. Durch die Einführung der Schreibweise $b = \tau \times \sigma$ für einen Block zu den Zeilenindizes aus τ und Spaltenindizes aus σ kann kürzer $M|_b$ geschrieben werden.

Zur Darstellung einer Matrix bzw. eines Matrixblocks können unterschiedliche Formate verwendet werden. Bei Verwendung der gewöhnlichen komponentenweisen Darstellung wird von einer vollbesetzten Matrix gesprochen. Darüber hinaus können in Abhängigkeit von den Eigenschaften der Matrix alternative Formate genutzt werden. Beispielsweise reichen für Diagonal-Matrizen die Einträge der Diagonalen oder für Zirkulante-Matrizen die Einträge der ersten Zeile der Matrix aus, um eine vollständige Darstellung der Matrix zu erhalten. Im Zusammenhang mit der $\mathcal{H}$-Matrix-Technik werden zur Darstellung der Matrixblöcke vollbesetzte Matrizen oder Rang-k-Matrizen verwendet. Letztere werden im nächsten Abschnitt eingeführt.

3.1.1 Niedrigrangmatrizen

Ein wesentlicher Bestandteil der $\mathcal{H}$-Matrix-Technik ist die Approximation von „geeigneten“ Matrixblöcken durch Niedrigrangmatrizen. Eine Strategie zur Identifikation der „geeigneten“ Blöcke wird in Abschnitt 3.1.2 beschrieben. Zunächst wird der Begriff der Niedrigrangmatrizen bzw. Rang-k-Matrizen eingeführt.

In der $\mathcal{H}$-Arithmetik werden Matrix-Addition, Matrix-Vektor- und Matrix-Matrix-Multiplikation sowie alle darauf basierenden Operationen auf Operationen in den Blöcken zurückgeführt. Der Effizienzvorteil der $\mathcal{H}$-Arithmetik beruht darauf, dass innerhalb der Teilblöcke die Matrizen entweder – im Fall kleiner Blöcke – durch vollbesetzte Matrizen oder – bei größeren Blöcken – durch eine Approximation mittels einer Matrix mit möglichst kleinem Rang $k \in \mathbb{N}_0$ dargestellt werden.

In diesem Zusammenhang spricht man von einer Rang-k-Matrix $M \in \mathbb{R}^{I\times J}$, wenn eine Darstellung

$$M = AB^T, \quad A \in \mathbb{R}^{I\times\{1,\dots,k\}}, B \in \mathbb{R}^{J\times\{1,\dots,k\}}, k \in \mathbb{N}_0 \tag{3.1}$$

existiert. Der Rang k muss dabei nicht exakt angenommen werden, sondern ist als Maximalrang der Matrix zu verstehen. Da im Allgemeinen k als klein angenommen wird, verwendet man auch den Begriff „Niedrigrangmatrix“. Bei der praktischen Umsetzung werden an Stelle der Matrix $M \in \mathbb{R}^{I\times J}$ die beiden Faktoren $A \in \mathbb{R}^{I\times\{1,\dots,k\}}$

und $B \in \mathbb{R}^{J \times \{1,\dots,k\}}$ gespeichert, auf deren Grundlage auch die Operationen in der $\mathcal{H}$-Arithmetik durchgeführt werden. Im Fall $k \ll \min\{|I|, |J|\}$ ist die Speicherersparnis dieser Darstellung mit $k(|I| + |J|)$ Einträgen gegenüber einer vollbesetzten Matrix mit $|I| \cdot |J|$ Einträgen offensichtlich. Auch bei der Durchführung von Matrixoperationen ist die Darstellung mittels Rang-k-Matrizen vorteilhaft. So ergibt sich beispielsweise für die Matrix-Vektor-Multiplikation ein Aufwand von $2k(|I| + |J|) - |I| - k$ Operationen bei der Verwendung von Rang-k-Matrizen gegenüber $2(|I| \cdot |J|) - |I|$ Operationen bei vollbesetzten Matrizen. Einen ausführlichen Vergleich der Kosten weiterer Operationen bei Verwendung von Rang-k- und $\mathcal{H}$-Matrizen an Stelle von vollbesetzten Matrizen findet man beispielsweise in [GH03].

3.1.2 Partitionierung

Da eine (gute) Niedrigrangapproximation der gesamten Matrix nur in den seltensten Fällen möglich ist, wird die Approximation lediglich in „geeigneten“ Blöcken der Matrix durchgeführt. Dazu muss die Matrix in Untermatrizen aufgeteilt und entschieden werden, in welchen dieser Blöcke eine Approximation vorgenommen werden soll. Die Aufteilung einer Matrix $M \in \mathbb{R}^{I \times J}$ in Teilblöcke entspricht der Partitionierung der Indexpaarmenge $I \times J$.

Definition 3.1.1 (Blockpartition) *Es seien I, J Indexmengen. Eine Teilmenge $P \subset \mathcal{P}(I \times J) \backslash \emptyset$ heißt Blockpartition oder kurz Partition von $I \times J$, wenn gilt*

für alle $b \in P$ ist $b = \tau \times \sigma$ mit $\tau \subset I, \sigma \subset J$ (Produktstruktur)

$I \times J = \bigcup_{b \in P} b$ (Vollständigkeit)

$b_1 \cap b_2 \neq \emptyset \Rightarrow b_1 = b_2$ für alle $b_1, b_2 \in P$ (Disjunktheit).

Zur Partitionierung einer Matrix existieren verschiedene Möglichkeiten, im Fall der $\mathcal{H}$-Matrizen wird eine hierarchische Partition verwendet. Diese ergibt sich aus der Aufteilung der Ausgangsmatrix in Teilblöcke, in denen die Aufteilung rekursiv fortgesetzt wird. Erst diese hierarchische Partitionierung ermöglicht die blockweise Formulierung der Matrixoperationen in der $\mathcal{H}$-Arithmetik.

Die Partition P soll dabei so konstruiert werden, dass die resultierenden Teilmatrizen $M|_b, b \in P$ entweder klein oder gut durch Rang-k-Matrizen zu approximieren sind. Für Blöcke, die durch Rang-k-Matrizen approximiert werden sollen, ergeben sich widersprüchliche Anforderungen. Zum einen sollten diese Blöcke möglichst groß gewählt werden, damit die Speicherersparnis und die Vorteile bei den Operationskosten im Vergleich zur Verwendung von vollbesetzten Matrizen größtmöglich ausfallen. Im Gegensatz dazu könnte zum Erzielen kleiner Approximationsfehler bei der Verwendung großer Blöcke ein sehr großer Rang k erforderlich sein. Bei der Konstruktion einer Partition sind diese beiden gegensätzlichen Anforderungen zu berücksichtigen und die gebildete Partition sollte aus einem Kompromiss der beiden Zielsetzungen hervorgehen.

Eine Methode, bei der diese Anforderungen erfüllt werden, ergibt sich durch die Konstruktion der Partition unter Verwendung einer Zulässigkeitsbedingung. Diese ermöglicht

es, auf effiziente Weise eine nach den oben angegebenen Kriterien „geeignete“ Partition zu konstruieren. Welche Matrixblöcke gut durch Rang-k-Matrizen approximiert werden können, hängt dabei wesentlich von der zugrunde liegenden Problemstellung ab. Daher sind die Strategie zur Partitionierung und die Auswahl der Zulässigkeitsbedingung ebenfalls von dieser abhängig bzw. sollten immer passend zur Problemstellung gewählt werden.

Bei der im Folgenden beschriebenen Konstruktion einer Partition von $I \times J$ wird zunächst für die Indexmengen I und J eine Hierarchie von Teilmengen in sogenannten Clusterbäumen $T(I)$ und $T(J)$ zusammengefasst. Durch Produktbildung der Knoten zweier Clusterbäume gleicher Stufe kann im Anschluss ein Blockclusterbaum $T(I \times J)$ zu der Indexpaarmenge $I \times J$ konstruiert werden. Die Verwendung der Zulässigkeitsbedingung ermöglicht es, zulässige Blöcke zu identifizieren, die dann nicht weiter aufgeteilt werden. Die Blätter von $T(I \times J)$ ergeben schließlich eine zulässige Partition.

Clusterbaum

Ausgangspunkt der Konstruktion einer Blockpartition der Indexpaarmenge $I \times J$ zu einer gegebenen Matrix $M \in \mathbb{R}^{I \times J}$ ist die Partitionierung der Indexmengen I und J. Dazu wird ein Clusterbaum gebildet, welcher eine Hierarchie von Teilmengen (Cluster) der Indexmenge enthält.

Definition 3.1.2 (Clusterbaum) *Ein Baum $T(I) = [V, E]$ mit Knoten V und Kanten E heißt Clusterbaum zur Indexmenge I, wenn die folgenden Bedingungen erfüllt sind:*

1) $I \in V$ ist die Wurzel von $T(I)$ und $v \neq \emptyset$ für alle $v \in V$

2) für alle $\tau \in V$ gilt entweder $S(\tau) = \emptyset$ oder $\dot{\bigcup}_{\sigma \in S(\tau)} \sigma = \tau$.

Die Menge der Söhne von τ wird mit $S(\tau) := \{t \in V : (\tau, t) \in E\}$ und die Menge der Blätter von $T(I)$ mit $\mathcal{L}(T(I)) := \{\tau \in V : S(\tau) = \emptyset\}$ bezeichnet.

Die Menge der Knoten V wird mit dem Clusterbaum $T(I)$ identifiziert, so dass an Stelle von $\tau \in V$ die Schreibweise $\tau \in T(I)$ verwendet wird.

Üblicherweise erfolgt die Konstruktion eines Clusterbaums durch rekursives Aufteilen der Indexmenge. Zu diesem Zweck wird die Sohnabbildung $S_I(\tau)$ eingeführt, welche die Aufteilung beschreibt. Meist werden die Cluster in zwei Teilmengen zerlegt, was zu einem binären Baum führt. Dies kann so lange fortgeführt werden, bis die Sohnmenge nur noch ein Element enthält. Es ist jedoch sinnvoll, die Aufteilung schon dann abzubrechen, wenn eine Bedingung der Form $|\tau| \geq n_{min}$ mit $n_{min} > 1$ erfüllt ist, da bei zu kleinen Blöcken die Verwendung von Rang-k-Matrizen keinen Vorteil gegenüber der vollbesetzten Darstellung liefert. Beispielsweise ist bei einer 2×2-Matrix keine Speicherersparnis durch die Verwendung von Rang-k-Matrizen zu erzielen.

Bei der Aufteilung der Indexmenge in Teilmengen können unterschiedliche Strategien angewandt werden. Die beiden gängigsten Vorgehensweisen haben zum Ziel, dass die

resultierenden Teilmengen entweder ungefähr gleich viele Elemente enthalten (kardinalitätsbasiert) oder dass das zugrunde liegende Rechengebiet in zwei gleich große Teilgebiete aufgeteilt wird (geometriebasiert). Bei Anwendung der zweiten Strategie muss eine Zuordnung der Indizes zu geometrischen Größen (z.B. zu Punkten in einem Gitter) möglich sein. Im Allgemeinen können die Mächtigkeiten der erzeugten Teilmengen bei der geometriebasierten Konstruktion stark voneinander abweichen.

Da für Finite-Differenzen-Diskretisierungen eine äußerst einfache geometriebasierte Konstruktion möglich ist, beschränken sich die weiteren Ausführungen auf diesen Fall. Zur geometriebasierten Konstruktion wird jedem Element $i \in \tau$ der Indexmenge eine Teilmenge $X_i \subset \mathbb{R}^d$ zugeordnet, wobei d die Dimension der Problemstellung bezeichnet. Zu dem Cluster $\tau \subset I$ wird der Träger von τ definiert als

$$X_\tau := \bigcup_{i \in \tau} X_i \subset \mathbb{R}^d, \tag{3.2}$$

wodurch sich jede Indexmenge mit einem geometrischen Bereich identifizieren lässt. Die Konstruktion des Clusterbaums erfolgt durch die rekursive (geometrische) Aufteilung von X_τ, aus der eine Aufteilung der Indexmenge τ resultiert. Außerdem lassen sich später der Durchmesser und die Distanz von Indexmengen über diesen Zusammenhang einführen und zur Auswertung einer Zulässigkeitsbedingung nutzen.

Zur Bildung eines Clusterbaums können jedoch prinzipiell beliebige andere Strategien verwendet werden. So wird beispielsweise in [GKLB08] für den speziellen Fall der $\mathcal{H}$-LU-Zerlegung eine auf Gebietszerlegung beruhende Clusterstrategie eingeführt. Dabei wird das zugrunde liegende Gebiet in zwei voneinander getrennte Bereiche und einen Separator aufgeteilt. Dies führt zu einer Dreiteilung der Indexmenge, so dass bei dieser Konstruktion ein ternärer Clusterbaum entsteht.

Zulässigkeitsbedingung

Eine Zulässigkeitsbedingung ist im Allgemeinen durch eine Boolsche Funktion

$$Adm : T(I) \times T(J) \to \{true, false\}$$

gegeben. Durch diese wird entschieden, ob bei der Konstruktion des Blockclusterbaums ein Block als zulässig gekennzeichnet werden kann oder weiter unterteilt werden muss.

Da sich die Standard-Zulässigkeitsbedingung aus der Anwendung für Finite-Element-Matrizen zu elliptischen Randwertproblemen im Wesentlichen auch für den Fall Finiter-Differenzen-Matrizen eignet, werden die weiteren Ausführungen anhand dieses konkreten Beispiels vorgenommen. Sie lässt sich unter Verwendung geometrischer Eigenschaften angeben. Durch die Einführung des Trägers von τ in (3.2) lassen sich der Durchmesser eines Clusters und der Abstand zweier Cluster definieren:

$$\operatorname{diam}(\tau) := \max\left\{\left\|x' - x''\right\| : x', x'' \in X_\tau\right\} \qquad \tau \subset I \tag{3.3}$$

$$\operatorname{dist}(\tau, \sigma) := \min\left\{\|x - y\| : x \in X_\tau, y \in X_\sigma\right\} \qquad \tau \subset I, \sigma \subset J. \tag{3.4}$$

Die verwendete Norm in (3.3) und (3.4) ist für gewöhnlich die euklidische Norm, sollte davon abweichend eine andere Norm verwendet werden, wird dies zusätzlich angegeben. Mit Hilfe dieser Größen wird die Standard-Zulässigkeitsbedingung für einen Block eingeführt.

Definition 3.1.3 (Standard-Zulässigkeitsbedingung) *Sei $\eta > 0$ und seien $\tau, \sigma \subset I$ Cluster. Der Block $b = \tau \times \sigma$ heißt η-zulässig, wenn*

$$\min\{\operatorname{diam}(\tau), \operatorname{diam}(\sigma)\} \leq \eta \operatorname{dist}(\tau, \sigma) \tag{3.5}$$

gilt.

Unter Verwendung dieser Zulässigkeitsbedingung kann aus den Elementen der Clusterbäume $T(I)$ und $T(J)$ eine (zulässige) Blockpartition von $I \times J$ erstellt werden. Dazu werden je zwei Cluster $\tau \in T(I)$ und $\sigma \in T(J)$ gleicher Stufe zur Bildung der Knoten des Blockclusterbaums in Form von Produkten $b = \tau \times \sigma$ verwendet. Dadurch können die Knoten des Blockclusterbaums mit Teilblöcken der Matrix identifiziert werden. Mit Hilfe der Zulässigkeitsbedingung für Cluster lassen sich die η-zulässigen Blöcke bestimmen. Ist ein Block nicht zulässig, wird er so lange weiter unterteilt, bis die resultierenden Teilblöcke entweder zulässig sind oder die Bedingung $\min\{|\tau|, |\sigma|\} \leq n_{min}$ mit $n_{min} > 1$ erfüllt ist.

Die Verwendung einer Zulässigkeitsbedingung zur Konstruktion eines Blockclusterbaums ist nicht zwingend erforderlich, sie ermöglicht jedoch die effiziente Konstruktion einer zulässigen Blockpartition. Da nach diesem Ansatz alle zulässigen Blöcke die Zulässigkeitsbedingung erfüllen, kann diese Eigenschaft ebenfalls in Beweisen theoretischer Resultate genutzt werden.

Neben Zulässigkeitsbedingungen, die auf der Verwendung geometrischer Größen beruhen, ist es ebenfalls möglich, diese in Abhängigkeit anderer Eigenschaften zu formulieren. So wird in [BF11] eine algebraische Zulässigkeitsbedingung für dünnbesetzte Matrizen eingeführt, zu deren Auswertung keine geometrischen Informationen erforderlich sind. Sie basiert ausschließlich auf dem Matrixgraphen, für den sich ebenfalls die Größen $\operatorname{diam}(\tau), \operatorname{diam}(\sigma)$ und $\operatorname{dist}(\tau, \sigma)$ mit Hilfe der Knoten des Matrixgraphen berechnen lassen. Dies ermöglicht die Anwendung der $\mathcal{H}$-Matrix-Technik für allgemeine dünnbesetzte Matrizen unabhängig von der zugrunde liegenden Diskretisierung. Daher spricht man in Anlehnung an algebraische Mehrgitterverfahren auch von einer algebraischen Zulässigkeitsbedingung.

Blockclusterbaum

Analog zum Clusterbaum enthält der Blockclusterbaum eine Hierarchie von Blockclustern. Hat man zu den Indexmengen I und J je einen Clusterbaum $T(I)$ und $T(J)$ mittels der Sohnabbildungen S_I und S_J gebildet, kann auf einfache Weise ein Blockclusterbaum $T(I \times J)$ von $I \times J$ konstruiert werden. Unter Verwendung einer beliebigen Zulässigkeitsbedingung *Adm* und dem Parameter $n_{min} > 1$ kann die folgende Konstruktion angegeben werden:

Konstruktion 3.1.4 (Blockclusterbaum)

1) Setze $I \times J$ als Wurzel von $T(I \times J)$

2) Starte die Rekursion mit $b = \tau \times \sigma$ für $\tau = I$ und $\sigma = J$

2a) Sei $b = \tau \times \sigma$. Definiere die Sohnabbildung durch

$$S_{I\times J}(\tau \times \sigma) := \begin{cases} \emptyset & \text{falls } Adm(\tau \times \sigma) = true \\ \emptyset & \text{falls } \min\{|\tau|, |\sigma|\} \leq n_{min} \\ \{\tau' \times \sigma' : \tau' \in S_I(\tau), \sigma' \in S_J(\sigma)\} & \text{sonst} \end{cases}$$

2b) Fahre mit 2a) für alle Söhne von b fort, wenn diese existieren

Die Blätter des auf diese Weise konstruierten Blockclusterbaums ergeben eine zulässige Partition P, das heißt eine Partitionierung der Indexpaarmenge $I \times J$, für die gilt

$$\text{entweder } Adm(b) = true \text{ oder } \min\{|\tau|, |\sigma|\} \leq n_{min} \text{ für alle } b = \tau \times \sigma \in P.$$

Für alle in dieser Arbeit vorkommenden zulässigen Partitionen wird eine Konstruktion dieser Art vorausgesetzt. Sie unterscheiden sich lediglich durch die Bestimmung der Clusterbäume und die verwendete Zulässigkeitsbedingung.

Sind die Clusterbäume $T(I)$ und $T(J)$ binäre Bäume, so entsteht mittels der Konstruktion 3.1.4 ein quaternärer Blockclusterbaum. Die Blöcke der zulässigen Partition P können in zwei Kategorien aufgeteilt werden: Kleine Blöcke, für die $\min\{|\tau|, |\sigma|\} \leq n_{min}$ gilt, werden durch vollbesetzte Matrizen dargestellt und Blöcke, welche die Zulässigkeitsbedingung *Adm* erfüllen, durch Rang-k-Matrizen approximiert. Die Blöcke der ersten Art bilden das Nahfeld

$$P^- := \{b \in P : \min\{|\tau|, |\sigma|\} \leq n_{min}\},$$

die übrigen das Fernfeld

$$P^+ := P \backslash P^-.$$

Die Menge $\mathcal{H}(k, P)$

Nach Einführung der Partition P lässt sich die Menge der Hierarchischen Matrizen zu P definieren:

Definition 3.1.5 *Seien I und J Indexmengen und P eine zulässige Partition von $I \times J$. Dann ist die Menge der hierarchischen Matrizen mit lokalem Rang $k \in \mathbb{N}_0$ und minimaler Blockgröße $n_{min} \in \mathbb{N}$ definiert durch*

$$\mathcal{H}(k, P) := \left\{M \in \mathbb{R}^{I\times J} : \forall\, \tau \times \sigma \in P : \operatorname{rang}(M|_{\tau\times\sigma}) \leq k \text{ oder } \min\{|\tau|, |\sigma|\} \leq n_{min}\right\}.$$

In dieser Definition ist der verwendete Rang k als konstant vorgegeben. Als Erweiterung kann eine lokale Rangverteilung $k(b)$ eingeführt werden, so dass verschiedene Ränge in den Blöcken $b \in P^+$ verwendet werden können. Auf diese Möglichkeit wird hauptsächlich bei der praktischen Umsetzung der $\mathcal{H}$-Matrix-Technik zurückgegriffen, bei der sie zu einem deutlichen Effizienzgewinn führen kann. Im Zusammenhang mit der theoretischen Untersuchung ist die Berücksichtigung lokaler Rangverteilungen jedoch problematisch. Daher wird in dieser Arbeit nur der Fall konstanter Ränge $k \in \mathbb{N}_0$ ohne Berücksichtigung lokaler Rangverteilungen behandelt.

Unter Verwendung von $\mathcal{H}$-Matrizen dieser Struktur lässt sich die Berechnung der Inversen oder der LU-Zerlegung einer Matrix approximativ durch Einführung der $\mathcal{H}$-Arithmetik in fast linearer Komplexität durchführen. Weitere Einzelheiten zu der $\mathcal{H}$-Arithmetik findet man beispielsweise in [Hac09] oder [Hac99, HK00b].

Approximationsfehler

Bei der Anwendung der $\mathcal{H}$-Matrix-Technik können zwei Arten von Approximationsfehlern auftreten: diejenigen, welche durch die Verwendung der $\mathcal{H}$-Arithmetik verursacht werden, und jene, die durch die Approximation einer Matrix durch eine $\mathcal{H}$-Matrix entstehen. Bei Verwendung der $\mathcal{H}$-Arithmetik werden die Operationen durch formatierte Operationen ersetzt, so dass sie nicht mehr exakt durchgeführt werden. Die Approximation einer Matrix $M \in \mathbb{R}^{I\times J}$ durch eine $\mathcal{H}$-Matrix $M_{\mathcal{H}} \in \mathcal{H}(k, P)$ mit $k \in \mathbb{N}_0$ und einer zulässigen Partition P zur Indexpaarmenge $I \times J$ kann durch Approximation der Blöcke $b \in P^+$ durch Rang-k-Matrizen erfolgen. In den (kleinen) Blöcken $b \in P^-$, in denen die Teilmatrizen als vollbesetzte Matrix gespeichert werden, können die Teilmatrizen ohne Approximation übernommen werden. Demnach tritt lediglich in den Blöcken $b \in P^+$ ein lokaler Fehler durch die Approximation mittels Rang-k-Matrizen auf. Für die Bestapproximation einer Matrix durch eine Rang-k-Matrix kann dabei folgendes Resultat angegeben werden:

Satz 3.1.6 ([Hac09, Satz 2.4.1]) *Die Matrix $M \in \mathbb{R}^{I\times J}$ habe die Singulärwertzerlegung $M = U\Sigma V^T$ (U, V orthogonal, Σ ist diagonal mit Singulärwerten $\sigma_i = \Sigma_{ii}$ in der Anordnung $\sigma_1 \geq \sigma_2 \geq \dots$). Die beiden Minimierungsaufgaben*

$$\min_{Rang(R)\leq k} \|M - R\|_2 \ \textit{und} \ \min_{Rang(R)\leq k} \|M - R\|_F$$

werden von

$$R := U\Sigma_k V^T \ \textit{mit} \ (\Sigma_k)_{ij} = \begin{cases} \sigma_i & \textit{für } i = j \leq \min\{k, |I|, |J|\}, \\ 0 & \textit{sonst}, \end{cases}$$

gelöst (Σ_k entsteht aus Σ, indem alle σ_i für $i > k$ durch null ersetzt werden). Der dabei auftretende Fehler ist

$$\|M - R\|_2 = \sigma_{k+1} \ \textit{bzw.} \ \|M - R\|_F = \sqrt{\sum_{i=k+1}^{\min(|I|,|J|)} \sigma_i^2}$$

(wobei $\sigma_{k+1} := 0$ für $k \geq \min\{|I|, |J|\}$ gesetzt sei).

Zur Approximation der Teilmatrizen in den zulässigen Blöcken $b \in P^+$ mittels Rang-k-Matrizen kann demnach die Bestapproximation verwendet werden, indem das Ergebnis der Singulärwertzerlegung auf den Rang k „gekürzt“ wird. Diese Vorgehensweise wird bei der Durchführung der numerischen Tests in Kapitel 6 genutzt.

Der globale Fehler lässt sich in Abhängigkeit der lokalen Fehler in den Matrixblöcken abschätzen bzw. direkt angeben. Dies hängt von der verwendeten Norm ab. In dieser Arbeit wird die Frobenius-Norm

$$\|M\|_F = \sqrt{\sum_{i \in I, j \in J} |M_{ij}|^2}$$

verwendet. Einzelheiten zum Gebrauch anderer Normen, insbesondere der Spektralnorm, finden sich in [Hac09]. Für die Frobenius-Norm gilt der einfache Zusammenhang

$$\|M_{\mathcal{H}}\|_F = \sqrt{\sum_{b \in P} \|M_{\mathcal{H}}|_b\|_F^2} \text{ für alle } M_{\mathcal{H}} \in \mathcal{H}(k, P).$$

Insbesondere lässt sich der globale Approximationsfehler mittels der lokalen Fehler durch

$$\|M - M_{\mathcal{H}}\|_F^2 = \sum_{b \in P} \|M|_b - M_{\mathcal{H}}|_b\|_F^2 \tag{3.6}$$

angeben. Ist es demnach möglich, den lokalen Fehler in allen Blöcken $b \in P$ durch

$$\|M|_b - M_{\mathcal{H}}|_b\|_F \leq \epsilon$$

abzuschätzen, so erhält man für den globalen Fehler

$$\|M - M_{\mathcal{H}}\|_F \leq \epsilon \sqrt{|P|},$$

wobei $|P|$ die Anzahl der Elemente der Partition P bezeichnet. Die Größenordnung von $|P|$ kann mit Hilfe der Konstanten C_{sp} abgeschätzt werden, die in Abschnitt 3.2.3 eingeführt wird. Dort wird eine entsprechende Abschätzung angegeben.

3.1.3 $\mathcal{H}$-Inverse

Im Zusammenhang mit $\mathcal{H}$-Matrizen und der Invertierung einer Matrix sind zwei Aspekte von Interesse: Die praktische Berechnung einer $\mathcal{H}$-Inversen in der $\mathcal{H}$-Arithmetik und die Frage nach der Existenz einer (guten) $\mathcal{H}$-Matrix Approximation der Inversen. Dabei ist die Berechnung einer $\mathcal{H}$-Inversen in der $\mathcal{H}$-Arithmetik nur dann sinnvoll, wenn eine $\mathcal{H}$-Matrix Approximation der Inversen existiert. Daher ist dies vor der Anwendung der $\mathcal{H}$-Matrix-Technik zu untersuchen.

Zu dieser Thematik konnten bereits mehrere Resultate erzielt werden. Eines wird in [Hac09] für symmetrische, positiv definite, wohlkonditionierte Matrizen angegeben.

Dieses lässt sich beispielsweise nutzen, um die Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen von Massematrizen bei Finite-Element-Diskretisierungen zu zeigen. Nach den Ausführungen in Abschnitt 2.1.1 ist die Diskretisierungsmatrix zum Modellproblem symmetrisch und positiv definit. Wie bei Steifigkeitsmatrizen im Kontext von Finite-Element-Diskretisierungen sind jedoch auch die Finite-Differenzen-Matrizen zum Modellproblem im Allgemeinen nicht wohlkonditioniert, so dass sich dieses Resultat in diesem Fall nicht verwenden lässt. Da speziell für Finite-Differenzen-Matrizen kein Resultat bekannt ist, wird für diese in Kapitel 4 ein eigenständiger Ansatz zum Modellproblem auf Grundlage der Vorgehensweise zu Finite-Element-Matrizen von elliptischen Randwertproblemen entwickelt.

Die praktische Berechnung der Inversen in der $\mathcal{H}$-Arithmetik lässt sich auf Grundlage der hierarchischen Partitionierung auf Operationen in den Blöcken der Matrix zurückführen. Stellt man eine reguläre Matrix M in der Blockgestalt

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \tag{3.7}$$

dar, lässt sich eine exakte Darstellung der Inversen in der Form

$$M^{-1} = \begin{bmatrix} M_{11}^{-1} + M_{11}^{-1} M_{12} S^{-1} M_{21} M_{11}^{-1} & -M_{11}^{-1} M_{12} S^{-1} \\ -S^{-1} M_{21} M_{11}^{-1} & S^{-1} \end{bmatrix} \tag{3.8}$$

mit dem Schur-Komplement $S := M_{22} - M_{21} M_{11}^{-1} M_{12}$ angeben. Zu beachten ist, dass diese Darstellung die Regularität von M_{11} voraussetzt, was in der Fortsetzung des Algorithmus der Regularität der entsprechenden Hauptuntermatrizen entspricht. Diese Bedingung ist beispielsweise für positiv definite Matrizen M und damit insbesondere für die Diskretisierungsmatrix aus dem Modellproblem erfüllt.

Nimmt man für M_{11} ebenfalls eine Blockstruktur wie in (3.7) an, kann die Darstellung aus (3.8) erneut zur Berechnung von M_{11}^{-1} verwendet werden. Auf diese Weise lassen sich die Blöcke der Inversen M^{-1} rekursiv mittels der Darstellung (3.8) bestimmen.

Bei exakter Berechnung der Matrix-Matrix-Multiplikation und der Matrix-Addition würde man in (3.8) die exakte Inverse erhalten. Zur Bestimmung der $\mathcal{H}$-Inversen werden die Operationen innerhalb der Blöcke durch formatierte Matrixoperationen für $\mathcal{H}$-Matrizen ersetzt, das heißt, dass die Operationen nicht mehr exakt, sondern approximativ durchgeführt werden. Diese Vorgehensweise ermöglicht die Berechnung der $\mathcal{H}$-Inversen in fast linearer Komplexität. Durch die Verwendung der formatierten Operationen stellt die auf diese Weise berechnete $\mathcal{H}$-Inverse jedoch nur eine Approximation der exakten Inversen dar.

Zu beachten ist, dass durch diese Vorgehensweise im Allgemeinen nicht die Bestapproximation der Inversen M^{-1} berechnet wird. Insbesondere ermöglicht daher der Beweis der Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen noch keine Aussage über die Genauigkeit der durch den beschriebenen Algorithmus berechneten $\mathcal{H}$-Matrix. Dies muss unabhängig von dem Ergebnis zur Existenz zusätzlich überprüft werden.

3.2 $\mathcal{H}$-Matrix-Technik für Finite-Differenzen-Matrizen

In Abschnitt 3.1 wurden die Grundlagen der $\mathcal{H}$-Matrix-Technik ohne Berücksichtigung eines besonderen Anwendungsfalls eingeführt, wobei die Ausführungen eine Übersicht bekannter Konzepte darstellten. Im Gegensatz dazu wird in diesem Abschnitt der spezielle Fall von Finite-Differenzen-Diskretisierungen behandelt, zu dem im Zusammenhang mit der $\mathcal{H}$-Matrix-Technik keine Veröffentlichungen bekannt sind. Daher werden zunächst die allgemeinen Überlegungen aus dem letzten Abschnitt für diese Problemstellung konkretisiert. Insbesondere wird in Abschnitt 3.2.1 eine geeignete Partitionierungsstrategie für Finite-Differenzen-Matrizen angegeben.

Außerdem ist zu klären, ob sich die Finite-Differenzen-Matrizen selbst in Form einer $\mathcal{H}$-Matrix darstellen lassen. Diese grundlegende Eigenschaft ist erforderlich, da bei der Berechnung der $\mathcal{H}$-Inversen oder $\mathcal{H}$-LU-Zerlegung die Ausgangsmatrix in Form einer $\mathcal{H}$-Matrix vorliegen muss. Diese Eigenschaft wird in Abschnitt 3.2.2 für Finite-Differenzen-Matrizen gezeigt. Ergänzt werden die Ausführungen durch Ergebnisse zur Komplexität für den Anwendungsfall von Finite-Differenzen-Matrizen in Abschnitt 3.2.3.

Zusätzlich wird in Abschnitt 3.2.4 das Konzept der diskreten separablen Entwicklung eingeführt. Diese lässt sich als Hilfsmittel zur Untersuchung der Approximierbarkeit von Matrixblöcken mittels Rang-k-Matrizen einsetzen. Dazu folgt in Abschnitt 3.2.5 eine kurze Darstellung zur Abschätzung der Fehler bei der Approximation von Matrizen durch $\mathcal{H}$-Matrizen unter Verwendung der diskreten separablen Entwicklung. Diese Ausführungen bilden die Grundlage der Untersuchungen zur Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen von Finite-Differenzen-Matrizen in Kapitel 4.

3.2.1 Partitionierung

In Abschnitt 3.1.2 wurde die Konstruktion einer zulässigen Partition unter Verwendung von Clusterbäumen und der Standard-Zulässigkeitsbedingung eingeführt. In den folgenden Abschnitten wird für den speziellen Fall von Finite-Differenzen-Diskretisierungen die Konstruktion des Clusterbaums beschrieben und eine Zulässigkeitsbedingung passend zum Modellproblem eingeführt. Auf dieser Grundlage lässt sich im Anschluss eine zulässige Partition bestimmen.

Die Überlegungen beschränken sich unter Berücksichtigung des Anwendungsfalls auf quadratische Matrizen $M \in \mathbb{R}^{I \times I}$ zur Indexmenge I.

Clusterbaum

Ein wesentlicher Bestandteil zur Bildung der zulässigen Partition ist die Konstruktion des Clusterbaums zur Indexmenge I. Bei Finite-Differenzen-Diskretisierungen auf einem Rechteckgitter Ω_h mit konstanter Schrittweite h liegt ein besonders einfacher Fall vor. Jedem Index $i \in I$ kann genau ein Gitterpunkt $X_i \in \Omega_h$ zugeordnet werden und wegen der konstanten Schrittweite h sind die Gitterpunkte regelmäßig verteilt.

Aufgrund dieser einfachen Struktur bietet es sich an, eine geometriebasierte Konstruktion des Clusterbaums vorzunehmen. Dazu wird das Ausgangsgitter in Richtung

der längsten Ausdehnung halbiert und damit in zwei Teilgitter unterteilt, woraus eine Aufteilung der Indexmenge in zwei Teilmengen resultiert. Sollten Gitterpunkte nicht eindeutig einem Teilgitter zuzuordnen sein, werden sie alle einem der beiden Gitter zugeordnet, um bei der Teilung ausschließlich Rechteckgitter zu erhalten. Aus diesem Grund reicht es aus, die diskrete Poincaré- und die diskrete Cacciopoli-Ungleichung aus den Abschnitten 2.2 und 2.3 nur für den Fall von Rechteckgittern anzugeben.

Da in jedem Teilungsschritt die konstruierten Teilgitter ungefähr die gleiche Größe besitzen (oder sogar exakt die gleiche Größe, falls keine Gitterpunkte vorkommen, die beiden Teilgittern zugeordnet werden könnten), ist die Anzahl der Gitterpunkte beider Teilgitter ebenfalls ungefähr gleich groß. Berücksichtigt man die Zuordnung der Gitterpunkte zu den Indizes, dann entspricht dies einer Aufteilung der Indexmenge in zwei ungefähr gleich große Teilmengen. Demnach resultiert in diesem Fall aus der geometriebasierten Konstruktion ein Clusterbaum, der auch die Anforderungen der kardinalitätsbasierten Konstruktion erfüllt.

Aus der Aufteilung der Indexmenge $\tau \subset I$ in zwei Teilmengen τ_1, τ_2 ergibt sich direkt eine Anordnung der Cluster und damit der Knoten von $T(I)$. Dazu werden in jedem Teilungsschritt die Indizes aus τ_1 (oder alternativ aus τ_2) an den Anfang von τ verschoben. Diese Anordnung wird auch als „interne Anordnung“ bezeichnet. Sie kommt insbesondere bei der numerischen Umsetzung zum Einsatz und kann ebenfalls als Anordnung bei der Berechnung einer $\mathcal{H}$-LU-Zerlegung verwendet werden.

Zulässigkeitsbedingung

Im Anwendungsfall von Finite-Element-Diskretisierungen für elliptische Randwertprobleme wird die Standard-Zulässigkeitsbedingung aus Definition 3.1.3 verwendet. Die dort angegebene Bedingung (3.5) wird für die Behandlung des Modellproblems durch die leicht angepasste Bedingung

$$\min\{\operatorname{diam}(\tau), \operatorname{diam}(\sigma)\} \leq \eta \left(\operatorname{dist}_\infty(\tau, \sigma) - h\right), \quad \eta > 0 \tag{3.9}$$

ersetzt, wobei $\operatorname{dist}_\infty(\tau, \sigma)$ die Auswertung von (3.4) bezüglich der Maximumsnorm bezeichnet. Die spezielle Wahl der Zulässigkeitsbedingung ergibt sich aus den Überlegungen in Abschnitt 4.1.3 zur Existenz einer $\mathcal{H}$-Matrix Approximation für die Inverse von Finite-Differenzen-Matrizen.

Die Berechnung der Größen $\operatorname{diam}(\tau), \operatorname{diam}(\sigma)$ und $\operatorname{dist}(\tau, \sigma)$ kann im Allgemeinen sehr aufwendig sein. Ein besonders einfacher Fall liegt vor, wenn die Träger zu den Clustern durch (achsenparallele) Quader gegeben sind. Für die Indexmengen $\tau, \sigma \subset I$ werden die Quader $Q_\tau = \Pi_{i=1}^d [a_i^\tau, b_i^\tau]$ und $Q_\sigma = \Pi_{i=1}^d [a_i^\sigma, b_i^\sigma]$ als Obermengen der Träger X_τ, X_σ eingeführt. Für diese lassen sich die Größen $\operatorname{diam}(Q_\tau)$ und $\operatorname{dist}_\infty(Q_\tau, Q_\sigma)$ exakt und effizient berechnen durch

$$\operatorname{diam}(Q_\tau) = \sqrt{\sum_{i=1}^d (b_i^\tau - a_i^\tau)^2} \tag{3.10}$$

und

$$\operatorname{dist}_\infty(Q_\tau, Q_\sigma) = \max_{i=1}^d \operatorname{dist}\left([a_i^\tau, b_i^\tau], [a_i^\sigma, b_i^\sigma]\right). \tag{3.11}$$

Der kleinste Quader Q_τ zur Indexmenge τ, für den $X_\tau \subset Q_\tau$ gilt, heißt Minimalquader $Q_{min}(X_\tau)$. Da durch die Konstruktion des Clusterbaums alle auftretenden Indexmengen $\tau, \sigma \subset I$ mit Rechteckgittern identifiziert werden können, stimmen die Größen $\operatorname{diam}(\tau), \operatorname{diam}(\sigma)$ aus (3.3) und $\operatorname{dist}_\infty(\tau, \sigma)$ aus (3.4) mit denen der entsprechenden Minimalquader überein:

$$\begin{aligned} \operatorname{diam}(\tau) &= \operatorname{diam}(Q_{min}(X_\tau)), \\ \operatorname{dist}_\infty(\tau, \sigma) &= \operatorname{dist}_\infty\left(Q_{min}(X_\tau), Q_{min}(X_\sigma)\right). \end{aligned}$$

Demnach können sie ebenfalls exakt und effizient nach (3.10) und (3.11) berechnet werden, was die effiziente Auswertung der Zulässigkeitsbedingung (3.9) ermöglicht.

