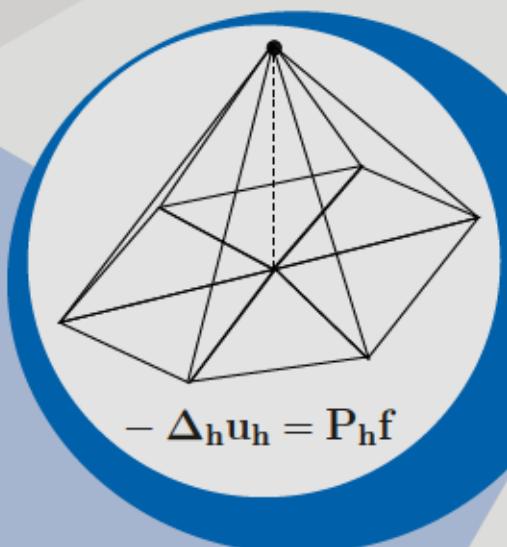


# Numerik 2

## Numerik partieller Differentialgleichungen

ROLF RANNACHER



HEIDELBERG  
UNIVERSITY PUBLISHING



## **NUMERIK 2**

Numerik partieller Differentialgleichungen

Lecture Notes **Mathematik**



# NUMERIK 2

## Numerik partieller Differentialgleichungen

Rolf Rannacher

Institut für Angewandte Mathematik  
Universität Heidelberg

HEIDELBERG  
UNIVERSITY PUBLISHING

## Über den Autor

Rolf Rannacher, Prof. i. R. für Numerische Mathematik an der Universität Heidelberg; Studium der Mathematik an der Universität Frankfurt am Main – Promotion 1974; Habilitation 1978 in Bonn; 1979/1980 Vis. Assoc. Prof. an der University of Michigan (Ann Arbor, USA), dann Professor in Erlangen und Saarbrücken – in Heidelberg seit 1988; Spezialgebiet „Numerik partieller Differentialgleichungen“, insbesondere „Methode der finiten Elemente“ mit Anwendungen in Natur- und Ingenieurwissenschaften; hierzu über 160 publizierte wissenschaftliche Arbeiten.

**Bibliografische Information der Deutschen Nationalbibliothek**  
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie. Detaillierte bibliografische Daten sind im Internet unter <http://dnb.ddb.de> abrufbar.



Dieses Werk ist unter der Creative Commons-Lizenz 4.0  
(CC BY-SA 4.0) veröffentlicht.

Die Online-Version dieser Publikation ist auf den Verlagswebseiten von HEIDELBERG UNIVERSITY PUBLISHING <http://heiup.uni-heidelberg.de> dauerhaft frei verfügbar (open access).  
urn: urn:nbn:de:bsz:16-heiup-book-281-3  
doi: <https://doi.org/10.17885/heiup.281.370>

Text © 2017, Rolf Rannacher

ISSN 2566-4816 (PDF)  
ISSN 2512-4455 (Print)

ISBN 978-3-946054-37-5 (PDF)  
ISBN 978-3-946054-38-2 (Softcover)

# Inhaltsverzeichnis

<b>Literaturverzeichnis</b>	<b>x</b>
<b>0 Einleitung</b>	<b>1</b>
0.1 Notation . . . . .	1
0.2 Ableitung von partiellen Differentialgleichungen . . . . .	3
0.3 Beispiele . . . . .	4
0.4 Numerische Methoden . . . . .	7
<b>1 Theorie partieller Differentialgleichungen</b>	<b>9</b>
1.1 Typeneinteilung . . . . .	10
1.2 Elliptische Probleme . . . . .	15
1.2.1 Existenz von Lösungen . . . . .	16
1.2.2 Eindeutigkeit von Lösungen . . . . .	21
1.2.3 Stetige Abhängigkeit der Lösungen von den Daten . . . . .	23
1.2.4 Regularität von Lösungen . . . . .	24
1.3 Hilfsmittel aus der Theorie von Funktionenräumen . . . . .	25
1.3.1 Sobolew-Räume . . . . .	25
1.3.2 Eigenschaften von Lebesgue- und Sobolew-Räumen . . . . .	29
1.3.3 Elemente der Spektraltheorie elliptischer Operatoren . . . . .	33
1.4 Parabolische Probleme . . . . .	34
1.5 Hyperbolische Probleme . . . . .	41
1.6 Übungen . . . . .	43
<b>2 Differenzen-Verfahren für elliptische Probleme</b>	<b>49</b>
2.1 Allgemeine Differenzenapproximationen . . . . .	49
2.1.1 Konsistenz . . . . .	51
2.2 Eigenschaften der Differenzengleichungen . . . . .	55
2.2.1 Das Konvergenzverhalten von Differenzenverfahren . . . . .	60
2.3 Lösungsaspekte . . . . .	65
2.3.1 Aufwandsanalyse: ein Beispiel . . . . .	70
2.4 Übungen . . . . .	72

<b>3 Finite-Elemente-Verfahren für elliptische Probleme</b>	<b>77</b>
3.1 Allgemeine Projektionsverfahren . . . . .	77
3.1.1 Beispiele von Galerkin-Ansatzräumen . . . . .	83
3.1.2 Diskretes Maximumprinzip für Finite-Elemente-Approximationen .	89
3.1.3 Approximation krummer Ränder . . . . .	91
3.2 Allgemeine Finite-Elemente-Ansätze . . . . .	97
3.3 Interpolation mit finiten Elementen . . . . .	108
3.4 A priori Fehleranalyse . . . . .	118
3.4.1 Punktweise Fehlerabschätzung . . . . .	120
3.5 Implementierungsaspekte . . . . .	127
3.5.1 Aufbau der Systemmatrizen und Vektoren . . . . .	128
3.5.2 Konditionierung der Systemmatrix . . . . .	130
3.5.3 Aufstellung der Systemmatrizen mit numerischer Integration . . . . .	134
3.6 A posteriori Fehleranalyse und Gittersteuerung . . . . .	142
3.6.1 Allgemeine a posteriori Fehlerabschätzung . . . . .	143
3.6.2 Spezielle a posteriori Fehlerschätzer . . . . .	147
3.6.3 Strategien zur Gittersteuerung . . . . .	155
3.6.4 Ein Testbeispiel . . . . .	160
3.7 Übungen . . . . .	162
<b>4 Lösung der FE-Gleichungen</b>	<b>173</b>
4.1 Krylow-Raum-Methoden . . . . .	173
4.1.1 Verfahren der konjugierten Richtungen (CG-Verfahren) . . . . .	175
4.1.2 CG-Verfahren für unsymmetrische und indefinit Probleme . . . . .	181
4.1.3 Vorkonditionierung (PCG-Verfahren) . . . . .	182
4.2 Mehrgitterverfahren . . . . .	184
4.2.1 Mehrgitteralgorithmus im Finite-Elemente-Kontext . . . . .	186
4.2.2 Konvergenz- und Aufwandsanalyse . . . . .	192
4.3 Übungen . . . . .	199
<b>5 Verfahren für parabolische Probleme</b>	<b>203</b>
5.1 Differenzenverfahren für parabolische Probleme . . . . .	207
5.1.1 Zeitschrittverfahren . . . . .	207

5.1.2	Stabilität und Konvergenz . . . . .	215
5.2	FE-Galerkin-Verfahren für parabolische Probleme . . . . .	224
5.2.1	A priori Konvergenzabschätzungen . . . . .	225
5.2.2	Fehlerkontrolle und Schrittweitensteuerung . . . . .	230
5.3	Verallgemeinerungen und Lösungsaspekte . . . . .	236
5.4	Übungen . . . . .	242
<b>6</b>	<b>Verfahren für hyperbolische Probleme</b>	<b>243</b>
6.1	Differenzenverfahren für die Wellengleichung . . . . .	243
6.2	Finite-Elemente-Verfahren für die Wellengleichung . . . . .	250
6.3	Lösungsaspekte . . . . .	252
6.4	Übungen . . . . .	252
<b>A</b>	<b>Lösungen der Übungsaufgaben</b>	<b>253</b>
A.1	Kapitel 1 . . . . .	253
A.2	Kapitel 2 . . . . .	267
A.3	Kapitel 3 . . . . .	274
A.4	Kapitel 4 . . . . .	298
A.5	Kapitel 5 . . . . .	305
A.6	Kapitel 6 . . . . .	306
<b>Index</b>		<b>309</b>



# Literaturverzeichnis

- [1] R. Rannacher: *Numerik 0: Einführung in die Numerische Mathematik*, Lecture Notes Mathematik, Heidelberg University Publishing, Heidelberg, 2017, <https://doi.org/10.17885/heip.206.281>.
- [2] R. Rannacher: *Numerik 1: Numerik Gewöhnlicher Differentialgleichungen*, Lecture Notes Mathematik, Heidelberg University Publishing, Heidelberg, 2017, <https://doi.org/10.17885/heip.281.370>.

## (I) Zur Theorie partieller Differentialgleichungen

- [3] A. Friedman: *Partial Differential Equations*, Holt, Rinehart und Winston 1970.
- [4] P. Grisvard: *Elliptic Problems in Nonsmooth Domains*, Pitman Publishing, Marshfield, Massachusetts, 1985.
- [5] G. Hellwig: *Partielle Differentialgleichungen*, B.G. Teubner 1960; english edition available.
- [6] J. Joos: *Partial Differential Equations*, Springer 2013.
- [7] M. Renardy, R. Rogers: *An Introduction to Partial Differential Equations*, Springer 1993.
- [8] W. R. Strauss: *Partial Differential Equations: An Introduction*, John Wiley 1992.
- [9] A. Tveito, R. Winther: *Introduction to Partial Differential Equations: A Computational Approach*, Springer 1998.
- [10] J. Wloka: *Partielle Differentialgleichungen, Sobolevräume und Randwertaufgaben*, B.G. Teubner 1982.