Blockclusterbaum

Da nur quadratische Matrizen $M \in \mathbb{R}^{I \times I}$ betrachtet werden, reicht zur Konstruktion des Blockclusterbaums $T(I \times I)$ ein Clusterbaum $T(I)$ aus, der wie beschrieben gebildet wird. Die Konstruktion 3.1.4 kann mit Hilfe von $T(I)$, der Vorgabe von n_{min} und unter Verwendung der Zulässigkeitsbedingung (3.9) erfolgen. Da $T(I)$ durch einen binären Clusterbaum gegeben ist, ergibt sich $T(I \times I)$ als quaternärer Baum. Die Blätter von $T(I \times I)$ bilden eine zulässige Partition P von $I \times I$.

Auf Grundlage dieser Partitionierung erfolgen alle weiteren Überlegungen. Deshalb wird vorausgesetzt, dass alle Partitionen, die im weiteren Verlauf auftreten, nach dieser Konstruktion gebildet werden. Sollten davon abweichende Konstruktionen vorgenommen werden, wird dies explizit angegeben. Dabei wird die allgemeine Vorgehensweise aus Konstruktion 3.1.4 nicht modifiziert, die Anpassungen beschränken sich in diesen Fällen auf die Bestimmung des Clusterbaums und die Einführung einer alternativen Zulässigkeitsbedingung.

3.2.2 $\mathcal{H}$-Matrix Eigenschaft von Finite-Differenzen-Matrizen

Zur Lösung linearer Gleichungssysteme unter Verwendung der $\mathcal{H}$-Matrix-Technik ist es erforderlich, dass die entsprechende Matrix durch eine $\mathcal{H}$-Matrix darstellbar bzw. zu approximieren ist. Nur in diesem Fall kann sie als Ausgangsmatrix zur Berechnung der $\mathcal{H}$-LU-Zerlegung oder der $\mathcal{H}$-Inversen verwendet werden. Um die $\mathcal{H}$-Matrix-Technik für Finite-Differenzen-Matrizen anwenden zu können, ist daher zu klären, ob die Diskretisierungsmatrix als $\mathcal{H}$-Matrix dargestellt werden kann.

Die zugrunde liegende Diskretisierung aus Abschnitt 2.1.1 führt zu einem 5-Punkte-Differenzenstern. Die Diskretisierungsmatrix $L_h \in \mathbb{R}^{I \times I}$ besitzt demnach maximal fünf Einträge pro Zeile und ist somit dünnbesetzt. Der Eintrag $(L_h)_{i,j}$ kann nur dann von null verschieden sein, wenn die zu den Indizes $i, j \in I$ gehörigen Gitterpunkte benachbart sind. Wird eine Partitionierung nach Abschnitt 3.2.1 konstruiert, erfüllen die zulässigen Blöcke $b = \tau \times \sigma \in P^+$ die Zulässigkeitsbedingung (3.9), wonach wegen $n_{min} > 1$ stets

$$\operatorname{dist}_\infty(X_\tau, X_\sigma) > h$$

gilt. Daraus lässt sich folgern, dass alle zulässigen Gitter X_τ und X_σ nicht benachbart sind, so dass man

$$L_h|_b = 0 \text{ für alle } b \in P^+$$

erhält. Demnach kann die Diskretisierungsmatrix bei Verwendung einer zulässigen Partition P, die mittels der oben beschriebenen Konstruktion unter Verwendung der Zulässigkeitsbedingung (3.9) erstellt wurde, durch eine $\mathcal{H}$-Matrix exakt dargestellt werden und es gilt sogar $L_h \in \mathcal{H}(0, P)$.

In Abbildung 3.1 ist exemplarisch eine Blockpartitionierung zu einer Diskretisierungsmatrix zusammen mit der entsprechenden Besetzungsstruktur (weiß) bei Verwendung der internen Anordnung angegeben. Die Blöcke des Nahfelds $b \in P^-$ sind in dunkelgrau und die des Fernfelds $b \in P^+$ in hellgrau dargestellt. Wie beschrieben, befinden sich alle Einträge der Matrix in Blöcken des Nahfelds.

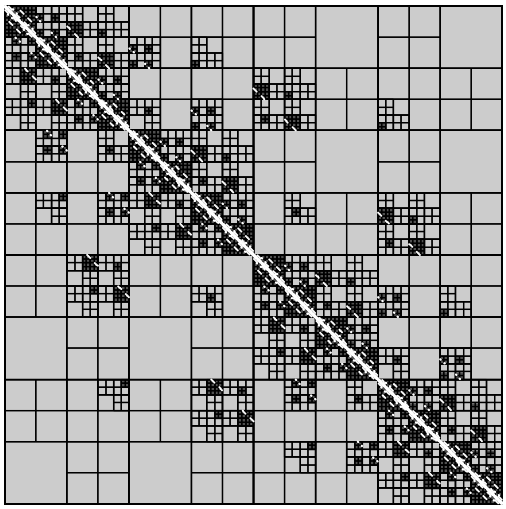


Abbildung 3.1: Blockpartitionierung einer Finite-Differenzen-Matrix

3.2.3 Komplexität

Der große Vorteil der $\mathcal{H}$-Matrix-Technik im Zusammenhang mit der Lösung linearer Gleichungssysteme ist die approximative Berechnung der $\mathcal{H}$-Inversen bzw. $\mathcal{H}$-LU-Zerlegung in fast linearer Komplexität $\mathcal{O}\left(\hat{n}\log^q(\hat{n})\right)$ mit $q > 0$ unabhängig von $\hat{n}$, wobei $\hat{n}$ der Parameter ist, der die Problemgröße beschreibt. Im Fall der Finite-Differenzen-Matrizen zu $n \times m$ Rechteckgittern gilt demnach $\hat{n} = nm$. Der Aufwand der Algorithmen lässt sich in Abhängigkeit des Speicherbedarfs der entsprechenden $\mathcal{H}$-Matrizen abschätzen, der wiederum von der verwendeten Partitionierung abhängt. Daher ist die Abschätzung des Speicherbedarfs der $\mathcal{H}$-Matrizen, die aus der oben angegebenen Partitionierung für den Fall von Finite-Differenzen-Matrizen resultieren, entscheidend für die Abschätzung des Aufwands zur Berechnung der $\mathcal{H}$-Inversen bzw. $\mathcal{H}$-LU-Zerlegung.

Zur Abschätzung des Speicherbedarfs kann die Konstante C_{sp} verwendet werden, die in der Dissertation [Gra01] eingeführt wurde. Sie wird zur Charakterisierung der „Schwachbesetztheit“ im Kontext der $\mathcal{H}$-Matrizen verwendet und ist für Clusterbäume $T(I), T(J)$ und die Partition P durch

$$C_{sp,l}(\tau, P) := |\{\sigma \in T(J) : \tau \times \sigma \in P\}| \text{ für } \tau \in T(I)$$
$$C_{sp,r}(\sigma, P) := |\{\tau \in T(I) : \tau \times \sigma \in P\}| \text{ für } \sigma \in T(J)$$

und

$$C_{sp}(P) := \max\left\{ \max_{\tau \in T(I)} C_{sp,l}(\tau, P), \max_{\sigma \in T(J)} C_{sp,r}(\sigma, P) \right\}$$

definiert. Die Konstante gibt demnach die maximale Anzahl der Blöcke zu einer Spalte bzw. Zeile der $\mathcal{H}$-Matrix bezüglich einer Indexmenge aus dem Clusterbaum an. Nähere Ausführungen zur Abschätzung des Speicherbedarfs und der Komplexität verschiedener Operationen mit Hilfe der Konstanten C_{sp} findet man beispielsweise in [Hac09],[GH03] oder [Beb08]. Unter Verwendung der Konstanten C_{sp} kann der Speicherbedarf $\mathcal{S}_{\mathcal{H}}(k, P)$ einer $\mathcal{H}$-Matrix, die zu den Clusterbäumen $T(I)$ und $T(J)$ konstruiert wurde, in der Form

$$\mathcal{S}_{\mathcal{H}}(k, P) \leq C_{sp}(P) \cdot \max\{n_{min}, k\} \cdot [(\operatorname{depth}(T(I)) + 1)\,|I| + (\operatorname{depth}(T(J)) + 1)\,|J|]$$

abgeschätzt werden.

Die Ausführungen zur Abschätzung der Konstanten C_{sp} im Fall der Partitionierung von Finite-Differenzen-Matrizen folgen den Überlegungen in [GH03]. Zur Vereinfachung wird die Abschätzung von C_{sp} exemplarisch für den zweidimensionalen Fall eines $n \times n$ Quadratgitters durchgeführt, das dementsprechend die Seitenlänge $(n-1)h$ besitzt. Zur Bildung der Teilmengen wird das Quadrat durch Halbierung der Seitenlängen in vier gleich große Teilquadrate eingeteilt. Die Indexmengen zur Bildung des Clusterbaums erhält man wie gewohnt, indem der Zusammenhang der Gitterpunkte zu den entsprechenden Indizes genutzt wird. Diese Vorgehensweise kann rekursiv fortgesetzt werden, so dass die Quadrate zur l-ten Stufe des Clusterbaums die Seitenlänge $(n-1)h2^{-l}$ besitzen und ein quarternärer Clusterbaum entsteht. Dies vereinfacht die Überlegungen, da die Durchmesser der resultierenden Quadrate und die Distanzen in Abhängigkeit des Levels l im Clusterbaum besonders einfach abgeschätzt werden können.

Das Quadrat der Stufe l, das die Gitterpunkte zur Indexmenge $\tau \in T^{(l)}(I)$ enthält, wird mit C_τ^l bezeichnet und das entsprechende Gitter mit $\Omega_\tau^l := C_\tau^l \cap \Omega_h$. Demnach gilt für die Gitter

$$\begin{aligned} \operatorname{diam}(\Omega_\tau^l) \leq \operatorname{diam}(C_\tau^l) &= \sqrt{2}(n-1)h2^{-l} \\ \operatorname{diam}(\Omega_\sigma^l) \leq \operatorname{diam}(C_\sigma^l) &= \sqrt{2}(n-1)h2^{-l} \\ \operatorname{dist}_\infty(\Omega_\tau^l, \Omega_\sigma^l) &\geq \operatorname{dist}_\infty(C_\tau^l, C_\sigma^l). \end{aligned} \tag{3.12}$$

Die weiteren Überlegungen basieren auf dem Ergebnis aus [GH03, Lemma 4.4], bei dem die Abschätzung

$$C_{sp} \leq 4 \max_{\tau \in T(J)} |\{\sigma \in T(I) : \tau \times \sigma \in T(I \times J) \backslash \mathcal{L}(T(I \times J)) \text{ und } \tau \times \sigma \text{ ist nicht zulässig}\}| \tag{3.13}$$

gezeigt wird, wobei $T(I \times J)$ ein nach Konstruktion 3.1.4 gebildeter Blockclusterbaum der Tiefe $p \geq 1$ ist und die der Konstruktion zugrunde liegenden Clusterbäume $T(I), T(J)$ auf die oben angegebene Weise gebildet sind. Zur Abschätzung von C_{sp} ist demnach zu $\tau \in T(J)$ die Anzahl der Cluster $\sigma \in T(I)$ abzuschätzen, für welche die Blöcke $\tau \times \sigma$ nicht zulässig sind.

Sind ein Cluster $\tau \in T^{(l)}(I)$ der Stufe l und das entsprechende Quadrat C_τ^l gegeben, so ist für $\sigma \in T^{(l)}(I)$ des gleichen Levels $\tau \times \sigma$ zulässig, wenn C_σ^l einen gewissen Mindestabstand zu C_τ^l besitzt. Die Menge der Cluster $\sigma \in T^{(l)}(I)$ der Stufe l, für die $\tau \times \sigma$ unzulässig ist, kann demnach durch die Anzahl der entsprechenden Quadrate C_σ^l abgeschätzt werden, die einen kleineren Abstand als diesen Mindestabstand zu C_τ^l besitzen.

Für die Cluster $\sigma \in T^{(l)}(I)$ mit $|\sigma| > n_{min}$, für die $\operatorname{dist}_\infty(C_\tau^l, C_\sigma^l) \geq \frac{\sqrt{2}}{\eta}(n-1)h2^{-l} + h$ gilt, erhält man mit (3.12) die Abschätzung

$$\begin{aligned}\operatorname{dist}_\infty(\Omega_\tau^l, \Omega_\sigma^l) - h \geq \operatorname{dist}_\infty(C_\tau^l, C_\sigma^l) - h &\geq \frac{\sqrt{2}}{\eta}(n-1)h2^{-l} \\ &\geq \frac{1}{\eta}\min\{\operatorname{diam}(\Omega_\tau^l), \operatorname{diam}(\Omega_\sigma^l)\},\end{aligned}$$

so dass die Blockcluster $\tau \times \sigma$ zulässig sind. Demnach kann die Anzahl der nicht zulässigen Cluster abgeschätzt werden, indem die Anzahl der Quadrate C_σ^l bestimmt wird, für die $\operatorname{dist}_\infty(C_\tau^l, C_\sigma^l) < \frac{\sqrt{2}}{\eta}(n-1)h2^{-l} + h$ gilt.

Für $\tau \in T^{(l)}(I)$ mit $|\tau| > n_{min}$ ist direkt ersichtlich, dass die Anzahl der Quadrate C_σ^l der Stufe l, deren Abstand zu dem Quadrat C_τ^l verschwindet, maximal 3^2 ist. Als Verallgemeinerung ergibt sich, dass die Anzahl der Quadrate der Stufe l, die einen geringeren Abstand als $j(n-1)h2^{-l}$ besitzen, maximal $(1+2j)^2$ beträgt.

Setzt man

$$j := \frac{\sqrt{2}}{\eta} + \frac{2^l}{n-1},$$

so erhält man

$$\begin{aligned}\left|\left\{\sigma \in T^{(l)}(I) : \operatorname{dist}_\infty(C_\tau^l, C_\sigma^l) < \frac{\sqrt{2}}{\eta}(n-1)h2^{-l} + h\right\}\right| &\leq \left(1 + 2\left(\frac{\sqrt{2}}{\eta} + \frac{2^l}{n-1}\right)\right)^2 \\ &\leq \left(3 + \frac{2\sqrt{2}}{\eta}\right)^2.\end{aligned}$$

Das bedeutet, dass zu einem Gitter Ω_τ^l der Stufe l die Anzahl der Gitter Ω_σ^l mit $\tau \times \sigma$ nicht zulässig durch $\left(3 + \frac{2\sqrt{2}}{\eta}\right)^2$ beschränkt ist. Demnach kann die Abschätzung (3.13) angewendet werden, woraus

$$C_{sp} \leq \left(6 + \frac{4\sqrt{2}}{\eta}\right)^2$$

folgt.

Um fast lineare Komplexität zur Abschätzung des Speicherbedarfs zu erhalten, muss darüber hinaus die Tiefe $\operatorname{depth}(T(I)), \operatorname{depth}(T(J))$ der Clusterbäume $T(I), T(J)$ abgeschätzt werden. Für die angegebene Konstruktion kann dies erfolgen, indem man das Level p bestimmt, auf dem für den Durchmesser der Quadrate $\operatorname{diam}(C_\tau^p) \leq \sqrt{2}h$ gilt. In diesem Fall enthalten die Quadrate der Stufe $p+1$ maximal einen Gitterpunkt, so dass keine weitere Aufteilung erfolgt. Dies lässt sich mit der Wahl

$$p = \log_2(n-1)$$

erreichen, da in diesem Fall

$$\operatorname{diam}(C_\tau^p) = \sqrt{2}(n-1)h2^{-p} = \sqrt{2}h$$

gilt. Somit erhält man

$$\operatorname{depth}(T(I)) \leq p+1 = 1 + \log_2(n-1)$$

und damit als Abschätzung für den Speicherbedarf der entsprechenden $\mathcal{H}$-Matrix die Abschätzung

$$\mathcal{S}_{\mathcal{H}}(k,P) \leq \left(6 + \frac{4\sqrt{2}}{\eta}\right)^2 \cdot \max\{n_{min}, k\} \cdot 2\left(2 + \frac{1}{2}\log_2(\hat{n})\right)|\hat{n}|\,,$$

wobei die Bezeichnung $\hat{n} := n^2 = |I| = |J|$ verwendet wurde. Insgesamt ergibt sich in diesem Fall also fast lineare Komplexität.

Mit Hilfe der Konstanten C_{sp} lässt sich darüber hinaus die Anzahl der Elemente einer Partition P abschätzen. Dies ist zur Bestimmung des globalen Fehlers in der Frobenius-Norm von Interesse, der bei der Approximation einer Matrix $M \in \mathbb{R}^{I\times I}$ durch eine $\mathcal{H}$-Matrix $M_{\mathcal{H}} \in \mathcal{H}(k,P)$ auftritt. Für diesen gilt nach den Ausführungen in Abschnitt 3.1.2

$$\|M - M_{\mathcal{H}}\|_F \leq \epsilon\sqrt{|P|},$$

wenn die lokalen Fehler in allen Blöcken $b \in P$ durch

$$\|M|_b - M_{\mathcal{H}}|_b\|_F \leq \epsilon$$

abgeschätzt werden können.

Nach [Hac09, Lemma 6.3.4] und [Hac09, Anmerkung 6.3.5] gilt für Partitionen P, die nach Konstruktion 3.1.4 zu dem Clusterbaum $T(I)$ und $n_{min} > 1$ gebildet wurden,

$$|P| \leq \left(\frac{4\,|I|}{n_{min}} - 1\right) C_{sp}.$$

3.2.4 Diskrete separable Entwicklung

Analog zur Definition der separablen Entwicklung (vgl. beispielsweise [Hac09]), kann eine diskrete separable Entwicklung für Gitterfunktionen eingeführt werden:

Definition 3.2.1 (Diskrete separable Entwicklung) *Kann für eine Gitterfunktion $g \in \mathcal{D}_h(X_h \times Y_h)$ mit den Gittern $X_h, Y_h \subset h\mathbb{Z}^d$ eine Darstellung der Form*

$$g(x,y) = \sum_{l=1}^{k} u_l(x) v_l(y) + R_k(x,y),\ x \in X_h, y \in Y_h \tag{3.14}$$

mit Gitterfunktionen $u_l \in \mathcal{D}_h(X_h)$ und $v_l \in \mathcal{D}_h(Y_h)$ angegeben werden, dann heißt die rechte Seite diskrete separable Entwicklung von g auf $X_h \times Y_h$ mit Restglied R_k.

Um den Zusammenhang zwischen der diskreten separablen Entwicklung und der Niedrigrangapproximation von Matrixblöcken zu verdeutlichen, seien Gitter X_h mit t Gitterpunkten $x_1, \ldots, x_t$ und Y_h mit s Gitterpunkten $y_1, \ldots, y_s$ gegeben. Die Werte der Gitterfunktion $g(x,y), x \in X_h, y \in Y_h$ können in einer $t \times s$ Matrix

$$\mathcal{G} = \begin{pmatrix} g(x_1,y_1) & g(x_1,y_2) & \cdots & g(x_1,y_s) \\ g(x_2,y_1) & g(x_2,y_2) & \cdots & g(x_2,y_s) \\ \vdots & \vdots & \ddots & \vdots \\ g(x_t,y_1) & g(x_t,y_2) & \cdots & g(x_t,y_s) \end{pmatrix}$$

dargestellt werden. Besitzt g eine diskrete separable Entwicklung der Form (3.14), so kann die Matrix $\mathcal{G}$ durch die Rang-k-Matrix

$$\mathcal{G}^k = AB^T$$

mit den Matrizen

$$A = \begin{pmatrix} u_1(x_1) & u_2(x_1) & \ldots & u_k(x_1) \\ u_1(x_2) & u_2(x_2) & \ldots & u_k(x_2) \\ \vdots & \vdots & & \vdots \\ u_1(x_t) & u_2(x_t) & \ldots & u_k(x_t) \end{pmatrix} \in \mathbb{R}^{\{1,\ldots,t\} \times \{1,\ldots,k\}}$$

und

$$B = \begin{pmatrix} v_1(y_1) & v_2(y_1) & \ldots & v_k(y_1) \\ v_1(y_2) & v_2(y_2) & \ldots & v_k(y_2) \\ \vdots & \vdots & & \vdots \\ v_1(y_s) & v_2(y_s) & \ldots & v_k(y_s) \end{pmatrix} \in \mathbb{R}^{\{1,\ldots,s\} \times \{1,\ldots,k\}}$$

approximiert werden. Der resultierende Fehler ergibt sich aus dem Restglied R_k der diskreten separablen Entwicklung. Wünschenswert ist ein Fehler, der exponentiell bezüglich k abnimmt, so dass man bereits bei der Verwendung von einem kleinen k eine gute Approximation erhält.

Definition 3.2.2 (Exponentielle Konvergenz) *Existieren Konstanten $c_1 \geq 0$ und $c_2, c_3 > 0$, so dass der Restterm der diskreten separablen Entwicklung* (3.14) *durch*

$$\|R_k\| \leq c_1 \exp(-c_2 k^{c_3})$$

abgeschätzt werden kann, heißt die diskrete separable Entwicklung exponentiell konvergent bezüglich der Norm $\|\cdot\|$.

Existiert demnach für die Gitterfunktion g eine diskrete separable Entwicklung g^k auf dem Gitter $X_h \times Y_h$ mit exponentieller Konvergenz, so kann ein Matrixblock, der durch die Auswertung von g auf dem Gitter $X_h \times Y_h$ gebildet wird, durch eine Rang-k-Matrix approximiert werden und der Fehler fällt exponentiell mit dem Rang k der Approximation. Ist umgekehrt an die Genauigkeit der Approximation die Forderung $\|R_k\| \leq \epsilon(k)$ gestellt, ist bei exponentieller Konvergenz die Verwendung von

$$k(\epsilon) = \left\lceil \left(\frac{1}{c_2} \log \frac{c_1}{\epsilon} \right)^{1/c_3} \right\rceil$$

erforderlich. Der Rang k (in der separablen Entwicklung entspricht k der Anzahl der Summanden), der benötigt wird, um eine Approximationsgüte ϵ zu erreichen, wächst beim Vorliegen von exponentieller Konvergenz nur logarithmisch, so dass auch bei Verwendung niedriger Ränge eine gute Approximation erzielt werden kann.

3.2.5 Fehlerabschätzung

Die Abschätzung des globalen Fehlers, der bei der Approximation einer Matrix $\mathcal{G} \in \mathbb{R}^{I \times I}$ durch eine $\mathcal{H}$-Matrix $\mathcal{G}_{\mathcal{H}} \in \mathcal{H}(k, P)$ zu einer Partition P von $I \times I$ resultiert, kann mit Hilfe der lokalen Fehler in den Matrixblöcken zu $b = \tau \times \sigma \in P$ erfolgen.

Gelingt es, die Existenz einer diskreten separablen Entwicklung für die zulässigen Gitter $X_\tau \times X_\sigma$ zu zeigen, dann lässt sich das Restglied zur Fehlerabschätzung verwenden. Im Hinblick auf die Ergebnisse in Abschnitt 4.2 sei deshalb für die folgenden Überlegungen vorausgesetzt, dass wie in Abschnitt 3.2.4 die Gitterfunktion $g(x, y), x \in X_\tau, y \in X_\sigma$ zum Matrixblock $\mathcal{G}|_b$ mit $X_\tau, X_\sigma \subset \Omega_h$ durch eine diskrete separable Entwicklung der Form

$$g^k(x, y) = \sum_{l=1}^{k} u_l(x) v_l(y), \quad x \in X_\tau, y \in X_\sigma \tag{3.15}$$

approximiert werden kann. Der dabei auftretende Fehler soll darüber hinaus eine Abschätzung der Form

$$\sum_{y \in X_\sigma} \left| g(x, y) - g^k(x, y) \right|^2 \leq \epsilon^2 \sum_{y \in D} |g(x, y)|^2 \tag{3.16}$$

für alle $x \in X_\tau$ erfüllen, wobei $D \subset h\mathbb{Z}^2$ zunächst durch ein beliebiges Gitter mit $\Omega_h \supset D \supset X_\sigma$ und $D \cap X_\tau = \emptyset$ gegeben sei.

Sind diese Bedingungen erfüllt, kann der Matrixblock zu zulässigem $b = \tau \times \sigma$ nach den Überlegungen in Abschnitt 3.2.4 durch eine Rang-k-Matrix approximiert werden und der resultierende lokale Fehler ergibt sich in der Frobenius-Norm durch

$$\begin{aligned}\left\|\mathcal{G}|_b - \mathcal{G}^k|_b\right\|_F^2 &= \sum_{x\in X_\tau}\sum_{y\in X_\sigma}\left|g(x,y) - g^k(x,y)\right|^2 \leq \epsilon^2 \sum_{x\in X_\tau}\sum_{y\in D}|g(x,y)|^2 \\ &\leq \epsilon^2 \left\|\mathcal{G}\right\|_F^2 . \end{aligned} \tag{3.17}$$

Definiert man die Matrix

$$\mathcal{G}_{\mathcal{H}} := \begin{cases}\mathcal{G}^k|_b & b \in P^+ \\ \mathcal{G}|_b & b \in P^-,\end{cases}$$

dann gilt für den Fehler, der bei der Approximation von $\mathcal{G}$ durch $\mathcal{G}_{\mathcal{H}}$ entsteht, die Abschätzung aus (3.17) für alle zulässigen Blöcke $b \in P^+$. In den Blöcken $b \in P^-$ treten keine Approximationsfehler auf. Der Übergang zum globalen Fehler kann nach (3.6) unter diesen Voraussetzungen angegeben werden durch

$$\left\|\mathcal{G} - \mathcal{G}_{\mathcal{H}}\right\|_F^2 \leq \epsilon^2 \left|P^+\right| \left\|\mathcal{G}\right\|_F^2 ,$$

wobei $|P^+|$ die Anzahl der Elemente $b \in P^+$ bezeichnet. Eine Abschätzung von $|P^+|$ kann, wie in Abschnitt 3.2.3 angegeben, mit Hilfe der Konstanten C_{sp} erfolgen.

Berücksichtigt man, dass die Einträge der Inversen von Finite-Differenzen-Matrizen durch die Werte der diskreten Greenschen Funktion g_h an den entsprechenden Gitterpunkten gegeben sind, können diese Überlegungen auf die Inverse bzw. g_h übertragen werden. Dies ermöglicht es, ein Ergebnis zur Approximierbarkeit der Inversen von Finite-Differenzen-Matrizen durch $\mathcal{H}$-Matrizen zu erhalten.

Dafür ist die Existenz einer diskreten separablen Entwicklung der diskreten Greenschen Funktion nachzuweisen, was der Existenz einer Niedrigrangapproximation in zulässigen Blöcken der Inversen entspricht. Damit der Approximationsfehler die gewünschte Form aufweist, ist die Gültigkeit einer Fehlerabschätzung der Form (3.16) für die diskrete Greensche Funktion mit $k = k(\epsilon) = \left\lceil \left(\frac{1}{c_2}\log\frac{c_1}{\epsilon}\right)^{1/c_3} \right\rceil$ ($c_1 \geq 0$ und $c_2, c_3 > 0$) zu zeigen. Dieses Ergebnis wird in Abschnitt 4.2 erzielt.

4 Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen von Finite-Differenzen-Matrizen

Die Existenz einer $\mathcal{H}$-Matrix Approximation konnte bereits für die Inverse verschiedener Matrizen gezeigt werden, beispielsweise für Finite-Element-Matrizen bei elliptischen Randwertproblemen und für dünnbesetzte, symmetrische, positiv definite, wohlkonditionierte Matrizen (vgl. Abschnitt 3.1.3). Ein Ergebnis speziell für Finite-Differenzen-Matrizen ist jedoch nicht bekannt. Daher wird für diesen Fall ein eigenständiger Ansatz eingeführt, dessen Entwicklung auf Grundlage der Ergebnisse zu Finite-Element-Matrizen aus [Hac09] bzw. [BH03] erfolgt. Die wesentlichen Ideen aus diesen Arbeiten lassen sich auf den Fall von Finite-Differenzen-Diskretisierungen übertragen. Die entsprechenden Resultate werden in den Abschnitten 4.1 und 4.2 angegeben.

Den Ausgangspunkt der Überlegungen bildet der Zusammenhang zwischen der diskreten Greenschen Funktion und den Einträgen der Inversen von Finite-Differenzen-Matrizen, der in Kapitel 2 dargestellt wurde. Nach den Ausführungen in Abschnitt 3.1.2 kann zur Abschätzung des globalen Fehlers, der bei der Approximation der Inversen durch eine $\mathcal{H}$-Matrix auftritt, der lokale Fehler in den Matrixblöcken genutzt werden. Da nur in den zulässigen Blöcken $b \in P^+$ eine Approximation vorgenommen wird, muss für diese Blöcke nachgewiesen werden, dass die entsprechenden Teilmatrizen der Inversen durch eine Rang-k-Matrix mit exponentieller Konvergenz approximiert werden können. Dies lässt sich nach den Ausführungen in Abschnitt 3.2.4 zeigen, indem die Existenz einer diskreten separablen Entwicklung der diskreten Greenschen Funktion mit exponentieller Konvergenz auf den entsprechenden zulässigen Gittern nachgewiesen wird.

Ein Resultat dieser Form für Gitterfunktionen ist bisher nicht bekannt, so dass ein eigenständiger methodischer Ansatz erarbeitet werden muss. Grundsätzlich sind dazu verschiedene Vorgehensweisen denkbar. Neben der Übertragung des Ansatzes, der für Finite-Element-Matrizen verwendet wurde, können auch andere Zusammenhänge genutzt werden, um die Existenz einer diskreten separablen Entwicklung der diskreten Greenschen Funktion mit exponentieller Konvergenz nachzuweisen. Daher wird zu Beginn von Abschnitt 4.1 kurz auf einige alternative Ideen eingegangen. Wie bereits bei der Einführung des Modellproblems ist jedoch auch in diesem Zusammenhang zu berücksichtigen, dass eine Verallgemeinerung des entsprechenden Ansatzes für die Problemstellung im Modell METRAS möglich sein sollte. Unter Beachtung dieser Zielsetzung stellen sich die alternativen Herangehensweisen als nicht geeignet heraus. Aus diesem Grund wird der Ansatz aus [Hac09] auf den Fall von Finite-Differenzen-Matrizen übertragen. Der große Vorteil dieser Vorgehensweise besteht darin, dass eine einfache

Verallgemeinerung des Ansatzes für andersartige Differenzenoperatoren möglich ist. Dazu muss die Gültigkeit einer diskreten Cacciopoli-Ungleichung (wie in Satz 2.3.2 bzw. Satz 2.3.3 für das Modellproblem) unter Berücksichtigung des entsprechenden Differenzenoperators gezeigt werden. Außer der Übertragung einer möglicherweise anderslautenden Konstanten in der Ungleichung sind keine weiteren Anpassungen erforderlich.

Mit Hilfe des Resultats zur Existenz einer diskreten separablen Entwicklung der diskreten Greenschen Funktion wird in Abschnitt 4.2 das Hauptresultat zur Approximation der Inversen von Finite-Differenzen-Matrizen mittels $\mathcal{H}$-Matrizen angegeben. In Abschnitt 4.3 wird als Ausblick kurz dargestellt, wie das Hauptresultat verwendet werden könnte, um die Existenz von $\mathcal{H}$-Matrix Approximationen der Faktoren einer LU-Zerlegung nachzuweisen.

4.1 Separable Approximation der diskreten Greenschen Funktion

Zum Nachweis der Existenz einer diskreten separablen Entwicklung der diskreten Greenschen Funktion mit exponentieller Konvergenz auf zulässigen Gittern können mehrere Ansätze verwendet werden. Eine mögliche Überlegung beruht darauf, die bekannten Ergebnisse zur Existenz einer separablen Entwicklung der Greenschen Funktion aus dem kontinuierlichen Fall zu nutzen. Berücksichtigt man den Zusammenhang der kontinuierlichen und der entsprechenden diskreten Problemstellung, so erhält man einen Kandidaten für die diskrete separable Entwicklung der diskreten Greenschen Funktion, indem man die Beschränkung der (kontinuierlichen) separablen Entwicklung auf die entsprechenden Gitter betrachtet. Auf diese Weise ließen sich die Ergebnisse aus der kontinuierlichen Problemstellung direkt zur Untersuchung der diskreten Problemstellung verwerten.

Um eine Fehlerabschätzung für diesen Fall zu erhalten, wäre jedoch zusätzlich eine Abschätzung des Fehlers erforderlich, der bei der Approximation der Greenschen Funktion durch die diskrete Greensche Funktion auftritt. Ein solches Resultat findet man beispielsweise in [Laa58] für die Standard-Diskretisierung des Laplace-Operators auf einem Rechteck (dies entspricht dem Modellproblem mit konstanten Koeffizienten $a = d = 1$). Weitere Ergebnisse dieser Form, insbesondere für andere Differenzenoperatoren und höherdimensionale Probleme, sind nicht bekannt, so dass eine Verallgemeinerung dieser Vorgehensweise nicht Erfolg versprechend ist. Daher wird dieser Ansatz nicht weiter verfolgt.

Das Ergebnis in [Laa58] beruht darauf, dass zu der untersuchten Problemstellung explizite Darstellungen der Greenschen und der diskreten Greenschen Funktion bekannt sind. Die explizite Darstellung der diskreten Greenschen Funktion besitzt die Form einer diskreten separablen Entwicklung und für den Restterm lässt sich exponentielle Konvergenz in der Maximums-Norm auf zulässigen Gittern nachweisen. In diesem Fall ließe sich demnach die explizite Darstellung der diskreten Greenschen Funktion zum Nachweis der Existenz einer diskreten separablen Entwicklung mit exponentieller Konvergenz nutzen. Von Vorteil ist bei dieser Vorgehensweise, im Gegensatz zum ersten Ansatz, dass die Untersuchung des Fehlers, der bei der Approximation der Greenschen durch die diskrete

Greensche Funktion auftritt, nicht mehr erforderlich ist. Eine Verallgemeinerung dieses Ansatzes scheitert jedoch erneut daran, dass nur in den seltensten Fällen eine explizite Darstellung der diskreten Greenschen Funktion angegeben werden kann. Daher wird in den folgenden Abschnitten der Ansatz aus [Hac09] verwendet.

4.1.1 Approximation von Gitterfunktionen

Das folgende Lemma wird analog zu [Hac09, Lemma 11.3.5] für den Fall von Gitterfunktionen auf einem Rechteckgitter formuliert. Das Ergebnis dient als Hilfsmittel für die weitere Untersuchung und lässt sich für diskret L-harmonische Funktionen unter Verwendung der diskreten Poincaré-Ungleichung für Rechteckgitter aus Satz 2.2.7 zeigen.

Lemma 4.1.1 *Sei $\Omega_h \subset h\mathbb{Z}^2$ ein Rechteckgitter mit konstanter Schrittweite h. Dann existiert für alle $k \in \mathbb{N}$ ein Unterraum $V_k \subset Z_h^L(\Omega_h)$ der Dimension $\dim V_k \leq k$ mit*

$$\operatorname{dist}_{L_h^2(\Omega_h)}(u, V_k) \leq \sqrt{2}\frac{\operatorname{diam}(\Omega_h)}{\sqrt{k}} \|\nabla_h u\|_{L_h^2(\Omega_h)}$$

für alle $u \in Z_h^L(\Omega_h)$.

Beweis Das Rechteckgitter Ω_h sei ohne Einschränkung durch ein $n \times m$ Gitter gegeben. Demnach besitzt es den Durchmesser

$$\operatorname{diam}(\Omega_h) = h\sqrt{(n-1)^2 + (m-1)^2}.$$

Die Gitterpunkte aus $\Gamma(\Omega_h)$ bilden ein Rechteck R, das die gleichen Seitenlängen wie Ω_h besitzt, und es gilt $\Omega_h \subset R$. Dieses Rechteck wird in k Teilrechtecke $R_i, i = 1, \ldots, k$ unterteilt. Geht man zunächst davon aus, dass $k = l^2$ mit $l \in \mathbb{N}$ gilt, dann können die Teilrechtecke gebildet werden, indem die Seiten des Rechtecks R in jeweils l Abschnitte unterteilt werden (vgl. Abbildung 4.1). Demnach besitzen die k Teilrechtecke R_i die Kantenlänge $\frac{(n-1)h}{l} \times \frac{(m-1)h}{l}$. Zur Bildung der Teilgitter wird $\Omega_i := R_i \cap \Omega_h$ gesetzt. Sollte die Zuordnung der Gitterpunkte zu mehreren Ω_i möglich sein (dieser Fall kann auftreten, wenn die Gitterpunkte auf dem Rand der R_i liegen), so werden diese Punkte alle einem der angrenzenden Teilgitter zugeordnet, so dass die Ω_i stets durch Rechteckgitter gegeben sind.

Für den Unterraum

$$W_k := \{v \in D_h(\Omega_h) : v \text{ konstant auf } \Omega_i \text{ für alle } i = 1, \ldots, k\}$$

gilt $\dim W_k \leq k$. Zur Approximation von u soll die Gitterfunktion $\bar{u} \in W_k$ verwendet werden, die durch $\bar{u}|_{\Omega_i} = \bar{u}_i$ mit dem Mittelwert

$$\bar{u}_i := \frac{1}{|\Omega_i|} \sum_{x \in \Omega_i} u(x)$$

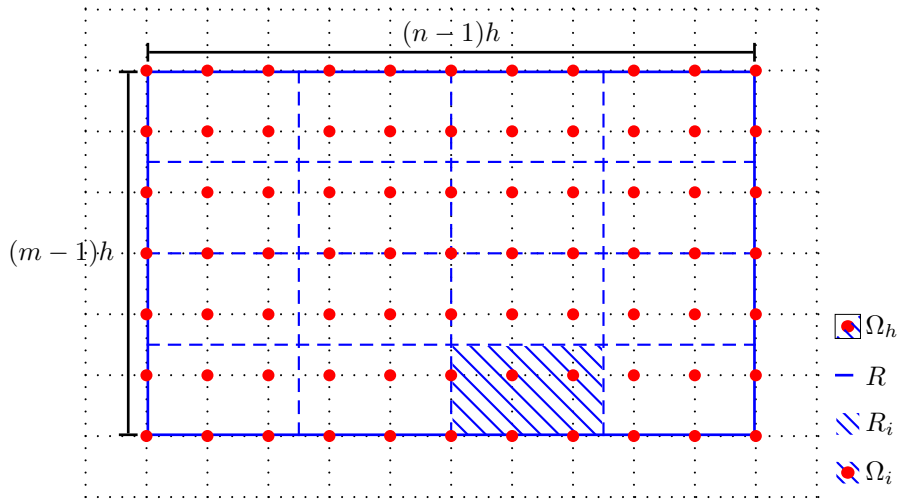


Abbildung 4.1: Rechteckeinteilung für $l = 4$

gegeben ist. Zur Abschätzung des Fehlers $\|u - \bar{u}_i\|_{L_h^2(\Omega_i)}$, der bei der Approximation einer Gitterfunktion $u \in \mathcal{D}_h(\Omega_h)$ auftritt, kann auf den Teilgittern Ω_i die diskrete Poincaré-Ungleichung aus Satz 2.2.7 mit

$$C_h = \frac{\operatorname{diam}(\Omega_i)}{\sqrt{2}}$$

verwendet werden.

Die Größe $\operatorname{diam}(\Omega_i)$ kann in Abhängigkeit des Gesamtdurchmessers $\operatorname{diam}(\Omega_h)$ durch

$$\operatorname{diam}(\Omega_i) \leq \operatorname{diam}(R_i) = \sqrt{\left(\frac{(n-1)h}{l}\right)^2 + \left(\frac{(m-1)h}{l}\right)^2} = \frac{\operatorname{diam}(\Omega_h)}{l}$$

abgeschätzt werden, so dass man mittels der diskreten Poincaré-Ungleichung auf allen Teilgittern $\Omega_i, i = 1, \ldots, k$ die Fehlerabschätzung

$$\begin{aligned} \|u - \bar{u}_i\|^2_{L_h^2(\Omega_i)} &\leq \left(\frac{\operatorname{diam}(\Omega_i)}{\sqrt{2}}\right)^2 \|\nabla_h u\|^2_{L_h^2(\Omega_i)} \\ &\leq \left(\frac{1}{\sqrt{2}}\frac{\operatorname{diam}(\Omega_h)}{l}\right)^2 \|\nabla_h u\|^2_{L_h^2(\Omega_i)} \end{aligned}$$

erhält.