## (II) Zur Numerik partieller Differentialgleichungen

- [11] O. Axelsson, V. A. Barker: *Finite Element Solution of Boundary Value Problems, Theory and Computation*, Academic Press 1984.
- [12] D. Braess: *Einführung in die Methode der Finiten Elemente*, Springer 1991.
- [13] S. C. Brenner, L. R. Scott: *The Mathematical Theory of Finite Element Methods*, Springer 1994.
- [14] F. Brezzi, M. Fortin: *Mixed and Hybrid Finite Element Methods*, Springer 1991.

- [15] P. G. Ciarlet: *The Finite Element Method for Elliptic Problems*, North-Holland 1978.
- [16] K. Eriksson, D. Estep, P. Hansbo, C. Johnson: *Computational Differential Equations*, Cambridge University Press 1996.
- [17] H. Goering, H.-G. Roos, L. Tobiska: *Finite-Elemente-Methode*, Akademie-Verlag 1993.
- [18] Ch. Großmann, H.-G. Roos: *Numerik partieller Differentialgleichungen*, B. G. Teubner 1992.
- [19] W. Hackbusch: *Multigrid Methods and Applications* Springer, 1985.
- [20] W. Hackbusch: *Theorie und Numerik elliptischer Differentialgleichungen*, B. G. Teubner 1986.
- [21] W. Hackbusch: *Iterative Lösung großer schwachbesetzter Gleichungssysteme*, B. G. Teubner 1991.
- [22] C. Johnson: *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press 1987.
- [23] P. Knabner, L. Angermann: *Numerik partieller Differentialgleichungen*, Springer 2000.
- [24] A. R. Mitchell: *Computational Methods in Partial Differential Equations*, John Wiley 1976.
- [25] A. R. Mitchell, R. Wait: *The Finite Element Method in Partial Differential Equations*, John Wiley 1977.
- [26] A. R. Mitchell, D. F. Griffiths: *The Finite Difference Method in Partial Differential Equations*, John Wiley 1980.
- [27] A. Quarteroni, A. Valli: *Numerical Approximation of Partial Differential Equations*, Springer 1994.
- [28] R. D. Richtmeyer, K. W. Morton: *Difference Methods for Initial Value Problems*, Interscience 1967.
- [29] H. R. Schwarz: *Methode der finiten Elemente*, B. G. Teubner 1991.
- [30] G. Strang, G. J. Fix: *An Analysis of the Finite Element Method*, Prentice-Hall 1973.
- [31] V. Thome: *Galerkin Finite Element Methods for Parabolic Problems*, Springer 1984.

# 0 Einleitung

Gegenstand dieses Textes sind numerische Algorithmen zur näherungsweisen Lösung von partiellen Differentialgleichungen. In der Regel lassen sich für die in der Praxis auftretenden partiellen Differentialgleichungen keine Lösungen in analytischer Form angeben. Man ist also auf numerische Approximation angewiesen. Dabei wird das kontinuierliche Ausgangsproblem durch ein „diskretes“, d. h. endlich dimensionales, ersetzt und dieses dann mit Hilfe des Computers gelöst.

## 0.1 Notation

Wir stellen zuächst einige der im folgenden verwendeten Begriffe und abkürzenden Bezeichnungen zusammen.

- i) *Unabhängige Variable:* Wir betrachten (offene beschränkte) Gebiete  $\Omega \subset \mathbb{R}^d$  für  $d = 1, 2, 3$  mit Rand  $\partial\Omega$ . Der äußere Normaleneinheitsvektor zu  $\partial\Omega$  ist  $n$ . Punkte im  $\mathbb{R}^d$  sind  $x = (x_1, \dots, x_d)^T$  oder speziell  $(x, y)^T$  im  $\mathbb{R}^2$  und  $(x, y, z)^T$  im  $\mathbb{R}^3$ . Die Zeitvariable ist  $t$ . Für  $d$ -dimensionale Vektoren  $a, b$  wird das übliche euklidische<sup>1</sup> Produkt mit  $(a, b)$  oder auch mit  $a \cdot b$  bezeichnet. Die euklidische Vektornorm ist  $\|a\| := (a, a)^{1/2}$  und die zugehörige natürliche Matrizennorm  $\|A\| := \max_{x \in \mathbb{R}^d} \{\|Ax\|, \|x\| = 1\}$ .
- ii) *Funktionen:* Wir betrachten i. Allg. skalare Funktionen  $u = u(x)$  oder  $u = u(x, t)$  für Argumente  $x \in \mathbb{R}^d$  bzw.  $t \in \mathbb{R}$ . In einigen Fällen treten auch vektorwertige Funktionen auf  $u = (u_1, \dots, u_d)^T$ , die i. Allg. wie skalare Funktionen bezeichnet werden. Analog werden Skalarprodukte und Normen über ein Gebiet  $\Omega$  unterschiedslos für skalare wie für vektorwertige Funktionen verwendet.
- iii) *Ableitungen:* Für Funktionen  $u(t)$ ,  $u(x)$  bzw.  $u(x, t)$  werden totale sowie partielle Ableitungen abgekürzt geschrieben als

$$d_t u := \frac{du}{dt}, \quad \partial_t u := \frac{\partial u}{\partial t}, \quad \partial_x u := \frac{\partial u}{\partial x}, \quad \partial_i u := \frac{\partial u}{\partial x_i}, \quad \text{u.s.w.}$$

und analog auch für höhere Ableitungen, z. B.:  $\partial_t^p u$  und  $\partial_x^q u$ . Mit dem Nabla-Operator  $\nabla$  werden der Gradient einer skalaren Funktion sowie die Divergenz einer Vektorfunktion geschrieben als  $\text{grad } u = \nabla u := (\partial_1 u, \dots, \partial_d u)^T$  und  $\text{div } u = \nabla \cdot u := \partial_1 u_1 + \dots + \partial_d u_d$ . Zu einem Vektor  $\beta \in \mathbb{R}^d$  wird die Ableitung in Richtung  $\beta$  mit  $\partial_\beta u := \beta \cdot \nabla u$  bezeichnet. Entsprechend ist z. B.  $\partial_n u = n \cdot \nabla u$  die Ableitung in Richtung der äußeren Normalen entlang des Gebietsrandes  $\partial\Omega$ . Kombination von Divergenz- und Gradientenbildung ergibt

---

<sup>1</sup>Euklid (ca. 355–290 v. Chr.): Griechischer Philosoph und Mathematiker; wirkte in Alexandria; sein mehrbändiges Lehrbuch „Die Elemente“ fasste die Grundlagen der klassischen Geometrie zusammen; von ihm stammt das klassische mathematische Ausdrucksschema „Voraussetzung - Behauptung - Beweis“.

den sog. „Laplace<sup>2</sup>-Operator“

$$\nabla \cdot (\nabla u) = \Delta u = \partial_1^2 u + \dots + \partial_d^2 u.$$

Mit dem Symbol  $\nabla^m u$  bezeichnen wir den Tensor aller partiellen Ableitungen der Ordnung  $m$  von  $u$ ; z. B. in zwei Dimensionen  $\nabla^2 u = (\partial_x^i \partial_y^j u)_{i+j=2}$ .

iv) *Integralsätze*: Für stückweise glatt berandete Gebiete  $\Omega \subset \mathbb{R}^d$  und hinreichend glatte Funktionen  $u, v$  gelten der klassische Integralsatz von Gauß<sup>3</sup>

$$\int_{\Omega} \nabla \cdot u \, dx = \int_{\partial\Omega} n \cdot u \, do, \quad (0.1.1)$$

sowie als Folgerung die Integralformel von Green<sup>4</sup>

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\partial\Omega} u \partial_n v \, do - \int_{\Omega} u \Delta v \, dx. \quad (0.1.2)$$

v) *Funktionenräume*: Auf Punktmengen  $\Omega \in \mathbb{R}^d$  verwenden wir die üblichen Vektorräume von stetigen bzw. stetig differenzierbaren Funktionen

$$C(\bar{\Omega}), \quad C^m(\Omega), \quad C_0^\infty(\Omega).$$

Der Raum  $C(\bar{\Omega})$  ist, versehen mit der Maximumsnorm

$$\|u\|_{\infty; \bar{\Omega}} := \max_{x \in \bar{\Omega}} |u(x)|,$$

vollständig, d. h. ein Banach<sup>5</sup>-Raum. Weiter ist  $L^2(\Omega)$  der Raum der auf  $\Omega$  messbaren und im Lebesgueschen Sinne quadratintegrablen Funktionen. Versehen mit dem Skalarprodukt und der zugehörigen Norm

$$(u, v)_{\Omega} := \int_{\Omega} u(x)v(x) \, dx, \quad \|u\|_{\Omega} = (u, u)_{\Omega}^{1/2},$$

ist  $L^2(\Omega)$  vollständig, d. h. ein Hilbert<sup>6</sup>-Raum. Wenn der Definitionsbereich  $\Omega$  aus dem Zusammenhang klar ist, wird er in der Bezeichnung von Skalarprodukten und Normen

<sup>2</sup>Pierre Simon Marquis de Laplace (1749–1827): Französischer Mathematiker und Astronom; Prof. in Paris; begründete u. a. die Wahrscheinlichkeitsrechnung.

<sup>3</sup>Carl Friedrich Gauß (1777–1855): bedeutender deutscher Mathematiker, Astronom und Physiker; wirkte in Göttingen.

<sup>4</sup>George Green (1793–1841): Englischer Mathematiker; Autodidakt und Besitzer einer Mühle; Beiträge zur Potentialtheorie.

<sup>5</sup>Stefan Banach (1892–1945): Polnischer Mathematiker; Prof. in Lvov; begründete die Funktionalanalysis.

<sup>6</sup>David Hilbert (1862–1943): bedeutender deutscher Mathematiker; wirkte in Königsberg und Göttingen; begründete u. a. den axiomatischen Aufbau der Mathematik; zum Wesen der Axiomatik (in der Geometrie) sagte er „Man muss jederzeit anstelle von Punkten, Geraden, Ebenen – Tische, Stühle, Bierseidel sagen können“.

meist weggelassen; z. B.:  $\|u\| = \|u\|_{\Omega}$ . Weitere Funktionenräume werden später an den Stellen eingeführt, wo sie gebraucht werden.

vi) *Ungleichungen:* Wir listen einige der im Folgenden häufig verwendeten Ungleichungen für Funktionen aus den oben definierten Funktionenräumen. Für Funktionen  $u, v \in L^2(\Omega)$  gilt die „Höldersche<sup>7</sup>“ Ungleichung<sup>8</sup>

$$\left| \int_{\Omega} u(x)v(x) dx \right| \leq \left( \int_{\Omega} |u(x)|^2 dx \right)^{1/2} \left( \int_{\Omega} |v(x)|^2 dx \right)^{1/2}, \quad (0.1.3)$$

bzw. in kompakter Schreibweise:  $|(u, v)|_{\Omega} \leq \|u\|_{\Omega} \|v\|_{\Omega}$ . Für Funktionen  $u \in C(\overline{\Omega}) \cap C^1(\Omega)$  mit den Eigenschaften  $u|_{\partial\Omega} = 0$  und  $|\nabla u| \in L^2(\Omega)$  gilt die „Poincarésche<sup>8</sup>“ Ungleichung<sup>9</sup>

$$\left( \int_{\Omega} |u(x)|^2 dx \right)^{1/2} \leq d_{\Omega} \left( \int_{\Omega} |\nabla u(x)|^2 dx \right)^{1/2}, \quad (0.1.4)$$

bzw. in kompakter Schreibweise:  $\|u\|_{\Omega} \leq d_{\Omega} \|\nabla u\|_{\Omega}$ , wobei  $d_{\Omega} := \text{diam}(\Omega)$ .

## 0.2 Ableitung von partiellen Differentialgleichungen

Partielle Differentialgleichungen werden meist als mathematische Modelle zur Beschreibung physikalischer Vorgänge abgeleitet. Ziel ist es, die Eigenschaften dieser Vorgänge durch die Gleichungen möglichst vollständig zu erfassen und davon ausgehend dann ihren Ablauf vorherzusagen. Bei der Konstruktion solcher Gleichungen geht man häufig nach recht formalen Regeln vor und verwendet Analogieschlüsse. Völlig unterschiedliche physikalische Prozesse lassen sich häufig durch Gleichungen sehr ähnlicher Gestalt beschreiben.

i) Harmonische (d. h. „schwingende“) Ausbreitungsvorgänge sind z. B. die Ausbreitung einer Wasserwelle (lokale Störung der Wasseroberfläche) oder eines Geräusches (lokale Störung der Luftdichte). Es erscheint sinnvoll, diese im einfachsten Fall durch eine Funktion  $u = u(x, t)$  im Ort  $x$  und der Zeit  $t$  der Form  $u(x, t) = \sin(x) \sin(t)$  zu beschreiben. Diese genügt der Differentialgleichung

$$\partial_t^2 u = \partial_x^2 u. \quad (0.2.5)$$

Es zeigt sich, dass diese sog. „Wellengleichung“ tatsächlich die in der Natur auftretenden Schwingungsvorgänge beschreibt. Sie ist der Prototyp einer „hyperbolischen“ Differentialgleichung.

ii) Andere Ausbreitungsvorgänge sind dadurch gekennzeichnet, dass lokale Störungen

---

<sup>7</sup>Ludwig Otto Hölder (1859–1937): Deutscher Mathematiker; Prof. in Tübingen; Beiträge zunächst zur Theorie der Fourier-Reihen und später vor allem zur Gruppentheorie; fand 1884 die nach ihm benannte Ungleichung.

<sup>8</sup>Jules Henri Poincaré (1854–1912): Französischer Mathematiker; Prof. an der École Polytechnique und der Sorbonne in Paris; eins der letzten mathematischen Universalgenies; fundamentale Beiträge zu allen Bereichen der Mathematik, zur Himmelsmechanik, Strömungsmechanik und Wissenschaftsphilosophie.

nicht schwingend, sondern „diffundierend“ und sich abschwächend fortgepflanzt werden, z. B.: Temperaturausbreitung in einem Leiter (Wärmeleitung) oder Verteilung einer Dichtekonzentration in einer Flüssigkeit (Stofftransport). Zur Beschreibung solcher Vorgänge dienen Funktionen der Form  $u(x, t) = \sin(x) e^{-t}$ . Diese genügen der Differentialgleichung

$$\partial_t u = \partial_x^2 u. \quad (0.2.6)$$

Es zeigt sich, dass diese sog. „Wärmeleitungsgleichung“ tatsächlich die in der Natur auftretenden Diffusionsvorgänge beschreibt. Sie ist der Prototyp einer „parabolischen“ Differentialgleichung. In Fall zweier Raumdimensionen,  $u = u(x, y, t)$ , lautet die Wärmeleitungsgleichung unter Berücksichtigung von externen Wärmequellen

$$\partial_t u = \partial_x^2 u + \partial_y^2 u = \Delta u + f. \quad (0.2.7)$$

iii) Der Grenzzustand für  $t \rightarrow \infty$  eines Diffusionsprozesses  $u(x, y, t)$ , z. B. beschrieben durch die Wärmeleitungsgleichung (0.2.7), ist als Lösung der sog. „Poisson<sup>9</sup>-Gleichung“

$$-\Delta u = -\partial_x^2 u - \partial_y^2 u = f, \quad (0.2.8)$$

gegeben. Diese ist der Prototyp einer „elliptischen“ Differentialgleichung.

## 0.3 Beispiele

Die folgenden Beispiele aus verschiedenen Wissenschaftsdisziplinen vermitteln einen Eindruck von der Vielfältigkeit der auftretenden Probleme.

### 1. Erhaltungsgleichungen

Die Grundgleichungen der (klassischen) Kontinuumsmechanik basieren auf dem physikalischen Grundprinzip der „Erhaltung“, d. h.: Zustandsgrößen wie z. B. Masse  $\rho(x, t)$ , innere Energie bzw. Temperatur  $T(x, t)$ , Impuls  $\rho(x, t)\vec{v}(x, t)$ , u.s.w., werden als Dichtefunktionen beschrieben, deren Integrale über beliebige bewegte „Kontrollvolumen“ sich beim Fehlen von äußeren Einflüssen nicht verändern. Für die zeitliche Veränderung der Masse  $m_V(t)$  eines solchen mit dem Geschwindigkeitsfeld  $v = (v_1, v_2, v_3)$  bewegten Volumens  $V(t)$  gilt (sog. „Reynoldsches<sup>10</sup> Transporttheorem“)

$$0 = d_t m_V(t) = d_t \int_{V(t)} \rho(x, t) dx = \int_{V(t)} \{\partial_t \rho + \nabla \cdot (\rho v)\} dx. \quad (0.3.9)$$

---

<sup>9</sup>Siméon Denis Poisson (1781–1840): Französischer Mathematiker und Physiker; Prof. in Paris; Beiträge zur mathematischen Formulierung der Physik, zum Magnetismus, zur Himmelsmechanik und Wahrscheinlichkeitsrechnung; einer der Begründer der Potentialtheorie.

<sup>10</sup>Osborn Reynolds (1842–1912): Englischer Ingenieur und Mathematiker; Prof. in Manchester; Beiträge zur Theorie des Elektro-Magnetismus und der Strömungslehre, Fundamente der Turbulenzbeschreibung und der hydrodynamischen Stabilität.

Da dies für beliebige Volumen  $V(t)$  gelten soll, ergibt sich für stetige Dichtefunktionen die folgende Erhaltungsgleichung 1. Ordnung (sog. „Kontinuitätsgleichung“):

$$\partial_t \rho + \nabla \cdot (\rho v) = 0. \quad (0.3.10)$$

Auf analogem Wege erhält man aus dem Erhaltungssatz für die Temperatur unter Berücksichtigung von Quelltermen und Wärmediffusion in einem ruhenden Medium ( $\vec{v} \equiv 0$ ) die folgende Erhaltungsgleichung 2. Ordnung (sog. „Wärmeleitungsgleichung“):

$$\partial_t T - \nabla \cdot (a \nabla T) = \nabla \cdot q, \quad (0.3.11)$$

Physikalische Anschauung erfordert, dass Lösungen zu diesen Gleichungen bei physikalisch sinnvollen Anfangs- und Randbedingungen stets positiv sind:  $\rho > 0$ ,  $T > 0$ . Wir werden sehen, dass dies tatsächlich der Fall ist. Die entsprechenden Erhaltungssätze für Impuls und Drehimpuls führen unter geeigneten zusätzlichen Annahmen zusammen mit der Kontinuitätsgleichung auf die bekannten „Navier<sup>11</sup>-Stokes<sup>12</sup>-Gleichungen“ für *inkompressible*, Newtonsche<sup>13</sup> Fluide mit konstanter Dichte und Temperatur ( $\rho \equiv \text{konst.}$ ,  $T \equiv \text{konst.}$ ):

$$\partial_t v - \nu \Delta v + v \cdot \nabla v + \nabla p = f, \quad \nabla \cdot v = 0. \quad (0.3.12)$$

Diese werden hier aber nur der Vollständigkeit halber formuliert und im Laufe dieses Textes nicht näher betrachtet.