Durch Summation ergibt sich die Fehlerabschätzung auf Ω_h für die stückweise konstante Gitterfunktion $\bar{u} \in W_k$

$$\|u - \bar{u}\|^2_{L_h^2(\Omega_h)} \leq \left(\frac{1}{\sqrt{2}}\frac{\operatorname{diam}(\Omega_h)}{l}\right)^2 \|\nabla_h u\|^2_{L_h^2(\Omega_h)}$$

und demnach

$$\operatorname{dist}_{L_h^2(\Omega_h)}(u, W_k) \leq \|u - \bar{u}\|_{L_h^2(\Omega_h)} \leq \frac{1}{\sqrt{2}}\frac{\operatorname{diam}(\Omega_h)}{l} \|\nabla_h u\|_{L_h^2(\Omega_h)}.$$

Ausgehend vom Ansatz $k = l^2$ kann man zu einem allgemeinen k gelangen, indem $l := \left\lfloor \sqrt{k} \right\rfloor \in \mathbb{N}$ gesetzt wird. Dann gilt $l^2 \leq k \leq (l+1)^2$. Das Ergebnis aus dem ersten Teil kann für $k' := l^2$ angewendet werden. Definiert man $W_k := W_{k'}$, dann gilt $\dim W_k = \dim W_{k'} \leq k' \leq k$. Da außerdem $\frac{1}{l} \leq \frac{2}{l+1} \leq \frac{2}{\sqrt{k}}$ erfüllt ist, ergibt sich für den allgemeinen Fall

$$\operatorname{dist}_{L_h^2(\Omega_h)}(u, W_k) \leq \sqrt{2}\frac{\operatorname{diam}(\Omega_h)}{\sqrt{k}} \left\|\nabla_h u\right\|_{L_h^2(\Omega_h)}.$$

Unter Verwendung der orthogonalen Projektion $\Pi : \mathcal{D}_h(\Omega_h) \to Z_h^L(\Omega_h)$ mittels des L_h^2-Skalarprodukts und der Definition $V_k := \Pi(W_k)$ behält die Abschätzung

$$\operatorname{dist}_{L_h^2(\Omega_h)}(u, V_k) \leq \sqrt{2}\frac{\operatorname{diam}(\Omega_h)}{\sqrt{k}} \left\|\nabla_h u\right\|_{L_h^2(\Omega_h)}$$

für $u \in Z_h^L(\Omega_h)$ ihre Gültigkeit, da für $u \in Z_h^L(\Omega_h)$

$$\left\|u - \Pi\bar{u}\right\|_{L_h^2(\Omega_h)} \leq \left\|\Pi(u - \bar{u})\right\|_{L_h^2(\Omega_h)} \leq \left\|u - \bar{u}\right\|_{L_h^2(\Omega_h)}$$

für alle $\bar{u} \in W_k$ gilt. ∎

4.1.2 Resultat für diskret L-harmonische Funktionen

Das Ziel der folgenden Untersuchung ist der Nachweis eines Resultats zur Approximierbarkeit diskret L-harmonischer Funktionen durch eine diskrete separable Entwicklung mit exponentieller Konvergenz. Da die diskrete Greensche Funktion zum Modellproblem auf zulässigen Gittern diskret L-harmonisch ist, kann das Resultat im Anschluss zum Nachweis der Existenz einer diskreten separablen Entwicklung der diskreten Greenschen Funktion mit exponentieller Konvergenz genutzt werden.

Im nächsten Abschnitt werden zunächst einige Vorüberlegungen angegeben, um spezielle Bezeichnungen einzuführen und einige grundlegende Überlegungen anzustellen, die für den Beweis des Resultats erforderlich sind.

Vorüberlegungen

Die Rechteckgitter K_h^0 und K_h^δ mit $\delta \geq 1$, letzteres wird mittels Konstruktion 2.3.1 für $l = \delta$ gebildet, stellen die Grundlage der weiteren Ausführungen dar. Um eine zweite Schachtelung von Rechteckgittern einzuführen, wird der Bereich zwischen $\Gamma(K_h^0)$ und $\Gamma(K_h^\delta)$ in $p \in \mathbb{N}$ Abschnitte (unabhängig von der Gitterstruktur) unterteilt, so dass sich Rechtecke $\tilde{K}_j, j = 0, \ldots, p$ mit

$$\tilde{K}_0 \subset \tilde{K}_1 \subset \cdots \subset \tilde{K}_p$$

ergeben.

Jeder der Abschnitte $\tilde{K}_{j+1} \backslash \tilde{K}_j$ kann mehr als eine Punktreihe enthalten und es können sich unterschiedlich viele Punktreihen in den einzelnen Abschnitten befinden. Es muss

lediglich sichergestellt werden, dass die Erweiterung von $\tilde{K}_j$ zu $\tilde{K}_{j+1}$ um mindestens eine Punktreihe erfolgt. Dies ist erforderlich, weil für die resultierenden Teilgitter die diskrete Cacciopoli-Ungleichung aus Satz 2.3.4 angewendet werden soll.

Die entsprechenden geschachtelten Rechteckgitter $K_j, j = 0, \ldots, p$ erhält man wie üblich aus

$$K_j := \tilde{K}_j \cap K_h^\delta, \quad j = 0, \ldots, p$$

(vgl. Abbildung 4.2). Für diese gilt

$$K_h^0 = K_0 \subset K_1 \subset \cdots \subset K_p = K_h^\delta.$$

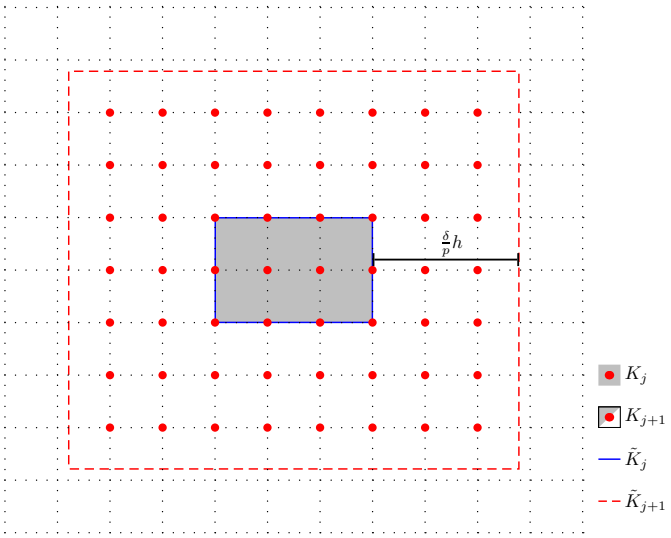


Abbildung 4.2: Geschachtelte Rechtecke

Berücksichtigt man

$$\frac{\operatorname{dist}_\infty(K_h, \Gamma(K_h^\delta))}{h} = \delta,$$

lässt sich durch die Anforderung

$$\delta \geq p$$

sicherstellen, dass mindestens eine Punktreihe pro Abschnitt vorkommt. Außerdem kann die Abschätzung

$$\operatorname{dist}_\infty(K_j, \Gamma(K_{j+1})) \geq \left\lfloor \frac{\delta}{p} \right\rfloor h$$

verwendet werden.

In dem später folgenden Resultat wird die Gültigkeit der Ungleichung

$$\operatorname{diam}(K_h) \leq \eta \operatorname{dist}_\infty(K_h, \Gamma(K_h^\delta)) \tag{4.1}$$

vorausgesetzt. Für die Anwendung des Satzes mit den zulässigen Mengen X_τ und X_σ wird

$$\delta := \frac{\operatorname{dist}_\infty(X_\tau, X_\sigma) - h}{h} \tag{4.2}$$

definiert. Demnach gilt $X_\tau^\delta \cap X_\sigma = \emptyset$ bzw. $X_\tau \cap X_\sigma^\delta = \emptyset$ und die Voraussetzung (4.1) ist wegen der Gültigkeit der Zulässigkeitsbedingung für X_τ und X_σ erfüllt. Denn im Fall von

$$\min\{\operatorname{diam}(X_\tau), \operatorname{diam}(X_\sigma)\} = \operatorname{diam}(X_\tau)$$

kann der Satz für $K_h = X_\tau$ angewendet werden, da nach der Zulässigkeitsbedingung

$$\operatorname{diam}(X_\tau) \leq \eta \left(\operatorname{dist}_\infty(X_\tau, X_\sigma) - h\right)$$

und damit

$$\operatorname{diam}(K_h) = \operatorname{diam}(X_\tau) \leq \eta \left(\operatorname{dist}_\infty(X_\tau, X_\sigma) - h\right) = \eta\delta h = \eta \operatorname{dist}_\infty(K_h, \Gamma(K_h^\delta))$$

gilt. Der Fall $\min\{\operatorname{diam}(X_\tau), \operatorname{diam}(X_\sigma)\} = \operatorname{diam}(X_\sigma)$ lässt sich analog für $K_h = X_\sigma$ behandeln.

Im Verlauf des Beweises wird das zunächst allgemein verwendete p durch $p := \left\lceil \log \frac{1}{\epsilon} \right\rceil$ konkretisiert. Nach den bereits angeführten Bedingungen muss in diesem Fall die Gültigkeit der Ungleichung

$$\delta \geq p = \left\lceil \log \frac{1}{\epsilon} \right\rceil$$

sichergestellt sein. Aus (4.1) erhält man

$$\delta h = \operatorname{dist}_\infty(K_h, \Gamma(K_h^\delta)) \geq \frac{1}{\eta} \operatorname{diam}(K_h).$$

Um zu gewährleisten, dass $\delta \geq p$ gilt, muss die Ungleichung

$$\frac{1}{h\eta} \operatorname{diam}(K_h) \geq p = \left\lceil \log \frac{1}{\epsilon} \right\rceil$$

erfüllt sein. Mit der Definition

$$\epsilon_0 := \exp\left(- \left\lceil \frac{1}{h\eta} \operatorname{diam}(K_h) \right\rceil \right)$$

erhält man unter Voraussetzung von $\epsilon_0 \leq \epsilon < 1$

$$p = \left\lceil \log \frac{1}{\epsilon} \right\rceil \leq \left\lceil \frac{1}{h\eta} \operatorname{diam}(K_h) \right\rceil \leq \lceil \delta \rceil = \delta,$$

wobei die letzte Gleichung gilt, da aus der Definition in (4.2) $\delta \in \mathbb{N}$ geschlossen werden kann.

Die Größe ϵ_0 hängt demnach von η und insbesondere von der Wahl der Größe n_{min} ab, weil durch die Vorgabe von n_{min} auch der minimal auftretende Durchmesser der

zulässigen Gitter beeinflusst wird. Da in der Regel $n_{min} > 1$ vorgegeben wird, gilt insbesondere $\min\{\operatorname{diam}(X_\tau), \operatorname{diam}(X_\sigma)\} > 0$ und damit $\epsilon_0 < 1$.

In dem Beweis von Lemma 4.1.2 wird außerdem eine Abschätzung der Durchmesser $\operatorname{diam}(K_j), j = 0, \ldots, p$ benötigt. Diese kann nach der angegebenen Konstruktion durch

$$\operatorname{diam}(K_j) \leq \operatorname{diam}(\tilde{K}_j) \leq \operatorname{diam}(K_h) + 2j\sqrt{2}\frac{\delta}{p}h \leq \operatorname{diam}(K_h) + 2\sqrt{2}\delta h$$

erfolgen. Zusammen mit der Voraussetzung $\operatorname{diam}(K_h) \leq \eta \operatorname{dist}_\infty(K_h, \Gamma(K_h^\delta)) = \eta\delta h$ ergibt sich

$$\operatorname{diam}(K_j) \leq \delta h \left(\eta + 2\sqrt{2}\right).$$

Bei der Kombination der Abschätzungen aus Lemma 4.1.1 und Satz 2.3.4 muss das Produkt der Konstanten aus den beiden Ungleichungen abgeschätzt werden. Dazu lassen sich folgende Ergebnisse nutzen:

- die Wahl von $k := \left\lceil (p\beta)^2 \right\rceil$ führt zu $\sqrt{k} \geq p\beta$
- $\operatorname{diam}(K_j) \leq \delta h \left(\eta + 2\sqrt{2}\right)$
- $\operatorname{dist}_\infty(K_j, \Gamma(K_{j+1})) \geq \left\lfloor \frac{\delta}{p} \right\rfloor h$
- $\frac{\delta}{p} \leq \left\lfloor \frac{\delta}{p} \right\rfloor + 1 \leq 2 \left\lfloor \frac{\delta}{p} \right\rfloor$.

Mittels dieser Überlegungen kann der resultierende Term abgeschätzt werden durch

$$\begin{aligned}
\sqrt{2}\frac{\operatorname{diam}(K_j)}{\sqrt{k}} \frac{\sqrt{\kappa}\, C_{caccio}}{\operatorname{dist}_\infty(K_j, \Gamma(K_{j+1}))} &\leq \sqrt{2}\sqrt{\kappa}\, C_{caccio}\frac{\delta}{p}\frac{h\left(\eta + 2\sqrt{2}\right)}{\beta}\frac{1}{\left\lfloor \frac{\delta}{p} \right\rfloor h} \\
&\leq 2\sqrt{2}\sqrt{\kappa}\, C_{caccio}\left\lfloor \frac{\delta}{p} \right\rfloor \frac{\left(\eta + 2\sqrt{2}\right)}{\beta}\frac{1}{\left\lfloor \frac{\delta}{p} \right\rfloor} \\
&= 2\sqrt{2}\sqrt{\kappa}\, C_{caccio}\frac{\left(\eta + 2\sqrt{2}\right)}{\beta}. \qquad (4.3)
\end{aligned}$$

Diskret L-harmonische Funktionen

Unter Verwendung der Ergebnisse aus dem vorangegangenen Abschnitt kann analog zu [Hac09, Lemma 11.3.6] das folgende Resultat zur Approximation von diskret L-harmonischen Funktionen durch eine diskrete separable Entwicklung mit exponentieller Konvergenz auf Rechteckgittern angegeben und bewiesen werden:

Lemma 4.1.2 *Seien Ω_h und $K_h \subset \Omega_h$ Rechteckgitter und es gelte*

$$\operatorname{diam}(K_h) \leq \eta \operatorname{dist}_\infty(K_h, \Gamma(K_h^\delta))$$

mit $\eta > 0$. Außerdem sei K_h^δ, wie in Konstruktion 2.3.1 beschrieben, mit $\delta \geq 1$ gebildet, es gelte $\Omega_h \backslash K_h^\delta \neq \emptyset$ und es sei

$$\epsilon_0 = \exp\left(-\left\lceil \frac{1}{h\eta}\operatorname{diam}(K_h)\right\rceil\right).$$

Dann existiert für jedes ϵ mit $\epsilon_0 \leq \epsilon < 1$ ein Unterraum $W \subset Z_h^L(K_h)$ mit der Dimension

$$\dim W \leq \left\lceil \log\frac{1}{\epsilon}\right\rceil + c^2\left\lceil \log\frac{1}{\epsilon}\right\rceil^3$$

mit $c = 2\sqrt{2}\sqrt{\kappa}\ C_{caccio}\left(\eta + 2\sqrt{2}\right)\exp(1)$ und es gilt

$$\operatorname{dist}_{L_h^2(K_h)}(u, W) \leq \epsilon \left\|u\right\|_{L_h^2(K_h^\delta \cap \Omega_h)} \quad \textit{für alle } u \in Z_{h,0}^L(K_h^\delta; \Omega_h). \tag{4.4}$$

Beweis Zunächst wird, wie in Abschnitt 4.1.2 beschrieben, eine Schachtelung von $p+1$ Rechteckgittern

$$K_h = K_0 \subset \mathrm{K}_1 \subset \cdots \subset K_p = K_h^\delta$$

konstruiert und die Bezeichnung $Z_j := Z_{h,0}^L(K_j; \Omega_h)$ eingeführt.

Aus der Anwendung von Lemma 4.1.1 für $K_j \cap \Omega_h$ ergibt sich die Existenz eines Unterraums $V_j \subset Z_j$ mit $\dim V_j \leq k$ und der Fehlerabschätzung

$$\operatorname{dist}_{L_h^2(K_j\cap\Omega_h)}(u, V_j) \leq \sqrt{2}\,\frac{\operatorname{diam}(K_j)}{\sqrt{k}}\left\|\nabla_h u\right\|_{L_h^2(K_j\cap\Omega_h)}$$

für alle $u \in Z_j$.

Aufgrund der Voraussetzung $\epsilon \geq \epsilon_0$ gilt nach den Vorüberlegungen in Abschnitt 4.1.2 für $p := \left\lceil \log\left(\frac{1}{\epsilon}\right)\right\rceil$

$$\delta \geq p,$$

so dass sich das Gitter K_{j+1} aus der Erweiterung des Gitters K_j um mindestens eine Punktreihe ergibt. Wegen $\Omega_h \backslash K_h^\delta \neq \emptyset$ gilt dies ebenfalls für die Erweiterung von $K_j \cap \Omega_h$ zu $K_{j+1} \cap \Omega_h$. Daher kann Satz 2.3.4 für $K_j \cap \Omega_h$ und $K_{j+1} \cap \Omega_h$ angewendet werden, so dass man

$$\left\|\nabla_h u\right\|_{L_h^2(K_j\cap\Omega_h)} \leq \frac{\sqrt{\kappa}\ C_{caccio}}{\operatorname{dist}_\infty(K_j, \Gamma(K_{j+1}))}\left\|u\right\|_{L_h^2(K_{j+1}\cap\Omega_h)}$$

für alle $u \in Z_{j+1}$ erhält.

Da für eine diskret L-harmonische Funktion $u \in Z_{j+1}$ ebenfalls $u \in Z_j$ gilt, lassen sich die beiden Abschätzungen kombinieren zu

$$\operatorname{dist}_{L_h^2(K_j\cap\Omega_h)}(u, V_j) \leq \sqrt{2}\frac{\sqrt{\kappa}\ C_{caccio}\operatorname{diam}(K_j)}{\sqrt{k}\operatorname{dist}_\infty(K_j, \Gamma(K_{j+1}))}\left\|u\right\|_{L_h^2(K_{j+1}\cap\Omega_h)}$$

für alle $u \in Z_{j+1}$.

Nach den Vorüberlegungen aus Abschnitt 4.1.2 und der Verwendung von $k := \left\lceil (\beta p)^2 \right\rceil$ mit einem zunächst frei wählbaren Faktor β kann die Abschätzung

$$\sqrt{2}\,\frac{\operatorname{diam}(K_j)}{\sqrt{k}}\,\frac{\sqrt{\kappa}\;C_{caccio}}{\operatorname{dist}_\infty(K_j,\Gamma(K_{j+1}))} \leq 2\sqrt{2}\sqrt{\kappa}\;C_{caccio}\frac{\left(\eta+2\sqrt{2}\right)}{\beta}$$

vorgenommen werden, womit man

$$\operatorname{dist}_{L_h^2(K_j\cap\Omega_h)}(u,V_j) \leq 2\sqrt{2}\sqrt{\kappa}\;C_{caccio}\frac{\left(\eta+2\sqrt{2}\right)}{\beta}\,\|u\|_{L_h^2(K_{j+1}\cap\Omega_h)}$$

für alle $u \in Z_{j+1}$ erhält.

Um nach p-maliger Anwendung der Ungleichung den Fehler ϵ zu erhalten, wird

$$\beta := 2\sqrt{2}\sqrt{\kappa}\;C_{caccio}\left(\eta+2\sqrt{2}\right)\epsilon^{-1/p}$$

gesetzt. Die resultierende Ungleichung lautet

$$\operatorname{dist}_{L_h^2(K_j\cap\Omega_h)}(u,V_j) \leq \epsilon^{1/p}\,\|u\|_{L_h^2(K_{j+1}\cap\Omega_h)}$$

für alle $u \in Z_{j+1}$.

Für alle $v_{j+1} \in Z_{j+1}$ existiert demnach eine Approximation $u_j \in V_j$ und der Fehler auf $K_j \cap \Omega_h$, gegeben durch $v_j := v_{j+1}|_{K_j\cap\Omega_h} - u_j \in Z_j$, lässt sich durch

$$\|v_j\|_{L_h^2(K_j\cap\Omega_h)} \leq \epsilon^{1/p}\,\|v_{j+1}\|_{L_h^2(K_{j+1}\cap\Omega_h)}$$

abschätzen.

Für $j = p-1$ ergibt sich daraus, dass für alle $u =: v_p \in Z_p$ eine Approximation $u_{p-1} \in Z_{p-1}$ existiert und der Fehler $v_{p-1} = v_p|_{K_{p-1}\cap\Omega_h} - u_{p-1} \in Z_{p-1}$ auf $K_{p-1}\cap\Omega_h$ kann durch

$$\|v_{p-1}\|_{L_h^2(K_{p-1}\cap\Omega_h)} \leq \epsilon^{1/p}\,\|v_p\|_{L_h^2(K_p\cap\Omega_h)}$$

abgeschätzt werden. Diese Argumentation lässt sich für $v_{p-1} \in Z_{p-1}$ und sukzessive bis zu $j = 0$ fortführen, so dass man die Existenz einer Approximation $u_0 \in Z_0$ von $v_p|_{K_0}$ mit dem Fehler $v_0 = v_1|_{K_0\cap\Omega_h} - u_0$ und der zugehörigen Abschätzung

$$\|v_0\|_{L_h^2(K_0\cap\Omega_h)} \leq \epsilon^{1/p}\,\|v_1\|_{L_h^2(K_1\cap\Omega_h)} \leq \epsilon\,\|v_p\|_{L_h^2(K_p\cap\Omega_h)} \tag{4.5}$$

erhält.

Folglich kann $u = v_p$ auf $K_0 = K_h \subset \Omega_h$ dargestellt werden als

$$u|_{K_h} = \sum_{j=0}^{p-1} u_j|_{K_h} + v_0$$

und für v_0 gilt auf K_h die Abschätzung (4.5). Der Unterraum $W \subset Z_h^L(K_h)$ kann also direkt angegeben werden als

$$W := \operatorname{span}\{u_j|_{K_h}, j = 0,\ldots,p-1\}$$

und es gilt die Abschätzung der Dimension

$$\dim W \leq kp = \left\lceil (\beta p)^2 \right\rceil p \leq p + \beta^2 p^3.$$

Durch die Wahl von $p := \lceil \log \frac{1}{\epsilon} \rceil$ erhält man $\epsilon^{-\frac{1}{p}} = \exp\left(\frac{\log \frac{1}{\epsilon}}{p}\right) \leq \exp(1)$ und mit

$$\beta = 2\sqrt{2}\sqrt{\kappa}\, C_{caccio} \left(\eta + 2\sqrt{2}\right) \epsilon^{-1/p} \leq 2\sqrt{2}\sqrt{\kappa}\, C_{caccio} \left(\eta + 2\sqrt{2}\right) \exp(1)$$

ergibt sich schließlich

$$\dim W \leq \left\lceil \log \frac{1}{\epsilon} \right\rceil + \left(2\sqrt{2}\sqrt{\kappa}\, C_{caccio} \left(\eta + 2\sqrt{2}\right) \exp(1)\right)^2 \left\lceil \log \frac{1}{\epsilon} \right\rceil^3. \qquad \blacksquare$$

4.1.3 Separable Approximation der diskreten Greenschen Funktion

Das Resultat aus Lemma 4.1.2 kann dazu genutzt werden, die Existenz einer diskreten separablen Entwicklung der diskreten Greenschen Funktion mit exponentieller Konvergenz auf zulässigen Gittern nachzuweisen. Dabei sei daran erinnert, dass die zulässige Partition so gebildet wurde, dass nur Rechteckgitter als zulässige Gitter $X_\tau, X_\sigma \subset \Omega_h$ vorkommen, welche die Zulässigkeitsbedingung (3.9) erfüllen. Definiert man, wie bereits in Abschnitt 4.1.2 vorweggenommen,

$$\delta := \frac{\operatorname{dist}_\infty(X_\tau, X_\sigma) - h}{h},$$

so gehört die diskrete Greensche Funktion zur Finite-Differenzen-Diskretisierung (2.10) für festes $x \in X_\tau$ zur Klasse der diskret L-harmonischen Funktionen auf $X_\sigma^\delta \cap \Omega_h$ bzw. für festes $y \in X_\sigma$ zu den diskret L-harmonischen Funktionen auf $X_\tau^\delta \cap \Omega_h$. Daher lässt sich Lemma 4.1.2 für die diskrete Greensche Funktion auf zulässigen Gittern verwenden, um die Existenz einer diskreten separablen Entwicklung mit exponentieller Konvergenz zu zeigen. Dieses Resultat kann analog zu [Hac09, Satz 11.3.8] angegeben werden.

Satz 4.1.3 *Seien Ω_h und $X_h, Y_h \subset \Omega_h$ Rechteckgitter mit*

$$\operatorname{diam}(Y_h) \leq \eta \left(\operatorname{dist}_\infty(X_h, Y_h) - h\right)$$

und Y_h^δ sei, wie in Konstruktion 2.3.1 beschrieben, mit $\delta := \frac{\operatorname{dist}_\infty(X_h, Y_h) - h}{h} \geq 1$ konstruiert. Die diskrete Greensche Funktion zur Problemstellung (2.10) sei mit g_h bezeichnet und es seien

$$\epsilon_0 = \exp\left(-\left\lceil \frac{1}{h\eta} \operatorname{diam}(Y_h) \right\rceil\right)$$

und c wie in Lemma 4.1.2 gegeben.

Dann existiert für jedes ϵ mit $\epsilon_0 \leq \epsilon < 1$ eine diskrete separable Entwicklung

$$g_h^k(x, y) = \sum_{l=1}^{k} u_l(x) v_l(y) \quad \text{mit } k \leq \left\lceil \log \frac{1}{\epsilon} \right\rceil + c^2 \left\lceil \log \frac{1}{\epsilon} \right\rceil^3$$

der diskreten Greenschen Funktion auf $X_h \times Y_h$ und es gilt die Fehlerabschätzung

$$\left\| g_h(x,\cdot) - g_h^k(x,\cdot) \right\|_{L_h^2(Y_h)} \leq \epsilon \left\| g_h(x,\cdot) \right\|_{L_h^2(Y_h^\delta \cap \Omega_h)} \quad \textit{für alle } x \in X_h.$$

Beweis Aus der Voraussetzung $\operatorname{diam}(Y_h) \leq \eta\,(\operatorname{dist}_\infty(X_h, Y_h) - h)$ erhält man zusammen mit der Definition von δ die Gültigkeit der Ungleichung

$$\operatorname{diam}(Y_h) \leq \eta\,(\operatorname{dist}_\infty(X_h, Y_h) - h) = \eta \operatorname{dist}_\infty(Y_h, \Gamma(Y_h^\delta)).$$

Daher sind die Voraussetzungen aus Lemma 4.1.2 für $K_h = Y_h, K_h^\delta = Y_h^\delta$ erfüllt. Unter Berücksichtigung der Definition von δ gilt $X_h \cap Y_h^\delta = \emptyset$, so dass die Funktion

$$g_x := g_h(x,\cdot)$$

für alle $x \in X_h$ diskret L-harmonisch auf $Y_h^\delta \cap \Omega_h$ ist.

Nach Lemma 4.1.2 existiert demnach ein Unterraum $W \subset Z_h^L(Y_h)$ mit der Dimension

$$k = \dim W \leq c^2 \left\lceil \log \frac{1}{\epsilon} \right\rceil^3 + \left\lceil \log \frac{1}{\epsilon} \right\rceil,$$

so dass für g_x eine Approximation $g_x^W \in W$ existiert, für die nach (4.4)

$$\left\| g_x^W - g_x \right\|_{L_h^2(Y_h)} \leq \epsilon \left\| g_x \right\|_{L_h^2(Y_h^\delta \cap \Omega_h)}$$

für alle $x \in X_h$ gilt.

Unter Verwendung einer Basis $\{v_1, \ldots, v_k\}$ von W lässt sich die Approximation durch

$$g_x^W = \sum_{l=1}^{k} u_l(x) v_l$$

mit von $x \in X_h$ abhängigen Koeffizienten $u_l(x)$ darstellen.

Demnach sind $u_l, l = 1, \ldots, k$ Gitterfunktionen auf X_h und $v_l, l = 1, \ldots, k$ Gitterfunktionen auf Y_h, so dass die diskrete separable Entwicklung durch

$$g_h^k(x,y) = \sum_{l=1}^{k} u_l(x) v_l(y)$$

angegeben werden kann und man erhält für den Fehler der Approximation

$$\left\| g_h(x,\cdot) - g_h^k(x,\cdot) \right\|_{L_h^2(Y_h)} \leq \epsilon \left\| g_h(x,\cdot) \right\|_{L_h^2(Y_h^\delta \cap \Omega_h)}$$

für alle $x \in X_h$. ∎

Bemerkung 4.1.4 *Die Aussage aus Satz 4.1.3 kann analog für den Fall*

$$\operatorname{diam}(X_h) \leq \eta \left(\operatorname{dist}_\infty(X_h, Y_h) - h\right)$$

und X_h^δ gezeigt werden. Denn unter dieser Voraussetzung gilt analog $X_h^\delta \cap Y_h = \emptyset$, so dass die Funktion

$$g_y := g_h(\cdot, y)$$

für alle $y \in Y_h$ diskret L-harmonisch auf $X_h^\delta \cap \Omega_h$ ist. Demnach ist nur die Abschätzung durch

$$\left\| g_h(\cdot, y) - g_h^k(\cdot, y) \right\|_{L_h^2(X_h)} \leq \epsilon \left\| g_h(\cdot, y) \right\|_{L_h^2(X_h^\delta \cap \Omega_h)}$$

für alle $y \in Y_h$ zu ersetzen.

4.2 Hauptresultat

Wie in Abschnitt 3.2.4 beschrieben, kann eine Matrix (gut) durch eine $\mathcal{H}$-Matrix approximiert werden, wenn die Approximation der entsprechenden Gitterfunktion auf zulässigen Gittern durch eine diskrete separable Entwicklung mit exponentieller Konvergenz möglich ist. Dieses Resultat wurde in Satz 4.1.3 für die diskrete Greensche Funktion auf zulässigen Gittern gezeigt. Nutzt man den Zusammenhang zwischen der diskreten Greenschen Funktion und der Inversen einer Finite-Differenzen-Matrix, kann ein Resultat zu deren Approximation mittels einer $\mathcal{H}$-Matrix angegeben werden.

Satz 4.2.1 *Sei $L_h \in \mathbb{R}^{I \times I}$ die Finite-Differenzen-Matrix zu der Problemstellung* (2.10) *auf einem Rechteckgitter Ω_h mit konstanter Schrittweite h und $P \subset T(I \times I)$ eine nach den Ausführungen in Abschnitt 3.2.1 mit Hilfe der Zulässigkeitsbedingung* (3.9) *konstruierte zulässige Partition. Außerdem seien c aus Lemma 4.1.2 und*

$$\epsilon_{0,max} := \max_{\tau \times \sigma \in P^+} \exp\left(- \left\lceil \frac{1}{h\eta} \min\{\operatorname{diam}(X_\tau), \operatorname{diam}(X_\sigma)\} \right\rceil \right)$$

gegeben.

Dann existiert für jedes ϵ mit $\epsilon_{0,max} \leq \epsilon < 1$ eine Matrix $L_{\mathcal{H},inv} \in \mathcal{H}(k, P)$, für die

$$\left\| L_h^{-1} - L_{\mathcal{H},inv} \right\|_F \leq \epsilon \sqrt{|P^+|} \left\| L_h^{-1} \right\|_F$$

mit

$$k \leq \left\lceil \log \frac{1}{\epsilon} \right\rceil + c^2 \left\lceil \log \frac{1}{\epsilon} \right\rceil^3$$

gilt.

Beweis Für beliebiges $b = \tau \times \sigma \in P^+$ ist nach Konstruktion 3.1.4 die Zulässigkeitsbedingung (3.9) erfüllt und es sind zwei Fälle zu unterscheiden: Entweder gilt $\min\{\operatorname{diam}(X_\tau), \operatorname{diam}(X_\sigma)\} = \operatorname{diam}(X_\sigma)$, so dass Satz 4.1.3 für die zulässigen Gitter $X_h = X_\tau$ und $Y_h = X_\sigma$ angewendet werden kann. Oder es wird alternativ für $\operatorname{diam}(X_\tau)$

das Minimum angenommen, so dass man nach Bemerkung 4.1.4 ein analoges Resultat für $X_h = X_\sigma$ und $Y_h = X_\tau$ verwenden kann. In beiden Fällen gilt

$$\epsilon_0 \leq \epsilon_{0,max}$$

für alle $b \in P^+$. Demnach erhält man für den ersten Fall nach Satz 4.1.3, dass für jedes ϵ mit $\epsilon_{0,max} \leq \epsilon < 1$ eine diskrete separable Entwicklung g_h^k der diskreten Greenschen Funktion g_h auf zulässigen Gittern $X_\tau \times X_\sigma$ mit

$$k \leq \left\lceil \log \frac{1}{\epsilon} \right\rceil + c^2 \left\lceil \log \frac{1}{\epsilon} \right\rceil^3$$

existiert und

$$\left\| g_h(x,\cdot) - g_h^k(x,\cdot) \right\|_{L_h^2(X_\sigma)} \leq \epsilon \left\| g_h(x,\cdot) \right\|_{L_h^2(X_\sigma^\delta \cap \Omega_h)} \tag{4.6}$$

für alle $x \in X_\tau$ gilt.

Im zweiten Fall erhält man nach Bemerkung 4.1.4 ein analoges Resultat mit der Abschätzung

$$\left\| g_h(\cdot,y) - g_h^k(\cdot,y) \right\|_{L_h^2(X_\tau)} \leq \epsilon \left\| g_h(\cdot,y) \right\|_{L_h^2(X_\tau^\delta \cap \Omega_h)} \tag{4.7}$$

für alle $y \in X_\sigma$.

Definiert man die Matrix $\mathcal{G}_h \in \mathbb{R}^{I \times I}$ durch

$$(\mathcal{G}_h)_{ij} = g_h(x^i, y^j),\ x^i, y^j \in \Omega_h,$$

so gilt nach den Ausführungen in Abschnitt 2.1.2 $L_h^{-1}|_b = h^2 \mathcal{G}_h|_b$ für alle $b \in P$.

Ersetzt man in den Blöcken $b \in P^+$ die Matrix $\mathcal{G}_h|_b$ durch die Approximation $\mathcal{G}_h^k|_b$, deren Einträge durch

$$\left(\mathcal{G}_h^k\right)_{ij} = g_h^k(x^i, y^j), x^i \in X_\tau, y^j \in X_\sigma$$

gegeben sind, so kann der lokale Fehler in der Frobenius-Norm für $\operatorname{diam}(X_\sigma) \leq \operatorname{diam}(X_\tau)$ nach (4.6) durch

$$\begin{aligned}
\left\| h^2 \mathcal{G}_h|_b - h^2 \mathcal{G}_h^k|_b \right\|_F^2 &= h^2 \sum_{x \in X_\tau} \sum_{y \in X_\sigma} \left| g_h(x,y) - g_h^k(x,y) \right|^2 \\
&\leq \epsilon^2 h^2 \sum_{x \in X_\tau} \sum_{y \in X_\sigma^\delta \cap \Omega_h} |g_h(x,y)|^2 \\
&\leq \epsilon^2 \left\| h^2 \mathcal{G}_h \right\|_F^2 = \epsilon^2 \left\| L_h^{-1} \right\|_F^2
\end{aligned}$$

abgeschätzt werden. Für $\operatorname{diam}(X_\sigma) \geq \operatorname{diam}(X_\tau)$ erhält man nach (4.7) analog

$$\begin{aligned}
\left\| h^2 \mathcal{G}_h|_b - h^2 \mathcal{G}_h^k|_b \right\|_F^2 &= h^2 \sum_{x \in X_\tau} \sum_{y \in X_\sigma} \left| g_h(x,y) - g_h^k(x,y) \right|^2 \\
&\leq \epsilon^2 h^2 \sum_{x \in X_\tau^\delta \cap \Omega_h} \sum_{y \in X_\sigma} |g_h(x,y)|^2 \\
&\leq \epsilon^2 \left\| h^2 \mathcal{G}_h \right\|_F^2 = \epsilon^2 \left\| L_h^{-1} \right\|_F^2 .
\end{aligned}$$

Außerdem kann nach den Ausführungen in Abschnitt 3.2.4 für den Matrixblock $\mathcal{G}_h^k|_b$ die Darstellung

$$\mathcal{G}_h^k|_b = AB^T$$

mit $A \in \mathbb{R}^{\tau\times\{1,\dots,k\}}, B \in \mathbb{R}^{\sigma\times\{1,\dots,k\}}$ als Rang-k-Matrix angegeben werden. Mit der Definition

$$L_{\mathcal{H},inv} := \begin{cases} L_h^{-1}|_b & b \in P^- \\ h^2\mathcal{G}_h^k|_b & b \in P^+ \end{cases}$$

gilt $L_{\mathcal{H},inv} \in \mathcal{H}(k,P)$ und für die Blöcke $b \in P^+$ kann der Approximationsfehler durch

$$\left\|L_h^{-1}|_b - L_{\mathcal{H},inv}|_b\right\|_F^2 \leq \epsilon^2 \left\|L_h^{-1}\right\|_F^2$$

abgeschätzt werden. Da in den Blöcken $b \in P^-$ nach Definition von $L_{\mathcal{H},inv}$ keine Approximation durchgeführt wird, erhält man für den globalen Fehler

$$\begin{aligned}\left\|L_h^{-1} - L_{\mathcal{H},inv}\right\|_F^2 &= \sum_{b\in P^+} \left\|L_h^{-1}|_b - L_{\mathcal{H},inv}|_b\right\|_F^2 \\ &\leq \left|P^+\right| \epsilon^2 \left\|L_h^{-1}\right\|_F^2 .\end{aligned}$$

∎

Eine Abschätzung der Größenordnung von $|P^+|$ kann wie in Abschnitt 3.2.3 unter Verwendung der Konstanten C_{sp} erfolgen.

Das Resultat aus Satz 4.2.1 ermöglicht es, eine Aussage zur Existenz einer $\mathcal{H}$-Matrix Approximation für den zweidimensionalen Fall von Finite-Differenzen-Matrizen zum Modellproblem zu treffen. Die exponentielle Konvergenz erhält man aus der Umkehrung von

$$k \leq \left\lceil \log \frac{1}{\epsilon} \right\rceil + c^2 \left\lceil \log \frac{1}{\epsilon} \right\rceil^3$$

zu

$$\epsilon \approx \exp\left(-c^{-\frac{2}{3}} k^{\frac{1}{3}}\right).$$

Die Größen in der Fehlerabschätzung sind von $c = 2\sqrt{2}\sqrt{\kappa}\ C_{caccio}\left(\eta + 2\sqrt{2}\right)\exp(1)$ abhängig, wobei

$$C_{caccio} = \begin{cases} 1 & \text{für konstante Koeffizienten } a,d \in \mathbb{R}^+ \\ 3\sqrt{2} & \text{für variable Koeffizienten } a(x_i,y_j), d(x_i,y_j) \in \mathcal{D}_h\left(\overline{\Omega}_h\right), a,d > 0 \end{cases}$$

gemäß der Abschätzungen in Satz 2.3.2 und Satz 2.3.3 gilt. Im Vergleich zum Ergebnis für Finite-Element-Matrizen, bei dem man $c_{FEM} = \frac{4\sqrt{2}}{\pi}\sqrt{\kappa}(\eta+2)\exp(1)$ erhält, fällt die Konstante c in Satz 4.2.1 deutlich größer aus. Dies ist insbesondere auf die größeren Konstanten in der diskreten Poincaré- und in der diskreten Cacciopoli-Ungleichung und auf die speziellen Gegebenheiten für Gitterfunktionen im diskreten Fall zurückzuführen. Die Größe c im theoretischen Resultat lässt vermuten, dass der zur Approximation verwendete Rang k sehr groß gewählt werden muss, um akzeptable Approximationsfehler zu erhalten. Dies lässt sich bei den numerischen Tests zum Modellproblem in Kapitel 6

in dieser Form nicht beobachten, da bereits bei Verwendung niedriger Ränge sehr kleine Fehler auftreten.

Sowohl beim Ergebnis für Finite-Element- als auch für Finite-Differenzen-Matrizen hängt c jedoch von dem Verhältnis

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}$$

des größten zum kleinsten Koeffizienten des Differenzenoperators L ab. Der Einfluss von κ auf den Fehlerverlauf steht daher im Mittelpunkt der numerischen Tests in Kapitel 6.

Dort wird zusätzlich überprüft, ob die zur Durchführung des Beweises von Lemma 4.1.2 erforderliche Bedingung $\epsilon_0 \leq \epsilon$ bei den numerischen Tests ebenfalls zu beobachten ist. Aufgrund dieser Voraussetzung konnte nur exponentielle Konvergenz für $\epsilon \geq \epsilon_0$ gezeigt werden. Bei der Durchführung der numerischen Tests ergeben sich jedoch für $k \to \infty$ exponentiell fallende Fehler.

Als Erweiterung des zweidimensionalen Resultats wird in Satz 5.4.1 ein Ergebnis für den dreidimensionalen Fall von Quadergittern angegeben. Dieses ergibt sich direkt aus der Erweiterung der Resultate für Rechteckgitter.

4.3 Ausblick: Schurkomplemente, LU-Faktoren

Das Resultat zur Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen von Finite-Differenzen-Matrizen zum Modellproblem kann als Grundlage für weitere Untersuchungen verwendet werden. Es lässt sich insbesondere für den Nachweis der Existenz einer $\mathcal{H}$-Matrix Approximation der Faktoren einer LU-Zerlegung nutzen.

Die Berechnung einer $\mathcal{H}$-LU-Zerlegung in fast linearer Komplexität ist zur direkten Lösung eines Gleichungssystems oder zum Einsatz als Präkonditionierer zur Beschleunigung eines iterativen Verfahrens neben der Verwendung der $\mathcal{H}$-Inversen ebenfalls von großem Interesse. Genau wie bei der Inversen ist auch in diesem Fall vor der Berechnung der $\mathcal{H}$-LU-Zerlegung zu untersuchen, ob eine $\mathcal{H}$-Matrix Approximation der LU-Faktoren existiert.

Zu dieser Fragestellung gibt es bereits mehrere Ergebnisse, unter anderem im Zusammenhang mit Finite-Element-Matrizen in [Beb07] oder auch für wohlkonditionierte, dünnbesetzte Matrizen in [BF11] oder [GKLB08]. Diese Resultate beruhen im Wesentlichen auf der Existenz einer $\mathcal{H}$-Matrix Approximation von verallgemeinerten Schurkomplementen zu der entsprechenden Matrix. Die übrigen Überlegungen zum Beweis der Existenz von $\mathcal{H}$-Matrix Approximationen der LU-Faktoren sind in diesen Fällen rein algebraischer Natur und lassen sich damit unabhängig von der zugrunde liegenden Problemstellung nutzen. Daher ließe sich dieser Teil für den Fall von Finite-Differenzen-Matrizen übernehmen, wenn es gelingt, die Existenz von $\mathcal{H}$-Matrix Approximationen der verallgemeinerten Schurkomplemente zu Finite-Differenzen-Matrizen nachzuweisen.

Der Beweis dieses Resultats beruht in den angegebenen Arbeiten wiederum auf einem entsprechenden Resultat zur Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen. Da in Satz 4.2.1 ein solches Ergebnis für Finite-Differenzen-Matrizen zum Modellproblem

gezeigt werden konnte, ließe sich dieses ebenfalls zum Nachweis der Existenz von verallgemeinerten Schurkomplementen zu Finite-Differenzen-Matrizen nutzen. Auf dieser Grundlage könnte im Anschluss analog zu der Vorgehensweise in den genannten Arbeiten der Beweis zur Existenz einer $\mathcal{H}$-Matrix Approximation der LU-Faktoren von Finite-Differenzen-Matrizen aufgebaut werden.

Diese grundlegenden Überlegungen werden an dieser Stelle nicht weiter ausgeführt, sie lassen jedoch vermuten, dass aufgrund des Resultats zur Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen auch die Verwendung der $\mathcal{H}$-LU-Zerlegung zur Lösung der Gleichungssysteme bei Finite-Differenzen-Diskretisierungen Erfolg versprechend ist.

5 Erweiterung für den dreidimensionalen Fall

In Kapitel 4 konnte die Existenz einer $\mathcal{H}$-Matrix Approximation für die Inverse von Finite-Differenzen-Matrizen zum zweidimensionalen Modellproblem gezeigt werden. Dazu wurde der Ansatz aus [Hac09] zu Finite-Element-Matrizen verwendet, um eine möglichst einfache Übertragung der Resultate für verwandte Problemstellungen zu ermöglichen. In diesem Kapitel wird die Erweiterung der Ergebnisse für den dreidimensionalen Fall vorgenommen. Diese ist insbesondere im Hinblick auf das Gleichungssystem im Modell METRAS von Interesse, da dieses im Allgemeinen aus der Diskretisierung einer dreidimensionalen Problemstellung hervorgeht.

Analog zum Modellproblem für den zweidimensionalen wird auch eines für den dreidimensionalen Fall angegeben, auf dessen Grundlage die theoretischen Untersuchungen erfolgen. Die in Kapitel 2 eingeführten Rechteckgitter lassen sich durch Erweiterung um eine z-Komponente direkt zu Quadergittern der Form

$$\Omega_h = \{(x_i, y_j, z_p) \in h\mathbb{Z}^3 : 1 \leq i \leq n, 1 \leq j \leq m, 1 \leq p \leq r\}$$

mit $n, m, r \in \mathbb{N}$ erweitern. Das diskrete Randwertproblem sei weiterhin durch (2.10) gegeben. Allerdings sei Ω_h durch ein Quadergitter und der zu betrachtende Differenzenoperator durch

$$Lu = (au_x)_{\bar{x}} + (du_y)_{\bar{y}} + (gu_z)_{\bar{z}} \tag{5.1}$$

mit ortsabhängigen Gitterfunktionen $a, d, g \in \mathcal{D}_h\left(\overline{\Omega}_h\right)$ und $a, d, g > 0$ gegeben. Daraus resultiert ein Differenzenstern mit sieben Einträgen, so dass man nach Elimination der Randpunkte eine dünnbesetzte Diskretisierungsmatrix $L_h \in \mathbb{R}^{I \times I}$ mit $I = \{1, \ldots, nmr\}$ erhält. Auch bei diesem Modellproblem ergibt sich für die konstanten Koeffizienten $a = d = g = 1$ der Standard 7-Punkte-Differenzenstern aus der direkten Diskretisierung des Laplace-Operators. Wie im zweidimensionalen Fall ist die Matrix für diese Problemstellung symmetrisch und positiv definit. Die diskrete Greensche Funktion kann analog zu den Ausführungen in Abschnitt 2.1.2 eingeführt werden, so dass erneut der bekannte Zusammenhang zur Inversen der Diskretisierungsmatrix besteht.

Die Konstruktion einer zulässigen Partition zu Finite-Differenzen-Matrizen in Abschnitt 3.2.1 ist so allgemein angegeben, dass sie sich direkt auf den Fall von Quadergittern übertragen lässt. Insbesondere sind auch im dreidimensionalen Fall nach dieser Konstruktion alle zulässigen Gitter durch Quadergitter gegeben. Dies erlaubt die Beschränkung der entsprechenden Resultate auf diese Gitterstrukturen. Die Zulässigkeitsbedingung (3.9) kann unverändert übernommen werden, da sie unabhängig von der Dimension der Problemstellung formuliert wurde. Die Überlegungen zur $\mathcal{H}$-Matrix Eigenschaft der Diskretisierungsmatrix aus Abschnitt 3.2.2 können analog erfolgen, so dass

man auch für die Diskretisierungsmatrix zum dreidimensionalen Modellproblem erneut $L_h \in \mathcal{H}(0, P)$ erhält.

Die Resultate zur Approximation von diskret L-harmonischen Funktionen bzw. der diskreten Greenschen Funktion mittels einer diskreten separablen Entwicklung mit exponentieller Konvergenz in Kapitel 4 lassen sich leicht auf den dreidimensionalen Fall von Quadergittern übertragen. Als wesentliche Hilfsmittel werden analog die diskrete Poincaré- und die diskrete Cacciopoli-Ungleichung für den Fall von Quadergittern benötigt. Diese werden als Erweiterung der Resultate für Rechteckgitter aus den Abschnitten 2.2 und 2.3 in den Abschnitten 5.1 und 5.2 angegeben. Die übrigen Überlegungen lassen sich ebenfalls problemlos übertragen.

Alle Resultate, die zum Beweis des Hauptresultats für den dreidimensionalen Fall (Satz 5.4.1) erforderlich sind, werden in Abschnitt 5.3 angegeben. Bis auf die Beweise der diskreten Poincaré- und der diskreten Cacciopoli-Ungleichung für den dreidimensionalen Fall, die aufwendiger sind, ist für die Erweiterung der übrigen Resultate lediglich die Anpassung von Konstanten in den verschiedenen Abschätzungen erforderlich. Die Ausführungen in den entsprechenden Beweisen beschränken sich daher auf die gegenüber den zweidimensionalen Resultaten veränderten Abschnitte, die übrigen Beweisschritte können analog erfolgen.

5.1 Diskrete Poincaré-Ungleichung

Zur Formulierung der diskreten Poincaré-Ungleichung im dreidimensionalen Fall kann der gleiche methodische Ansatz verwendet werden, der für die Erweiterung von dem eindimensionalen auf den zweidimensionalen Fall entwickelt wurde. Die Ergebnisse aus dem folgenden Abschnitt werden daher analog zu den Ergebnissen aus dem zweidimensionalen Fall angegeben.

Lemma 5.1.1 *Sei* $\Omega_h := \{(x_i, y_j, z_p) \in h\mathbb{Z}^3 : 1 \le i \le n, 1 \le j \le m, 1 \le p \le r\}$, $n, m, r \in \mathbb{N}$ *ein* $n \times m \times r$ *Quadergitter mit konstanter Schrittweite* h *und* $u \in \mathcal{D}_h(\Omega_h)$ *eine Gitterfunktion auf* Ω_h *mit dem Mittelwert* $\bar{u}$*. Dann gilt*

$$u(x_i, y_l, z_p) - \bar{u} = \frac{h}{nmr} \sum_{j=1}^{n} \sum_{k=1}^{m} \sum_{\kappa=1}^{n+m+r-3} \tilde{c}(\kappa, i, l, p) \partial_\kappa^{j,l,p,k} u$$

für alle $(x_i, y_l, z_p) \in \Omega_h$ *mit*

$$\tilde{c}(\kappa, i, l, p) = \begin{cases} \frac{r}{n} c_1(\kappa, i) & 1 \le \kappa \le n-1 \\ \frac{r}{m} c_2(\kappa - n + 1, l) & n \le \kappa \le n+m-2 \\ c_3(\kappa - n - m + 2, p) & n-1+m \le \kappa \le n+m+r-3, \end{cases}$$

$$\partial_\kappa^{j,l,p,k} u = \begin{cases} u_x(\kappa h, y_l, z_p) & 1 \le \kappa \le n-1 \\ u_y(x_j, (\kappa - n + 1)h, z_p) & n \le \kappa \le n+m-2 \\ u_z(x_j, y_k, (\kappa - n - m + 2)h) & n-1+m \le \kappa \le n+m+r-3, \end{cases}$$

$$c_1(\kappa,i)=\begin{cases}\kappa & \kappa\le i-1\\ -(n-\kappa) & \kappa\ge i,\end{cases} \tag{5.2}$$

$$c_2(\kappa,l)=\begin{cases}\kappa & \kappa\le l-1\\ -(m-\kappa) & \kappa\ge l\end{cases} \tag{5.3}$$

und

$$c_3(\kappa,p)=\begin{cases}\kappa & \kappa\le p-1\\ -(r-\kappa) & \kappa\ge p.\end{cases} \tag{5.4}$$

Außerdem lässt sich abschätzen:

$$|u(x_i,y_l,z_p)-\bar{u}|^2\le\frac{h^2}{(nmr)^2}nm\left(\sum_{\kappa=1}^{n+m+r-3}\tilde{c}^2(\kappa,i,l,p)\right)\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{\kappa=1}^{n+m+r-3}\left|\partial_\kappa^{j,l,p,k}u\right|^2.$$

Beweis Der Beweis ergibt sich wie im zweidimensionalen Fall aus der Darstellung

$$\begin{aligned}u(x_i,y_l,z_p)-u(x_j,y_k,z_q)&=(u(x_i,y_l,z_p)-u(x_j,y_l,z_p))+(u(x_j,y_l,z_p)-u(x_j,y_k,z_p))\\&\quad+(u(x_j,y_k,z_p)-u(x_j,y_k,z_q))\end{aligned}$$

und der Verwendung von Lemma 2.2.1. Demnach erhält man für die drei Summanden

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}(u(x_i,y_l,z_p)-u(x_j,y_l,z_p))&=\frac{h}{n}\sum_{\kappa=1}^{n-1}c_1(\kappa,i)u_x(\kappa h,y_l,z_p),\\ \frac{1}{m}\sum_{k=1}^{m}(u(x_j,y_l,z_p)-u(x_j,y_k,z_p))&=\frac{h}{m}\sum_{\kappa=1}^{m-1}c_2(\kappa,l)u_y(x_j,\kappa h,z_p)\end{aligned}$$

und

$$\frac{1}{r}\sum_{q=1}^{r}(u(x_j,y_k,z_p)-u(x_j,y_k,z_q))=\frac{h}{r}\sum_{\kappa=1}^{r-1}c_3(\kappa,p)u_z(x_j,y_k,\kappa h).$$

Analog zum zweidimensionalen Fall gilt damit

$$\begin{aligned}u(x_i,y_l,z_p)-\bar{u}=\frac{h}{nmr}\sum_{j=1}^{n}\sum_{k=1}^{m}\Bigg(&\sum_{\kappa=1}^{n-1}\frac{r}{n}c_1(\kappa,i)u_x(\kappa h,y_l,z_p)\\&+\sum_{\kappa=1}^{m-1}\frac{r}{m}c_2(\kappa,l)u_y(x_j,\kappa h,z_p)+\sum_{\kappa=1}^{r-1}c_3(\kappa,p)u_z(x_j,y_k,\kappa h)\Bigg),\end{aligned}$$

woraus der erste Teil der Behauptung folgt. Durch dreimalige Anwendung der Ungleichung von Cauchy-Schwarz erhält man den zweiten Teil der Behauptung. ∎

Analog zum Beweis der zweidimensionalen diskreten Poincaré-Ungleichung werden folgende Lemmata benötigt, um später eine Abschätzung der Konstanten C_h zu ermöglichen:

Lemma 5.1.2 *Sei $\tilde{c}$ wie in Lemma 5.1.1 mit $n, m, r \in \mathbb{N}$ gegeben, dann gilt*

$$\sum_{i=1}^{n}\sum_{l=1}^{m}\sum_{p=1}^{r}\sum_{\kappa=1}^{n+m+r-3} \tilde{c}^2(\kappa,i,l,p) = \frac{r^2}{6}\left(mr(n^2-1)+nr(m^2-1)+nm(r^2-1)\right).$$

Beweis Durch direkte Berechnung erhält man

$$\begin{aligned}\sum_{i=1}^{n}\sum_{l=1}^{m}\sum_{p=1}^{r}\sum_{\kappa=1}^{n+m+r-3} \tilde{c}^2(\kappa,i,l,p) &= \sum_{i=1}^{n}\sum_{l=1}^{m}\sum_{p=1}^{r}\left(\sum_{\kappa=1}^{n-1}\frac{r^2}{n^2}c_1^2(\kappa,i)\right.\\ &\quad\left.+\sum_{\kappa=1}^{m-1}\frac{r^2}{m^2}c_2^2(\kappa,l)+\sum_{\kappa=1}^{r-1}c_3^2(\kappa,p)\right)\\ &= mr\left(\frac{r^2}{n^2}\sum_{i=1}^{n}\sum_{\kappa=1}^{n-1}c_1^2(\kappa,i)\right)+nr\left(\sum_{l=1}^{m}\sum_{\kappa=1}^{m-1}\frac{r^2}{m^2}c_2^2(\kappa,l)\right)\\ &\quad+nm\left(\sum_{p=1}^{r}\sum_{\kappa=1}^{r-1}c_3^2(\kappa,p)\right),\end{aligned}$$

woraus die Behauptung folgt. ∎

Lemma 5.1.3 *Sei $\tilde{c}$ wie in Lemma 5.1.1 mit $n, m, r \in \mathbb{N}$ gegeben, dann gilt*

$$\begin{aligned}\max_{p\in[1,r]}\left(\sum_{\kappa=1}^{n+m+r-3}\tilde{c}^2(\kappa,i,l,p)\right) &= \frac{1}{6}(2r^3-3r^2+r)+\sum_{\kappa=1}^{n-1}\frac{r^2}{n^2}c_1^2(\kappa,i)\\ &\quad+\sum_{\kappa=1}^{m-1}\frac{r^2}{m^2}c_2^2(\kappa,l), \qquad (5.5)\\ \max_{l\in[1,m]}\max_{p\in[1,r]}\left(\sum_{\kappa=1}^{n+m+r-3}\tilde{c}^2(\kappa,i,l,p)\right) &= \frac{r^2}{m^2}\frac{1}{6}(2m^3-3m^2+m)+\frac{1}{6}(2r^3-3r^2+r)\\ &\quad+\sum_{\kappa=1}^{n-1}\frac{r^2}{n^2}c_1^2(\kappa,i), \qquad (5.6)\\ \max_{p\in[1,r]}\sum_{l=1}^{m}\left(\sum_{\kappa=1}^{n+m+r-3}\tilde{c}^2(\kappa,i,l,p)\right) &= m\left(\frac{1}{6}(2r^3-3r^2+r)+\sum_{\kappa=1}^{n-1}\frac{r^2}{n^2}c_1^2(\kappa,i)\right)\\ &\quad+\frac{r^2}{6}(m^2-1). \qquad (5.7)\end{aligned}$$

Beweis Nach der Definition von $\tilde{c}$ erhält man

$$\left(\sum_{\kappa=1}^{n+m+r-3} \tilde{c}^2(\kappa,i,l,p)\right) = \sum_{\kappa=1}^{n-1} \frac{r^2}{n^2} c_1^2(\kappa,i) + \sum_{\kappa=1}^{m-1} \frac{r^2}{m^2} c_2^2(\kappa,l) + \sum_{\kappa=1}^{r-1} c_3^2(\kappa,p).$$

Die dritte Summe dieses Ausdrucks kann geschrieben werden als

$$\begin{aligned}\sum_{\kappa=1}^{r-1} c_3^2(\kappa,p) &= \sum_{\kappa=1}^{p-1} \kappa^2 + \sum_{\kappa=p}^{r-1} (r-\kappa)^2 \\ &= p(pr - r^2 - r) + \frac{1}{6}(2r^3 + 3r^2 + r) =: \gamma(p),\end{aligned}$$

so dass man

$$\max_{p\in[1,r]} \sum_{\kappa=1}^{r-1} c_3^2(\kappa,p) = \gamma(1) = \gamma(r) = \frac{1}{6}(2r^3 - 3r^2 + r) \tag{5.8}$$

erhält. Daraus folgt (5.5) und es kann zum Beweis von (5.6) genutzt werden, indem auf analoge Weise die Darstellung

$$\sum_{\kappa=1}^{m-1} \frac{r^2}{m^2} c_2^2(\kappa,l) = \frac{r^2}{m^2}\left(l(lm - m^2 - m) + \frac{1}{6}(2m^3 + 3m^2 + m)\right) =: \alpha(l) \tag{5.9}$$

hergeleitet wird, aus der wiederum

$$\max_{l\in[1,m]} \sum_{\kappa=1}^{m-1} \frac{r^2}{m^2} c_2^2(\kappa,l) = \alpha(1) = \alpha(m) = \frac{r^2}{m^2}\frac{1}{6}(2m^3 - 3m^2 + m)$$

folgt. Zusammengefasst ergibt dies die zweite Behauptung (5.6). Die dritte Behauptung (5.7) erhält man aus der Kombination von (5.8) und (5.9) zusammen mit

$$\sum_{l=1}^{m}\sum_{\kappa=1}^{m-1} \frac{r^2}{m^2} c_2^2(\kappa,l) = \frac{r^2}{m^2}\frac{1}{6} m^2(m^2-1) = \frac{r^2}{6}(m^2-1). \tag{5.10}$$

■

Lemma 5.1.4 *Es gelten die Ungleichungen*

$$n\left(2m - 3 + \frac{1}{m}\right) + n\left(2r - 3 + \frac{1}{r}\right) + (n^2-1) \le 3\left((n-1)^2 + (m-1)^2 + (r-1)^2\right),$$

$$m\left(2r - 3 + \frac{1}{r}\right) + m\left(n - \frac{1}{n}\right) + (m^2-1) \le 3\left((n-1)^2 + (m-1)^2 + (r-1)^2\right)$$

und

$$r\left(n - \frac{1}{n}\right) + r\left(m - \frac{1}{m}\right) + (r^2-1) \le 3\left((n-1)^2 + (m-1)^2 + (r-1)^2\right)$$

für alle $n, m, r \in \mathbb{N}$ *mit* $2 \le n \le m \le r$.

Beweis Die Behauptungen lassen sich mittels vollständiger Induktion über r für $2 \le n \le m \le r$ zeigen. ■

Diskrete Poincaré-Ungleichung für Gitterfunktionen mit Mittelwert Null im dreidimensionalen Fall

Mit Hilfe der vorangegangenen Lemmata lässt sich, erneut analog zum zweidimensionalen Fall, die dreidimensionale diskrete Poincaré-Ungleichung für Gitterfunktionen auf Quadergittern zeigen:

Satz 5.1.5 *Sei $\Omega_h \subset h\mathbb{Z}^3$ ein Quadergitter mit konstanter Schrittweite h. Dann gilt*

$$\|u - \bar{u}\|_{L_h^2(\Omega_h)} \leq C_h \|\nabla_h u\|_{L_h^2(\Omega_h)} \quad \textit{für alle } u \in \mathcal{D}_h(\Omega_h) \tag{5.11}$$

mit der Konstanten

$$C_h = \frac{\operatorname{diam}(\Omega_h)}{\sqrt{2}}.$$

Beweis Das Quadergitter sei ohne Einschränkung durch ein $n \times m \times r$ Quadergitter $\Omega_h := \{(x_i, y_j, z_p) \in h\mathbb{Z}^3 : 1 \leq i \leq n, 1 \leq j \leq m, 1 \leq p \leq r\}$ mit $n, m, r \in \mathbb{N}$ gegeben und es gelte $n \leq m \leq r$. Für den ein- und zweidimensionalen Fall gilt die Behauptung nach den Ergebnissen aus Satz 2.2.2 und Satz 2.2.7, daher kann im Folgenden $n, m, r \geq 2$ angenommen werden.

Nach dem Resultat aus Lemma 5.1.1 gilt für Gitterfunktionen $u \in \mathcal{D}_h(\Omega_h)$ die Abschätzung

$$|u(x_i, y_l, z_p) - \bar{u}|^2 \leq \frac{h^2}{(nmr)^2} nm \left(\sum_{\kappa=1}^{n+m+r-3} \tilde{c}^2(\kappa, i, l, p) \right) \sum_{j=1}^{n} \sum_{k=1}^{m} \sum_{\kappa=1}^{n+m+r-3} \left| \partial_\kappa^{j,l,p,k} u \right|^2 \tag{5.12}$$

für alle $(x_i, y_l, z_p) \in \Omega_h$.

Summiert man (5.12) über i, l und p und multipliziert mit h^3, erhält man die Abschätzung

$$\begin{aligned}
\|u - \bar{u}\|_{L_h^2(\Omega_h)}^2 &\leq h^3 \frac{h^2}{nmr^2} \sum_{i=1}^{n} \sum_{l=1}^{m} \sum_{p=1}^{r} \left(\sum_{\kappa=1}^{n+m+r-3} \tilde{c}^2(\kappa, i, l, p) \right) \sum_{j=1}^{n} \sum_{k=1}^{m} \sum_{\kappa=1}^{n+m+r-3} \left| \partial_\kappa^{j,l,p,k} u \right|^2 \\
&= h^3 \frac{h^2}{nmr^2} \sum_{i=1}^{n} \sum_{l=1}^{m} \sum_{p=1}^{r} \left(\sum_{\kappa=1}^{n+m+r-3} \tilde{c}^2(\kappa, i, l, p) \right) \sum_{j=1}^{n} \sum_{k=1}^{m} \sum_{\kappa=1}^{n-1} |u_x(\kappa h, y_l, z_p)|^2 \\
&+ h^3 \frac{h^2}{nmr^2} \sum_{i=1}^{n} \sum_{l=1}^{m} \sum_{p=1}^{r} \left(\sum_{\kappa=1}^{n+m+r-3} \tilde{c}^2(\kappa, i, l, p) \right) \sum_{j=1}^{n} \sum_{k=1}^{m} \sum_{\kappa=1}^{m-1} |u_y(x_j, \kappa h, z_p)|^2 \\
&+ h^3 \frac{h^2}{nmr^2} \sum_{i=1}^{n} \sum_{l=1}^{m} \sum_{p=1}^{r} \left(\sum_{\kappa=1}^{n+m+r-3} \tilde{c}^2(\kappa, i, l, p) \right) \sum_{j=1}^{n} \sum_{k=1}^{m} \sum_{\kappa=1}^{r-1} |u_z(x_j, y_k, \kappa h)|^2 .
\end{aligned}$$

Unter Verwendung von Lemma 5.1.3 und Lemma 5.1.2 erhält man daraus

$$\begin{aligned}
\|u-\bar{u}\|^2_{L^2_h(\Omega_h)} &\le h^3\frac{h^2}{r^2}\sum_{i=1}^{n}\max_{l\in[1,m]}\max_{p\in[1,r]}\left(\sum_{\kappa=1}^{n+m+r-3}\tilde{c}^2(\kappa,i,l,p)\right)\\
&\qquad\cdot\sum_{\kappa=1}^{n-1}\sum_{l=1}^{m}\sum_{p=1}^{r}|u_x(\kappa h,y_l,z_p)|^2\\
&\quad+h^3\frac{h^2}{nr^2}\sum_{i=1}^{n}\sum_{l=1}^{m}\max_{p\in[1,r]}\left(\sum_{\kappa=1}^{n+m+r-3}\tilde{c}^2(\kappa,i,l,p)\right)\sum_{j=1}^{n}\sum_{\kappa=1}^{m-1}\sum_{p=1}^{r}|u_y(x_j,\kappa h,z_p)|^2\\
&\quad+h^3\frac{h^2}{nmr^2}\frac{r^2}{6}\left(mr(n^2-1)+nr(m^2-1)+nm(r^2-1)\right)\\
&\qquad\cdot\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{\kappa=1}^{r-1}|u_z(x_j,y_k,\kappa h)|^2\\
&=h^3\frac{h^2}{6}\left(n\left(2m-3+\frac{1}{m}\right)+n\left(2r-3+\frac{1}{r}\right)+n^2-1\right)\\
&\qquad\cdot\sum_{\kappa=1}^{n-1}\sum_{l=1}^{m}\sum_{p=1}^{r}|u_x(\kappa h,y_l,z_p)|^2\\
&\quad+h^3\frac{h^2}{6}\left(m\left(2r-3+\frac{1}{r}\right)+m\left(n-\frac{1}{n}\right)+m^2-1\right)\\
&\qquad\cdot\sum_{j=1}^{n}\sum_{\kappa=1}^{m-1}\sum_{p=1}^{r}|u_y(x_j,\kappa h,z_p)|^2\\
&\quad+h^3\frac{h^2}{6}\left(r\left(n-\frac{1}{n}\right)+r\left(m-\frac{1}{m}\right)+r^2-1\right)\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{\kappa=1}^{r-1}|u_z(x_j,y_k,\kappa h)|^2\,.
\end{aligned}$$

Nach Lemma 5.1.4 und mit $\operatorname{diam}(\Omega_h)^2 = h^2\left((n-1)^2+(m-1)^2+(r-1)^2\right)$ ergibt sich insgesamt

$$\begin{aligned}
\|u-\bar{u}\|^2_{L^2_h(\Omega_h)} \le \operatorname{diam}(\Omega_h)^2\frac{1}{2}\Bigg(&h^3\sum_{\kappa=1}^{n-1}\sum_{l=1}^{m}\sum_{p=1}^{r}|u_x(\kappa h,y_l,z_p)|^2\\
&+h^3\sum_{j=1}^{n}\sum_{\kappa=1}^{m-1}\sum_{p=1}^{r}|u_y(x_j,\kappa h,z_p)|^2+h^3\sum_{j=1}^{n}\sum_{l=1}^{m}\sum_{\kappa=1}^{r-1}|u_z(x_j,y_l,\kappa h)|^2\Bigg)\,,
\end{aligned}$$

woraus die Behauptung folgt. ∎

5.2 Diskrete Cacciopoli-Ungleichung

Die Vorgehensweise aus Konstruktion 2.3.1 zur Bildung der geschachtelten Rechteckgitter kann für den dreidimensionalen Fall analog erfolgen, indem das Ausgangsgitter

sukzessive um die Randpunkte erweitert wird. Der Differenzenoperator sei durch (5.1) gegeben und die Bezeichnungen

$$\lambda_{min} := \min_{x \in \overline{\Omega}_h} \{a(x), d(x), g(x)\} \text{ und } \lambda_{max} := \max_{x \in \overline{\Omega}_h} \{a(x), d(x), g(x)\}$$

analog zum zweidimensionalen Fall eingeführt.

Die beiden Ansätze zum Beweis der diskreten Cacciopoli-Ungleichung für konstante und variable Koeffizienten lassen sich direkt auf den dreidimensionalen Fall übertragen, wobei sich die Konstanten in der Ungleichung durch $C_{caccio} = 1$ für konstante Koeffizienten und $C_{caccio} = 3\sqrt{3}$ für variable Koeffizienten ergeben. Zunächst wird erneut ein Resultat für konstante Koeffizienten mit einem Ansatz analog zu den Ausführungen aus [CFL28] bzw. Satz 2.3.2 angegeben:

Satz 5.2.1 *Sei $K_h \subset h\mathbb{Z}^3$ ein Quadergitter mit konstanter Schrittweite h und der Differenzenoperator L von der Form (5.1) mit konstanten Koeffizienten $a, d, g \in \mathbb{R}^+$. Aus $K_h^0 := K_h$ wird, wie in Konstruktion 2.3.1 beschrieben, eine Schachtelung bis zum Gitter $K_h^l, l \geq 1$ gebildet. Dann gilt*

$$\|\nabla_h u\|_{L_h^2(K_h)} \leq C_{caccio} \frac{\sqrt{\kappa}}{\operatorname{dist}_\infty(K_h, \Gamma(K_h^l))} \|u\|_{L_h^2(K_h^l)}$$

mit $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$ und $C_{caccio} = 1$ für alle $u \in Z_h^L(K_h^l)$.

Beweis Die Bezeichnungen aus (2.20) können analog für Quadergitter eingeführt und für die dreidimensionale Problemstellung um $\partial^{z-}K_h, \partial^{z+}K_h, \Gamma^{z-}(K_h)$ und $\Gamma^{z+}(K_h)$ ergänzt werden. Damit gilt analog zum zweidimensionalen Fall die Abschätzung

$$\begin{aligned}
\|\nabla_h u\|_{L_h^2(K_h)}^2 &= h^3 \sum_{K_h \setminus \Gamma^{x+}(K_h)} u_x^2 + h^3 \sum_{K_h \setminus \Gamma^{y+}(K_h)} u_y^2 + h^3 \sum_{K_h \setminus \Gamma^{z+}(K_h)} u_z^2 \\
&\leq \frac{1}{\lambda_{min}} \frac{1}{2} \left(h^3 \sum_{\partial^{x-}K_h \cup K_h \setminus \Gamma^{x+}(K_h)} a u_x^2 + h^3 \sum_{K_h} a u_x^2 + h^3 \sum_{\partial^{y-}K_h \cup K_h \setminus \Gamma^{y+}(K_h)} d u_y^2 \right. \qquad (5.13) \\
&\quad \left. + h^3 \sum_{K_h} d u_y^2 + h^3 \sum_{\partial^{z-}K_h \cup K_h \setminus \Gamma^{z+}(K_h)} g u_z^2 + h^3 \sum_{K_h} g u_z^2 \right). \qquad (5.14)
\end{aligned}$$

Mit den Bezeichnungen $u^{x+}(x_i, y_j, z_p) = u(x_{i+1}, y_j, z_p), u^{x-}(x_i, y_j, z_p) = u(x_{i-1}, y_j, z_p)$ (analog für die y- und z-Richtung) erhält man für die Summen aus (5.13) und (5.14) mittels partieller Summation

$$h^3 \sum_{\partial^{x-}K_h \cup K_h \setminus \Gamma^{x+}(K_h)} a u_x^2 = -h^3 \sum_{K_h} (a u_x)_{\bar{x}} u - h^2 \sum_{\partial^{x-}K_h} a u_x u + h^2 \sum_{\Gamma^{x+}(K_h)} a u_x u \quad (5.15)$$

und analog

$$h^3 \sum_{K_h} a u_x^2 = -h^3 \sum_{K_h} (a u_x)_{\bar{x}} u + h^2 \sum_{\Gamma^{x+}(K_h)} a u_x u^{x+} - h^2 \sum_{\partial^{x-}K_h} a u_x u^{x+}. \quad (5.16)$$

Auf die gleiche Weise ergeben sich die Darstellungen

$$h^3 \sum_{\partial^{y-} K_h \cup K_h \setminus \Gamma^{y+}(K_h)} du_y^2 = -h^3 \sum_{K_h} (du_y)_{\bar{y}}\, u - h^2 \sum_{\partial^{y-} K_h} du_y u + h^2 \sum_{\Gamma^{y+}(K_h)} du_y u, \quad (5.17)$$

$$h^3 \sum_{K_h} du_y^2 = -h^3 \sum_{K_h} (du_y)_{\bar{y}} u + h^2 \sum_{\Gamma^{y+}(K_h)} du_y u^{y+} - h^2 \sum_{\partial^{y-} K_h} du_y u^{y+}, \quad (5.18)$$

$$h^3 \sum_{\partial^{z-} K_h \cup K_h \setminus \Gamma^{z+}(K_h)} g u_z^2 = -h^3 \sum_{K_h} (g u_z)_{\bar{z}}\, u - h^2 \sum_{\partial^{z-} K_h} g u_z u + h^2 \sum_{\Gamma^{z+}(K_h)} g u_z u \quad (5.19)$$

und

$$h^3 \sum_{K_h} g u_z^2 = -h^3 \sum_{K_h} (g u_z)_{\bar{z}} u + h^2 \sum_{\Gamma^{z+}(K_h)} g u_z u^{z+} - h^2 \sum_{\partial^{z-} K_h} g u_z u^{z+}. \quad (5.20)$$

Die Addition der Summen (5.15)-(5.20) führt zu dem folgenden Resultat:

$$\begin{aligned}
&h^3 \sum_{\partial^{x-} K_h \cup K_h \setminus \Gamma^{x+}(K_h)} a u_x^2 + h^3 \sum_{K_h} a u_x^2 + h^3 \sum_{\partial^{y-} K_h \cup K_h \setminus \Gamma^{y+}(K_h)} du_y^2 + h^3 \sum_{K_h} du_y^2 \\
&\qquad + h^3 \sum_{\partial^{z-} K_h \cup K_h \setminus \Gamma^{z+}(K_h)} g u_z^2 + h^3 \sum_{K_h} g u_z^2 \\
&= -2h^3 \sum_{K_h} \left[(a u_x)_{\bar{x}} + (du_y)_{\bar{y}} + (g u_z)_{\bar{z}} \right] u \\
&\quad - h^2 \sum_{\partial^{x-} K_h} a u_x u + h^2 \sum_{\Gamma^{x+}(K_h)} a u_x u + h^2 \sum_{\Gamma^{x+}(K_h)} a u_x u^{x+} - h^2 \sum_{\partial^{x-} K_h} a u_x u^{x+} \quad (5.21) \\
&\quad - h^2 \sum_{\partial^{y-} K_h} du_y u + h^2 \sum_{\Gamma^{y+}(K_h)} du_y u + h^2 \sum_{\Gamma^{y+}(K_h)} du_y u^{y+} - h^2 \sum_{\partial^{y-} K_h} du_y u^{y+} \quad (5.22) \\
&\quad - h^2 \sum_{\partial^{z-} K_h} g u_z u + h^2 \sum_{\Gamma^{z+}(K_h)} g u_z u + h^2 \sum_{\Gamma^{z+}(K_h)} g u_z u^{z+} - h^2 \sum_{\partial^{z-} K_h} g u_z u^{z+}. \quad (5.23)
\end{aligned}$$

Die Ausdrücke (5.21), (5.22) und (5.23) lassen sich jeweils kombinieren zu

$$\begin{aligned}
&- h^2 \sum_{\partial^{x-} K_h} a u_x u + h^2 \sum_{\Gamma^{x+}(K_h)} a u_x u + h^2 \sum_{\Gamma^{x+}(K_h)} a u_x u^{x+} - h^2 \sum_{\partial^{x-} K_h} a u_x u^{x+} \\
&\qquad = h \sum_{\partial^{x-} K_h} a \left(u^2 - \left(u^{x+}\right)^2 \right) + h \sum_{\Gamma^{x+}(K_h)} a \left(\left(u^{x+}\right)^2 - u^2 \right),
\end{aligned}$$

$$\begin{aligned}
&- h^2 \sum_{\partial^{y-} K_h} du_y u + h^2 \sum_{\Gamma^{y+}(K_h)} du_y u + h^2 \sum_{\Gamma^{y+}(K_h)} du_y u^{y+} - h^2 \sum_{\partial^{y-} K_h} du_y u^{y+} \\
&\qquad = h \sum_{\partial^{y-} K_h} d \left(u^2 - \left(u^{y+}\right)^2 \right) + h \sum_{\Gamma^{y+}(K_h)} d \left(\left(u^{y+}\right)^2 - u^2 \right)
\end{aligned}$$

und

$$
\begin{aligned}
-h^2 \sum_{\partial^{z-} K_h} g u_z u + h^2 \sum_{\Gamma^{z+}(K_h)} g u_z u + h^2 \sum_{\Gamma^{z+}(K_h)} g u_z u^{z+} - h^2 \sum_{\partial^{z-} K_h} g u_z u^{z+} \\
= h \sum_{\partial^{z-} K_h} g \left(u^2 - \left(u^{z+} \right)^2 \right) + h \sum_{\Gamma^{z+}(K_h)} g \left(\left(u^{z+} \right)^2 - u^2 \right).
\end{aligned}
$$

Beachtet man, dass

$$
\sum_{\partial^{x-} K_h} a \left(u^{x+} \right)^2 = \sum_{\Gamma^{x-}(K_h)} a u^2 \text{ und } \sum_{\Gamma^{x+}(K_h)} a \left(u^{x+} \right)^2 = \sum_{\partial^{x+} K_h} a u^2
$$

(analog für y- und z-Richtung) gilt, kann die Summe der Ausdrücke (5.21), (5.22) und (5.23) demnach abgeschätzt werden durch

$$
\begin{aligned}
&-h^2 \sum_{\partial^{x-} K_h} a u_x u + h^2 \sum_{\Gamma^{x+}(K_h)} a u_x u + h^2 \sum_{\Gamma^{x+}(K_h)} a u_x u^{x+} - h^2 \sum_{\partial^{x-} K_h} a u_x u^{x+} \\
&\qquad - h^2 \sum_{\partial^{y-} K_h} d u_y u + h^2 \sum_{\Gamma^{y+}(K_h)} d u_y u + h^2 \sum_{\Gamma^{y+}(K_h)} d u_y u^{y+} - h^2 \sum_{\partial^{y-} K_h} d u_y u^{y+} \\
&\qquad - h^2 \sum_{\partial^{z-} K_h} g u_z u + h^2 \sum_{\Gamma^{z+}(K_h)} g u_z u + h^2 \sum_{\Gamma^{z+}(K_h)} g u_z u^{z+} - h^2 \sum_{\partial^{z-} K_h} g u_z u^{z+} \\
&\qquad \leq h \left[\sum_{\partial K_h} c_{adg} u^2 - \sum_{\Gamma(K_h)} c_{adg} u^2 \right] \\
&\qquad = h \left[\sum_{\Gamma(K_h^1)} c_{adg} u^2 - \sum_{\Gamma(K_h^0)} c_{adg} u^2 \right]
\end{aligned}
$$

mit

$$
c_{adg}(x_i, y_j, z_p) := \begin{cases} a & \text{für } (x_i, y_j, z_p) \in \partial^{x+} K_h \cup \partial^{x-} K_h \cup \Gamma^{x+}(K_h) \cup \Gamma^{x-}(K_h) \\ d & \text{für } (x_i, y_j, z_p) \in \partial^{y+} K_h \cup \partial^{y-} K_h \cup \Gamma^{y+}(K_h) \cup \Gamma^{y-}(K_h) \\ g & \text{für } (x_i, y_j, z_p) \in \partial^{z+} K_h \cup \partial^{z-} K_h \cup \Gamma^{z+}(K_h) \cup \Gamma^{z-}(K_h). \end{cases}
$$

Auf Grundlage dieses Ergebnisses lassen sich alle weiteren Überlegungen aus dem Beweis von Satz 2.3.2 auf den Fall von Quadergittern übertragen. ∎

Auch der Beweis der diskreten Cacciopoli-Ungleichung im Fall variabler Konstanten lässt sich für den dreidimensionalen Fall erweitern, so dass man das folgende Resultat erhält:

Satz 5.2.2 *Sei $K_h \subset h\mathbb{Z}^3$ ein Quadergitter mit konstanter Schrittweite h und der Differenzenoperator L von der Form* (5.1) *mit ortsabhängigen Koeffizienten $a, d, g \in \mathcal{D}_h\left(\overline{\Omega}_h\right)$*

mit $a, d, g > 0$ gegeben. Aus $K_h^0 := K_h$ wird, wie in Konstruktion 2.3.1 beschrieben, eine Schachtelung bis zum Gitter $K_h^l, l \geq 1$ gebildet. Dann gilt

$$\|\nabla_h u\|_{L_h^2(K_h)} \leq C_{caccio} \frac{\sqrt{\kappa}}{\operatorname{dist}_\infty(K_h, \Gamma(K_h^l))} \|u\|_{L_h^2(K_h^l)}$$

mit $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$ und $C_{caccio} = 3\sqrt{3}$ für alle $u \in Z_h^L(K_h^l)$.