## 2. Variationsgleichungen

Die Grundgleichungen der (klassischen) Elastizitätstheorie basieren auf dem physikalischen Prinzip der „Energieminimierung“, d. h.: Der von einem elastischen Körper unter statischer äußerer Belastung eingenommene Zustand ist so bestimmt, dass er die potentielle Gesamtenergie des Systems minimiert. Zur Beschreibung eines elastischen Systems werden in der *linearen* Theorie die folgenden Größen verwendet: der Verschiebungsvektor  $u$ , der Verzerrungstensor  $\varepsilon(u) := \frac{1}{2}(\nabla u + \nabla u^T)$ , der flächenorientierte (symmetrische) Spannungstensor  $\sigma$  und die Volumenkraft  $f$ . Die Spannungen  $\sigma$  sind die Reaktionskräfte des Körpers auf von außen erzwungene Verzerrungen  $\varepsilon(u)$  und werden idealisierend in einer linearen Beziehung angenommen, dem sog. (elastischen) „Materialgesetz“,  $\sigma = A\varepsilon(u)$ , mit dem symmetrischen und positiv definiten Elastizitätstensor  $A$ . Die potentielle Gesamtenergie des belasteten, elastischen Körpers schreibt sich dann in der Form:

---

<sup>11</sup>Claude (Louise Marie Henri) Navier (1785–): Französischer Bauingenieur und Mathematiker; Prof. an der École Polytechnique in Paris; Beiträge zum Brückenbau (erste Theorie der Hängebrücke), Elastizitätstheorie und Strömungsmechanik.

<sup>12</sup>Sir Georg Gabriel Stokes (1819–1903): Englischer Mathematiker und Physiker; Prof. in Cambridge; Beiträge zur Differential- und Integralrechnung, zur Hydrodynamik und zur Theorie des Lichts, Spektralanalyse und Fluoreszenz.

<sup>13</sup>Isaac Newton (1643–1727): Englischer Physiker und Mathematiker; Professor an der Universität Cambridge; entwickelte u. a. die Grundlagen der klassischen Mechanik und der Differentialrechnung.

$$E(u) = \frac{1}{2} \int_{\Omega} \sigma : \varepsilon(u) \, dx - \int_{\Omega} f u \, dx = \frac{1}{2} \int_{\Omega} A \varepsilon(u) : \varepsilon(u) \, dx - \int_{\Omega} f u \, dx. \quad (0.3.13)$$

Die sich unter der äußeren Belastung einstellende Verschiebung in einen neuen Gleichgewichtszustand  $u_*$  verleiht dann  $E(\cdot)$  einen minimalen Wert unter allen zulässigen Verschiebungen, d. h. solchen mit denselben Randwerten. Für einen solchen optimalen Zustand  $u_*$  gilt dann notwendig

$$\frac{d}{d\varepsilon} E(u_* + \varepsilon\varphi)|_{\varepsilon=0} = 0,$$

für beliebige „zulässige“ Variationen  $\varphi$ . Auswertung dieser Beziehung liefert die sog. „Variationsgleichung“

$$(A \varepsilon(u), \varepsilon(\varphi))_{\Omega} = (f, \varphi)_{\Omega} \quad \forall \text{ „zulässigen“ } \varphi. \quad (0.3.14)$$

Dabei bedeutet „zulässig“ für eine Testfunktion  $\varphi$ , dass sie hinreichend glatt ist und entlang des Randes  $\partial\Omega$  verschwindet. Nimmt man an, dass das Minimum  $u$  hinreichend glatt ist, so folgt durch partielle Integration

$$(\nabla \cdot A \varepsilon(u) + f, \varphi)_{\Omega} = 0 \quad \forall \text{ „zulässigen“ } \varphi,$$

und hieraus das Differentialgleichungssystem 2. Ordnung (in kompakter sowie ausführlicher Schreibweise)

$$-\nabla \cdot A \varepsilon(u) = f \Leftrightarrow -\sum_{i=1}^d \sum_{k,l=1}^d \partial_i A_{ijkl} \varepsilon_{kl}(u) = f_j \quad (j = 1, \dots, d). \quad (0.3.15)$$

Der wohl einfachste (mehrdimensionale) Spezialfall eines elastisch deformierten Körpers ist die „eingespannte Membran“ in einem (beschränkten) Gebiet  $\Omega \subset \mathbb{R}^2$  (Trommelfell). Hier wird eine Belastung nur in vertikaler Richtung zugelassen  $\vec{f} = (0, 0, f)$  und entsprechend auch nur die Auslenkung in vertikaler Richtung  $\vec{u} = (0, 0, u)$  berücksichtigt. Der Elastizitätstensor ist diagonal und wird hier der Einfachheit halber zu  $A = aI$  gesetzt. Dann nimmt das obige Funktional der potentiellen Gesamtenergie die Form an

$$E(u) = \frac{1}{2} a \|\nabla u\|_{\Omega}^2 - (f, u)_{\Omega}, \quad (0.3.16)$$

und die zugehörige Differentialgleichung („Poisson-Gleichung“) lautet

$$-a \Delta u = f, \quad \text{in } \Omega. \quad (0.3.17)$$

Da die Membran am Rand eingespannt sein soll, muss weiter  $u|_{\partial\Omega} = 0$  sein.

In einer elastischen Membran wirken als Gegenkräfte zur Belastung reine Federkräfte (in vertikaler Richtung) und keine Biegemomente. Dies ist begründet durch die Vernachlässigung der Dicke der Membran. Flache Körper mit (konstanter) positiver, aber geringer Dicke werden als „Platten“ bezeichnet. Im Rahmen der linearen „Kirchhoffschén

Plattentheorie“ werden zwar Biegemomente berücksichtigt, aber nur vertikale Verschiebungen zugelassen („Kirchhoff-Hypothese“). Im Fall kleiner Verschiebungen (gegenüber der Plattendicke) erhält dann die potentielle Gesamtenergie die Form

$$E(u) = \frac{1}{2}D\|\Delta u\|_{\Omega}^2 + (1 - \sigma)D\left\{(\partial_1^2 u, \partial_2^2 u)_{\Omega} - \|\partial_1 \partial_2 u\|_{\Omega}^2\right\} - (f, u)_{\Omega},$$

mit Materialparametern  $\sigma \in (0, 1)$ ,  $D > 0$ . Die zulässigen Funktionen müssen in diesem Fall quadratintegrierbare zweite Ableitungen besitzen. Durch den Variationsansatz wie oben erhält man als notwendige (und hinreichende) Bedingung für ein Energieminimum die sog. „Plattengleichung“

$$D\Delta^2 u = f. \quad (0.3.18)$$

Aus naheliegenden Gründen wird der Operator 4. Ordnung  $\Delta^2$  auch „biharmonischer Operator“ genannt. Ist die Platte am Gebietsrand „eingespannt“, so müssen die Randbedingungen  $u|_{\partial\Omega} = 0$ ,  $\partial_n u|_{\partial\Omega} = 0$  erfüllt sein. Wir merken an, dass ein identisches Modell bei der Beschreibung viskoser, inkompressibler Strömungen in zwei Raumdimensionen als Gleichung für die sog. „Stromfunktion“  $\varphi$  auftritt; das Geschwindigkeitsfeld ergibt sich dabei durch Rotationsbildung  $u := (\partial_y \varphi, -\partial_x \varphi)^T$ .

### 3. Schwingungsgleichungen

Wenn der elastische Körper unter Belastung zeitliche Schwingungen ausführen kann, so wird die zugehörige Auslenkung eine Funktion des Orts und der Zeit,  $u(x, t)$ , und genügt im Rahmen der linearen Theorie der sog. „elastische Schwingungsgleichung“

$$\partial_t^2 u - \nabla \cdot A \varepsilon(u) = f. \quad (0.3.19)$$

Die Gleichung für die frei schwingende Membran ist die klassische „Wellengleichung“

$$\partial_t^2 u = a \Delta u. \quad (0.3.20)$$

Diese beschreibt allgemein die Ausbreitung von Wellen in schwingfähigen Medien (z. B. Schallwelle in einem Gas, Auslenkung einer elastischen Membran oder die Amplitude eines elektrischen Feldes). Sie wird üblicherweise zusammen mit „Randbedingungen“  $u|_{\partial\Omega} = 0$  sowie „Anfangsbedingungen“  $u|_{t=0} = u^0$ ,  $\partial_t u|_{t=0} = u^1$  betrachtet.

## 0.4 Numerische Methoden

Aus den oben skizzierten Wegen zur Herleitung von partiellen Differentialgleichungen gewinnt man unmittelbar Ansätze zu deren Diskretisierung.

## 1. Differenzenverfahren

Ersetzt man die Ableitungen in den Differentialgleichungen (0.2.5), (0.2.6), (0.2.7) und (0.2.8) durch geeignete Differenzenquotienten bezüglich eines regulären Punktegitters, gewinnt man lineare Gleichungen für die zugehörigen „diskreten“ Funktionswerte. Dies ist eine sog. „Differenzenapproximation“ der Differentialgleichung. Die Eigenschaften solcher Differenzengleichungen, d. h. ihre Lösbarkeit und Approximationsgüte, werden weiter unten eingehend behandelt.

## 2. Finite-Volumen-Verfahren

Die lokale Erhaltungseigenschaft (0.3.9) führt dadurch auf ein numerisches Verfahren, dass man für die Lösung einen bzgl. einer endlichen Zerlegung des Gebiets  $\Omega$  in Teilvolumina (z. B. Dreiecke oder Vierecke) stückweise konstanten Ansatz macht und die Gültigkeit der Erhaltungseigenschaften dafür auf jedem der sog. „Kontrollvolumen“, fordert. Dies führt auf ein algebraisches Gleichungssystem für die zugehörigen Zellmittelwerte. Dieser Diskretisierungsansatz wird „Finite-Volumen-Methode“ genannt. Derartige Methoden finden hauptsächlich im Bereich der numerischen Strömungsmechanik Anwendung. Wegen ihrer eingeschränkten Anwendbarkeit und schwierigen Analysierbarkeit werden wir uns mit dieser Methodenklasse in dieser Vorlesung nicht weiter befassen.

## 3. Variationsmethoden („Methode der finiten Elemente“)

Die Variationsgleichung (0.3.14) führt durch Einschränkung auf einen endlich dimensionalen Teilraum  $V_h$  des Lösungsraums des kontinuierlichen Variationsproblems auf eine diskrete Variationsaufgabe

$$(A\varepsilon(u_h), \varepsilon(\varphi_h))_\Omega = (f, \varphi_h)_\Omega \quad \forall \text{ „zulässigen“ } \varphi_h \in V_h. \quad (0.4.21)$$

Dies ist äquivalent zu einem linearen (quadratischen) Gleichungssystem für die Entwicklungskoeffizienten der diskreten Lösung  $u_h$  bzgl. einer geeignet gewählten Basis des Ansatzraumes  $V_h$ . Sind die Funktionen in  $V_h$  stückweise polynomial bzgl. einer Zerlegung des Gebiets  $\Omega$  z. B. in Dreiecke oder Vierecke gewählt, liegt eine sog. „Finite-Elemente-Methode“ vor. Diese Methoden werden unten sehr ausführlich behandelt.

# 1 Theorie partieller Differentialgleichungen

In diesem Kapitel wird eine Einführung in die Theorie der partiellen Differentialgleichungen gegeben, soweit sie für deren numerische Behandlung relevant ist. Wir betrachten zunächst lineare Differentialgleichungen zweiter Ordnung der Form

$$Lu := - \sum_{i,j=1}^d a_{ij} \partial_i \partial_j u + \sum_{j=1}^d a_j \partial_j u + au = f, \quad (1.0.1)$$

mit gegebenen Koeffizientenfunktionen  $a_{ij}$ ,  $a_j$ ,  $a$  und rechter Seite  $f$ . Wenn diese Funktionen nicht zusätzlich von der unbekannten Lösung  $u$  abhängen, nennt man die Gleichung „linear“. Wegen der Vertauschbarkeit der Reihenfolge der Ableitungen kann o.B.d.A.  $a_{ij} = a_{ji}$  angenommen werden. Für allgemeine nichtlineare Gleichungen

$$F(x, u, \nabla u, \nabla^2 u) = 0 \quad (1.0.2)$$

gibt es keine einheitliche Lösungstheorie. Wir beschränken uns im Folgenden daher im Wesentlichen auf lineare Probleme.

Differentialgleichungen werden in der Regel auf (beschränkten oder halbbeschränkten) Gebieten  $\Omega \subset \mathbb{R}^d$  betrachtet. Dazu kommen dann noch Bedingungen entlang des Randes  $\partial\Omega$ . Die geeignete Wahl dieser „Randbedingungen“ ist eine sehr delikate Sache und erfordert eingehende Berücksichtigung der speziellen Eigenschaften des Differentialoperators. Diesen wird der nächste Abschnitt gewidmet sein.

Damit eine Differentialgleichung mit den zugehörigen Randbedingungen ein sinnvolles Modell eines realen physikalischen Vorgangs ist, sind eine Reihe von Forderungen zu stellen:

- i) *Existenz* von Lösungen in einem möglicherweise verallgemeinerten Sinne; unter einer „klassischen“ Lösung versteht man eine solche, für die alle auftretenden Ableitungen im Gebietsinnern im strengen Sinne definiert sind und die bis an den Rand stetig ist, d. h.:  $u \in C^2(\Omega) \cap C(\bar{\Omega})$ .
- ii) *Eindeutigkeit* der Lösungen möglicherweise unter Hinzunahme von weiteren physikalisch motivierten Bedingungen.
- iii) *Stetige Abhängigkeit* von den Daten wegen der meist inexakten Verfügbarkeit von Koeffizienten und Randdaten in den physikalischen Modellen; Lösungen sollten sich unter kleinen Datenstörungen auch nur wenig ändern.

Eine Aufgabe, welche diesen Minimalforderungen genügt, nennt man „wohl-gestellt“ (im Sinne von Hadamard<sup>1</sup>).

Bei Anfangs- oder Randwertaufgaben gewöhnlicher Differentialgleichungen war die

---

<sup>1</sup>Jacque Salomon Hadamard (1865–1963): Französischer Mathematiker; Prof. in Bordeaux und Paris; viele wichtige Beiträge zur komplexen Analysis und speziellen Funktionen, zur analytischen Zahlentheorie, zur Variationsrechnung und zu den Differentialgleichungen der mathematischen Physik.

Regularität der Lösung kein besonderes Thema, da sich die Regularität der Daten direkt auf die entsprechende der Lösung überträgt. Bei partiellen Differentialgleichungen ist dies nicht immer der Fall und bedarf für verschiedene Typen von Differentialgleichungen gesonderter Untersuchung.

## 1.1 Typeneinteilung

Partielle Differentialgleichungen lassen sich in drei Haupttypen einteilen: die „elliptischen“, die „parabolischen“ und die „hyperbolischen“ Gleichungen. Wir werden das dieser Unterteilung zugrunde liegende Prinzip anhand einer leicht überschaubaren Situation erläutern. Dies sind die linearen, skalaren Gleichungen 2. Ordnung in zwei Variablen:

$$Lu = a_{11}\partial_x^2 u + 2a_{12}\partial_x\partial_y u + a_{22}\partial_y^2 u + a_1\partial_x u + a_2\partial_y u + au = f$$

mit konstanten Koeffizienten  $a_{ij}$ . Dabei sollen nicht alle drei Koeffizienten  $a_{11}, a_{12}, a_{22}$  der Ableitungen zweiter Ordnung gleichzeitig Null sein. Diese Gleichung wird auf einem Gebiet  $\Omega \subset \mathbb{R}^2$  betrachtet.

Ausgangspunkt ist ein direkter Lösungsansatz, wie er auch bei gewöhnlichen Differentialgleichungen angewendet werden kann. Für die Anfangswertaufgabe

$$u'(t) = f(t, u(t)), \quad t \geq 0, \quad u(0) = u^0,$$

erhält man aus der Vorgabe  $u(0) = u^0$  durch sukzessives Differenzieren von  $f(t, x)$  Formeln für alle Ableitungen von  $u$ :

$$u^{(i)}(0) = \frac{d^{i-1}}{dt^{i-1}} f(0, u^0) =: f^{(i-1)}(0, u^0), \quad i = 1, 2, 3, \dots$$

Wenn die Ableitungen  $f^{(i-1)}$  nicht zu schnell wachsen (für  $f(t, x) = \sin(x)$  sind sie z. B. gleichmäßig beschränkt), konvergiert die Taylor<sup>2</sup>-Reihe

$$u(t) = u^0 + \sum_{i=1}^{\infty} \frac{t^i}{i!} f^{(i-1)}(0, u^0)$$

für alle  $t \geq 0$  absolut und stellt die (eindeutige) Lösung der Anfangswertaufgabe dar.

Wir versuchen, diese Konstruktion auf partielle Differentialgleichungen zu übertragen. Dazu sei  $\Gamma$  ein Jordan<sup>3</sup>-Kurvenstück in  $\Omega$  mit beliebig oft differenzierbarer Parametrisierung  $\Gamma = \{(x(\tau), y(\tau)), \tau \in [0, 1]\}$ . Entlang  $\Gamma$  seien für die Lösung  $u(x, y)$  der Differentialgleichung die Funktionswerte  $u$  sowie ihre Ableitungen  $\partial_n u$  in Normalenrichtung n

---

<sup>2</sup>Brook Taylor (1685–1731): Englischer Mathematiker und Schüler Newtons; die nach ihm benannte Reihenentwicklung war im Kern bereits Gregory, Newton, Leibniz und Johann Bernoulli bekannt.