Beweis Wie bereits beschrieben, werden für den Beweis nur die wesentlichen Veränderungen im Vergleich zum zweidimensionalen Fall angegeben.

Das Quadergitter K_h sei durch ein $n \times m \times r$ Gitter gegeben und die Abschneidefunktion η auf den dreidimensionalen Fall erweitert, so dass weiterhin $|\eta_x|, |\eta_y|, |\eta_z| \leq \frac{1}{lh}$ gilt. Dann erhält man

$$\begin{aligned}\|\nabla_h u\|_{L_h^2(K_h)}^2 &= h^3 \sum_{i=1}^{n-1} \sum_{j=1}^{m} \sum_{p=1}^{r} u_x^2 + h^3 \sum_{i=1}^{n} \sum_{j=1}^{m-1} \sum_{p=1}^{r} u_y^2 + h^3 \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{p=1}^{r-1} u_z^2 \\ &\leq \frac{1}{4} \frac{1}{\lambda_{min}} h^3 \sum_{i=-l+1}^{n+l-1} \sum_{j=-l+1}^{m+l-1} \sum_{p=-l+1}^{r+l-1} \left\| \begin{pmatrix} a^{\frac{1}{2}} \left(\eta(x_{i+1}, y_j, z_p) + \eta\right) u_x \\ d^{\frac{1}{2}} \left(\eta(x_i, y_{j+1}, z_p) + \eta\right) u_y \\ g^{\frac{1}{2}} \left(\eta(x_i, y_j, z_{p+1}) + \eta\right) u_z \end{pmatrix} \right\|^2 .\end{aligned}$$

Erneut ergibt sich

$$\begin{aligned}&h^3 \sum_{i=-l+1}^{n+l-1} \sum_{j=-l+1}^{m+l-1} \sum_{p=-l+1}^{r+l-1} \left(\eta(x_i, y_j, z_{p+1}) + \eta\right)^2 u_z g u_z \\ &\quad = h^3 \sum_{i=-l+1}^{n+l-1} \sum_{j=-l+1}^{m+l-1} \sum_{p=-l+1}^{r+l-1} \Big\{ 4(n^2 u)_z g u_z - g u_z \eta_z \Big[u \left(\eta(x_i, y_j, z_{p+1}) + 3\eta\right) \\ &\qquad\qquad + u(x_i, y_j, z_{p+1}) \left(3\eta(x_i, y_j, z_{p+1}) + \eta\right) \Big] \Big\}\end{aligned}$$

und man erhält analoge Ergebnisse für die übrigen Summen zu au_x und du_y.

Mittels partieller Summation kann (unter Berücksichtigung von $\eta = 0$ auf $\Gamma(K_h^l)$) die Darstellung

$$h^3 \sum_{i=-l+1}^{n+l-1} \sum_{j=-l+1}^{m+l-1} \sum_{p=-l+1}^{r+l-1} (n^2 u)_z g u_z = -h^3 \sum_{i=-l+1}^{n+l-1} \sum_{j=-l+1}^{m+l-1} \sum_{p=-l+1}^{r+l-1} (n^2 u) \left(g u_z\right)_{\bar{z}}$$

angegeben werden, die sich ebenfalls für die Summen mit au_x und du_y ergibt. Für Gitterfunktionen $u \in Z_h^L(K_h^l)$ gilt demnach

$$\begin{aligned}&h^3 \sum_{i=-l+1}^{n+l-1} \sum_{j=-l+1}^{m+l-1} \sum_{p=-l+1}^{r+l-1} \left\{ (n^2 u)_x a u_x + (n^2 u)_y d u_y + (n^2 u)_z g u_z \right\} \\ &\qquad = h^3 \sum_{i=-l+1}^{n+l-1} \sum_{j=-l+1}^{m+l-1} \sum_{p=-l+1}^{r+l-1} (n^2 u) \left[(a u_x)_{\bar{x}} + (d u_y)_{\bar{y}} + (g u_z)_{\bar{z}} \right] = 0.\end{aligned}$$

Außerdem kann erneut abgeschätzt werden durch

$$\left|-gu_z\eta_z\Big[u\left(\eta(x_i,y_j,z_{p+1})+3\eta\right)+u(x_i,y_j,z_{p+1})\left(3\eta(x_i,y_j,z_{p+1})+\eta\right)\Big]\right| \\ \leq 3\frac{\sqrt{\lambda_{max}}}{lh}\left|g^{\frac{1}{2}}u_z\right|\left|\eta(x_i,y_j,z_{p+1})+\eta\right|\left(|u(x_i,y_j,z_{p+1})|+|u|\right)$$

und die Abschätzung lässt sich analog für die Terme mit au_x und du_y durchführen. Somit erhält man insgesamt

$$h^3\sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1}\sum_{p=-l+1}^{r+l-1}\left\|\begin{pmatrix}a^{\frac{1}{2}}\left(\eta(x_{i+1},y_j,z_p)+\eta\right)u_x\\ d^{\frac{1}{2}}\left(\eta(x_i,y_{j+1},z_p)+\eta\right)u_y\\ g^{\frac{1}{2}}\left(\eta(x_i,y_j,z_{p+1})+\eta\right)u_z\end{pmatrix}\right\|^2$$

$$\leq 3\frac{\sqrt{\lambda_{max}}}{lh}h^3\sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1}\sum_{p=-l+1}^{r+l-1}\begin{pmatrix}\left|a^{\frac{1}{2}}u_x\right|\left|\eta(x_{i+1},y_j,z_p)+\eta\right|\\ \left|d^{\frac{1}{2}}u_y\right|\left|\eta(x_i,y_{j+1},z_p)+\eta\right|\\ \left|g^{\frac{1}{2}}u_z\right|\left|\eta(x_i,y_j,z_{p+1})+\eta\right|\end{pmatrix}^T\begin{pmatrix}|u|\\ |u|\\ |u|\end{pmatrix}$$

$$+3\frac{\sqrt{\lambda_{max}}}{lh}h^3\sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1}\sum_{p=-l+1}^{r+l-1}\begin{pmatrix}\left|a^{\frac{1}{2}}u_x\right|\left|\eta(x_{i+1},y_j,z_p)+\eta\right|\\ \left|d^{\frac{1}{2}}u_y\right|\left|\eta(x_i,y_{j+1},z_p)+\eta\right|\\ \left|g^{\frac{1}{2}}u_z\right|\left|\eta(x_i,y_j,z_{p+1})+\eta\right|\end{pmatrix}^T\begin{pmatrix}|u(x_{i+1},y_j,z_p)|\\ |u(x_i,y_{j+1},z_p)|\\ |u(x_i,y_j,z_{p+1})|\end{pmatrix}$$

$$\leq 6\sqrt{3}\frac{\sqrt{\lambda_{max}}}{lh}\left(h^3\sum_{i=-l+1}^{n+l-1}\sum_{j=-l+1}^{m+l-1}\sum_{p=-l+1}^{r+l-1}\left\|\begin{pmatrix}\left|a^{\frac{1}{2}}u_x\right|\left|\eta(x_{i+1},y_j,z_p)+\eta\right|\\ \left|d^{\frac{1}{2}}u_y\right|\left|\eta(x_i,y_{j+1},z_p)+\eta\right|\\ \left|g^{\frac{1}{2}}u_z\right|\left|\eta(x_i,y_j,z_{p+1})+\eta\right|\end{pmatrix}\right\|^2\right)^{\frac{1}{2}} \\ \cdot\|u\|_{L_h^2(K_h^l)}.$$

Fasst man die Ergebnisse zusammen, ergibt sich

$$\|\nabla_h u\|_{L_h^2(K_h)}\leq\sqrt{\frac{\lambda_{max}}{\lambda_{min}}}\frac{3\sqrt{3}}{lh}\|u\|_{L_h^2(K_h^l)}$$

für alle $u\in Z_h^L(\Omega_h)$, woraus mit $\operatorname{dist}_\infty(K_h,\Gamma(K_h^l))=hl$ die Behauptung folgt. ∎

Spezielle diskrete Cacciopoli-Ungleichung

Wie für Rechteckgitter wird auch für Quadergitter eine spezielle Formulierung der diskreten Cacciopoli-Ungleichung benötigt.

Satz 5.2.3 *Seien Ω_h und $K_h\subset\Omega_h$ Quadergitter mit der konstanten Schrittweite h und der Differenzenoperator L von der Form (5.1) mit ortsabhängigen Koeffizienten $a,d,g\in\mathcal{D}_h\left(\overline{\Omega}_h\right)$, $a,d,g>0$ gegeben. Aus $K_h^0:=K_h$ wird, wie in Konstruktion 2.3.1*

beschrieben, eine Schachtelung bis zum Gitter $K_h^l, l \geq 1$ *mit* $\Omega_h \backslash K_h^l \neq \emptyset$ *gebildet. Dann gilt*

$$\|\nabla_h u\|_{L_h^2(K_h)} \leq C_{caccio} \frac{\sqrt{\kappa}}{\operatorname{dist}_\infty(K_h, \Gamma(K_h^l))} \|u\|_{L_h^2(K_h^l \cap \Omega_h)}$$

mit $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$ *und* $C_{caccio} = 3\sqrt{3}$ *für alle* $u \in Z_{h,0}^L(K_h^l; \Omega_h)$.

Beweis Der Beweis kann analog zu dem Beweis von Satz 2.3.4 erfolgen. ∎

5.3 Approximationsresultate

Zunächst wird das Resultat aus Lemma 4.1.1 auf den dreidimensionalen Fall erweitert, um im Anschluss ein Ergebnis zur Approximation von diskret L-harmonischen Funktionen mittels einer diskreten separablen Entwicklung zu erhalten.

Lemma 5.3.1 *Sei* $\Omega_h \subset h\mathbb{Z}^3$ *ein Quadergitter mit konstanter Schrittweite* h. *Dann existiert für alle* $k \in \mathbb{N}$ *ein Unterraum* $V_k \subset Z_h^L(\Omega_h)$ *der Dimension* $\dim V_k \leq k$ *mit*

$$\operatorname{dist}_{L_h^2(\Omega_h)}(u, V_k) \leq \sqrt{2} \frac{\operatorname{diam}(\Omega_h)}{\sqrt[3]{k}} \|\nabla_h u\|_{L_h^2(\Omega_h)}$$

für alle $u \in Z_h^L(\Omega_h)$.

Beweis Der Beweis kann analog zum Beweis von Lemma 4.1.1 durchgeführt werden, lediglich die Konstruktion der Rechteckteilgitter ist folgendermaßen zu ersetzen:

Das Quadergitter Ω_h sei durch ein $n \times m \times r$ Quadergitter mit $n, m, r \in \mathbb{N}$ gegeben. Demnach besitzt es den Durchmesser

$$\operatorname{diam}(\Omega_h) = h\sqrt{(n-1)^2 + (m-1)^2 + (r-1)^2}.$$

Die Gitterpunkte aus $\Gamma(\Omega_h)$ bilden einen Quader Q, welcher die gleichen Seitenlängen wie Ω_h besitzt und es gilt $\Omega_h \subset Q$. Dieser Quader wird in k Teilquader $Q_i, i = 1, \ldots, k$ unterteilt. Geht man zunächst davon aus, dass $k = l^3$ mit $l \in \mathbb{N}$ gilt, dann können die Teilquader gebildet werden, indem die Seiten des Quaders Q in jeweils l Abschnitte unterteilt werden. Demnach besitzen die k Teilquader Q_i die Kantenlänge $\frac{(n-1)h}{l} \times \frac{(m-1)h}{l} \times \frac{(r-1)h}{l}$. Zur Bildung der Teilgitter wird $\Omega_i := Q_i \cap \Omega_h$ gesetzt. Sollte die Zuordnung der Gitterpunkte zu mehreren Ω_i möglich sein (dieser Fall kann auftreten, wenn die Gitterpunkte auf dem Rand der Q_i liegen), so werden diese Punkte alle einem angrenzenden Teilgitter zugeordnet, so dass die Ω_i stets durch Quadergitter gegeben sind.

Die übrigen Beweisschritte lassen sich analog ausführen, wobei nach Satz 5.1.5 in der Poincaré-Ungleichung weiterhin die Konstante

$$C_h = \frac{\operatorname{diam}(\Omega_i)}{\sqrt{2}}$$

verwendet werden kann. Bei dem Übergang von $k = l^3$ zu einem allgemeinen k wird $l := \left\lfloor \sqrt[3]{k} \right\rfloor \in \mathbb{N}$ verwendet, so dass $l^3 \leq k \leq (l+1)^3$ und $\frac{1}{l} \leq \frac{2}{l+1} \leq \frac{2}{\sqrt[3]{k}}$ erfüllt ist. Die Abschätzung ergibt sich dann zu

$$\operatorname{dist}_{L_h^2(\Omega_h)}(u, W_k) \leq 2\frac{1}{\sqrt{2}} \frac{\operatorname{diam}(\Omega_h)}{\sqrt[3]{k}} \left\|\nabla_h u\right\|_{L_h^2(\Omega_h)}. \qquad \blacksquare$$

Zum Beweis des Hauptresultats für den dreidimensionalen Fall wird außerdem ein Ergebnis analog zu Lemma 4.1.2 benötigt.

Lemma 5.3.2 *Seien Ω_h und $K_h \subset \Omega_h$ Quadergitter und es gelte*

$$\operatorname{diam}(K_h) \leq \eta \operatorname{dist}_\infty(K_h, \Gamma(K_h^\delta))$$

mit $\eta > 0$. Außerdem sei K_h^δ, wie in Konstruktion 2.3.1 beschrieben, mit $\delta \geq 1$ gebildet, es gelte $\Omega_h \backslash K_h^\delta \neq \emptyset$ und es sei

$$\epsilon_0 = \exp\left(-\left\lceil \frac{1}{h\eta} \operatorname{diam}(K_h) \right\rceil\right).$$

Dann existiert für jedes ϵ mit $\epsilon_0 \leq \epsilon < 1$ ein Unterraum $W \subset Z_h^L(K_h)$ mit der Dimension

$$\dim W \leq \left\lceil \log \frac{1}{\epsilon} \right\rceil + c^3 \left\lceil \log \frac{1}{\epsilon} \right\rceil^4$$

mit $c = 2\sqrt{2}\sqrt{\kappa}\, C_{caccio} \left(\eta + 2\sqrt{3}\right) \exp(1)$ und es gilt

$$\operatorname{dist}_{L_h^2(K_h)}(u, W) \leq \epsilon \left\|u\right\|_{L_h^2(K_h^\delta \cap \Omega_h)} \quad \textit{für alle } u \in Z_{h,0}^L(K_h^\delta; \Omega_h).$$

Beweis Der Beweis lässt sich analog zum Beweis von Lemma 4.1.2 führen, es müssen lediglich die Größen in den Abschätzungen an den dreidimensionalen Fall angepasst werden. Führt man die übrigen Beweisschritte mit $k := \left\lceil (\beta p)^3 \right\rceil$ durch und berücksichtigt

$$\operatorname{diam}(K_j) \leq \operatorname{diam}(\tilde{K}_j) \leq \operatorname{diam}(K_h) + 2j\sqrt{3}\frac{\delta}{p}h \leq \operatorname{diam}(K_h) + 2\sqrt{3}\delta h$$

sowie das Resultat aus Lemma 5.3.1, dann erhält man

$$c = 2\sqrt{2}\sqrt{\kappa}\, C_{caccio} \left(\eta + 2\sqrt{3}\right) \exp(1).$$

Für die Abschätzung der Dimension ergibt sich

$$\dim W \leq \left\lceil (\beta p)^3 \right\rceil p \leq p + \beta^3 p^4. \qquad \blacksquare$$

Mit Hilfe dieses Resultats lässt sich die Existenz einer diskreten separablen Entwicklung der diskreten Greenschen Funktion mit exponentieller Konvergenz zur dreidimensionalen Problemstellung auf Quadergittern zeigen:

Satz 5.3.3 *Seien Ω_h und $X_h, Y_h \subset \Omega_h$ Quadergitter mit*

$$\operatorname{diam}(Y_h) \leq \eta \left(\operatorname{dist}_\infty(X_h, Y_h) - h\right)$$

und Y_h^δ sei, wie in Konstruktion 2.3.1 beschrieben, mit $\delta := \frac{\operatorname{dist}_\infty(X_h,Y_h)-h}{h} \geq 1$ gebildet. Die diskrete Greensche Funktion zur Problemstellung (2.10) *mit dem Differenzenoperator* (5.1) *sei mit g_h bezeichnet und es seien*

$$\epsilon_0 = \exp\left(-\left\lceil \frac{1}{h\eta} \operatorname{diam}(Y_h) \right\rceil\right)$$

und c wie in Lemma 5.3.2 gegeben.

Dann existiert für jedes ϵ mit $\epsilon_0 \leq \epsilon < 1$ eine diskrete separable Entwicklung

$$g_h^k(x,y) = \sum_{l=1}^{k} u_l(x) v_l(y) \quad \text{mit } k \leq \left\lceil \log \frac{1}{\epsilon} \right\rceil + c^3 \left\lceil \log \frac{1}{\epsilon} \right\rceil^4$$

der diskreten Greenschen Funktion auf $X_h \times Y_h$ und es gilt die Fehlerabschätzung

$$\left\| g_h(x,\cdot) - g_h^k(x,\cdot) \right\|_{L_h^2(Y_h)} \leq \epsilon \left\| g_h(x,\cdot) \right\|_{L_h^2(Y_h^\delta \cap \Omega_h)} \quad \text{für alle } x \in X_h.$$

Beweis Der Beweis kann analog zu dem Beweis von Satz 4.1.3 geführt werden, es müssen lediglich die Größen aus Lemma 5.3.2 übernommen werden. ∎

Die Aussage aus Bemerkung 4.1.4 gilt analog für den dreidimensionalen Fall.

5.4 Hauptresultat

Mittels der Ergebnisse aus den letzten Abschnitten kann direkt das Hauptresultat zur Existenz einer $\mathcal{H}$-Matrix Approximation von Finite-Differenzen-Matrizen zu dem dreidimensionalen Modellproblem auf einem Quadergitter angegeben werden.

Satz 5.4.1 *Sei $L_h \in \mathbb{R}^{I \times I}$ die Finite-Differenzen-Matrix zu der Problemstellung* (2.10) *mit dem Differenzenoperator* (5.1) *auf einem Quadergitter Ω_h mit konstanter Schrittweite h und $P \subset T(I \times I)$ eine nach den Ausführungen in Abschnitt 3.2.1 mit Hilfe der Zulässigkeitsbedingung* (3.9) *konstruierte zulässige Partition. Außerdem seien c aus Lemma 5.3.2 und*

$$\epsilon_{0,max} := \max_{\tau \times \sigma \in P^+} \exp\left(-\left\lceil \frac{1}{h\eta} \min\{\operatorname{diam}(X_\tau), \operatorname{diam}(X_\sigma)\} \right\rceil\right)$$

gegeben.

Dann existiert für jedes ϵ mit $\epsilon_{0,max} \leq \epsilon < 1$ eine Matrix $L_{\mathcal{H},inv} \in \mathcal{H}(k,P)$, für die

$$\left\| L_h^{-1} - L_{\mathcal{H},inv} \right\|_F \leq \epsilon \sqrt{|P^+|} \left\| L_h^{-1} \right\|_F$$

mit

$$k \leq \left\lceil \log \frac{1}{\epsilon} \right\rceil + c^3 \left\lceil \log \frac{1}{\epsilon} \right\rceil^4$$

gilt.

Beweis Der Beweis lässt sich analog zu dem Beweis von Satz 4.2.1 führen, es müssen lediglich die Größen aus Satz 5.3.3 übernommen werden. ■

6 Numerik

Die theoretischen Resultate zur Existenz einer $\mathcal{H}$-Matrix Approximation für die Inverse von Finite-Differenzen-Matrizen aus Kapitel 4 und 5 dienen als Grundlage verschiedener numerischer Tests. Das Ziel aller Berechnungen ist die (numerische) Überprüfung, ob sich die Inverse von Finite-Differenzen-Matrizen gut durch eine $\mathcal{H}$-Matrix aus der Menge $\mathcal{H}(k, P)$ approximieren lässt. Eine gute Approximation ist dann gegeben, wenn der Fehler exponentiell mit dem verwendeten Rang k fällt.

Zur Durchführung der Tests muss eine zulässige Partition P der zugrunde liegenden Indexpaarmenge bestimmt werden. Deren Berechnung kann nach den Angaben in Abschnitt 3.2.1 mit Hilfe der Zulässigkeitsbedingung (3.9) für Finite-Differenzen-Matrizen erfolgen. Die resultierende Partition wird im weiteren Verlauf als „gewöhnliche Partition“ bezeichnet. Die Ergebnisse der numerischen Tests werden zeigen, dass bei Verwendung der gewöhnlichen Partitionierungsstrategie in einigen Fällen die Approximationsfehler mit wachsendem Rang nur sehr langsam kleiner werden. Da die Approximation und damit auch der Fehler von der eingesetzten Partition abhängen, könnte der Einsatz einer alternativen Partition zu besseren Resultaten führen. In Abschnitt 6.1 wird für diese Fälle eine modifizierte Partitionierungsstrategie entwickelt, durch deren Einsatz sich deutlich bessere Ergebnisse erzielen lassen. Diese wird zur Unterscheidung von der gewöhnlichen Partition P als modifizierte Partition P_{mod} bezeichnet. Die abweichenden Strategien zur Bestimmung der modifizierten Partitionierung werden auf Grundlage theoretischer Überlegungen eingeführt. Eine umfassende Analyse erfolgt in diesen Fällen jedoch nicht, so dass die Vorgehensweisen und Empfehlungen allein auf den Resultaten der numerischen Tests beruhen.

Die praktische Durchführung der numerischen Tests erfolgt stets auf die gleiche Weise: Zunächst wird die Diskretisierungsmatrix zum jeweiligen Testproblem gebildet und im Anschluss die Inverse der Matrix numerisch berechnet. Um für diese eine $\mathcal{H}$-Matrix Approximation mit festem Rang k zur vorgegebenen Partition P zu bestimmen, müssen die Teilmatrizen der Inversen zu den zulässigen Blöcken $b \in P^+$ durch Rang-k-Matrizen approximiert werden. Dies erfolgt durch die Berechnung der Singulärwertzerlegung und anschließendes „Kürzen“ auf den vorgegebenen Rang k. Die Einträge der Inversen zu den Blöcken $b \in P^-$ werden ohne Approximation übernommen. Für die auf diese Weise berechnete Approximation wird der relative Approximationsfehler in der Frobenius-Norm zu den Rängen $k = 1, \ldots, 5$ berechnet, so dass sich die Fehlerentwicklung in Abhängigkeit von k untersuchen lässt.

Die effiziente Berechnung einer approximativen $\mathcal{H}$-Inversen in der $\mathcal{H}$-Arithmetik ist in diesem Zusammenhang nicht von Interesse. Aus diesem Grund werden weder Laufzeiten noch Speicheranforderungen, sondern nur die Approximationsfehler angegeben. Eine Abschätzung der Speicheranforderungen kann für den speziellen Fall von Finite-

Differenzen-Matrizen wie in Abschnitt 3.2.3 erfolgen.

Im Anschluss an die Einführung der modifizierten Partitionierungsstrategie in Abschnitt 6.1 folgen in den Abschnitten 6.1.1 und 6.1.2 numerische Tests zum zwei- und dreidimensionalen Modellproblem, bei denen die neu eingeführte modifizierte Partition zum Einsatz kommt. Für die Inversen der Matrizen zum Modellproblem konnte in Satz 4.2.1 bzw. Satz 5.4.1 die Existenz einer $\mathcal{H}$-Matrix Approximation nachgewiesen werden. Deshalb steht bei der numerischen Untersuchung in diesem Zusammenhang die Abhängigkeit des Fehlers von der Größe $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$ im Mittelpunkt. Da λ_{min} bzw. λ_{max} durch den kleinsten bzw. größten Koeffizienten des Differenzenoperators auf dem zugrunde liegenden Gitter gegeben sind, hängt κ von der Wahl des entsprechenden Differenzenoperators ab. Die Größe von κ hat direkten Einfluss auf die Konstante c in den Sätzen 4.2.1 und 5.4.1 und damit – nach den theoretischen Resultaten – auch auf den Fehlerverlauf bei der Approximation (vgl. Abschnitt 4.2). Daher wird untersucht, ob sich diese Abhängigkeit ebenfalls bei den numerischen Tests beobachten lässt. Im Zuge dieser Untersuchungen erfolgt ein Vergleich der Ergebnisse bei Verwendung der gewöhnlichen und der modifizierten Partition.

Zusätzlich dazu werden im Anschluss in Abschnitt 6.2 weitere Problemstellungen eingeführt, aus deren Diskretisierungen andere Finite-Differenzen-Matrizen hervorgehen. Die Auswahl beschränkt sich dabei auf Probleme, die sich durch die Erweiterung des Modellproblems um einen Konvektionsterm in Abschnitt 6.2.1, die Berücksichtigung von Neumann-Randwerten in Abschnitt 6.2.2, die Diskretisierung der Wärmeleitungsgleichung in Abschnitt 6.2.3 sowie die Verwendung der Koeffizienten aus dem Modell METRAS in Abschnitt 6.2.4 ergeben. Im Gegensatz zum Modellproblem liegen für diese Problemstellungen keine theoretischen Ergebnisse zur Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen vor. Unabhängig davon kann jedoch numerisch überprüft werden, ob sich die Inversen dieser Matrizen unter Verwendung der eingeführten Partitionierungsstrategie gut durch $\mathcal{H}$-Matrizen approximieren lassen. Im Zuge der numerischen Tests zu diesen Problemstellungen kann erneut beobachtet werden, dass sich die erzielten Ergebnisse bei Verwendung der gewöhnlichen Partition zum Teil deutlich verschlechtern. In diesen Fällen lassen sich die Erkenntnisse zur Konstruktion einer modifizierten Partition zum Modellproblem aus Abschnitt 6.1 nutzen, um für die Probleme aus Abschnitt 6.2 ebenfalls Anpassungen der Partitionierungsstrategie vorzunehmen. Der Einsatz der auf diese Weise modifizierten Partitionen führt auch in diesen Fällen zu deutlich besseren Ergebnissen.

Alle Tests wurden für unterschiedliche Gittergrößen durchgeführt, wobei auch die Fälle $n \neq m$ bei $n \times m$ Rechteckgittern (und analog für Quadergitter im dreidimensionalen Fall) Berücksichtigung fanden. Die Tests lieferten für unterschiedliche Problemgrößen qualitativ vergleichbare Ergebnisse, so dass die Resultate im zweidimensionalen Fall exemplarisch für Gitter mit $n = m = 128$, also für eine Matrix der Dimension $nm = 16384$, angegeben werden. Im dreidimensionalen Fall wird für das Modellproblem $n = m = r = 32$ gewählt, woraus eine Matrix der Dimension $nmr = 32768$ resultiert. Analoges gilt für die Wahl der Konstanten η aus der Zulässigkeitsbedingung und n_{min}, die in den Testproblemen durch $\eta = 1$ und $n_{min} = 32$ vorgegeben werden.

In dieser Arbeit dienen die numerischen Tests der Überprüfung der theoretischen Er-

gebnisse zur Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen von Finite-Differenzen-Matrizen. Für den Einsatz von $\mathcal{H}$-Matrizen in der Praxis sei darauf hingewiesen, dass zwar in der Theorie sowohl für die Berechnung einer $\mathcal{H}$-Inversen als auch für die Berechnung einer $\mathcal{H}$-LU-Zerlegung fast lineare Komplexität nachgewiesen werden kann, die entsprechende Konstante für die $\mathcal{H}$-Inverse jedoch deutlich größer ausfällt. Daher ist bei der praktischen Anwendung der $\mathcal{H}$-Matrix-Technik die $\mathcal{H}$-LU-Zerlegung vorzuziehen. Insbesondere steht für die Berechnung der $\mathcal{H}$-LU-Zerlegung eine spezielle auf Gebietszerlegung beruhende Clusterstrategie zur Verfügung. Durch deren Einsatz erhält man eine Partitionierung der Faktoren der LU-Zerlegung mit großen zulässigen Blöcken, die durch Nullmatrizen gegeben sind. Dadurch kann ein deutlicher Effizienzgewinn erzielt werden (vgl. [GKLB09]). Abschließende numerische Tests zur Lösung des Gleichungssystems im Modell METRAS unter Berücksichtigung dieser Strategien finden im Rahmen dieser Arbeit nicht statt. Die Effizienz der $\mathcal{H}$-Matrix-Technik zur Lösung linearer Gleichungssysteme konnte jedoch für andere Problemstellungen durch numerische Tests bestätigt werden. Zur Eignung der $\mathcal{H}$-Matrix-Technik zur Lösung linearer Gleichungssysteme ist beispielsweise in [GHK08] ein Vergleich mit anderen „etablierten“ Verfahren zu finden. Dort wird für dünnbesetzte Matrizen die $\mathcal{H}$-LU-Zerlegung auf Grundlage einer algebraischen Clusterstrategie berechnet und als Präkonditionierer eingesetzt.

6.1 Eine modifizierte Partitionierungsstrategie

Wie die Ergebnisse der numerischen Tests zur Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen zum zweidimensionalen Modellproblem in Abschnitt 6.1.1 zeigen werden, ist für einige Testprobleme der Einsatz der $\mathcal{H}$-Matrix-Technik unter Verwendung der gewöhnlichen Partitionierungsstrategie nicht geeignet. Diese Problematik ist deshalb von besonderem Interesse, weil die Testprobleme, bei denen dieses Verhalten auftritt, in Anlehnung an die Problemstellung im Modell METRAS gewählt wurden.

Die spezielle Form der Koeffizienten des Differenzenoperators im Modell METRAS ergibt sich unter anderem daraus, dass das Modellgitter auf ein Rechengitter mit konstanter Schrittweite $h = 1$ transformiert wird. Die Vorgabe unterschiedlicher Schrittweiten $h^x(x_i)$ und $h^y(y_j)$ im Modellgebiet entspricht im Modellproblem den Koeffizienten $a(x_i, y_j) = a(x_i)$ und $d(x_i, y_j) = d(y_j)$, die nur von einer Koordinatenrichtung abhängen.

Im einfachsten Fall von konstanten, jedoch unterschiedlichen Schrittweiten $h^x \neq h^y$ führt dies zu den Koeffizienten $a(x_i) = a = (h^x)^{-2}$ und $d(y_j) = d = (h^y)^{-2}$. Im Modell METRAS können sich die Schrittweiten stark voneinander unterscheiden, weshalb in Abschnitt 6.1.1 numerische Tests mit konstanten, aber unterschiedlichen Schrittweiten $h^x \gg h^y$ bzw. $h^y \gg h^x$ durchgeführt werden. Für diese erhält man $d \gg a$ bzw. $a \gg d$, woraus wiederum $\kappa = \frac{\lambda_{max}}{\lambda_{min}} = \frac{\max\{a,d\}}{\min\{a,d\}} \gg 1$ resultiert. Im Einklang mit den theoretischen Ergebnissen aus Abschnitt 4.2 kann bei der Durchführung der entsprechenden numerischen Tests unter Verwendung der gewöhnlichen Partition eine Verschlechterung des Fehlerverlaufs für wachsendes κ beobachtet werden. Daher stellt sich die Frage, ob die Strategie zur Berechnung der zulässigen Partition so angepasst werden kann, dass die Verwendung einer modifizierten Partition in diesen Fällen zu besseren Ergebnissen

führt. Andernfalls wäre der Einsatz der $\mathcal{H}$-Matrix-Technik nicht Erfolg versprechend.

Beachtet man den Zusammenhang zwischen den Koeffizienten und den unterschiedlichen Schrittweiten im Modellgitter, so wird ersichtlich, dass diese bei der Konstruktion der gewöhnlichen Partition nicht berücksichtigt werden. Diese erfolgt ausschließlich auf Grundlage des numerischen Gitters, unabhängig von den Koeffizienten des Differenzenoperators. Dieser Zusammenhang liefert jedoch einen Ansatz zur Modifikation der Partitionierungsstrategie, indem bei der Berechnung der zulässigen Partition die unterschiedlichen Schrittweiten im Modellgitter berücksichtigt werden. Bei der geometriebasierten Berechnung des Clusterbaums wird dazu das Gitter mit den Schrittweiten h^x und h^y verwendet, die sich direkt aus den Koeffizienten a und d ergeben. Die Berechnung von $\operatorname{diam}(\tau), \operatorname{diam}(\sigma)$ und $\operatorname{dist}(\tau, \sigma)$ in der Zulässigkeitsbedingung kann ebenfalls ohne großen Aufwand analog angepasst werden. Die Zulässigkeitsbedingung lässt sich in diesem Fall in der Form

$$\min\{\operatorname{diam}^{(ad)}(\tau), \operatorname{diam}^{(ad)}(\sigma)\} \leq \eta \left(\operatorname{dist}_{\infty}^{(ad)}(\tau, \sigma) - h^{\infty} \right) \tag{6.1}$$

angeben, wobei h^{∞} die Schrittweite in der Richtung bezeichnet, in der die Distanz in $\operatorname{dist}_{\infty}(\tau, \sigma)$ gemessen wird. Die Kennzeichnung *ad* soll verdeutlichen, dass die Berechnung unter Berücksichtigung der Koeffizienten des Differenzenoperators erfolgt.

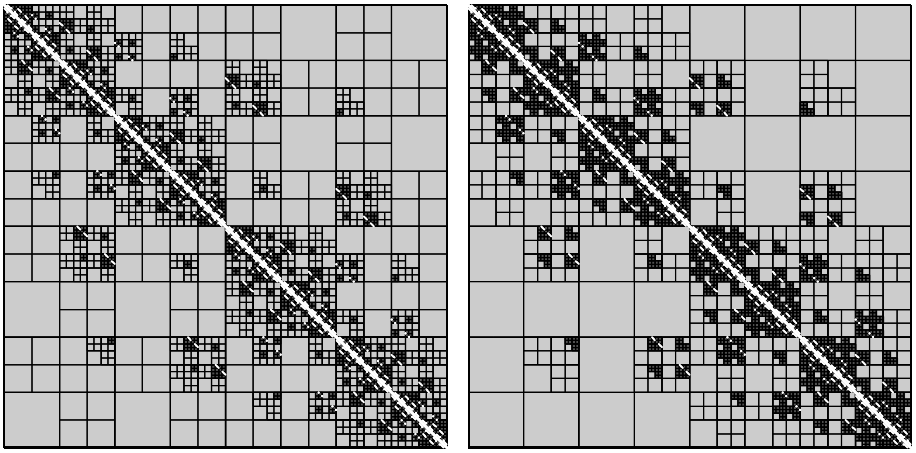
Bei Verwendung dieser Zulässigkeitsbedingung ist es möglich, dass bei stark voneinander abweichenden Schrittweiten h^x und h^y Indexmengen τ, σ als zulässig gekennzeichnet werden, obwohl die entsprechenden Gitter nur wenige Gitterpunkte voneinander entfernt sind. Bei der Durchführung der numerischen Tests unter Verwendung der neuen Bedingung (6.1) hat sich herausgestellt, dass genau in diesen als zulässig gekennzeichneten Blöcken deutlich größere lokale Fehler als in den anderen zulässigen Blöcken auftreten. Um dies zu vermeiden, wird eine weitere Modifikation der Zulässigkeitsbedingung durch die Abfrage der Bedingung

$$\frac{\operatorname{dist}_{\infty}(\tau, \sigma)}{h^{\infty}} \geq \delta_0 \tag{6.2}$$

mit $\delta_0 \in \mathbb{N}$ eingeführt, die sicherstellt, dass zulässige Blöcke einen Mindestabstand von δ_0 Gitterpunkten zueinander besitzen.

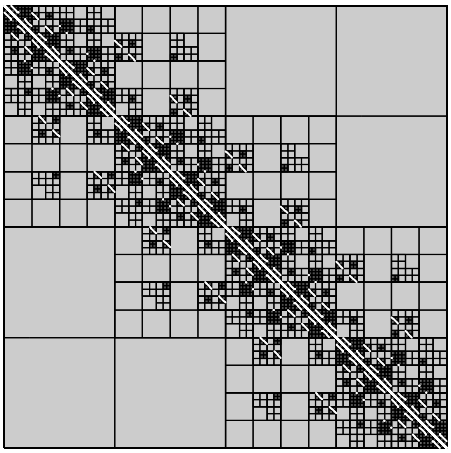
Diese Modifikation wirkt sich nur auf die Zulässigkeit kleiner Blöcke ($|\tau|, |\sigma|$ klein) aus, da für große Blöcke, auch unter Berücksichtigung der unterschiedlichen Schrittweiten, aus der Bedingung (6.1) meist bereits die Gültigkeit von (6.2) folgt. Die Struktur der Partition bleibt somit nach der Einführung der zusätzlichen Bedingung weitestgehend erhalten. Bei den numerischen Tests wird $\delta_0 = 3$ verwendet.

Der große Vorteil der eingeführten Strategie ist, dass unter Berücksichtigung der unterschiedlichen Schrittweiten die Berechnung der modifizierten Partition für beliebige Konstellationen von (konstanten) Koeffizienten erfolgen kann. Insbesondere ergibt sich für den Fall $a = d$ die gleiche Partition wie beim Einsatz der gewöhnlichen Partitionierungsstrategie. Für wachsendes κ unterscheiden sich die Partitionen jedoch deutlich voneinander. In Abbildung 6.1 sind exemplarisch modifizierte Partitionen für Testprobleme mit wachsendem κ zu den Schrittweiten $h^y = 1$ und $h^x = 2^{\alpha}, \alpha \in \{0, 1, 2, 3, 4, 5\}$ für $n = m = 64$ dargestellt. Dies entspricht den Koeffizienten $d = 1$ und $a = 2^{-2\alpha}$, so dass man $\kappa = 2^{2\alpha}$ erhält.