<sup>3</sup>Marie Ennemond Camille Jordan (1838–): Französischer Mathematiker; Prof. in Paris; Beiträge zur Algebra, Gruppentheorie, Analysis und Topologie.

zu  $\Gamma$  vorgegeben. Dies entspricht der Tatsache, dass wir es mit einer Differentialgleichung 2. Ordnung zu tun haben. Mit  $u$  und  $\partial_n u$  ist der ganze Gradient  $\nabla u = (\partial_x u, \partial_y u)^T$  entlang  $\Gamma$  bekannt. Wir wollen versuchen, aus diesen Vorgaben alle weiteren Ableitungen von  $u$  entlang  $\Gamma$  zu bestimmen, um damit wieder einen Taylor-Reihenansatz für  $u$  in einer Umgebung von  $\Gamma$  zu machen. Zu diesem Zweck führen wir die Abkürzungen ein:

$$p := \partial_x u, \quad q := \partial_y u, \quad r := \partial_x^2 u, \quad s := \partial_x \partial_y u, \quad t := \partial_y^2 u.$$

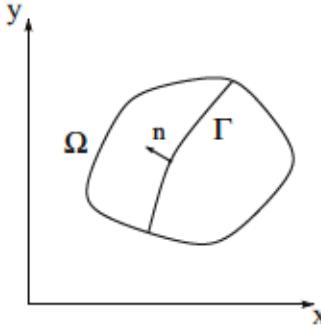


Abbildung 1.1: Konfiguration der allgemeinen partiellen Differentialgleichung

Differentiation von  $p$  und  $q$  entlang  $\Gamma$ , d. h. bzgl. des Parameters  $\tau$ , ergibt

$$\begin{aligned}\partial_\tau p &= \partial_x p \partial_\tau x + \partial_y p \partial_\tau y = r \partial_\tau x + s \partial_\tau y, \\ \partial_\tau q &= \partial_x q \partial_\tau x + \partial_y q \partial_\tau y = s \partial_\tau x + t \partial_\tau y.\end{aligned}$$

mit den bekannten tangentialen Ableitungen  $\partial_\tau x$  und  $\partial_\tau y$  entlang  $\Gamma$ . Zusammen mit der Differentialgleichung  $Lu = f$  ergibt dies ein  $3 \times 3$ -Gleichungssystem für die drei gesuchten Ableitungen  $r, s, t$ :

$$\begin{aligned}a_{11}r + 2a_{12}s + a_{22}t &= f - a_1 p - a_2 q - au \\ \partial_\tau x r + \partial_\tau y s &= \partial_\tau p \\ \partial_\tau x s + \partial_\tau y t &= \partial_\tau q\end{aligned}$$

mit entlang  $\Gamma$  bekannter rechter Seite. Die Determinante der Koeffizientenmatrix  $B$  erhält man durch Entwicklung nach der ersten Zeile zu

$$\det B = a_{11} \partial_\tau y^2 - 2a_{12} \partial_\tau x \partial_\tau y + a_{22} \partial_\tau x^2.$$

Wir unterscheiden jetzt zwei Fälle.

i) **Fall  $\det B \neq 0$  entlang ganz  $\Gamma$ :**

In diesem Fall sind alle zweiten Ableitungen  $r, s, t$  von  $u$  durch Vorgabe von  $u, \partial_n u$  entlang  $\Gamma$  (eindeutig) bestimmbar. Durch weitere Differentiation des Gleichungssystems nach  $x$  und  $y$  erhält man wieder ein System für die dritten Ableitungen  $\partial_x r, \partial_x s, \partial_x t$

sowie  $\partial_y r, \partial_y s, \partial_y t$  jeweils mit derselben Koeffizientenmatrix. Durch weiteres Differenzieren lassen sich so alle höheren Ableitungen von  $u$  entlang  $\Gamma$  bestimmen. Durch den Reihenansatz

$$u(x, y) = \sum_{i+j \geq 0} \frac{(x - x_0)^i (y - y_0)^j}{(i + j)!} \partial_x^i \partial_y^j u(x_0, y_0)$$

bzgl. eines Punktes  $(x_0, y_0) \in \Gamma$  erhält man dann in einer Umgebung der Kurve  $\Gamma$  eine Lösung der Differentialgleichung, die auf  $\Gamma$  die vorgegebenen Werte annimmt. Diese nennt man Lösung der „Cauchyschen<sup>4</sup> Anfangswertaufgabe“ der Differentialgleichung bzgl. der „Anfangskurve“  $\Gamma$ .

ii) Fall  $\det B = 0$  in einem Punkt  $(x_0, y_0) \in \Gamma$ :

Die quadratische Gleichung

$$a_{11} \partial_\tau y^2 - 2a_{12} \partial_\tau x \partial_\tau y + a_{22} \partial_\tau x^2 = 0$$

bestimmt gewisse Richtungen  $\partial_\tau y / \partial_\tau x = dy/dx$  bzw.  $\partial_\tau x / \partial_\tau y = dx/dy$  von Kurven (mit Graph  $y = y(x)$  oder  $x = x(y)$ ) durch den Punkt  $(x_0, y_0)$ . Zu deren Bestimmung sei etwa angenommen, dass  $a_{11} \neq 0$  und  $\partial_\tau x \neq 0$ . Dann besitzt die Gleichung

$$\left(\frac{dy}{dx}\right)^2 - \frac{2a_{12}}{a_{11}} \left(\frac{dy}{dx}\right) + \frac{a_{22}}{a_{11}} = 0$$

die Lösungen

$$\left(\frac{dy}{dx}\right)_{+/-} = \frac{a_{12}}{a_{11}} \pm \frac{1}{a_{11}} \sqrt{a_{12}^2 - a_{11}a_{22}}.$$

Diese entsprechen Steigungen von Kurven durch den Punkt  $(x_0, y_0) \in \Gamma$ , entlang welcher die höheren Ableitungen von  $u$  sich nicht aus den Vorgaben entlang  $\Gamma$  bestimmen lassen. Entlang dieser kritischen, auch „charakteristisch“ genannten Kurven (sog. „Charakteristiken“ des Differentialoperators  $L$ ) lässt sich also die Lösung der Differentialgleichung nicht aus den obigen Vorgaben konstruieren. Entlang solcher Kurven können Unstetigkeiten in der Lösung oder ihres Gradienten auftreten. Es ist also sehr wichtig, die Existenz von Charakteristiken und deren Gestalt für den zu betrachtenden Differentialoperator vor Ansatz eines numerischen Verfahrens genau zu bestimmen. Offensichtlich hängt die Existenz von Charakteristiken allein von den Koeffizienten der höchsten Ableitungen des Operators  $L$ , d. h. seinem sog. „Hauptteil“  $a_{11} \partial_x^2 u + 2a_{12} \partial_x \partial_y u + a_{22} \partial_y^2 u$ , ab. Diesem wird die quadratische Form

$$q(x, y) := a_{11}x^2 + 2a_{12}xy + a_{22}y^2$$

zugeordnet. Die Gleichung  $q(x, y) = 0$  beschreibt Kegelschnitte in der  $(x, y)$ -Ebene:

$$a_{12}^2 - a_{11}a_{22} \begin{cases} < 0 : & \text{Ellipse,} \\ = 0 : & \text{Parabel,} \\ > 0 : & \text{Hyperbel.} \end{cases}$$

---

<sup>4</sup>Augustin Louis Cauchy (1789–1857): Ingenieur, Physiker und bedeutendster französischer Mathematiker seiner Zeit; wirkte an der École Polytechnique und der Sorbonne in Paris; gilt als Begründer der modernen Analysis und der Funktionentheorie.

Von dieser rein formalen Charakterisierung stammen die obigen Bezeichnungen für die drei Typen von partiellen Differentialgleichungen. Die Klassifikation eines Differentialoperators als „elliptisch“, „parabolisch“ oder „hyperbolisch“ wird für jeden einzelnen Punkt  $(x_0, y_0)$  separat vorgenommen. Im Falle variabler Koeffizienten  $a_{ij} = a_{ij}(x, y)$  oder im nichtlinearen Fall  $a_{ij}(u(x, y))$  kann der Typ einer Gleichung also im Lösungsgebiet wechseln. Wir werden im Folgenden nur Gleichungen eines einheitlichen Typs betrachten; in vielen Anwendungen spielt aber gerade der Typwechsel eine wichtige Rolle.

Als Nächstes wollen wir die prototypischen Vertreter von (linearen) elliptischen, parabolischen und hyperbolischen Differentialgleichungen ableiten. Dies wird uns erneut auf die obige Typenunterteilung führen. Dazu schreiben wir den Hauptteil  $L_0$  des Differentialoperators in Matrix-Vektor-Form:

$$L_0 = a_{11}\partial_x^2 + 2a_{12}\partial_x\partial_y + a_{22}\partial_y^2 = \begin{pmatrix} \partial_x \\ \partial_y \end{pmatrix}^T \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} \partial_x \\ \partial_y \end{pmatrix} = \nabla^T A \nabla$$

Die symmetrische Matrix  $A$  besitzt zwei reelle Eigenwerte  $\lambda, \mu$  und ein zugehöriges Orthonormalsystem von Eigenvektoren  $\{\xi, \eta\}$ . Mit der Spaltenmatrix  $Q := [\xi, \eta]$  gilt

$$QQ^T = I, \quad Q^T A Q = D = \text{diag}(\lambda, \mu).$$

Damit können wir schreiben:

$$\begin{aligned} L_0 &= \nabla^T Q D Q^T \nabla = (Q^T \nabla)^T D (Q^T \nabla) \\ &= \begin{pmatrix} \xi_1 \partial_x + \xi_2 \partial_y \\ \eta_1 \partial_x + \eta_2 \partial_y \end{pmatrix}^T \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix} \begin{pmatrix} \xi_1 \partial_x + \xi_2 \partial_y \\ \eta_1 \partial_x + \eta_2 \partial_y \end{pmatrix} \end{aligned}$$

bzw.

$$L_0 = \lambda(\xi_1 \partial_x + \xi_2 \partial_y)^2 + \mu(\eta_1 \partial_x + \eta_2 \partial_y)^2,$$

oder mit den Richtungsableitungen  $\partial_\xi = \xi \cdot \nabla$  und  $\partial_\eta = \eta \cdot \nabla$ :

$$L_0 = \lambda \partial_\xi^2 + \mu \partial_\eta^2.$$

Die Eigenwerte erhält man als Nullstellen des charakteristischen Polynoms

$$\begin{aligned} \det(A - zI) &= (a_{11} - z)(a_{22} - z) - a_{12}^2 = z^2 - (a_{11} + a_{22})z + a_{11}a_{22} - a_{12}^2 \\ &= (z - \lambda)(z - \mu) = z^2 - (\lambda + \mu)z + \lambda\mu. \end{aligned}$$

Durch Koeffizientenvergleich findet man („Vietascher<sup>5</sup> Wurzelsatz“).

$$\lambda + \mu = a_{11} + a_{22}, \quad \lambda\mu = a_{11}a_{22} - a_{12}^2.$$

---

<sup>5</sup>Francois Viete, lat. Franciscus Vieta (1540–1603): Französischer Mathematiker; Arbeiten über algebraische Gleichungen und sphärische Trigonometrie; gab trigonometrische Tafeln heraus und führte die systematische Buchstabenrechnung ein.

a) „Elliptischer“ Fall  $a_{12}^2 - a_{11}a_{22} < 0$ :

Beide Eigenwerte  $\lambda, \mu$  sind ungleich Null und haben dasselbe Vorzeichen. Die Lösungen der charakteristischen Gleichung sind nicht reell, d. h.: Es existieren keine charakteristischen Kurven durch den Punkt  $(x_0, y_0)$ . In diesem Fall ist der Konstruktionsprozess für die höheren Ableitungen von  $u$  durchführbar. Die Normalform eines elliptischen Operators  $L$  ist im Fall  $\lambda = \mu = 1$ :

$$Lu = \partial_\xi^2 u + \partial_\eta^2 u + \psi(\xi, \eta, u, \partial_\xi u, \partial_\eta u)$$

Der „Hauptteil“ dieses Operators ist also gerade der „Laplace-Operator“  $\Delta$ , der sich somit als prototypischer Vertreter elliptischer Differentialoperatoren 2. Ordnung erweist. Wir werden uns daher im Folgenden hauptsächlich mit der zugehörigen Poisson-Gleichung beschäftigen:

$$\partial_x^2 u + \partial_y^2 u = f. \quad (1.1.3)$$

b) „Parabolischer“ Fall:  $a_{12}^2 - a_{11}a_{22} = 0$

Einer der Eigenwerte ist Null; der zweite ist dann notwendig ungleich Null. Es existiert genau eine charakteristische Richtung im Punkt  $(x_0, y_0)$  mit der Steigung  $dy/dx = a_{12}/a_{11}$ . Die Normalform eines parabolischen Operators  $L$  ist im Fall  $\lambda = 1, \mu = 0$ :

$$Lu = \partial_\xi^2 u + \psi(\xi, \eta, u, \partial_\xi u, \partial_\eta u).$$

Der Hauptteil dieses Operators ist im linearen Fall gerade der sog. „Wärmeleitungsoperator“, der prototypische Vertreter parabolischer Differentialoperatoren 2. Ordnung. Wir werden uns daher im Folgenden mit der zugehörigen Wärmeleitungsgleichung beschäftigen:

$$\partial_t u - \partial_x^2 u = f. \quad (1.1.4)$$

c) „Hyperbolischer“ Fall:  $a_{12}^2 - a_{11}a_{22} > 0$

Beide Eigenwerte sind ungleich Null, haben aber verschiedene Vorzeichen. Es existieren zwei charakteristische Richtungen im Punkt  $(x_0, y_0)$  mit den Steigungen  $(dy/dx)_\pm = a_{12}/a_{11} \pm a_{11}^{-1}\sqrt{a_{12}^2 - a_{11}a_{22}}$ . Die Normalform eines hyperbolischen Differentialoperators  $L$  ist im Fall  $\lambda = 1, \mu = -1$ :

$$Lu = \partial_\xi^2 u - \partial_\eta^2 u + \psi(\xi, \eta, u, \partial_\xi u, \partial_\eta u).$$

Der Hauptteil dieses Operators ist der sog. „Wellenoperator“, der prototypische Vertreter hyperbolischer Differentialoperatoren 2. Ordnung. Wir werden uns daher im Folgenden mit der zugehörigen Wellengleichung beschäftigen:

$$\partial_t^2 u - \partial_x^2 u = f. \quad (1.1.5)$$

Wir haben gesehen, dass die „Cauchysche Anfangswertaufgabe“ durch Reihenansatz lösbar ist, wenn die „Anfangskurve“  $\Gamma$  nirgends mit einer Charakteristik des Differential-

operators zusammenfällt. Andernfalls kann die Situation eintreten, dass zu beiden Seiten der Kurve  $\Gamma$  eine Lösung existiert, diese aber nicht auf  $\Gamma$  stetig-differenzierbar fortsetzbar ist. Im Folgenden werden wir die für die drei Gleichungstypen geeigneten Randbedingungen diskutieren und dabei ganz unterschiedliche Ergebnisse erhalten.

## 1.2 Elliptische Probleme

Wir haben gesehen, dass für (im ganzen Lösungsgebiet  $\Omega$ ) elliptische Differentialoperatoren die „Cauchysche Anfangswertaufgabe“ für jede (analytische) „Anfangskurve“  $\Gamma \subset \Omega$  lösbar ist. Die Verallgemeinerung dieser Aussage für nichtlineare Differentialoperatoren der Art

$$L(u) = \partial_x^2 u - F(x, y, u, \partial_x u, \partial_y u, \partial_y^2 u) \quad (1.2.6)$$

ist der berühmte Satz von Cauchy-Kovalevskaya<sup>6</sup>. Dieser sehr allgemeine Existenzsatz für elliptische Differentialgleichungen ist aber für die Praxis nur von geringer Bedeutung. Die über einen lokalen Reihenansatz konstruierte Lösung  $u$  hängt nämlich i. Allg. nicht stetig von den vorgegebenen Anfangswerten entlang  $\Gamma$  ab. Dies ist aber eine unverzichtbare Bedingung an ein physikalisch sinnvolles Modell.

**Beispiel 1.1:** In der Halbebene  $\Omega = \{(x, y) \in \mathbb{R}^2 : x > 0\}$  seien entlang der Randkurve  $\Gamma = \{(0, y) \in \mathbb{R}^2\}$  die Randwerte  $u(0, y) = u_0^0(y) = 0$ ,  $\partial_x u(0, y) = u_0^1(y) = 0$  gegeben. Die zugehörige Lösung der Poisson-Gleichung  $\Delta u = 0$  ist  $u \equiv 0$ . Mit  $\varepsilon > 0$  seien die Randdaten nun gestört zu  $u_\varepsilon^0(y) = 0$ ,  $u_\varepsilon^1(y) = \varepsilon \sin(y/\varepsilon)$ , wobei  $\lim_{\varepsilon \rightarrow 0} u_\varepsilon^1(y) = 0$ . Die zugehörige gestörte Lösung der Poisson-Gleichung (nachrechnen!)

$$u_\varepsilon(x, y) = \varepsilon^2 \sin(y/\varepsilon) \sinh(x/\varepsilon), \quad \sinh(z) = \frac{1}{2}(e^z - e^{-z}),$$

konvergiert aber für  $\varepsilon \rightarrow 0$  nicht gegen Null. Es zeigt sich, dass in diesem Fall entlang der Anfangskurve  $\Gamma$  nicht gleichzeitig Werte für  $u$  und  $\partial_n u$  vorgegeben werden dürfen, wenn man an physikalisch sinnvollen Lösungen interessiert ist.

Wir haben gesehen, dass man bei der Wahl von Randbedingungen für elliptische Operatoren vorsichtig sein muss, wenn das resultierende Randwertproblem wohl-gestellt sein soll. Sei also  $\Omega \subset \mathbb{R}^2$  ein beschränktes Gebiet mit hinreichend glattem Rand  $\partial\Omega$ . Wir wollen dabei Ränder mit einer glatten Parametrisierung (mindestens zweimal stetig differenzierbar) oder ein Polygonegebiet (mit endlich vielen Ecken) zulassen. Als prototypischen Modellfall betrachten wir die „Poisson-Gleichung“

---

<sup>6</sup>Sofia Vasilyevna Kovalevskaya (1850–1891): Russische Mathematikerin, eine der ersten Frauen mit Universitätskarriere; 1869 Studium in Heidelberg als „Gasthörerin“, da hier für Frauen ein offizielles Universitätsstudium noch nicht möglich war; ab 1871 Studium in Berlin bei Weierstrass und danach in Göttingen; eine ihrer ersten Veröffentlichungen enthält den nach ihr benannten „Existenzsatz“; ab 1884 Stelle als Privatdozentin in Stockholm; leistete Beiträge zur Analysis und zur Theorie von Differentialgleichungen der Physik.

$$-\Delta u = f \quad \text{auf } \Omega. \quad (1.2.7)$$

Es gibt drei Typen von Randbedingungen und zugehörige Randwertaufgaben („RWAn“):

- a) *Dirichletsche<sup>7</sup> Randbedingungen* („1. RWA“):  $u = g$  auf  $\partial\Omega$ .
- b) *Neumannsche<sup>8</sup> Randbedingungen* („2. RWA“):  $\partial_n u = g$  auf  $\partial\Omega$ .
- c) *Robinsche<sup>9</sup> Randbedingungen* („3. RWA“):  $\partial_n u + \alpha u = g$  auf  $\partial\Omega$ .

Die Randfunktionen  $g$  werden i. Allg. als glatt und  $\alpha \geq 0$  angenommen. Alle diese RWAn sind, wie wir zum Teil zeigen werden, unter geeigneten Zusatzbedingungen an die Daten wohl gestellt.