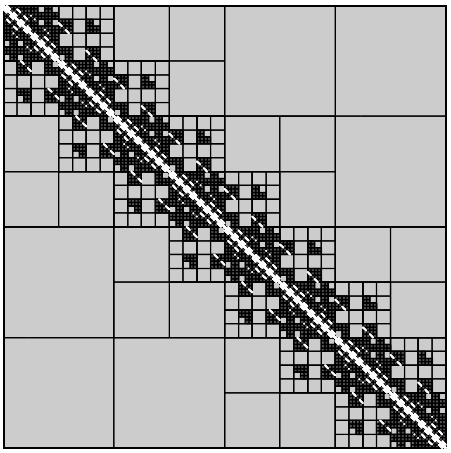


(a) Partition für $a = 1, d = 1$

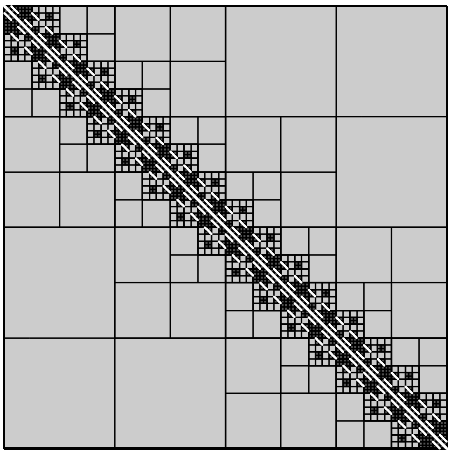
(b) Partition für $a = 2^{-2}, d = 1$



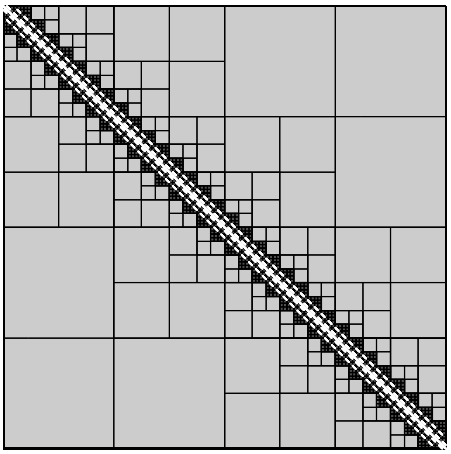
(c) Partition für $a = 2^{-4}, d = 1$



(d) Partition für $a = 2^{-6}, d = 1$



(e) Partition für $a = 2^{-8}, d = 1$

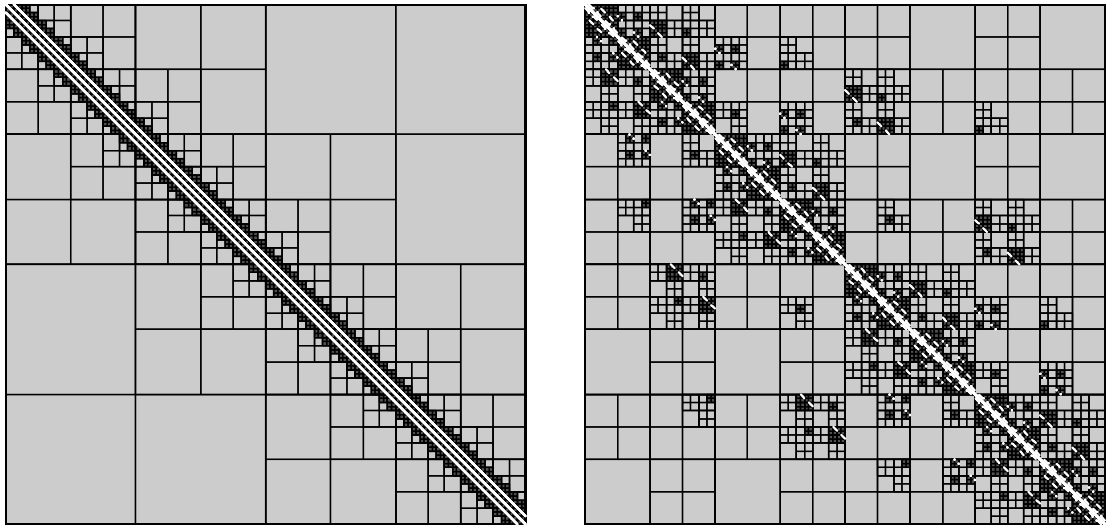


(f) Partition für $a = 2^{-10}, d = 1$

Abbildung 6.1: Modifizierte Partitionen für $n = m = 64$

Im Vergleich zur gewöhnlichen Partition ($h^x = a = 1, h^y = d = 1$) fällt auf, dass bei der modifizierten Partition mit wachsendem κ die Größe der zulässigen Blöcke zunimmt. Dies ist für die Speicherplatzanforderungen und die effiziente Durchführung der Operationen in der $\mathcal{H}$-Arithmetik von großem Vorteil. Der Einsatz der modifizierten Partition wäre demnach bereits dann dem der gewöhnlichen Partition vorzuziehen, wenn die Approximationsfehler in beiden Fällen vergleichbare Größenordnungen aufweisen. Genau dieses Verhalten wird sich bei den numerischen Tests zum Modellproblem in den Abschnitten 6.1.1 und 6.1.2 für kleine κ ergeben.

Da sich im Zusammenhang mit dem Modellproblem insbesondere der Fall $\kappa \to \infty$ als problematisch herausstellen wird, folgen dazu weitere Überlegungen. Zunächst wird die besondere Struktur der modifizierten Partition für $\kappa \to \infty$ anhand eines Testproblems mit stark voneinander abweichenden Koeffizienten $a = 2^{-14}, d = 1$ für die Problemgröße $n = m = 64$ näher betrachtet. In Abbildung 6.2 ist deren Struktur im Vergleich zur gewöhnlichen Partition dargestellt. In diesem Fall weichen die Schrittweiten so stark voneinander ab ($h^x = 2^7, h^y = 1$), dass bei der Konstruktion des Clusterbaums die maximale Ausdehnung der Teilgitter bei jedem Teilungsschritt stets in x-Richtung vorliegt. Dies führt dazu, dass die Aufteilung in zwei Teilgitter stets in dieser Richtung vorgenommen wird. Erst wenn eindimensionale Gitter auftreten und damit die maximale Ausdehnung des Gitters in y-Richtung vorliegt, wird die Teilung in dieser Richtung vorgenommen. Durch diese Vorgehensweise entspricht die interne Anordnung, die sich während der Durchführung der modifizierten Clusterstrategie ergibt, der lexikographischen Anordnung in vertikaler Richtung (dies ist auch an der dargestellten Besetzungsstruktur der Diskretisierungsmatrix in Abbildung 6.2(a) zu erkennen).



(a) Partition nach neuer Clusterstrategie

(b) Partition nach gewöhnlicher Clusterstrategie

Abbildung 6.2: Partition für $a = 2^{-14}$, $d = 1$

Durch diese Überlegungen konnte verdeutlicht werden, wie sich die Berücksichtigung

der unterschiedlichen Schrittweiten auf die Konstruktion und damit ebenfalls auf die besondere Struktur der modifizierten Partition auswirkt. Allerdings ist noch unklar, ob sich der Einsatz der modifizierten Partition für das Modellproblem mit konstanten Koeffizienten unter besonderer Berücksichtigung von $\kappa \to \infty$ eignet.

Zu dieser Problematik können weitere theoretische Überlegungen erfolgen, die zeigen werden, dass die modifizierte Partitionierungsstrategie insbesondere für den Einsatz bei großem κ sehr gut geeignet ist. Betrachtet man die Diskretisierungsmatrix, die sich im Grenzfall mit den Koeffizienten $d = 1, a = 0$ ergibt, hat der resultierende Differenzenstern die Form

$$\frac{1}{h^2}\begin{bmatrix} 0 & 1 & 0 \\ 0 & -2 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Dies entspricht der Standard-Diskretisierung des eindimensionalen Laplace-Operators, wobei das zugrunde liegende (eindimensionale) Gitter durch die m Gitterpunkte des Rechteckgitters in vertikaler Richtung gegeben ist. Demnach ergeben sich für das $n \times m$ Rechteckgitter insgesamt n Probleme der Dimension m. Nummeriert man die Koordinaten lexikographisch in vertikaler Richtung, so erhält man als Diskretisierungsmatrix eine Blockdiagonalmatrix mit n Blöcken, welche jeweils die Gestalt von eindimensionalen Diskretisierungsmatrizen auf einem Gitter mit m Punkten besitzen. Die Diskretisierungsmatrix besitzt demnach die Form

$$L_h^{(a=0)} = \begin{pmatrix} A & & & \\ & A & & \\ & & \ddots & \\ & & & A \end{pmatrix} \in \mathbb{R}^{I \times I} \tag{6.3}$$

mit $I = \{1, \ldots, nm\}$. Die Diagonalblöcke sind durch die tridiagonale Diskretisierungsmatrix zum eindimensionalen Laplace-Operator gegeben:

$$A = \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{\{1,\ldots,m\}\times\{1,\ldots,m\}}.$$

Da die Matrix $L_h^{(a=0)}$ in diesem Fall eine Tridiagonalmatrix ist, kann nach [Hac09, Proposition 3.9.1] die Inverse $\left(L_h^{(a=0)}\right)^{-1}$ sogar exakt als $\mathcal{H}$-Matrix dargestellt werden.

Insbesondere enthält die Inverse von Blockdiagonalmatrizen ebenfalls nur Einträge in den Diagonalblöcken, so dass außerhalb der Diagonalblöcke keine Einträge vorhanden sind. In den entsprechenden Blöcken ist folglich die exakte Darstellung mittels einer Matrix vom Rang 0 möglich. Die Einträge innerhalb der Diagonalblöcke sind durch die der Inversen des eindimensionalen Problems (zur tridiagonalen Matrix A) gegeben, für die eine exakte Darstellung als $\mathcal{H}$-Matrix existiert.

Die entsprechende modifizierte Partition $P_{mod}^{a=0}$ für diesen Grenzfall ergibt sich für die Konstellation $d \gg a$. Nach den bisherigen Ausführungen erhält man demnach eine modifizierte Partition von der Struktur, wie sie in Abbildung 6.2(a) dargestellt ist. Diese zeichnen zwei wesentliche Eigenschaften aus: Es liegen große zulässige Blöcke außerhalb der Diagonalen vor und die resultierende interne Anordnung, die sich bei der Berechnung des Clusterbaums ergibt, entspricht der lexikographischen Anordnung in vertikaler Richtung. Beim Einsatz der modifizierten Partition $P_{mod}^{a=0}$ für diese Problemstellung ergibt sich nach der internen Anordnung die Diskretisierungsmatrix aus (6.3), deren Inverse $\left(L_h^{(a=0)}\right)^{-1}$ ebenfalls durch eine Blockdiagonalmatrix gegeben ist. Bei der Approximation der Inversen mittels einer $\mathcal{H}$-Matrix aus der Menge $\mathcal{H}(k, P_{mod}^{a=0})$ treten demnach außerhalb der Diagonalblöcke keine Approximationsfehler auf und es könnten in diesen Blöcken sogar Rang-k-Matrizen mit dem Rang $k = 0$ verwendet werden. Nach diesen Ausführungen ist zu erwarten, dass der Einsatz der modifizierten Partition für den Grenzfall $\kappa \to \infty$ sehr gut geeignet ist und zu kleinen Approximationsfehlern führt.

Die Überlegungen lassen sich analog auf den Fall $a = 1, d = 0$ mit dem Differenzenstern

$$\frac{1}{h^2}\begin{bmatrix} 0 & 0 & 0 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

übertragen, wobei in diesem Fall eine Anordnung der Gitterpunkte entlang der horizontalen Koordinatenachsen zu dem entsprechenden Ergebnis führt. Bei Verwendung der modifizierten Partitionierungsstrategie ergibt sich genau diese Anordnung. Der Einsatz der modifizierten Partitionierungsstrategie kann demnach für die Grenzfälle ebenfalls aufgrund rein theoretischer Überlegungen motiviert werden.

Die Resultate der numerischen Tests unter Verwendung der modifizierten Partitionierungsstrategie werden in Abschnitt 6.1.1 für das zweidimensionale und in Abschnitt 6.1.2 für das dreidimensionale Modellproblem angegeben. Dabei lässt sich insbesondere für den Grenzfall $\kappa \to \infty$ das erwartete Verhalten beobachten.

Erweiterung für variable Schrittweiten

Als Erweiterung der bisher betrachteten Problemstellung können die konstanten Koeffizienten durch ortsabhängige Koeffizienten $a(x_i), d(y_j)$ ersetzt werden. Dies entspricht dem Übergang von festen Schrittweiten h^x, h^y zu variablen Schrittweiten $h^x(x_i), h^y(y_j)$. Die Anpassung der Clusterstrategie kann direkt auf diesen Fall übertragen werden. Sollte $\min(h^x(x_i)) \gg \max(h^y(y_j))$ bzw. $\min(h^y(y_j)) \gg \max(h^x(x_i))$ gelten, dann wird sich unabhängig von den variablen Schrittweiten eine Partition ergeben, bei der im Zuge der Konstruktion des Clusterbaums die Aufteilung der Gitter vorzugsweise in eine Koordinatenrichtung erfolgt. Nur wenn die Schrittweiten ähnliche Größen aufweisen, spiegeln sich variable Schrittweiten in der Partitionierung wieder.

Für variierende Schrittweiten ist zusätzlich zu beachten, dass die Größe h^∞ aus der Bedingung (6.1) nicht direkt ersichtlich ist. Bei der praktischen Durchführung kann eine Mittelung der variablen Schrittweiten erfolgen, die zur Berechnung der Distanz

verwendet werden. Dies stellt insbesondere sicher, dass keine benachbarten Gitter als zulässig gekennzeichnet werden, da bei dieser Vorgehensweise auch für den Fall, dass die Gitter nur einen Gitterpunkt voneinander entfernt sind, weiterhin $\text{dist}_{\infty}^{(ad)}(\tau,\sigma) - h^{\infty} = 0$ gilt.

6.1.1 Ergebnisse zum 2D Modellproblem

In Abschnitt 4.2 konnte gezeigt werden, dass zur Finite-Differenzen-Matrix, die aus dem zweidimensionalen Modellproblem (2.10) auf einem $n \times m$ Rechteckgitter mit dem Differenzenoperator (2.11) hervorgeht, eine $\mathcal{H}$-Matrix Approximation der Inversen existiert. Nach den dort erzielten Ergebnissen nimmt der Fehler für wachsenden Rang k in der Form

$$\epsilon \approx \exp\left(-c^{-\frac{2}{3}} k^{\frac{1}{3}}\right)$$

exponentiell ab, wobei $c = 2\sqrt{2}\sqrt{\kappa}\, C_{caccio}\left(\eta + 2\sqrt{2}\right)\exp(1)$ gilt und demnach der Fehlerverlauf von der Größe

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}$$

abhängt. Dabei sind λ_{min} und λ_{max} durch den kleinsten bzw. größten Koeffizienten des entsprechenden Differenzenoperators auf dem Rechteckgitter gegeben. Für wachsendes κ ist demnach eine Verschlechterung des Fehlerverlaufs zu erwarten.

Unter Berücksichtigung dieses Zusammenhangs werden zwei Klassen von Testproblemen eingeführt, in denen die Koeffizienten des Differenzenoperators (2.11) so gewählt werden, dass sich die Abhängigkeit des Approximationsfehlers von der Größe κ untersuchen lässt. Zunächst werden variable Koeffizienten $a(x_i, y_j), d(x_i, y_j)$ durch zufällige Werte aus dem Intervall $[2^{-2\alpha}, 1], \alpha \geq 0$ vorgegeben. Dabei sollen die Werte $2^{-2\alpha}$ und 1 an mindestens einem Punkt angenommen werden, so dass sich $\kappa = 2^{2\alpha}$ ergibt und bei wachsendem α ebenfalls κ zunimmt. Die zweite Klasse von Testproblemen beschränkt sich auf den Fall konstanter Koeffizienten $a(x_i, y_j) = a$ und $d(x_i, y_j) = d$. Für diese gilt $\kappa = \frac{a}{d}$ für $a \geq d$ bzw. $\kappa = \frac{d}{a}$ für $d \geq a$. Je stärker die Koeffizienten voneinander abweichen, desto größer fällt κ aus.

In Abschnitt 6.1 wurde bereits erwähnt, dass die Auswahl der zweiten Klasse von Testproblemen unter besonderer Berücksichtigung der Gegebenheiten im Modell METRAS erfolgt. In dem Modell können unterschiedliche Schrittweiten im Modellgitter vorgegeben werden. Da das Modellgitter auf ein Rechengitter mit äquidistanter Schrittweite transformiert wird, ergeben sich die entsprechenden Koeffizienten des Differenzenoperators. In Anlehnung an diesen Zusammenhang wurden im Modellproblem die Koeffizienten a und d eingeführt. Wenn die Schrittweiten im Modellgitter stark voneinander abweichen, gilt dies ebenfalls für die entsprechenden Koeffizienten im Differenzenoperator und es ergibt sich $\kappa \gg 1$. Daher findet der Fall von Koeffizienten $a \gg d$ bzw. $d \gg a$ bei den numerischen Tests besondere Berücksichtigung.

Zufällige Koeffizienten

Die ersten numerischen Tests werden für zufällige Koeffizienten durchgeführt, die durch $a(x_i, y_j), d(x_i, y_j) \in [2^{-2\alpha}, 1]$ mit $\alpha \in \{0, 1, 2, 4, 8, 16\}$ gegeben sind. Der Fall $\alpha = 0$ entspricht der Standard-Diskretisierung des Laplace-Operators. In Tabelle 6.1 sind die relativen Approximationsfehler in der Frobenius-Norm zu den Rängen $k = 1, \ldots, 5$ in Abhängigkeit von α angegeben.

Tabelle 6.1: Numerische Ergebnisse für $a(x_i, y_j), d(x_i, y_j) \in [2^{-2\alpha}, 1]$

α	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
0	6.7556e-03	9.4278e-04	4.6858e-05	1.1390e-05	6.5521e-07
1	6.7566e-03	9.5102e-04	4.7324e-05	1.1470e-05	6.5970e-07
2	6.8751e-03	9.6309e-04	4.8761e-05	1.1825e-05	7.0218e-07
4	6.8182e-03	9.6830e-04	4.8883e-05	1.1872e-05	6.9060e-07
8	6.8405e-03	9.6670e-04	4.8506e-05	1.1661e-05	6.5745e-07
16	6.9031e-03	9.7657e-04	4.8995e-05	1.1714e-05	6.7763e-07

In allen Fällen kann wie erwartet ein exponentiell fallender Fehler beobachtet werden. Auffällig ist, dass die Fehler unabhängig von α nahezu identisch sind und nicht, wie nach Satz 4.2.1 zu erwarten, eine Verschlechterung des Fehlerverhaltens für wachsendes α zu beobachten ist. Die Abhängigkeit von der Größe κ kann daher für diese numerischen Testprobleme nicht festgestellt werden.

Da die Koeffizienten in diesem Fall durch zufällige Werte vorgegeben wurden, scheint eine Approximation von Finite-Differenzen-Matrizen zum Modellproblem für eine Vielzahl von Problemstellungen unabhängig von κ Erfolg versprechend. Wie sich anhand der folgenden Testprobleme herausstellen wird, ist es jedoch möglich, die Koeffizienten so zu wählen, dass für wachsendes κ ein deutlich schlechteres Ergebnis zu beobachten ist.

Konstante Koeffizienten $a \neq d$

Die zweite Klasse von Testproblemen ergibt sich durch die Vorgabe von konstanten Koeffizienten $d = 1$ und $a = 2^{-2\alpha}$, $\alpha \geq 0$. Dies entspricht den Schrittweiten $h^y = 1$ und $h^x = 2^\alpha$ und man erhält $\kappa = 2^{2\alpha}$. Bei wachsendem α wird demnach auch κ größer. Bei der Untersuchung der auf diese Weise konstruierten Testprobleme kann bei Verwendung der gewöhnlichen Partitionierung für wachsendes κ ein deutlich schlechteres Fehlerverhalten als für die Testprobleme mit zufällig gewählten Koeffizienten beobachtet werden. Zum Vergleich ist in Abbildung 6.3, exemplarisch für den Fall $\alpha = 8$, der Fehler für konstante Koeffizienten $a = 2^{-16}, d = 1$ und für zufällige Koeffizienten aus dem Intervall $[2^{-16}, 1]$ dargestellt. Obwohl κ unter diesen Gegebenheiten in beiden Problemstellungen übereinstimmt, kann im Vergleich ein deutlich schlechteres Fehlerverhalten für das Problem mit konstanten Koeffizienten $d \gg a$ festgestellt werden.

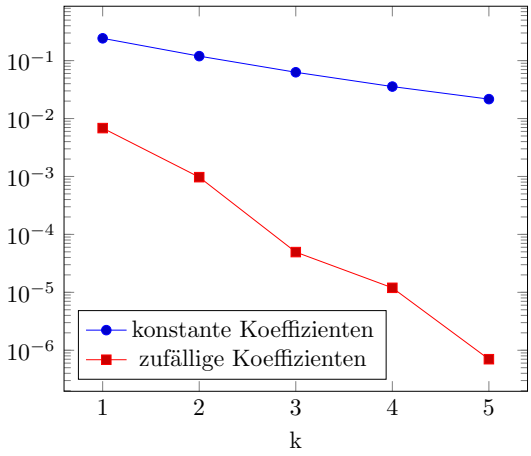


Abbildung 6.3: Relative Fehler für $\kappa = 2^{16}$

Wie bereits in Abschnitt 6.1 erwähnt, ist der Einsatz der $\mathcal{H}$-Matrix-Technik unter Verwendung der gewöhnlichen Partition nach den numerischen Ergebnissen für diese Klasse von Testproblemen im Fall $\kappa \to \infty$ nicht geeignet. Es konnte jedoch eine modifizierte Partitionierungsstrategie angegeben werden, deren Einsatz nach den theoretischen Überlegungen für den Grenzfall $\kappa \to \infty$ zu besseren Ergebnissen führen müsste. Das Ziel der weiteren numerischen Tests ist es, die Eignung der modifizierten Partition P_{mod} für verschiedene Konstellationen von Koeffizienten zu untersuchen. Dazu werden die Approximationsfehler bei Verwendung der gewöhnlichen Partition P und der modifizierten Partition P_{mod} für eine Vielzahl von Testproblemen miteinander verglichen.

In diesem Zusammenhang können zwei besondere Fälle hervorgehoben werden. Zum einen stimmen im Fall $a = d$ die Partitionen P und P_{mod} überein, so dass auch für diesen Fall die Verwendung der modifizierten Partition keine Einschränkung darstellt. Die zweite Besonderheit ergibt sich im Fall $a \gg d$ bzw. $d \gg a$, bei dem die Aufteilung der Gitter mittels geometriebasierter Clusterung vorzugsweise in einer Koordinatenrichtung stattfindet. Dies ergibt eine Partition von der Struktur, wie sie in Abbildung 6.2(a) dargestellt ist. Die Partitionen besitzen für kleine κ demnach eine vergleichbare Struktur, wohingegen für große κ deutliche Unterschiede zu erkennen sind. Daher wird untersucht, ob bzw. in welcher Form sich auch die Approximationsfehler für wachsendes κ verändern.

Zunächst werden für den Fall $\kappa \gg 1$ numerische Tests durchgeführt, um zu untersuchen, wie sich der Einsatz der Partitionen P und P_{mod} für wachsendes κ auswirkt. Die Koeffizienten werden für die Tests durch $a = 2^{-2\alpha}, \alpha \in \{10, 11, 12, 13, 14\}$ und $d = 1$ vorgegeben. Die Approximationsfehler sind in Tabelle 6.2 zu finden. Es bestätigt sich das Ergebnis, dass für wachsendes α bzw. κ eine Verschlechterung des Fehlers bei Verwendung der gewöhnlichen Partition beobachtet werden kann. Im Gegensatz dazu werden die Fehler für wachsendes κ beim Einsatz der modifizierten Partition immer kleiner. Demnach kann für die numerischen Ergebnisse genau das Verhalten beobachtet werden, das nach den theoretischen Überlegungen aus Abschnitt 6.1 zur Eignung der modifizierten Partition im Grenzfall $\kappa \to \infty$ zu erwarten war.

Tabelle 6.2: Numerische Ergebnisse für $a = 2^{-2\alpha}, d = 1$

α	Partition	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
10	P	3.7238e-01	2.6045e-01	1.8572e-01	1.3610e-01	1.0186e-01
	P_{mod}	1.4511e-03	8.3948e-05	8.2636e-06	1.2002e-06	2.3303e-07
11	P	4.2901e-01	3.3669e-01	2.6521e-01	2.1107e-01	1.6900e-01
	P_{mod}	2.1845e-04	8.2912e-06	6.5084e-07	8.3102e-08	1.4919e-08
12	P	4.7701e-01	4.0470e-01	3.4170e-01	2.8814e-01	2.4198e-01
	P_{mod}	2.3095e-05	6.3753e-07	4.3228e-08	5.1133e-09	8.7830e-10
13	P	5.1407e-01	4.5730e-01	4.0347e-01	3.5332e-01	3.0620e-01
	P_{mod}	1.8922e-06	4.2472e-08	2.6454e-09	3.0004e-10	5.0337e-11
14	P	5.3933e-01	4.9256e-01	4.4576e-01	3.9929e-01	3.5277e-01
	P_{mod}	1.3375e-07	2.6614e-09	1.5831e-10	1.7563e-11	2.9109e-12

Die bisherigen numerischen Tests haben ergeben, dass sich die modifizierte Partitionierungsstrategie für die beiden besonderen Konstellationen mit $a = d$ und $a \gg d$ bzw. $d \gg a$ eignet. Ergänzend dazu werden weitere Tests für die Koeffizienten $d \geq a$ mit $a = 2^{-2\alpha}, \alpha \in \{1, 3, 5, 7, 9\}$ und $d = 1$ durchgeführt. Dies ermöglicht eine Beurteilung, ob die Verwendung der resultierenden Partitionen, die sich beim Übergang von den Koeffizienten $a = d$ zu $d \gg a$ ergeben, ebenfalls zu guten Ergebnissen führt. Die Resultate zu diesen Testproblemen sind in Tabelle 6.3 angegeben.

Der Einsatz der gewöhnlichen Partition führt für kleine α zu einem guten Fehlerverlauf, der ein ähnliches Verhalten wie der Fehler bei Verwendung der modifizierten Partition aufweist. Es ist jedoch ebenfalls erkennbar, dass sich bei wachsendem κ der Fehlerverlauf im Fall der gewöhnlichen Partition immer weiter verschlechtert. Der Einsatz der modifizierten Partition führt im Vergleich zur gewöhnlichen Partition zu kleineren Fehlern oder zu Fehlern mit ähnlichen Größenordnungen. Bei Verwendung der modifizierten Partition lässt sich für kleine α ebenfalls eine leichte Verschlechterung des Fehlers für wachsendes α beobachten, bevor für große α die bereits beobachtete deutliche Verbesserung eintritt. Unter Berücksichtigung der günstigeren Struktur der modifizierten Partition ist auch in diesen Fällen die Verwendung der modifizierten Partition von Vorteil.

Die numerischen Tests zu dieser Klasse von Testproblemen haben ergeben, dass die $\mathcal{H}$-Matrix-Technik unter Einsatz der gewöhnlichen Partition für den Fall $\kappa \to \infty$ nicht geeignet ist, da die Approximationsfehler für wachsenden Rang nur leicht abnehmen. Es konnte jedoch eine modifizierte Partition eingeführt werden, durch deren Einsatz insbesondere für Probleme mit $\kappa \to \infty$ deutlich bessere Ergebnisse erzielt wurden. Hinzu kommt, dass die Struktur der modifizierten Partition durch die größeren zulässigen Blöcke gegenüber der gewöhnlichen Partition von Vorteil ist. Aus diesem Grund ist der Einsatz der modifizierten Partition auch für kleine κ vorteilhaft, da die Fehler in diesem

Tabelle 6.3: Numerische Ergebnisse für $a = 2^{-2\alpha}, d = 1$

α	Partition	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
1	P	7.2614e-03	9.6349e-04	6.3805e-05	1.8481e-05	1.1929e-06
	P_{mod}	8.4827e-03	1.8375e-03	1.0217e-04	3.2920e-05	2.7310e-06
3	P	2.3689e-02	2.3872e-03	3.5454e-04	1.2125e-04	2.8317e-05
	P_{mod}	1.9066e-02	2.5446e-03	3.5237e-04	1.4016e-04	2.1855e-05
5	P	7.4596e-02	1.5375e-02	3.7515e-03	1.0750e-03	3.7620e-04
	P_{mod}	3.9795e-02	6.4232e-03	1.3682e-03	3.8861e-04	1.2650e-04
7	P	1.7750e-01	6.8805e-02	2.9353e-02	1.3792e-02	7.1356e-03
	P_{mod}	4.5983e-02	9.6009e-03	2.4479e-03	7.1346e-04	2.3117e-04
9	P	3.0924e-01	1.8528e-01	1.1563e-01	7.5786e-02	5.1825e-02
	P_{mod}	6.4911e-03	6.0025e-04	7.9809e-05	1.4050e-05	3.0949e-06

Fall ähnliche Größenordnungen besitzen.

6.1.2 Ergebnisse zum 3D Modellproblem

Die Finite-Differenzen-Matrix zum dreidimensionalen Modellproblem resultiert aus der Problemstellung (2.10) mit einem $n \times m \times r$ Quadergitter Ω_h. Der Differenzenoperator ist in diesem Fall durch (5.1) mit entsprechenden Koeffizienten $a, d, g \in \mathcal{D}_h\left(\overline{\Omega}_h\right)$ gegeben. Analog zu den zweidimensionalen Testproblemen werden die Koeffizienten so gewählt, dass die Abhängigkeit des Fehlers von der Größe κ untersucht werden kann. Dazu eignen sich ebenfalls die beiden Klassen von Testproblemen aus dem zweidimensionalen Fall.

Für die erste Klasse von Testproblemen werden erneut ortsabhängige Koeffizienten $a(x_i, y_j, z_k)$, $d(x_i, y_j, z_k)$, $g(x_i, y_j, z_k)$ durch zufällige Werte aus dem Intervall $[2^{-2\alpha}, 1]$ mit $\alpha \in \{0, 1, 2, 4, 8, 16\}$ vorgegeben. Die theoretischen Ergebnisse aus Kapitel 5 lassen auch in diesem Fall, unabhängig von α, exponentielle Konvergenz erwarten. Erneut ist nach den theoretischen Resultaten für wachsendes α bzw. κ mit einer Verschlechterung des Fehlerverlaufs zu rechnen. Die Ergebnisse zu diesen ersten numerischen Tests für die dreidimensionale Problemstellung sind in Tabelle 6.4 angegeben.

Bei allen Testproblemen ist wie erwartet exponentielle Konvergenz zu beobachten. Wie im zweidimensionalen Fall kann darüber hinaus keine Verschlechterung der Ergebnisse für wachsendes α bzw. κ beobachtet werden. Demnach lassen sich unabhängig von κ für diese Klasse von Testproblemen gute Ergebnisse erzielen.

Analog zum zweidimensionalen Problem wird eine zweite Klasse von Testproblemen mit konstanten Koeffizienten $a(x_i, y_j, z_k) = a, d(x_i, y_j, z_k) = d$ und $g(x_i, y_j, z_k) = g$ untersucht, die erneut im Zusammenhang mit der Einführung unterschiedlicher Schrittweiten h^x, h^y und h^z interpretiert werden können. Im Hinblick auf das Modell METRAS,

Tabelle 6.4: Numerische Ergebnisse für $a(x_i, y_j, z_k), d(x_i, y_j, z_k), g(x_i, y_j, z_k) \in [2^{-2\alpha}, 1]$

α	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
0	7.9871e-03	2.9828e-03	8.2446e-04	2.8279e-04	1.6883e-04
1	7.9552e-03	2.9770e-03	8.2224e-04	2.8376e-04	1.6829e-04
2	7.9484e-03	2.9831e-03	8.3468e-04	2.8693e-04	1.7128e-04
4	7.9275e-03	2.9676e-03	8.3378e-04	2.8546e-04	1.7017e-04
8	7.9276e-03	2.9427e-03	8.3077e-04	2.8598e-04	1.7107e-04
16	7.9041e-03	2.9900e-03	8.3338e-04	2.8674e-04	1.6954e-04

für das häufig $h^x \approx h^y$ und entweder $h^x \approx h^z$ oder $h^x \gg h^z$gilt, werden für diese Klasse von Testproblemen Koeffizienten mit $a = d$ und $g \geq a$ vorgegeben. Der Fall $a = d = g = 1$ entspricht erneut der Standard-Diskretisierung des dreidimensionalen Laplace-Operators.

Wie für das zweidimensionale Modellproblem können bei Verwendung der gewöhnlichen Partition für diese Klasse von Testproblemen für $\kappa \to \infty$ deutlich schlechtere numerische Ergebnisse als im Fall zufälliger Koeffizienten beobachtet werden. Die Überlegungen und Modifikationen der Partitionierungsstrategie aus Abschnitt 6.1, die im zweidimensionalen Fall zu einer Verbesserung der Ergebnisse führten, können direkt auf den dreidimensionalen Fall übertragen werden. Die Clusterstrategie und die Berechnung der Größen $\operatorname{diam}(\tau), \operatorname{diam}(\sigma)$ und $\operatorname{dist}(\tau, \sigma)$ lassen sich analog für die dreidimensionale Problemstellung so anpassen, dass die unterschiedlichen Schrittweiten des Modellgitters berücksichtigt werden. Die modifizierte Zulässigkeitsbedingung ergibt sich demnach zu

$$\min\{\operatorname{diam}^{(adg)}(\tau), \operatorname{diam}^{(adg)}(\sigma)\} \leq \eta \left(\operatorname{dist}_{\infty}^{(adg)}(\tau, \sigma) - h^{\infty} \right). \tag{6.4}$$

Auch in diesem Fall wird die Zulässigkeitsbedingung um die Überprüfung von

$$\frac{\operatorname{dist}_{\infty}(\tau, \sigma)}{h^{\infty}} \geq \delta_0 \tag{6.5}$$

erweitert, die sicherstellt, dass bei stark voneinander abweichenden Schrittweiten der Abstand der zulässigen Gitter nicht weniger als δ_0 Gitterpunkte beträgt.

Zunächst wird wie im zweidimensionalen Fall untersucht, ob der Einsatz der auf diese Weise konstruierten modifizierten Partition P_{mod} für den Fall $g \gg a = d$ zu besseren Ergebnissen führt. Dazu werden die Koeffizienten $g = 1$ und $a = d = 2^{-2\alpha}$ mit $\alpha \in \{10, 11, 12, 13, 14\}$ gewählt. Die Ergebnisse sind in Tabelle 6.5 angegeben.

Erneut ist zu erkennen, dass der Einsatz der gewöhnlichen Partition zu einem schlechten Fehlerverlauf führt und sich durch die Verwendung der modifizierten Partition deutlich bessere Ergebnisse erzielen lassen, für die wie im zweidimensionalen Fall eine Verbesserung des Fehlers für wachsendes κ zu beobachten ist. Der Fehlerverlauf weist jedoch

Tabelle 6.5: Numerische Ergebnisse für $g = 1, a = d = 2^{-2\alpha}$

α	Partition	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
10	P	2.8557e-01	2.6990e-01	2.5346e-01	2.3952e-01	2.2583e-01
	P_{mod}	2.1467e-05	1.2219e-06	6.4845e-07	4.3312e-09	3.1347e-09
11	P	3.0675e-01	2.9329e-01	2.7922e-01	2.6616e-01	2.5296e-01
	P_{mod}	2.8482e-06	9.8248e-08	5.9623e-08	2.2363e-10	1.6096e-10
12	P	3.1919e-01	3.0698e-01	2.9427e-01	2.8184e-01	2.6908e-01
	P_{mod}	2.8018e-07	5.4477e-09	3.5617e-09	1.2619e-11	8.9608e-12
13	P	3.2601e-01	3.1446e-01	3.0248e-01	2.9042e-01	2.7795e-01
	P_{mod}	2.2551e-08	2.3626e-10	1.6062e-10	7.5562e-13	5.3404e-13
14	P	3.2959e-01	3.1839e-01	3.0678e-01	2.9492e-01	2.8262e-01
	P_{mod}	1.6120e-09	8.9944e-12	6.2395e-12	4.6437e-14	3.2816e-14

bei Verwendung von P_{mod} bei wachsendem Rang k ein „sprunghaftes“ Verhalten auf. Zum Teil ist bei der Vergrößerung des Rangs nur eine vergleichsweise kleine Verbesserung des Fehlers und in anderen Fällen eine Reduzierung des Fehlers um einen Faktor in der Größenordnung 10^{-2} zu erkennen. Die Ursache dieses Phänomens konnte nicht näher bestimmt werden. Trotz dieses „unregelmäßigen“ Abfalls des Fehlers ergeben sich für die Partition P_{mod} deutlich bessere Ergebnisse als für die gewöhnliche Partition. Demnach ist auch in diesem Fall der Einsatz der modifizierten Partition von Vorteil.

Im dreidimensionalen Fall ist ebenfalls zu überprüfen, ob die Verwendung der modifizierten Partitionierungsstrategie bei der Konstellation von Koeffizienten geeignet ist, die sich für den Übergang von dem Fall $a = d = g$ zu $a = d, g \gg a$ ergibt. Dazu werden weitere Tests mit $a = d = 2^{-2\alpha}, g = 1$ mit $\alpha \in \{1, 3, 5, 7, 9\}$ durchgeführt. Die Ergebnisse zu diesen Testproblemen sind in Tabelle 6.6 zu finden.

Erneut lässt sich beobachten, dass die Verwendung der modifizierten Partition von Vorteil ist, je größer die Schrittweitenunterschiede ausfallen. Für kleines α liefert auch die gewöhnliche Partition akzeptable Ergebnisse, allerdings verschlechtert sich der Fehlerverlauf für wachsendes α deutlich. Für diese Fälle ist der Einsatz der gewöhnlichen Partition demnach nicht geeignet. Auch bei diesen Testproblemen konnte teilweise ein sprunghafter Verlauf der Fehler bei Verwendung der Partition P_{mod} beobachtet werden.

Vergleicht man die Fehler aus Tabelle 6.6 für $\alpha = 1$, dann fällt auf, dass die Fehler beim Einsatz der modifizierten Partition im Vergleich zur gewöhnlichen Partition leicht größer ausfallen. Bei der Durchführung weiterer numerischer Tests mit kleinem κ zu anderen Problemgrößen (insbesondere für stark unterschiedliche Größen von n, m und r) ließ sich dieses Verhalten ebenfalls beobachten, wobei der Unterschied der Fehler zum Teil deutlich größer ausfiel. In allen Fällen lieferte jedoch auch der Einsatz der modifizierten Partition akzeptable Resultate, so dass deren Verwendung weiterhin unproblematisch

Tabelle 6.6: Numerische Ergebnisse für $g = 1, a = d = 2^{-2\alpha}$

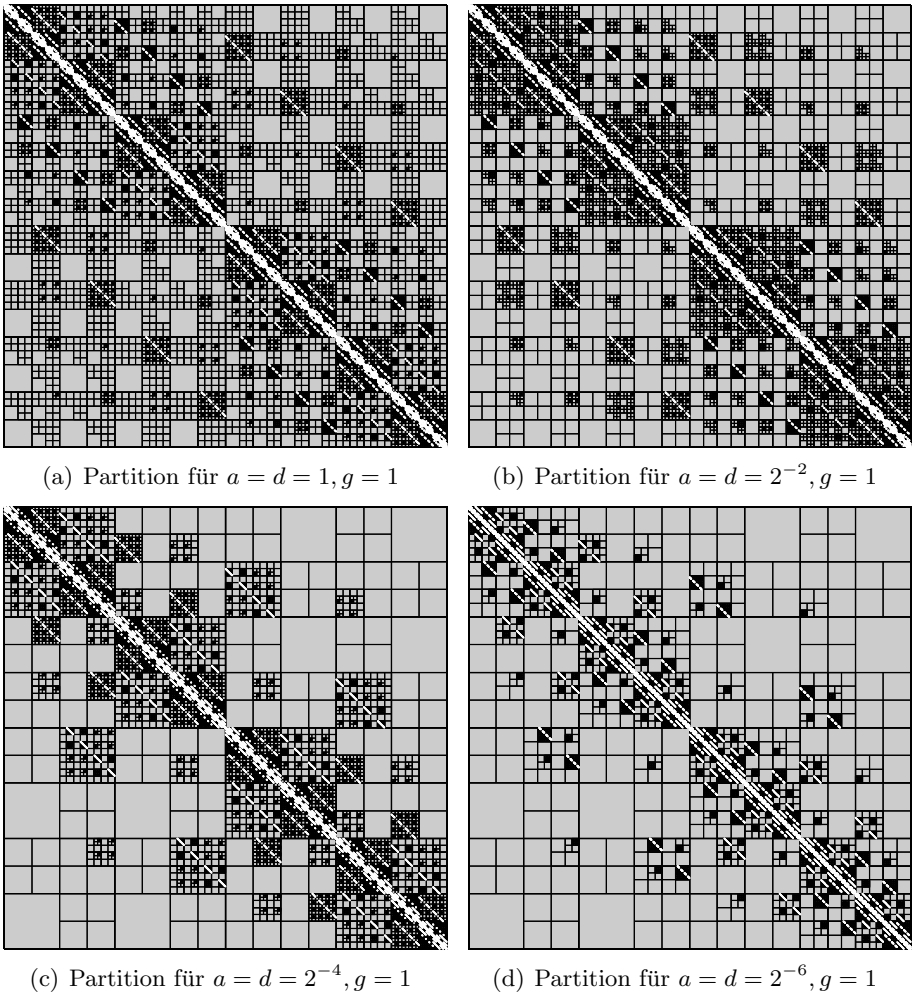
α	Partition	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
1	P	6.4558e-03	3.8051e-03	8.5521e-04	2.5419e-04	1.3651e-04
	P_{mod}	1.2363e-02	6.3897e-03	2.2417e-03	5.8685e-04	3.6267e-04
3	P	2.0802e-02	1.2324e-02	3.6100e-03	2.3317e-03	1.5119e-03
	P_{mod}	1.2102e-02	6.9034e-03	1.6669e-03	9.6278e-04	5.5996e-04
5	P	6.8833e-02	4.7651e-02	2.6884e-02	2.1094e-02	1.6898e-02
	P_{mod}	7.2032e-03	2.9651e-03	9.4208e-04	5.8793e-04	2.0095e-04
7	P	1.5887e-01	1.3252e-01	1.0578e-01	9.3063e-02	8.2526e-02
	P_{mod}	1.1517e-03	2.2576e-04	6.0428e-05	2.3025e-05	1.0798e-05
9	P	2.5297e-01	2.3385e-01	2.1384e-01	1.9911e-01	1.8530e-01
	P_{mod}	1.1171e-04	9.5504e-06	4.0978e-06	8.7629e-08	6.0154e-08

ist. Zu einem tieferen Verständnis dieser Problematik bei kleinem κ sind jedoch weitere theoretische und numerische Untersuchungen erforderlich, die zur Einführung weiterer Verbesserungen der Partitionierungsstrategie in diesem Fall beitragen könnten. Diese Problematik wird an dieser Stelle nicht weiter untersucht.

Um auch im dreidimensionalen Fall die Struktur der resultierenden Partitionen darzustellen, sind in Abbildung 6.4 exemplarisch für $n = m = r = 16$ die modifizierten Partitionen angegeben, die sich beim Übergang von den Koeffizienten $a = d = g = 1$ zu $a = d = 2^{-6}, g = 1$ ergeben. Wie im zweidimensionalen Fall entspricht die modifizierte Partition zunächst der gewöhnlichen und zeichnet sich für wachsendes κ dadurch aus, dass die zulässigen Blöcke im Vergleich deutlich größer ausfallen. Berücksichtigt man, dass auch für kleine α die Fehler beim Einsatz der beiden Partitionen vergleichbar sind, so ist die Verwendung der modifizierten Partition in allen Fällen von Vorteil.

Wie für das zweidimensionale Modellproblem ergeben die numerischen Tests im dreidimensionalen Fall ebenfalls, dass sich die Approximationsfehler bei der $\mathcal{H}$-Matrix Approximation der Inversen unter Verwendung der gewöhnlichen Partition in Abhängigkeit von den Koeffizienten zum Teil stark verschlechtern können. Die Einführung der modifizierten Partitionierungsstrategie ermöglicht es jedoch, deutlich bessere Ergebnisse zu erzielen. Darüber hinaus ist deren Verwendung durch die größeren zulässigen Blöcke gegenüber der gewöhnlichen Partition von Vorteil. Daher ist nach den numerischen Ergebnissen auch im dreidimensionalen Fall die Verwendung der modifizierten Partitionierungsstrategie für Problemstellungen dieser Art zu empfehlen.

Aufgrund dieser vielversprechenden Ergebnisse werden in Abschnitt 6.2.4 Testmatrizen eingeführt, die sich bei der Verwendung des Differenzenoperators aus dem Modell METRAS in Kombination mit Dirichlet Randwerten ergeben. Die Überlegungen für konstante Koeffizienten a, d, g lassen sich wie im zweidimensionalen Fall direkt für variable



(a) Partition für $a = d = 1, g = 1$

(b) Partition für $a = d = 2^{-2}, g = 1$

(c) Partition für $a = d = 2^{-4}, g = 1$

(d) Partition für $a = d = 2^{-6}, g = 1$

Abbildung 6.4: Modifizierte Partitionen für $n = m = r = 16$

Koeffizienten $a(x_i), d(y_j), g(z_k)$ übertragen. Weitere Einzelheiten, die für die speziellen Gegebenheiten im Modell METRAS berücksichtigt werden müssen, sind in Abschnitt 6.2.4 angegeben.

6.2 Numerische Ergebnisse zu weiteren Testproblemen

6.2.1 Konvektions-Diffusionsgleichung

Der Diskretisierung der Konvektions-Diffusionsgleichung mittels Finiter-Differenzen liegt das zweidimensionale Randwertproblem

$$\begin{aligned} -\epsilon \Delta u + b \cdot \nabla u &= f && \text{in } \Omega \\ u &= 0 && \text{auf } \partial\Omega \end{aligned}$$

für ein Rechteck Ω mit der Konvektionsrichtung $b = (b_1, b_2), b_1, b_2 \in \mathbb{R}, \|b\| = 1$ und $\epsilon > 0$ zugrunde. Führt man ein Rechteckgitter Ω_h mit konstanter Schrittweite $h > 0$ ein, dann lässt sich zur Diskretisierung mittels Finiter-Differenzen der Differenzenstern

$$\frac{1}{h^2}\begin{bmatrix} & -\epsilon & \\ -\epsilon & 4\epsilon & -\epsilon \\ & -\epsilon & \end{bmatrix} + \frac{1}{h}\begin{bmatrix} & b_2^- & \\ -b_1^+ & |b_1| + |b_2| & b_1^- \\ & -b_2^+ & \end{bmatrix} \tag{6.6}$$

mit $b_i^+ := \max\{0, b_i\}$ und $b_i^- := \min\{0, b_i\}, i \in \{1, 2\}$ verwenden, um für alle $\epsilon > 0$ eine M-Matrix zu erhalten (vgl. [Hac86, Bemerkung 10.2.6]). Nach Elimination der Randwerte resultiert daraus eine dünnbesetzte, im Allgemeinen nicht symmetrische und nicht wohlkonditionierte Diskretisierungsmatrix.

Zur Anwendbarkeit der $\mathcal{H}$-Matrix-Technik im Zusammenhang mit der Konvektions-Diffusionsgleichung sind bereits Ergebnisse für den Fall von Finite-Element-Diskretisierungen bekannt. In [LB03] wurde die Approximierbarkeit der Inversen von Finite-Element-Matrizen durch eine $\mathcal{H}$-Matrix bei konstanter Konvektion für den konvektionsdominanten Fall ($\epsilon \to 0$) untersucht. Die grundlegenden Überlegungen erfolgen dort für eine Diskretisierung der Problemstellung mit der Konvektionsrichtung $b = (1, 0)$. In diesem Fall stimmt die Diskretisierungsmatrix, die sich aus der dort angegebenen Finite-Element-Diskretisierung ergibt, mit der Matrix überein, die aus der Finite-Differenzen-Diskretisierung unter Verwendung des Differenzensterns (6.6) resultiert.

Zur Approximation der Inversen dieser Matrix mittels einer $\mathcal{H}$-Matrix wurde in [LB03] festgestellt, dass sich die Ergebnisse für $\epsilon \to 0$ zum Teil deutlich verschlechtern. Die Einführung einer modifizierten Strategie zur Partitionierung und einer alternativen Zulässigkeitsbedingung, die unter Berücksichtigung von b und ϵ formuliert wurden, ermöglichte die Konstruktion von Partitionen, deren Einsatz zu besseren Ergebnissen führte.

Da die Finite-Differenzen-Matrix insbesondere für die Konvektionsrichtung $b = (1, 0)$ mit der Finite-Element-Matrix übereinstimmt, sollte sich auch in diesem Fall die Verwendung der gewöhnlichen Partition als ungeeignet herausstellen. Die Ergebnisse numerischer Tests zur Finite-Differenzen-Matrix mit $b = (1, 0)$ und $\epsilon \to 0$ sind in Tabelle 6.7 angegeben. Wie erwartet ist eine deutliche Verschlechterung der Ergebnisse für $\epsilon \to 0$ festzustellen, so dass auch für diese Problemstellung die Einführung einer modifizierten Partitionierung erforderlich ist.

Bei der Durchführung numerischer Tests ließen sich für die Konvektionsrichtungen $b = (1, 0)$ und $b = (0, 1)$ für $\epsilon \to 0$ besonders schlechte Ergebnisse beobachten. Da sich in diesen Fällen die gleichen Matrizen wie für die in [LB03] betrachtete Finite-Element-Diskretisierung ergeben, können die dort erzielten Ergebnisse auch für Finite-Differenzen-Matrizen genutzt werden. Vergleicht man die modifizierten Partitionen aus [LB03] mit den in Abschnitt 6.1 eingeführten Partitionen P_{mod}, so lassen sich einige Übereinstimmungen feststellen.

Für $b = (1, 0)$ und $\epsilon \to 0$ wird die Partition aus [LB03] gebildet, indem unter Berücksichtigung der Konvektionsrichtung die Aufteilung der Cluster parallel zur x-Achse erfolgt. Dies stimmt mit der Strategie zur Bildung der modifizierten Partition aus Abschnitt 6.1 für die Koeffizienten $a \gg d$ überein, wobei in diesem Fall die unterschiedlichen

Tabelle 6.7: Numerische Ergebnisse für $b = (1, 0), \epsilon = 2^{-\alpha}$

α	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
0	1.4259e-01	2.6752e-02	4.5568e-03	7.9139e-04	1.6523e-04
1	2.2407e-01	6.7123e-02	1.7766e-02	4.3070e-03	1.0214e-03
3	4.0173e-01	2.2990e-01	1.2388e-01	6.3537e-02	3.1120e-02
5	5.4242e-01	4.2757e-01	3.2865e-01	2.4827e-01	1.8513e-01
7	6.6987e-01	6.1007e-01	5.5156e-01	4.9571e-01	4.4322e-01

Schrittweiten $h^x \ll h^y$ zu dieser Art der Aufteilung führen. Analog erhält man für die Konvektionsrichtung $b = (0, 1)$ und $\epsilon \to 0$ die gleiche Partition wie für $a \ll d$, bei der die Aufteilung der Cluster stets parallel zur y-Achse vorgenommen wird.

Aufgrund der Übereinstimmung der Modifikationen aus [LB03] mit denen aus Abschnitt 6.1 für die oben beschriebenen Konstellationen bietet es sich an, die Strategie aus Abschnitt 6.1 so zu modifizieren, dass sie auch für den Fall der Konvektions-Diffusionsgleichung verwendet werden kann. An Stelle der Berücksichtigung der unterschiedlichen Schrittweiten bzw. der Koeffizienten a, d des Differenzenoperators ist für die Konvektions-Diffusionsgleichung die Berücksichtigung der Größen b und ϵ erforderlich. Definiert man zur Durchführung der modifizierten Partitionierungsstrategie für diesen Anwendungsfall anstatt der Koeffizienten a, d die neuen Koeffizienten

$$\hat{a} := (|b_1| + \epsilon)^2 \text{ und } \hat{d} := (|b_2| + \epsilon)^2, \tag{6.7}$$

so kann sie direkt zur Konstruktion einer modifizierten Partition $\hat{P}_{mod}$ für die Finite-Differenzen-Matrix zur Konvektions-Diffusionsgleichung genutzt werden. Insbesondere ergibt sich durch diese Konstruktion eine Partition, die für $\epsilon \to 0$ im Fall $b = (1, 0)$ wie gewünscht der Konstellation $\hat{a} \gg \hat{d}$ und für $b = (0, 1)$ analog der Konstellation $\hat{d} \gg \hat{a}$ entspricht. Nach den Ergebnissen aus [LB03] sollte diese Strategie insbesondere für die Fälle $b = (1, 0), b = (0, 1)$ und $\epsilon \to 0$ gute Ergebnisse liefern. Darüber hinaus kann die Partitionierungsstrategie für beliebige andere Konvektionsrichtungen b angewendet werden, wobei die Eignung der resultierenden Partitionen in diesen Fällen noch mittels numerischer Tests zu überprüfen ist.

Es sei darauf hingewiesen, dass sich die Überlegungen ebenfalls auf den Fall übertragen lassen, bei dem statt der Standard-Diskretisierung des Laplace-Operators ein allgemeiner Differenzenoperator vom Typ (2.11) aus dem Modellproblem mit variablen Koeffizienten verwendet wird. Für die Durchführung der modifizierten Partitionierung können unter Berücksichtigung der Koeffizienten a und d aus dem Differenzenoperator die angepassten Koeffizienten

$$\tilde{a} := \left(|b_1| + \epsilon\sqrt{a}\right)^2 \text{ und } \tilde{d} = \left(|b_2| + \epsilon\sqrt{d}\right)^2$$

gewählt werden. Für $b = (0, 0)$ und $\epsilon = 1$ entspricht diese Vorgehensweise der in Abschnitt 6.1 und für $\epsilon \to 0$ dominieren wiederum die Größen b_1 und b_2, so dass auch in

diesem Fall das gewünschte Verhalten erzielt wird.

In Tabelle 6.8 sind Ergebnisse zur Approximation der Inversen für verschiedene Konvektionsrichtungen b angegeben. Die Fehler sind dabei für die gewöhnliche Partitionierung P und die modifizierte Partitionierung $\hat{P}_{mod}$ zu den neuen Koeffizienten $\hat{a}, \hat{d}$ aus (6.7) und $\epsilon = 2^{-\alpha}, \alpha \in \{1, 3, 5\}$ angegeben. Für den Spezialfall $b = \frac{1}{\sqrt{2}}(1, 1)$ gilt für die Koeffizienten $\hat{a} = \hat{d}$, so dass die modifizierte mit der gewöhnlichen Partition übereinstimmt, weshalb jeweils nur ein Ergebnis angegeben ist.

Tabelle 6.8: Numerische Ergebnisse für $\epsilon = 2^{-\alpha}$

$b \cdot \|b\|$	α	Part.	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
$(1,0)$	1	P	2.2407e-01	6.7123e-02	1.7766e-02	4.3070e-03	1.0214e-03
		$\hat{P}_{mod}$	5.1750e-02	5.1608e-03	7.1105e-04	1.2155e-04	1.8787e-05
	3	P	4.0173e-01	2.2990e-01	1.2388e-01	6.3537e-02	3.1120e-02
		$\hat{P}_{mod}$	3.8143e-02	3.2235e-03	4.4471e-04	7.8558e-05	1.1355e-05
	5	P	5.4242e-01	4.2757e-01	3.2865e-01	2.4827e-01	1.8513e-01
		$\hat{P}_{mod}$	1.0711e-02	1.1964e-03	1.5703e-04	2.3692e-05	4.5261e-06
$(0,1)$	1	P	1.5269e-01	3.2910e-02	6.8860e-03	1.4501e-03	3.5106e-04
		$\hat{P}_{mod}$	5.1750e-02	5.1608e-03	7.1105e-04	1.2155e-04	1.8787e-05
	3	P	3.2523e-01	1.4802e-01	6.5067e-02	2.8089e-02	1.1738e-02
		$\hat{P}_{mod}$	3.8143e-02	3.2235e-03	4.4471e-04	7.8558e-05	1.1355e-05
	5	P	4.9503e-01	3.5065e-01	2.4002e-01	1.6204e-01	1.0903e-01
		$\hat{P}_{mod}$	1.0711e-02	1.1964e-03	1.5703e-04	2.3692e-05	4.5261e-06
$(1,1)$	1	P	9.1041e-02	1.3550e-02	1.9986e-03	2.6960e-04	3.6205e-05
	3	P	1.3128e-01	2.8230e-02	6.0215e-03	1.1887e-03	2.1713e-04
	5	P	1.4840e-01	3.6344e-02	8.8799e-03	2.0335e-03	4.2892e-04
$(-1,2)$	1	P	9.5473e-02	1.5750e-02	2.6811e-03	4.7094e-04	7.7774e-05
		$\hat{P}_{mod}$	9.9716e-02	1.6439e-02	2.7892e-03	4.8243e-04	7.9321e-05
	3	P	1.4536e-01	3.5266e-02	8.7332e-03	2.1999e-03	5.4277e-04
		$\hat{P}_{mod}$	1.4943e-01	3.6058e-02	8.8715e-03	2.2182e-03	5.4510e-04
	5	P	1.6851e-01	4.7040e-02	1.3364e-02	3.8557e-03	1.1002e-03
		$\hat{P}_{mod}$	1.4252e-01	2.8496e-02	5.4025e-03	9.5291e-04	1.5347e-04

Die numerischen Ergebnisse aus Tabelle 6.8 zeigen, dass durch den Einsatz der modifizierten Partition mit den Koeffizienten $\hat{a}, \hat{d}$ für alle Testprobleme vergleichbare oder bessere Ergebnisse als beim Einsatz der gewöhnlichen Partition erzielt werden können. Insbesondere ist für die problematischen Konvektionsrichtungen $b = (1, 0)$ und $b = (0, 1)$

für $\epsilon \to 0$ eine Verschlechterung der Ergebnisse bei Verwendung der gewöhnlichen Partition und eine Verbesserung im Fall der modifizierten Partition zu beobachten. Außerdem sollte, wie schon in Abschnitt 6.1 festgestellt, berücksichtigt werden, dass die modifizierte Partitionierungsstrategie zu deutlich größeren zulässigen Blöcken führt. Dies ist für die Speicheranforderungen und die Durchführung der Operationen in der $\mathcal{H}$-Arithmetik von Vorteil, so dass der Einsatz der modifizierten Partition auch bei vergleichbaren Fehlern vorzuziehen ist.

Allerdings lässt sich ebenfalls beobachten, dass sich die Ergebnisse für die Konvektionsrichtung $b = \frac{1}{\sqrt{5}}(-1, 2)$, die in [LB03] als problematisch eingestuft wurde, auch bei Verwendung der gewöhnlichen Partition nur in einem geringeren Maße verschlechtern, als es dort festgestellt wurde. Dieses unterschiedliche Verhalten könnte darauf zurückgeführt werden, dass sich die Diskretisierungsmatrizen, die aus der Finite-Element-Diskretisierung in [LB03] und der oben beschriebenen Finite-Differenzen-Diskretisierung hervorgehen, beim Auftreten anderer Konvektionsrichtungen voneinander unterscheiden. Obwohl der Einsatz der gewöhnlichen Partition auch in diesem Fall zunächst gute Ergebnisse liefert, fällt der Fehler bei Verwendung der modifizierten Partition für den Fall $\epsilon \to 0$ erneut kleiner aus. Die Unterschiede sind jedoch deutlich geringer als für die Konvektionsrichtungen $b = (1, 0)$ und $b = (0, 1)$.

Eine Besonderheit tritt bei der Konvektionsrichtung $b = \frac{1}{\sqrt{2}}(1, 1)$ auf. Wie bereits erwähnt, stimmen in diesem Fall die Partitionen P und $\hat{P}_{mod}$ überein. Sollte sich demnach die gewöhnliche Partition als ungeeignet herausstellen, könnte auch durch den Einsatz der modifizierten Partition keine Verbesserung erzielt werden. Die numerischen Tests zeigen jedoch, dass bei der Verwendung der gewöhnlichen Partition ein guter Fehlerverlauf beobachtet werden kann. Insgesamt ist nur eine leichte Verschlechterung der Ergebnisse für $\epsilon \to 0$ zu beobachten. Dieses Verhalten konnte ebenfalls bei der Durchführung weiterer Tests mit deutlich kleineren Werten für ϵ festgestellt werden, bei denen die Approximationsfehler im Vergleich zum Ergebnis aus Tabelle 6.8 mit $\epsilon = 2^{-5}$ nahezu unverändert blieben.

Insgesamt können durch die Verwendung der modifizierten Partition aus Abschnitt 6.1 mit den angepassten Koeffizienten aus (6.7) gute Ergebnisse erzielt werden. Zusätzlich zu den Tests mit den Konvektionsrichtungen aus Tabelle 6.8 wurden weitere Problemstellungen mit einer Vielzahl anderer Konvektionsrichtungen untersucht, die alle vergleichbare Ergebnisse lieferten. Den numerischen Resultaten zufolge scheint der Einsatz der $\mathcal{H}$-Matrix-Technik auch im Zusammenhang mit der Diskretisierung der Konvektions-Diffusionsgleichung mittels Finiter-Differenzen-Verfahren geeignet zu sein. Erneut sollte jedoch die gewöhnliche Partitionierungsstrategie durch die oben angegebenen Modifikationen ergänzt werden, um für beliebige Konvektionsrichtungen, insbesondere für den konvektionsdominanten Fall, eine Verschlechterung der Approximation zu verhindern.

6.2.2 Neumann-Randwerte

Neben der Verwendung anderer Differenzenoperatoren wirkt sich auch die Einführung einer alternativen Randbedingung im Modellproblem auf die resultierende Diskretisierungsmatrix aus. Im Hinblick auf die Gegebenheiten im Modell METRAS wird deshalb

numerisch untersucht, ob die Verwendung von Neumann- statt Dirichlet-Randwerten die Ergebnisse zur Approximation der Inversen durch eine $\mathcal{H}$-Matrix negativ beeinflusst.

Beim Auftreten von Neumann-Randwerten ist, im Gegensatz zu Dirichlet-Randwerten, auch eine Diskretisierung der Ableitung in der Randbedingung erforderlich. Die Normalableitungen am Rand werden im folgenden Testproblem durch einseitige Diskretisierungen mittels Rückwärtsdifferenzen vorgenommen. Außerdem beschränkt sich die Untersuchung erneut auf ein $n \times m$ Rechteckgitter $\Omega_h = \{(x_i, y_j) \in h\mathbb{Z}^2 : 1 \leq i \leq n, 1 \leq j \leq m\}$. Das zugrunde liegende diskrete Randwertproblem ist durch

$$\begin{aligned} -\Delta_h u &= f && \text{in } \Omega_h \\ u_{\bar{n}} &= \varphi && \text{auf } \partial'\Omega_h \end{aligned}$$

mit $\partial'\Omega_h := \partial\Omega_h \backslash \{(0,0), (0,(m+1)h), ((n+1)h, 0), ((n+1)h, (m+1)h)\}$, $f \in \mathcal{D}_h(\Omega_h)$ und $\varphi \in \mathcal{D}_h(\partial'\Omega_h)$ gegeben. Auch in diesem Fall lassen sich die Randpunkte eliminieren, so dass man ein Gleichungssystem der Form

$$L_h u = q_h \tag{6.8}$$

mit $L_h \in \mathbb{R}^{I \times I}, q_h \in \mathbb{R}^I$ und $I = \{1, \ldots, nm\}$ erhält.

Die Matrix L_h ist singulär, so dass das Gleichungssystem im Allgemeinen nicht lösbar ist. Es lässt sich jedoch analog zur kontinuierlichen Problemstellung eine Lösbarkeitsbedingung, unter Verwendung der Kurzschreibweise $\mathbb{1} = (1, 1, \ldots, 1)^T$, angeben:

Satz 6.2.1 ([Hac86, Satz 4.7.3]) *Das Gleichungssystem* (6.8) *ist genau dann lösbar, wenn*

$$-h^2 \sum_{x \in \Omega_h} f(x) = h \sum_{x \in \partial'\Omega_h} \varphi(x) \tag{6.9}$$

gilt. Je zwei Lösungen von (6.8) *können sich nur um eine Konstante unterscheiden:* $u^1 - u^2 = c\mathbb{1}, c \in \mathbb{R}$.

Um im Fall der Lösbarkeit von (6.8) eine eindeutige Lösung zu erhalten, ist es erforderlich, eine Normierung der Lösung einzuführen. Dazu kann an einem Gitterpunkt x_0 der Wert $u(x_0) = 0$ gesetzt werden. Dies entspricht dem Streichen einer Zeile und Spalte des Gleichungssystems. Das resultierende Gleichungssystem ist lösbar und es ergibt sich eine symmetrische M-Matrix ([Hac86, Satz 4.7.4]). Alternativ kann eine Normierung der Lösung in der Form $\sum_{x \in \Omega_h} u(x) = \sigma$ mit $\sigma \in \mathbb{R}$ vorgegeben werden. Dieser Ansatz führt zu dem erweiterten Gleichungssystem

$$\hat{L}_h \hat{u} = \hat{q}_h$$
$$\text{mit } \hat{L}_h = \begin{pmatrix} L_h & \mathbb{1} \\ \mathbb{1}^T & 0 \end{pmatrix}, \hat{u} = \begin{pmatrix} u \\ \lambda \end{pmatrix} \text{ und } \hat{q}_h = \begin{pmatrix} q_h \\ \sigma \end{pmatrix}.$$

Die Matrix $\hat{L}_h \in \mathbb{R}^{\hat{I} \times \hat{I}}$ mit $\hat{I} = \{1, \ldots, nm+1\}$ ist regulär. Im Fall von $\lambda = 0$ ist u Lösung von (6.8), erfüllt die Lösbarkeitsbedingung (6.9) und die Normierung ist durch $\sum_{x \in \Omega_h} u(x) = \sigma$ gegeben. Für $\lambda \neq 0$ kann u als Lösung von (6.8) zur korrigierten rechten

Seite $\tilde{q}_h = q_h - \lambda \mathbb{1}$ interpretiert werden, wobei in diesem Fall für $\tilde{f}(x) := f(x) - \lambda$ und φ die Lösbarkeitsbedingung (6.9) erfüllt ist ([Hac86, Satz 4.7.5]).

Im Modell METRAS resultiert die Matrix aus einer Problemstellung mit homogenen Neumann-Randwerten und es ist zusätzlich eine Normierung der Lösung vorgegeben. Daher wird als Diskretisierungsmatrix des Testproblems die erweiterte Matrix $\hat{L}_h$ verwendet, so dass man eine reguläre Matrix $\hat{L}_h \in \mathbb{R}^{\hat{I} \times \hat{I}}$ erhält.

Die Partitionierung kann grundsätzlich nach der gleichen Strategie wie bei Dirichlet-Randwerten erfolgen, es ist jedoch zu beachten, dass die Gleichung zum Index $nm + 1$ aus der Erweiterung des Gleichungssystems stammt. Dementsprechend lässt sich diesem Index kein Gitterpunkt zuordnen. Dies wäre jedoch für die gewöhnliche geometriebasierte Konstruktion des Clusterbaums erforderlich. Daher wird im ersten Schritt der Konstruktion des Clusterbaums die Indexmenge aufgeteilt in die zu Gitterpunkten gehörigen Indizes und den zusätzlichen Index, der aus der Erweiterung des Gleichungssystems resultiert.

Im Hinblick auf die praktische Berechnung der $\mathcal{H}$-Inversen in der $\mathcal{H}$-Arithmetik sollte diese Vorgehensweise zusätzlich angepasst werden. Denn um die Durchführbarkeit garantieren zu können, müssten alle Hauptuntermatrizen invertierbar sein. Wie bereits beschrieben, ist jedoch die Matrix $\hat{L}_h|_{I' \times I'}$ mit $I' := \hat{I} \backslash \{nm + 1\}$ singulär. Um sicherzustellen, dass die $\mathcal{H}$-Invertierung durchführbar ist, kann die Partition so angepasst werden, dass zusätzlich zum Index $nm + 1$ ein weiterer Index im ersten Schritt der Konstruktion des Clusterbaums separiert wird, bevor die übliche geometriebasierte Konstruktion für die $nm - 1$ übrigen Indizes beginnt. Die Blöcke zu den separierten Indizes werden als nicht zulässig gekennzeichnet, so dass sie durch vollbesetzte Matrizen dargestellt werden. Da die aus diesem Ansatz resultierende Matrix $\hat{L}_h|_{I'' \times I''}$ mit $I'' := \hat{I} \backslash \{nm + 1, \gamma\}$ mit einem Index $\gamma \in \{1, \ldots, nm\}$ symmetrisch und irreduzibel diagonaldominant mit positiven Diagonalelementen ist, gilt nach Kriterium 2.1.4, dass sie positiv definit ist. Demnach kann mittels dieser Partitionierungsstrategie die Durchführbarkeit der $\mathcal{H}$-Invertierung garantiert werden.

Die Berechnung der Partition zur Matrix $\hat{L}_h|_{I'' \times I''}$ mit der Indexmenge I'' kann wie gewöhnlich erfolgen, da allen Indizes aus I'' ein Gitterpunkt aus Ω_h zugeordnet werden kann. Die Berechnung des Approximationsfehlers erfolgt wie üblich auf Grundlage der Partitionierung und der Approximation durch Rang-k-Matrizen in den zulässigen Blöcken.

Um einen Vergleich der Problemstellungen mit Dirichlet- und Neumann-Randwerten zu ermöglichen, wird für die numerischen Tests analog zum Modellproblem die diskrete Problemstellung

$$\begin{aligned} -Lu &= f && \text{in } \Omega_h \\ u_{\bar{n}} &= 0 && \text{auf } \partial' \Omega_h \end{aligned}$$

mit einem allgemeinen Differenzenoperator L vom Typ (2.11) und den Koeffizienten $a, d \in \mathcal{D}_h\left(\overline{\Omega}_h\right)$, $a, d > 0$ betrachtet. Wie bei den Tests zum Modellproblem werden die Konstellationen mit zufällig vorgegebenen Koeffizienten $a(x_i, y_j), d(x_i, y_j) \in [2^{-2\alpha}, 1]$, $\alpha \geq 0$ und konstanten voneinander abweichenden Koeffizienten $d \gg a$ bzw. $a \gg d$

gewählt, um das Fehlerverhalten bei wachsendem κ beurteilen zu können. Die modifizierte Partitionierungsstrategie aus Abschnitt 6.1 kann für den Fall konstanter Koeffizienten direkt übernommen werden, so dass Ergebnisse zu diesem Problem sowohl bei Verwendung der gewöhnlichen Partition P als auch beim Einsatz der modifizierten Partition P_{mod} angegeben werden.

Wie im Fall von Dirichlet-Randwerten werden zunächst variable Koeffizienten durch zufällige Werte aus dem Intervall $[2^{-2\alpha}, 1]$ vorgegeben. Die Ergebnisse zu den numerischen Tests mit $\alpha \in \{0, 1, 2, 4, 8, 16\}$ sind Tabelle 6.9 zu entnehmen. Ebenso wie im Fall von Dirichlet-Randwerten kann unabhängig von der Größe $\kappa = 2^{2\alpha}$ für alle Testprobleme exponentielle Konvergenz festgestellt werden. Die Fehler weisen ähnliche Größenordnungen wie für das Modellproblem mit Dirichlet-Randwerten auf. Insbesondere lässt sich auch bei der Vorgabe von Neumann-Randwerten für diese Klasse von Testproblemen nicht beobachten, dass sich die Fehlerverläufe mit wachsendem κ verschlechtern.

Tabelle 6.9: Numerische Ergebnisse für $a, d \in [2^{-2\alpha}, 1]$

α	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
0	2.8360e-02	4.6910e-03	1.2991e-04	4.6656e-05	1.6526e-06
1	2.8357e-02	4.6840e-03	1.3043e-04	4.6671e-05	1.6654e-06
2	2.8508e-02	4.7480e-03	1.3238e-04	4.7104e-05	1.7046e-06
4	2.8604e-02	4.7739e-03	1.3537e-04	4.7991e-05	1.7519e-06
8	2.8592e-02	4.7615e-03	1.3163e-04	4.6451e-05	1.6694e-06
16	2.8533e-02	4.7347e-03	1.3283e-04	4.6706e-05	1.6951e-06

Treten stark voneinander abweichende konstante Koeffizienten $a \gg d$ bzw. $d \gg a$ auf, kann jedoch bei Verwendung der gewöhnlichen Partition wie im Modellproblem eine Verschlechterung der Ergebnisse beobachtet werden. Die Ergebnisse der Tests mit konstanten Koeffizienten $d = 1$ und $a = 2^{-2\alpha}$ mit $\alpha \in \{1, 3, 5, 7, 9\}$ sind in Tabelle 6.10 bei Verwendung der gewöhnlichen Partition P und der modifizierten Partition P_{mod} dargestellt.

Die Ergebnisse unterscheiden sich nur leicht von denen zum Modellproblem mit Dirichlet-Randwerten. Die Verwendung der modifizierten Partitionierung führt erneut für $\kappa \to \infty$ zu einer deutlich besseren Approximation als bei Verwendung der gewöhnlichen Partition. Doch auch in den Fällen, bei denen die Fehler ähnliche Größenordnungen besitzen, ist der Einsatz der modifizierten Partition von Vorteil, weil die zulässigen Blöcke im Vergleich zur gewöhnlichen Partitionierung größer ausfallen. Daher ist nach den numerischen Resultaten auch beim Auftreten von Neumann-Randwerten die Verwendung der modifizierten Partition zu empfehlen.

Da bei den numerischen Tests ähnliche Ergebnisse wie bei Dirichlet-Randwerten beobachtet werden konnten, deuten die Resultate darauf hin, dass sich die $\mathcal{H}$-Matrix-Technik ebenfalls im Zusammenhang mit Finite-Differenzen-Matrizen einsetzen lässt, die sich aus

Tabelle 6.10: Numerische Ergebnisse für $d = 1, a = 2^{-2\alpha}$

α	Partition	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
1	P	2.4934e-02	4.2741e-03	1.3871e-04	5.5251e-05	2.4651e-06
	P_{mod}	2.5442e-02	5.1249e-03	1.9181e-04	6.1665e-05	3.4543e-06
3	P	2.4827e-02	1.1857e-03	2.8477e-04	1.1160e-04	1.7007e-05
	P_{mod}	2.0692e-02	9.2080e-03	4.9626e-04	8.3466e-05	1.3883e-05
5	P	4.0133e-02	2.7855e-03	5.4480e-04	1.2635e-04	4.8929e-05
	P_{mod}	3.2111e-02	4.0108e-03	7.1976e-04	2.2008e-04	5.7792e-05
7	P	4.9848e-02	6.8877e-03	2.1017e-03	7.1024e-04	2.9478e-04
	P_{mod}	2.8088e-02	1.5396e-03	3.6001e-04	8.4057e-05	2.1482e-05
9	P	5.4457e-02	1.0619e-02	4.4143e-03	2.0847e-03	1.1271e-03
	P_{mod}	2.6782e-02	2.8255e-04	1.4349e-05	1.0310e-06	1.3553e-07

der Diskretisierung des Modellproblems mit Neumann-Randwerten ergeben.

6.2.3 Wärmeleitungsgleichung

Die Diskretisierungsmatrix des Modellproblems tritt ebenfalls im Kontext der Diskretisierung des zweidimensionalen Anfangs-Randwertproblems für die Wärmeleitungsgleichung auf, das durch

$$\begin{aligned} u_t - \Delta u &= 0 && \text{in } \Omega \times (0,T) \\ u &= 0 && \text{auf } \partial\Omega \times (0,T) \\ u &= u_0 && \text{für } t = 0 \end{aligned}$$

mit einem Rechteck Ω und $T > 0$ gegeben ist.

Zur Diskretisierung mittels Finiter-Differenzen wird wie gewöhnlich ein zweidimensionales Rechteckgitter mit konstanter Gitterweite h eingeführt, das um eine Diskretisierung in der Zeit mit der Zeitschrittweite q ergänzt wird.

Die Diskretisierung der Ortsableitungen kann wie bekannt mit der Standard-Diskretisierung des Laplace-Operators erfolgen. Eine Erweiterung für allgemeinere Differenzenoperatoren vom Typ (2.11) ist auch in diesem Fall unproblematisch. Mit der Diskretisierungsmatrix L_h, die sich aus der Diskretisierung der Ortsableitungen ergibt, kann als allgemeiner Ansatz zur Diskretisierung der vollständigen Gleichung die θ-Methode angegeben werden. Diese ergibt sich mittels einseitiger Diskretisierung der zeitlichen Ableitung und Mittelung der Ortsableitungen zu dem neuen und alten Zeitlevel (gekennzeichnet mittels $p+1$ und p) durch

$$\frac{u^{p+1} - u^p}{q} = \theta L_h u^{p+1} + (1-\theta) L_h u^p$$

mit $0 \leq \theta \leq 1$. Dieser Ansatz erfordert in jedem Zeitschritt die Lösung des Gleichungssystems

$$(I - q\theta L_h)\, u^{p+1} = (I + q(1-\theta)L_h)\, u^p$$

mit der Systemmatrix $(I - q\theta L_h)$. Diese Vorschrift beinhaltet für $\theta = 0$ ein explizites Verfahren, für die anderen Fälle ergeben sich implizite Verfahren. Für $\theta = \frac{1}{2}$ erhält man das Crank-Nicolson-Verfahren, das zur Diskretisierungsmatrix $A_h = \left(I - q\frac{1}{2}L_h\right)$ führt, welche die Grundlage der folgenden Untersuchungen bildet.

Die Einführung impliziter Methoden erfolgt, um die Verwendung großer Zeitschrittweiten zu ermöglichen. Daher werden zunächst numerische Tests zur Approximation der Inversen von A_h durch eine $\mathcal{H}$-Matrix für die Standard-Diskretisierung des Laplace-Operators ($a = d = 1$) unter Verwendung unterschiedlicher Zeitschrittweiten durchgeführt. Dabei wird das Verhältnis $\lambda = \frac{q}{h^2} = 2^\alpha$ mit $\alpha \in \{-1, 1, 3, 5, 7, 10, 20\}$ gewählt, um die Auswirkung wachsender Zeitschrittweiten beurteilen zu können. Die Ergebnisse zu diesen Tests sind in Tabelle 6.11 zu finden.

Tabelle 6.11: Numerische Ergebnisse für $\theta = \frac{1}{2}$ und $\lambda = 2^\alpha$

α	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
-1	1.1152e-14	3.5894e-15	9.1199e-16	2.0119e-16	4.2653e-17
1	5.0624e-09	7.4073e-10	8.0543e-11	7.7440e-12	8.6955e-13
3	9.0244e-06	6.8127e-07	6.0290e-08	8.2132e-09	1.5656e-09
5	4.0688e-04	2.9723e-05	3.2651e-06	6.9022e-07	6.9723e-08
7	2.5140e-03	2.5905e-04	2.2169e-05	4.8973e-06	4.2175e-07
10	6.4272e-03	8.5537e-04	5.1520e-05	1.2268e-05	8.0891e-07
20	6.7574e-03	9.4301e-04	4.6882e-05	1.1396e-05	6.5566e-07

Die resultierenden Approximationsfehler, die bei den numerischen Tests auftreten, fallen insbesondere im Vergleich zu den vorherigen Problemstellungen bei allen Tests sehr klein aus. Selbst für $\lambda = 2^{20}$ erhält man eine sehr gute Approximation und je kleiner λ wird, desto kleiner werden auch die Fehler. Die Ergebnisse dieser Tests deuten darauf hin, dass sich die $\mathcal{H}$-Matrix-Technik zur Lösung der resultierenden Gleichungssysteme unabhängig von den gewählten Schrittweiten eignet. Auch die Verwendung der allgemeinen θ-Methode für $\theta \neq \frac{1}{2}$ sollte unter diesen Voraussetzungen zu guten Ergebnissen führen.

Ersetzt man die Standard-Diskretisierung des Laplace-Operators durch einen allgemeinen Differenzenoperator der Form (2.11), so stellt sich die Frage, ob für $\kappa \to \infty$ wie beim Modellproblem eine Verschlechterung der Ergebnisse auftritt. Daher werden zusätzliche Tests für $d = 1$ und $a = 2^{-2\alpha}, \alpha \in \{1, 3, 5, 7, 9\}$ durchgeführt, um zu untersuchen, ob auch im Zusammenhang mit der Diskretisierungsmatrix zur Wärmeleitungsgleichung eine Verschlechterung der numerischen Ergebnisse zu beobachten ist. Analog zum Modellproblem kann in Abhängigkeit der Koeffizienten auch in diesem Fall eine modifizierte

Partition P_{mod} berechnet werden. Die Ergebnisse der Tests sind in Tabelle 6.12 für die Verwendung der gewöhnlichen Partition P und der modifizierten Partition P_{mod} mit $\lambda = 2^{10}$ zusammengefasst.

Tabelle 6.12: Numerische Ergebnisse für $d = 1, a = 2^{-2\alpha}$ und $\lambda = 2^{10}$

α	Partition	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
1	P	6.9857e-03	8.5213e-04	7.1590e-05	2.1060e-05	1.5407e-06
	P_{mod}	9.1945e-03	1.8437e-03	1.2928e-04	4.0371e-05	3.8107e-06
3	P	2.2884e-02	2.6033e-03	4.1170e-04	1.2399e-04	3.6649e-05
	P_{mod}	2.1934e-02	2.9548e-03	4.6776e-04	1.8722e-04	3.5388e-05
5	P	6.7199e-02	1.7054e-02	4.7168e-03	1.4746e-03	5.4762e-04
	P_{mod}	3.8628e-02	8.2944e-03	1.9497e-03	5.2558e-04	1.7401e-04
7	P	1.4009e-01	6.7942e-02	3.3951e-02	1.8216e-02	1.0620e-02
	P_{mod}	3.4685e-02	1.0847e-02	3.4095e-03	1.1263e-03	3.9616e-04
9	P	2.1653e-01	1.5741e-01	1.1422e-01	8.4184e-02	6.2888e-02
	P_{mod}	2.7265e-03	4.8187e-04	9.0716e-05	1.9624e-05	4.9255e-06

Wie beim Modellproblem ist ebenso bei der Diskretisierung der Wärmeleitungsgleichung zu beobachten, dass sich die Ergebnisse bei Verwendung der gewöhnlichen Partition für $\kappa \to \infty$ deutlich verschlechtern, so dass sich die $\mathcal{H}$-Matrix-Technik unter diesen Gegebenheiten nicht zur Lösung der resultierenden Gleichungssysteme eignet. Durch den Einsatz der modifizierten Partition können jedoch auch für diese Testprobleme deutlich bessere Ergebnisse erzielt werden. Da wie in den vorangegangenen Tests darüber hinaus die modifizierte Partition größere zulässige Blöcke als die gewöhnliche Partition aufweist, sollte bei der Diskretisierung der Wärmeleitungsgleichung beim Auftreten eines allgemeinen Differenzenoperators vom Typ (2.11) stets die modifizierte Partitionierungsstrategie eingesetzt werden.

Die Ergebnisse der numerischen Tests deuten demnach darauf hin, dass sich die $\mathcal{H}$-Matrix-Technik auch zur Lösung der resultierenden Gleichungssysteme bei der Diskretisierung der Wärmeleitungsgleichung mittels Finiter-Differenzen-Verfahren einsetzen lässt. Es sollten jedoch wie bei den vorherigen Testproblemen die Modifikationen der Partitionierungsstrategie aus Abschnitt 6.1 berücksichtigt werden, um eine Verschlechterung der Ergebnisse bei Verwendung der gewöhnlichen Partitionierungsstrategie zu vermeiden.

6.2.4 Koeffizienten aus dem Modell METRAS

In Abschnitt 6.1.2 wurde für das dreidimensionale Modellproblem eine modifizierte Partitionierungsstrategie beschrieben, welche im Hinblick auf die besonderen Eigenschaf-

ten des Gleichungssystems im Modell METRAS eingeführt wurde. Auf Grundlage der dort erzielten Ergebnisse werden weitere numerische Tests für die speziellen Finite-Differenzen-Matrizen durchgeführt, die sich bei der Verwendung des Differenzenoperators bzw. des Differenzensterns aus dem Modell METRAS ergeben. Die Einträge des Differenzensterns sind in [SBL$^+$96] angegeben. Für die numerischen Tests werden im Unterschied zu der Konstellation im Modell METRAS Dirichlet- statt Neumann-Randwerte vorgegeben, um die in Abschnitt 6.2.2 beschriebene Problematik im Zusammenhang mit Neumann-Randwerten zu umgehen. Nach Elimination der Randpunkte erhält man eine dünnbesetzte Diskretisierungsmatrix mit maximal 15 Einträgen pro Zeile.

Die spezielle Struktur der Diskretisierungsmatrix resultiert aus den Besonderheiten im Modell METRAS, die bereits bei der Einführung des Modellproblems in Abschnitt 2.1.1 dargestellt wurden. Die Einträge der Matrix sind im Wesentlichen von der Gestalt des zugrunde liegenden Oberflächenprofils und von den variablen Schrittweiten im Modellgitter abhängig. Da dieser Zusammenhang zwischen den Koeffizienten des Differenzenoperators und dem Modellgitter bekannt ist, sollte nach den Ergebnissen in Abschnitt 6.1.2 zum dreidimensionalen Modellproblem die Partitionierung unter Berücksichtigung der Koeffizienten bzw. der Größen im Modellgitter erfolgen.

Bei der Anpassung der Partitionierungsstrategie ist zu beachten, dass den Berechnungen in diesem Fall kein Quadergitter wie im Modellproblem zugrunde liegt, sondern im Modell METRAS ein bodenfolgendes, krummliniges und in vertikaler Richtung nicht orthogonales Koordinatensystem eingesetzt wird. Durch die Verwendung dieses speziellen Koordinatensystems variiert der Abstand zweier Koordinatenebenen für konstante η-Koordinaten in Abhängigkeit von der x- und y-Koordinate bzw. in Abhängigkeit von dem Bodenprofil $z_s(x, y)$ (vgl. Abbildung 2.2, in der exemplarisch Koordinatenebenen für konstante η-Koordinate angegeben sind). Dies führt dazu, dass auch die Gitterweite in z-Richtung von den x- und y-Koordinaten abhängt.

Darüber hinaus ist in Bodennähe meist eine höhere Auflösung des Modells als am oberen Rand erwünscht, so dass die Gitterweiten in z-Richtung häufig so gewählt werden, dass sie in Bodennähe sehr klein sind und sich mit zunehmender Höhe deutlich vergrößern. Diese beiden Besonderheiten müssen bei der Konstruktion der Partition berücksichtigt werden.

Die wesentlichen Ideen zur Modifizierung der Partitionierungsstrategie, die in Abschnitt 6.1 angegeben wurden, lassen sich jedoch für diese Problemstellung übernehmen. So wird zur geometriebasierten Konstruktion des Clusterbaums und zur Auswertung der Größen $\operatorname{diam}(\tau), \operatorname{diam}(\sigma)$ und $\operatorname{dist}_\infty(\tau, \sigma)$ aus der Zulässigkeitsbedingung die Struktur des Modellgitters berücksichtigt. Dabei ist zu beachten, dass die resultierenden Teilgitter in diesem Fall nicht mehr durch achsenparallele Quader gegeben sind, sondern krummlinige Berandungen besitzen können. In horizontaler Richtung ist das verwendete Koordinatensystem orthogonal, so dass zur Berechnung der Größen in x- und y-Richtung keine Anpassungen vorgenommen werden müssen. Eine Strategie zur Berücksichtigung variabler Gitterweiten in diesen Richtungen wurde bereits in Abschnitt 6.1 angegeben.

In vertikaler Richtung muss die Partitionierungsstrategie jedoch an die spezifischen Gegebenheiten angepasst werden. Dazu sind sowohl die (geometriebasierte) Berechnung des Clusterbaums als auch die Bestimmung der Größen $\operatorname{diam}(\tau), \operatorname{diam}(\sigma)$ und $\operatorname{dist}_\infty(\tau, \sigma)$

zur Auswertung der Zulässigkeitsbedingung zu modifizieren.

Bei der geometriebasierten Konstruktion des Clusterbaums wird nach den Ausführungen in Abschnitt 3.2.1 das vorliegende Gitter in Richtung der maximalen Ausdehnung in zwei Teilgitter aufgeteilt. Bei der Bestimmung der maximalen Ausdehnung in z-Richtung ist dabei zu beachten, dass die Schrittweiten in dieser Richtung von den x- und y-Koordinaten abhängen können. Aus diesem Grund wird als maximale Ausdehnung des Gitters in z-Richtung das arithmetische Mittel über alle Abstände in z-Richtung im entsprechenden Teilgitter verwendet.

Sollte sich herausstellen, dass die Aufteilung des Gitters in z-Richtung erforderlich ist, muss zusätzlich berücksichtigt werden, dass sich die Gitterweiten in dieser Richtung mit zunehmender Höhe stark vergrößern. Würde die Aufteilung des Gitters in diesem Fall geometriebasiert durch Halbierung des z-Achsenabschnitts und anschließender Zuordnung der Gitterpunkte zu den entsprechenden Teilgittern erfolgen, wäre es möglich, dass sich die Anzahl der Gitterpunkte der resultierenden Teilgitter stark voneinander unterscheidet. Dies könnte folglich zu unbalancierten Bäumen führen. Daher erfolgt die Aufteilung in diesem Fall nicht geometriebasiert, sondern kardinalitätsbasiert. Dazu wird nicht die Strecke in z-Richtung, sondern die Anzahl der Gitterpunkte in z-Richtung unabhängig von den Schrittweiten halbiert. Auf diese Weise ergeben sich zwei Teilgitter mit ungefähr gleich vielen Gitterpunkten. Diese Anpassungen ermöglichen die Berechnung eines balancierten Clusterbaums unter Berücksichtigung der Gegebenheiten im Modell METRAS.

Eine ähnliche Problematik, bei der die geometriebasierte Konstruktion des Clusterbaums nicht gleichzeitig zu einem kardinalitätsbasierten Clusterbaum führt, wird in [HK00a] beschrieben. Dort erfolgen Überlegungen für den Einsatz der $\mathcal{H}$-Matrix-Technik im Zusammenhang mit graduierten Gittern und es wird ebenfalls eine modifizierte Partitionierungsstrategie eingeführt.

Neben der Bestimmung des Clusterbaums muss auch die Berechnung der Größen $\operatorname{diam}(\tau), \operatorname{diam}(\sigma)$ und $\operatorname{dist}_\infty(\tau, \sigma)$ im Zusammenhang mit der Auswertung der Zulässigkeitsbedingung angepasst werden. Zur Berechnung der Durchmesser $\operatorname{diam}(\tau)$ und $\operatorname{diam}(\sigma)$ ist ebenfalls die Bestimmung des Abstands in vertikaler Richtung erforderlich. Dazu wird, wie oben beschrieben, über die (ortsabhängigen) Abstände in z-Richtung gemittelt. Bei der Berechnung von $\operatorname{dist}_\infty(\tau, \sigma)$ ist erneut lediglich die Bestimmung der Größe in z-Richtung anzupassen. Die Verwendung von achsenparallelen Bounding-Boxes ist in diesem Fall aufgrund der krummlinigen Berandungen der Teilgitter problematisch. Daher wird auch die Bestimmung des Abstands in z-Richtung als Mittelwert der Abstände bestimmt, wobei sich die Berechnung auf die durch τ und σ vorgegebenen horizontalen Gitterbereiche beschränkt und nicht für das gesamte Gitter ausgeführt wird.

Unter Berücksichtigung dieser Modifikationen kann die Konstruktion einer modifizierten Partition P_{MET} in Abhängigkeit von den speziellen Gegebenheiten im Modell METRAS erfolgen. Die beschriebene Vorgehensweise führt wie im Modellproblem dazu, dass die Struktur der resultierenden Partitionen von den Koeffizienten bzw. von dem zugrunde liegenden Oberflächenprofil und den vorgegebenen Schrittweiten im Gitter abhängt. Für die numerischen Tests werden daher verschiedene Oberflächenprofile und variierende Schrittweiten in z-Richtung vorgegeben. Unter Berücksichtigung der Konstellation im

Modell METRAS, für die meist $h^x \approx h^y$ gilt, werden die Schrittweiten in horizontaler Richtung stets durch $h^x = h^y$ gewählt. Für die Schrittweiten in z-Richtung werden die beiden Fälle $h^x \approx h^z$ oder $h^x \gg h^z$ unterschieden, wobei in Bodennähe kleine Schrittweiten vorgegeben werden, die sich mit zunehmender Höhe deutlich vergrößern.

Analog zu den numerischen Tests zum Modellproblem in Abschnitt 6.1.2 wird die Approximation auf Grundlage der gewöhnlichen Partition P (unabhängig von den Koeffizienten) sowie zum Vergleich unter Verwendung der modifizierten Partition P_{MET} berechnet, bei deren Konstruktion die Struktur des Modellgitters berücksichtigt wird. Sollte sich ein ähnliches Verhalten wie im Modellproblem ergeben, dann sind im Fall $h^x \approx h^y \approx h^z$ vergleichbare Ergebnisse für beide Partitionen zu erwarten. Für $h^x \approx h^y$ und $h^x \gg h^z$ sollte die Verwendung der Partition P_{MET} im Vergleich zur gewöhnlichen Partition P von Vorteil sein.

Das erste Testproblem wird für den Fall $h^x = h^y \approx h^z$ aufgestellt und ist in Anlehnung an ein Testproblem aus [Sch07] gewählt. Das Oberflächenprofil ist in allgemeiner Form durch

$$z_s^B(x,y) := H\frac{L^2}{L^2 + (x-x_c)^2 + (y-y_c)^2} \tag{6.10}$$

gegeben und kann als Oberfläche eines einzelnen Bergs angesehen werden. Das Modellgebiet ist hierbei in der Form $-620 \leq x, y \leq 620$ und $0 \leq z \leq 2000$ gegeben und es wird das Oberflächenprofil (6.10) mit $H = 100, L = 1000$ und $x_c = y_c = 0$ verwendet. Mit der Wahl $h^x = h^y = 40$ erhält man ein horizontales 32×32 Gitter. In z-Richtung werden ebenfalls insgesamt 32 Gitterpunkte vorgegeben, wobei die Schrittweite von $h^z \approx 13$ an der Oberfläche bis zu $h^z \approx 122$ am oberen Rand des Modellgebiets zunimmt.

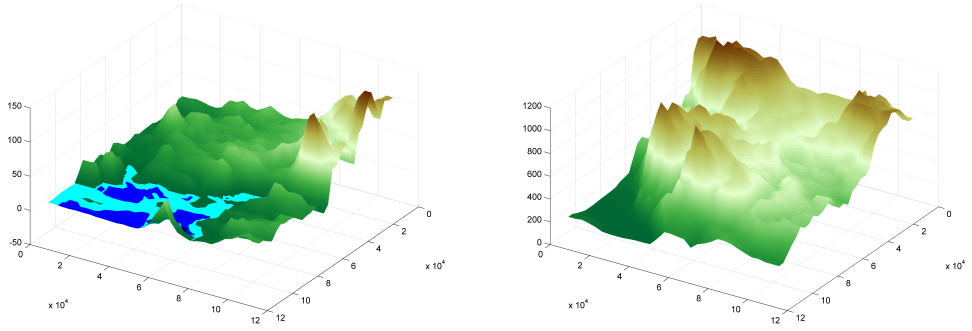
Die allgemeine Form des Oberflächenprofils für das zweite Testproblem ist aus [SV84] für den Fall $h^x \gg h^z$ übernommen und durch

$$z_s^{GK}(x,y) := \frac{D - B\cos\left(\frac{2\pi y}{C}\right)}{1 + \frac{x^2}{A^2}} \tag{6.11}$$

gegeben, was einer gestreckten Gebirgskette entspricht. Das Modellgebiet ist in diesem Fall durch $-64000 \leq x, y \leq 64000$ und $0 \leq z \leq 12000$ vorgegeben und die Parameter im Oberflächenprofil (6.11) werden durch $A = 15000, B = 800, C = 128000$ und $D = 3000$ gewählt. In horizontaler Richtung ergibt sich durch die Vorgabe der Gitterweite $h^x = h^y = 4000$ erneut ein 32×32 Gitter und zwischen den 32 Gitterpunkten in z-Richtung nimmt die Gitterweite von $h^z \approx 75$ in Bodennähe bis $h^z \approx 710$ am oberen Rand des Gitters zu.

Zusätzlich zu diesen beiden Beispielen werden zwei Testprobleme eingeführt, deren Oberflächenprofile sich aus SRTM-Daten ergeben. Die Oberflächenprofile z_s^{EM} und z_s^{SA} entsprechen dem Gebiet der Elbmündung und einem Teilgebiet der Schwäbischen Alb, wobei zur Bildung der Profile die Datensätze N53E009 (z_s^{EM}) und N48E008 (z_s^{SA}) verwendet wurden. Dazu sind die Daten auf ein horizontales 32×32 Gitter reduziert worden, für das sich eine Schrittweite von $h^x = h^y \approx 3666$ ergibt. In vertikaler Richtung wurde erneut für $0 \leq z \leq 20000$ ein Gitter mit 32 Gitterpunkten und Schrittweiten von $h^z \approx 130$

bis $h^z \approx 1250$ verwendet. In Abbildung 6.5 sind (interpolierte) Darstellungen der Oberflächenprofile z_s^{EM} und z_s^{SA} angegeben. Die Ergebnisse zu den vier unterschiedlichen numerischen Testproblemen sind in Tabelle 6.13 dargestellt.



(a) z_s^{EM} (N53E009) (b) z_s^{SA} (N48E008)

Abbildung 6.5: Oberflächenprofile

Tabelle 6.13: Numerische Ergebnisse für die Koeffizienten aus dem Modell METRAS

z_s	Partition	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
z_s^B	P	2.2255e-02	5.6821e-03	2.1550e-03	1.2186e-03	7.5303e-04
	P_{MET}	6.1646e-03	3.1518e-03	8.0194e-04	2.4780e-04	1.3079e-04
z_s^{GK}	P	1.2742e-01	1.0916e-01	9.1009e-02	8.1807e-02	7.3936e-02
	P_{MET}	4.2378e-04	6.5190e-05	2.1243e-05	1.0851e-05	4.6439e-06
z_s^{SA}	P	6.8303e-02	5.3306e-02	3.9234e-02	3.4088e-02	2.9878e-02
	P_{MET}	2.5383e-03	4.2078e-04	2.0789e-04	6.2089e-05	3.2813e-05
z_s^{EM}	P	6.6371e-02	5.1625e-02	3.7778e-02	3.2781e-02	2.8690e-02
	P_{MET}	2.6494e-03	4.6007e-04	2.2701e-04	6.6925e-05	3.4235e-05

Die Fehlerverläufe in der Tabelle zeigen, dass auch für die numerischen Testprobleme, die sich auf Grundlage der Koeffizienten aus dem Modell METRAS ergeben, insgesamt gute Ergebnisse erzielt werden können. Wie erwartet, kann für das Problem mit dem Oberflächenprofil z_s^B, für das $h^x = h^y \approx h^z$ gilt, festgestellt werden, dass sich die Fehler bei Verwendung der Partition P nur leicht im Vergleich zur Partition P_{MET} unterscheiden. Die Partition P_{MET} liefert jedoch auch für diese Konstellation ein besseres Ergebnis. Bei den anderen Problemstellungen, für die $h^x = h^y \gg h^z$ gilt, können ähnliche Ergebnisse wie beim dreidimensionalen Modellproblem beobachtet werden. Der Einsatz der gewöhnlichen Partition führt bei diesen Testproblemen zu deutlich schlechteren Er-

gebnissen als die Verwendung der modifizierten Partition P_{MET}. Zusätzlich sind die zulässigen Blöcke der Partition P_{MET}, genau wie bei der Partition P_{mod} aus Abschnitt 6.1.2, größer als die der gewöhnlichen Partition. Daher ist für diese Testprobleme die Verwendung der Partition P_{MET} stets von Vorteil.

Die numerischen Ergebnisse zu diesen Testproblemen lassen noch kein abschließendes Urteil darüber zu, ob sich die $\mathcal{H}$-Matrix-Technik zur Lösung der linearen Gleichungssysteme im Modell METRAS eignet. Es konnte jedoch eine modifizierte Partitionierungsstrategie speziell für diese Problemstellung angegeben werden, deren Einsatz für die konkreten Testprobleme sehr gute Ergebnisse lieferte. Die numerischen Ergebnisse deuten demnach darauf hin, dass auch bei der Verwendung des Differenzensterns aus dem Modell METRAS eine $\mathcal{H}$-Matrix Approximation der Inversen existiert. Berücksichtigt man zusätzlich die positiven Ergebnisse aus Abschnitt 6.2.2, die für das Modellproblem mit Neumann-Randwerten erzielt wurden, lässt sich vermuten, dass auch für die linearen Gleichungssysteme im Modell METRAS eine $\mathcal{H}$-Matrix Approximation der Inversen existiert und die $\mathcal{H}$-Matrix-Technik zu deren Lösung geeignet ist.

7 Fazit

Betrachtet man die erzielten Resultate zur Existenz einer $\mathcal{H}$-Matrix Approximation für die Inverse von Finite-Differenzen-Matrizen, lässt sich zusammenfassend feststellen, dass sowohl die theoretischen Ergebnisse zu den Modellproblemen aus Kapitel 4 und 5 als auch die numerischen Ergebnisse aus Kapitel 6 positiv ausfallen.

Den Ausgangspunkt der Untersuchungen bildete die Fragestellung, ob die $\mathcal{H}$-Matrix-Technik zur Beschleunigung der Lösung linearer Gleichungssysteme im Modell METRAS eingesetzt werden kann. Dazu war zunächst zu untersuchen, ob für die Inverse der entsprechenden Diskretisierungsmatrizen eine (gute) $\mathcal{H}$-Matrix Approximation existiert. Da die Diskretisierungsmatrizen im Modell METRAS jedoch aus einer speziellen Finite-Differenzen-Diskretisierung hervorgehen, ließ sich zum Nachweis keines der bekannten Resultate – wie z.B. für wohlkonditionierte, positiv definite Matrizen oder für Finite-Element-Matrizen – nutzen.

Darüber hinaus sind für Finite-Differenzen-Matrizen auch im Allgemeinen keine Veröffentlichungen im Zusammenhang mit $\mathcal{H}$-Matrizen bekannt. Aus diesem Grund mussten für diesen speziellen Anwendungsfall zunächst grundlegende Überlegungen angestellt werden. Die Einführung der wesentlichen Grundlagen und die Untersuchung der entsprechenden allgemeinen Fragestellungen zur Verwendung der $\mathcal{H}$-Matrizen bei Finite-Differenzen-Verfahren erfolgte in Kapitel 3.

Aus dem gleichen Grund war es erforderlich, einen eigenständigen Ansatz zum Nachweis der Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen von Finite-Differenzen-Matrizen zu entwickeln. Die theoretischen Untersuchungen wurden daher nicht direkt für das komplexe Gleichungssystem aus dem Modell METRAS, sondern auf Grundlage eines vereinfachten Modellproblems durchgeführt, dessen Auswahl jedoch in Anlehnung an die charakteristischen Eigenschaften der Problemstellung im Modell METRAS erfolgte. Das Hauptaugenmerk bei der Erarbeitung des neuen methodischen Ansatzes richtete sich darauf, dass sich die Vorgehensweise möglichst einfach für die Untersuchung anderer Problemstellungen, insbesondere der aus dem Modell METRAS, erweitern bzw. auf diese übertragen lässt.

Aufbauend auf den in Kapitel 3 eingeführten Grundlagen wurde daher im Anschluss der methodische Ansatz zum Beweis der Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen von Finite-Element-Matrizen aus [BH03] bzw. [Hac09] auf den Fall von Finite-Differenzen-Matrizen übertragen, da dieser Ansatz die genannten Anforderungen erfüllt. Bei der Realisierung ergab sich die Problematik, dass diskrete Varianten der Poincaré- und der Cacciopoli-Ungleichung für Gitterfunktionen benötigt wurden. Diese Resultate waren bisher in der speziellen Form für Gitterfunktionen nicht bekannt, so dass sie hergeleitet und bewiesen wurden. Dies erfolgte für Rechteckgitter in Satz 2.2.7 bzw. Satz 2.3.4 und für Quadergitter in Satz 5.1.5 bzw. Satz 5.2.3. Darauf aufbauend, wurde in Kapitel

4 für das zweidimensionale sowie in Kapitel 5 für das dreidimensionale Modellproblem die Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen der Diskretisierungsmatrizen gezeigt.

Die erzielten theoretischen Ergebnisse konnten bei der Durchführung vielfältiger numerischer Tests zum Modellproblem bestätigt werden. Dabei ließ sich bei einigen Testproblemen feststellen, dass der Approximationsfehler von der Gestalt der Koeffizienten des Differenzenoperators abhing. Im Einklang mit den theoretischen Ergebnissen ergab sich in diesen Fällen eine deutliche Verschlechterung des Fehlerverlaufs. Unter diesen Gegebenheiten wäre der Einsatz der $\mathcal{H}$-Matrix-Technik wenig Erfolg versprechend. Da diese Schwierigkeiten insbesondere im Zusammenhang mit Problemstellungen auftraten, bei denen die Koeffizienten des Differenzenoperators im Hinblick auf das Gleichungssystem im Modell METRAS gewählt wurden, war die weitere Untersuchung dieser Problematik erforderlich.

Unter Berücksichtigung des Zusammenhangs zwischen den Koeffizienten des Differenzenoperators und den unterschiedlichen Schrittweiten im Modellgitter ist es gelungen, eine alternative Partitionierungsstrategie zu entwickeln, durch deren Einsatz sich deutlich bessere Ergebnisse erzielen ließen. Ein weiterer Vorteil der neu eingeführten modifizierten Partition ist durch die, im Vergleich zur gewöhnlichen Partition, größeren zulässigen Blöcke gegeben. Daraus resultierten bei Verwendung der modifizierten statt der gewöhnlichen Partition geringere Speicheranforderungen und somit geringere Kosten bei der Durchführung der Operationen in der $\mathcal{H}$-Arithmetik.

Zusätzlich wurde numerisch untersucht, ob sich die Inversen von Finite-Differenzen-Matrizen zu weiteren Problemstellungen, für die im Gegensatz zum Modellproblem kein theoretisches Resultat bekannt ist, durch $\mathcal{H}$-Matrizen approximieren lassen. Bei den numerischen Tests zur Diskretisierung der Wärmeleitungsgleichung auf einem Rechteckgitter mittels der θ-Methode und der Diskretisierung der Problemstellung, die sich durch die Einführung von Neumann- statt Dirichlet-Randwerten im Modellproblem ergibt, konnten gute Resultate erzielt werden. Sie sind vergleichbar mit denen zum Modellproblem, wobei erneut eine Verschlechterung der Fehler beim Auftreten bestimmter Konstellationen der Koeffizienten des Differenzenoperators beobachtet wurde. In diesen Fällen ließen sich durch den Einsatz der modifizierten Partition aus dem Modellproblem ebenfalls deutlich bessere Ergebnisse erzielen.

Darüber hinaus wurden numerische Tests für die Inverse der Diskretisierungsmatrix zur Konvektions-Diffusionsgleichung mit konstanter Konvektion durchgeführt, bei denen eine Verschlechterung der Ergebnisse für den Fall $\epsilon \to 0$ zu beobachten war. Auf Grundlage einer alternativen Partitionierung, die in [LB03] für Finite-Element-Matrizen beschrieben ist, ließ sich die modifizierte Partitionierungsstrategie, die zuvor für das Modellproblem eingeführt wurde, so anpassen, dass sie ebenfalls im Zusammenhang mit der Konvektions-Diffusionsgleichung eingesetzt werden konnte. Diese Vorgehensweise führte auch für diese Problemstellung zu einer deutlichen Verbesserung der Ergebnisse.

Unabhängig von diesen positiven Ergebnissen kann die grundlegende Fragestellung nach der Anwendbarkeit der $\mathcal{H}$-Matrix-Technik zur Lösung der Gleichungssysteme im Modell METRAS noch nicht abschließend beantwortet werden. Das Modellproblem wurde zwar unter Berücksichtigung der speziellen Gegebenheiten im Modell METRAS auf-

gestellt, aber die vorgenommenen Vereinfachungen haben zur Folge, dass auf Grundlage der erzielten Resultate zum Modellproblem keine theoretische Aussage zur Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen der Diskretisierungsmatrizen im Modell METRAS getroffen werden kann. Durch die Erarbeitung des methodischen Ansatzes für Finite-Differenzen-Matrizen anhand des Modellproblems wurde jedoch ein wesentlicher Beitrag zur Untersuchung dieser Fragestellung geleistet. Darauf aufbauend lassen sich analoge Resultate für ähnliche Problemstellungen mit andersartigen Differenzenoperatoren, insbesondere für die Problemstellung aus dem Modell METRAS, herleiten. Dazu kann die Vorgehensweise, die für das Modellproblem entwickelt wurde, im Wesentlichen übernommen werden. Viele Resultate und Überlegungen – wie die diskrete Poincaré-Ungleichung oder der Zusammenhang zwischen der Inversen und der diskreten Greenschen Funktion – gelten unabhängig von dem Differenzenoperator in der Problemstellung. Daher ist der Aufwand zur Übertragung der Ergebnisse im Wesentlichen auf den Beweis einer diskreten Cacciopoli-Ungleichung unter Berücksichtigung des speziellen Differenzenoperators beschränkt, der insbesondere für die Problemstellung im Modell METRAS jedoch aufwendiger ausfällt. Insofern liefern die erzielten Ergebnisse grundlegende Erkenntnisse zur Durchführung weiterführender theoretischer Untersuchungen.

Darüber hinaus deuten die numerischen Ergebnisse aus Abschnitt 6.2.4 zur Diskretisierungsmatrix, die sich bei Verwendung des Differenzenoperators aus dem Modell METRAS (ergänzt durch Dirichlet-Randwerte) ergibt, darauf hin, dass sich die $\mathcal{H}$-Matrix-Technik auch zur Beschleunigung der Lösung dieses Gleichungssystems eignet. Dabei ließ sich ein dem dreidimensionalen Modellproblem vergleichbares Verhalten beobachten, bei dem eine Verschlechterung der Approximation für bestimmte Konstellationen der Koeffizienten auftrat. Da für das Modellproblem in diesen Fällen bereits eine modifizierte Partitionierungsstrategie eingeführt wurde, die zu einer Verbesserung der Ergebnisse führte, konnten diese Erkenntnisse ebenfalls für die speziellen Gegebenheiten im Zusammenhang mit dem Gleichungssystem im Modell METRAS verwendet werden. Durch den Einsatz der auf diese Weise modifizierten Partition ließen sich auch für die Problemstellung mit dem Differenzenoperator aus dem Modell METRAS deutlich bessere Ergebnisse erzielen.

Demnach lieferten die numerischen Tests zum Modellproblem bereits erste Resultate, die im Zusammenhang mit der Anwendung der $\mathcal{H}$-Matrix-Technik zur Lösung der Gleichungssysteme im Modell METRAS berücksichtigt werden sollten. Insbesondere die Ergebnisse zur Konstruktion einer geeigneten Partition sollten aufgrund der erzielten Ergebnisse beachtet werden, um eine mögliche Verschlechterung der Fehler bei Verwendung der gewöhnlichen Partition zu vermeiden.

Aufbauend auf den neuen theoretischen Ergebnissen ist die Erweiterung der Resultate zur Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen für den speziellen Fall des Gleichungssystems im Modell METRAS möglich. Da die Berechnung der $\mathcal{H}$-Inversen in der Praxis nicht so effizient durchgeführt werden kann wie die der $\mathcal{H}$-LU-Zerlegung, sind vor dem praktischen Einsatz der $\mathcal{H}$-Matrix-Technik im Modell METRAS weitere theoretische und numerische Untersuchungen zur $\mathcal{H}$-LU-Zerlegung von Interesse. Der theoretische Nachweis der Existenz einer $\mathcal{H}$-Matrix Approximation der Faktoren der LU-Zerlegung kann auf Grundlage der Ergebnisse zur Inversen vorgenommen werden, wes-

halb dieser Aspekt im Mittelpunkt der Untersuchungen stand. Die numerischen Tests zur Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen sind analog für die LU-Zerlegung durchzuführen. Diese sind durch Tests zu ergänzen, die eine praxisnahe Beurteilung der Leistungsfähigkeit und Einsatzmöglichkeit der $\mathcal{H}$-LU-Zerlegung zur Beschleunigung der Simulation im Modell METRAS ermöglichen.

Die grundlegenden Erkenntnisse zur Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen von Finite-Differenzen-Matrizen rechtfertigen die weitere theoretische Untersuchung und insbesondere die Durchführung weiterer numerischer Tests. Der Einsatz der $\mathcal{H}$-Matrix-Technik zur Beschleunigung der Lösung der Gleichungssysteme im Modell METRAS scheint nach den erzielten Ergebnissen eine Erfolg versprechende Alternative bzw. Ergänzung zu den eingesetzten Lösungsverfahren darzustellen.

Literaturverzeichnis

[Beb03] Bebendorf, M.: A note on the Poincaré inequality for convex domains. In: *J. Anal. Appl.* 22 (2003), S. 751–756

[Beb07] Bebendorf, M.: Why Finite Element Discretizations Can Be Factored by Triangular Hierarchical Matrices. In: *SIAM Journal on Numerical Analysis* 45 (2007), Nr. 4, S. 1472–1494

[Beb08] Bebendorf, M.: *Hierarchical matrices: A means to efficiently solve elliptic boundary value problems ; with 53 tables.* Berlin und Heidelberg : Springer, 2008

[BF11] Bebendorf, M. ; Fischer, T.: On the purely algebraic data-sparse approximation of the inverse and the triangular factors of sparse matrices. In: *Numerical Linear Algebra with Applications* 18 (2011), Nr. 1, S. 105–122

[BH03] Bebendorf, M. ; Hackbusch, W.: Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients. In: *Numerische Mathematik* 95 (2003), Nr. 1, S. 1–28

[CFL28] Courant, R. ; Friedrichs, K. ; Lewy, H.: Über die partiellen Differenzengleichungen der mathematischen Physik. In: *Mathematische Annalen* 100 (1928), S. 32–74

[GH03] Grasedyck, L. ; Hackbusch, W.: Construction and Arithmetics of $\mathcal{H}$-Matrices. In: *Computing* 70 (2003), Nr. 4, S. 295–334

[GHK08] Grasedyck, L. ; Hackbusch, W. ; Kriemann, R.: Performance of $\mathcal{H}$-LU preconditioning for sparse matrices. In: *Computational Methods in Applied Mathematics* 8 (2008), Nr. 4, S. 336–349

[GKLB08] Grasedyck, L. ; Kriemann, R. ; Le Borne, S.: Parallel black box $\mathcal{H}$ - LU preconditioning for elliptic boundary value problems. In: *Computing and Visualization in Science* 11 (2008), Nr. 4-6, S. 273–291

[GKLB09] Grasedyck, L. ; Kriemann, R. ; Le Borne, S.: Domain decomposition based $\mathcal{H}$-LU preconditioning. In: *Numerische Mathematik* 112 (2009), Nr. 4, S. 565–600

[Gra01] Grasedyck, L.: *Theorie und Anwendung Hierarchischer Matrizen.* Kiel, Christian-Albrechts-Universität zu Kiel, Diss., 2001

[Hac86] Hackbusch, W.: *Theorie und Numerik elliptischer Differentialgleichungen: Mit zahlreichen Beispielen und Übungsaufgaben.* Stuttgart : Teubner, 1986

[Hac99] Hackbusch, W.: A Sparse Matrix Arithmetic Based on $\mathcal{H}$-Matrices. Part I: Introduction to $\mathcal{H}$-Matrices. In: *Computing* 62 (1999), Nr. 2, S. 89–108

[Hac09] Hackbusch, W.: *Hierarchische Matrizen: Algorithmen und Analysis.* Berlin und Heidelberg : Springer, 2009

[HK00a] Hackbusch, W. ; Khoromskij, B. N.: $\mathcal{H}$-matrix approximation on graded meshes. In: Whiteman, J. R. (Hrsg.): *The mathematics of finite elements and applications X.* Oxford : Elsevier, 2000, S. 307–316

[HK00b] Hackbusch, W. ; Khoromskij, B. N.: A Sparse $\mathcal{H}$-Matrix Arithmetic. Part II: Application to Multi-Dimensional Problems. In: *Computing* 64 (2000), Nr. 1, S. 21–47

[Laa58] Laasonen, P.: On the Solution of Poisson's Difference Equation. In: *Journal of the ACM* 5 (1958), Nr. 4, S. 370–382

[LB03] Le Borne, S.: $\mathcal{H}$-matrices for Convection-diffusion Problems with Constant Convection. In: *Computing* 70 (2003), Nr. 3, S. 261–274

[SBL+96] Schlünzen, K.H. ; Bigalke, K. ; Lüpkes, C. ; Niemeier, U. ; Salzen, K. von: *Concept and realization of the mesoscale transport- and fluid-model 'METRAS': METRAS Techn. Rep. 5.* Hamburg, 1996

[Sch07] Schröder, G.: *Development and test of a multiple grids option in a meso-scale model.* Hamburg, Universität Hamburg, Diss., 2007

[Sül91] Süli, E.: Convergence of Finite Volume Schemes for Poisson's Equation on Nonuniform Meshes. In: *SIAM Journal on Numerical Analysis* 28 (1991), Nr. 5, S. 1419–1430

[SV84] Schumann, U. ; Volkert, H.: Three-dimensional mass- and momentum-consistent Helmholtz-equation in terrain-following coordinates. In: Hackbusch, W. (Hrsg.): *Efficient solutions of elliptic systems.* Braunschweig : Vieweg, 1984, S. 109–131

[Tem77] Temam, R.: *Navier-Stokes equations: Theory and numerical analysis.* Amsterdam and New York : North-Holland Pub. Co., 1977

Zusammenfassung

Die Technik der Hierarchischen Matrizen ($\mathcal{H}$-Matrizen) ermöglicht die Berechnung einer approximativen $\mathcal{H}$-Inversen oder $\mathcal{H}$-LU-Zerlegung in fast linearer Komplexität und kann auf diese Weise zur effizienten Lösung linearer Gleichungssysteme eingesetzt werden. Vor der Verwendung der $\mathcal{H}$-Matrix-Technik ist zu untersuchen, ob eine $\mathcal{H}$-Matrix Approximation der Inversen bzw. der Faktoren der LU-Zerlegung existiert. Resultate dieser Form konnten bereits für diverse Matrizen (z.B. für Finite-Element-Matrizen) gezeigt werden. Für Gleichungssysteme, die aus der Diskretisierung partieller Differentialgleichungen mittels Finiter-Differenzen-Verfahren resultieren, sind jedoch keine Veröffentlichungen zum Einsatz der $\mathcal{H}$-Matrix-Technik bekannt. Mit der Zielsetzung die Anwendbarkeit der $\mathcal{H}$-Matrix-Technik für Finite-Differenzen-Matrizen aus dem meteorologischen Transport- und Strömungsmodell METRAS zu untersuchen, wird für ein zwei- und ein dreidimensionales Modellproblem die Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen von Finite-Differenzen-Matrizen nachgewiesen.

Dazu wird der methodische Ansatz für Finite-Element-Matrizen auf den Fall von Finite-Differenzen-Matrizen übertragen. Zu diesem Zweck ist die Gültigkeit einer diskreten Poincaré- und einer diskreten Cacciopoli-Ungleichung für Gitterfunktionen nachzuweisen, welche in der erforderlichen Form bisher nicht bekannt waren. Diese werden für den Fall von Rechteck- bzw. Quadergittern bewiesen und können unabhängig von dieser Arbeit auch in anderen Zusammenhängen verwendet werden.

Die Ergebnisse zur Existenz einer $\mathcal{H}$-Matrix Approximation der Inversen von Finite-Differenzen-Matrizen werden mittels numerischer Tests bestätigt. Bei in Anlehnung an das Gleichungssystem aus dem Modell METRAS aufgestellten Testproblemen lässt sich im Einklang mit den theoretischen Ergebnissen jedoch eine Verschlechterung des Fehlerverlaufs in Abhängigkeit von einem Parameter feststellen. Für diese Fälle wird eine modifizierte Partitionierungsstrategie vorgestellt, deren Verwendung zu deutlich besseren Ergebnissen führt.

Lebenslauf

entfällt aus datenschutzrechtlichen Gründen