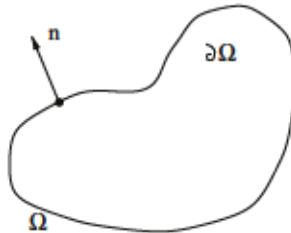


Abbildung 1.2: Konfiguration der elliptischen Randwertaufgabe

### 1.2.1 Existenz von Lösungen

Die Frage nach der Existenz von Lösungen der 1. RWA ist wesentlich schwieriger als bei den RWAn gewöhnlicher Differentialgleichungen. Als Vorbereitung für analoge Argumente im Zusammenhang mit den Diskretrisierungsverfahren wollen wir zwei völlig unterschiedliche Zugänge zu dieser Frage diskutieren, den „klassischen“, potentialtheoretischen und den „modernen“, funktionalanalytischen. Der Einfachheit halber wird nur der Fall „homogener“ Dirichlet-Randbedingungen  $u|_{\partial\Omega} = 0$  betrachtet.

#### i) Potentialtheoretische Methode:

Zunächst ist der Begriff einer „klassischen“ Lösung für die Dirichletschen Randbedingung zu präzisieren. Wir verstehen darunter eine Funktion  $u \in C^2(\Omega) \cap C(\bar{\Omega})$ , welche im Innern von  $\Omega$  der Differentialgleichung und entlang des Randes  $\partial\Omega$  der Randbedingung genügt.

<sup>7</sup>Johann Peter Gustav Lejeune Dirichlet (1805–1859): Geb. in Duren (damals bei Frankreich; wirkte in Berlin und Göttingen (Nachfolger von Gauss); wichtige Beiträge zur Zahlentheorie, Analysis und Differentialgleichungen („Dirichletsches Prinzip“).

<sup>8</sup>Carl Gottfried Neumann (1832–1925): Deutscher Mathematiker; seit 1858 Privatdozent und seit 1863 apl. Prof. in Halle; nach Professuren in Basel und Tübingen ab 1868 in Leipzig; lieferte Beiträge zur Theorie der (partiellen) Differential- und Integralgleichungen, insbesondere zum Dirichlet-Problem. Die „Neumann-Randbedingungen“ sowie die „Neumann-Reihe“ sind nach ihm benannt; gründete zusammen mit Alfred Clebsch die Zeitschrift Mathematische Annalen.

<sup>9</sup>Victor Gustave Robin (1855–1897): Französischer Mathematiker; lehrte an der Sorbonne in Paris; Beiträge zur Potentialtheorie und Thermodynamik; hat die nach ihm benannte 3. Randbedingung anscheinend selbst gar nicht benutzt.

Ferner soll (aus physikalischen Gründen) ihr Gradient (möglicherweise im uneigentlichen Riemannschen<sup>10</sup> Sinne) quadratintegrierbar sein:  $|\nabla u| \in L^2(\Omega)$ . Zur Konstruktion solcher klassischer Lösungen postulieren wir zunächst die Existenz einer Funktion  $G(x, y)$  auf  $\overline{\Omega} \times \overline{\Omega}$ ,

$$G \in C^2(\{\Omega \times \Omega\} \setminus \{x = y\}) \cap C(\{\overline{\Omega} \times \overline{\Omega}\} \setminus \{x = y\}),$$

mit den Eigenschaften

$$-\Delta_x G(\cdot, y) = 0 \quad \text{in } \Omega \setminus \{y\}, \quad G(\cdot, y) = 0 \quad \text{auf } \partial\Omega \setminus \{y\}, \quad (1.2.8)$$

für beliebiges festes  $y \in \overline{\Omega}$ . Für  $x = y$  habe  $G(x, y)$  eine dimensionsabhängige Singularität, die so beschaffen ist, dass sich  $-\Delta_x G(\cdot, y)$  wie die Distribution  $\delta(\cdot, y)$  verhält, d. h.: Für  $v \in C(\overline{\Omega})$  gilt mit den Kugelumgebungen  $B_\varepsilon = B_\varepsilon(x) := \{y \in \Omega : |y - x| \leq \varepsilon\}$ ,  $\varepsilon > 0$ :

$$x \in \Omega : \lim_{\varepsilon \rightarrow 0} \int_{\Omega \setminus B_\varepsilon} -\Delta_x G(x, y) v(y) dy = v(x), \quad (1.2.9)$$

$$x \in \partial\Omega : \lim_{\varepsilon \rightarrow 0} \int_{\partial\Omega \setminus B_\varepsilon} \partial_n G(x, y) v(y) dy = v(x). \quad (1.2.10)$$

Eine solches  $G(x, y)$  wird „Greensche Funktion (1. Art)“ genannt. Wir machen damit den Lösungsansatz

$$u(x) := \int_{\Omega} G(x, y) f(y) dy + \int_{\partial\Omega} \partial_n G(x, y) g(y) do_y.$$

Die formulierten Eigenschaften der Greenschen Funktion erlauben es, zu zeigen, dass dieser Ansatz tatsächlich eine klassische Lösung der 1. RWA liefert. Diese Rechnung ist aufwendig und kann z. B. im Buch von Hellwig [5] nachgelesen werden.

Die Konstruktion einer Greenschen Funktion für allgemeine Gebiete im  $\mathbb{R}^d$  ist schwer. Im Fall  $d = 1$  ist aber eine explizite Konstruktion möglich und für  $d = 2$  folgt ihre Existenz mit Hilfe des Riemannschen Abbildungssatzes aus der Theorie komplexer Funktionen (siehe Hellwig [5]). Auch für sehr spezielle Konfigurationen, wie z. B. Halbebenen oder Kreise, lässt sich hier die Greensche Funktion explizit angeben.

**Beispiele:** (i) Fall  $\mathbb{R}^1$ : In einer Dimension lautet die 1. RWA des Laplace-Operators auf dem Gebiet  $\Omega = (0, 1)$  mit homogenen Randbedingungen

$$-u''(x) = f(x), \quad x \in (0, 1), \quad u(0) = u(1) = 0.$$

Wir suchen eine Lösung  $G(x, y)$  der RWA

$$-G''_x(x, y) = \delta(x - y), \quad x \in (0, 1), \quad G(0, y) = G(1, y) = 0, \quad y \in (0, 1) \setminus \{x\}.$$

---

<sup>10</sup>Bernhard Riemann (1826–1866): Deutscher Mathematiker; Prof. in Göttingen als Nachfolger Dirichlets; Mitbegründer der Funktionentheorie und der modernen Geometrie; einer der bedeutendsten Mathematiker seiner Zeit, von großem Einfluss auch auf die theoretische Physik.

Deren erste Ableitung muss für  $x = y$  einen Sprung der Höhe 1 aufweisen. Zweimalige Integration ergibt

$$G'_x(x, y) = \begin{cases} c_1, & 0 \leq x \leq y, \\ c_1 - 1, & y \leq x \leq 1, \end{cases}$$

und

$$G(x, y) = \begin{cases} c_1 x + c_2, & 0 \leq x \leq y, \\ (c_1 - 1)x + c_3, & y \leq x \leq 1. \end{cases}$$

Die Stetigkeit von  $G(x, y)$  bei  $x = y$  bedingt  $c_1 y + c_2 = (c_1 - 1)y + c_3$  bzw.  $c_3 = c_2 + y$  und folglich

$$G(x, y) = \begin{cases} c_1 x + c_2, & 0 \leq x \leq y, \\ (c_1 - 1)x + c_2 + y, & y \leq x \leq 1. \end{cases}$$

Die Randbedingungen  $G(0, y) = G(1, y) = 0$  erfordern  $c_2 = 0$  und  $c_1 - 1 + c_2 + y = 0$  und folglich  $c_2 = 0$  sowie  $c_1 = 1 - y$ . Die Greensche Funktion lautet also

$$G(x, y) = \begin{cases} (1 - y)x, & 0 \leq x \leq y, \\ y(1 - x), & y \leq x \leq 1. \end{cases}$$

Nach Konstruktion hat die Greensche Funktion die Symmetrieeigenschaft  $G(x, y) = G(y, x)$  und für  $x \neq y$  gilt  $G'_x(x, y) = G'_y(y, x)$  sowie  $G''_x(x, y) = G''_y(y, x) = 0$ . Für eine Funktion  $v \in C_0^\infty(0, 1)$  (Raum der „Testfunktionen“) gilt dann im Distributions Sinn:

$$\begin{aligned} - \int_0^1 G''_x(x, y)v(y) dy &= - \int_0^1 G''_y(y, x)v(y) dy \\ &:= - \int_0^1 G(y, x)v''(y) dy \\ &= - \int_0^1 G(x, y)v''(y) dy \\ &= \int_0^1 G'_y(x, y)v'(y) dy - G(0, y)v'(0) + G(1, y)v'(1) \\ &= c_1 \int_0^y v'(x) dx + (c_1 - 1) \int_y^1 v'(x) dx \\ &= c_1(v(y) - v(0)) + (c_1 - 1)(v(1) - v(y)) \\ &= v(y), \end{aligned}$$

d. h.: Die Funktion  $G(x, y)$  hat die Eigenschaften einer Greenschen Funktion.

(ii) Fall  $\mathbb{R}^2$ : Auf dem Kreis  $\Omega := \{x \in \mathbb{R}^2 : |x| < R\}$  ist durch

$$\begin{aligned} G(x, y) &= -\frac{1}{2\pi} \left\{ \log(|x - y|) + \log\left(\frac{R}{|x|}\right) - \log\left(\left|\frac{R^2}{|x|^2}x - y\right|\right) \right\}, \quad x \neq 0, \\ G(x, y) &= -\frac{1}{2\pi} \left\{ \log(|y|) - \log(R) \right\}, \quad x = 0, \end{aligned}$$

eine Greensche Funktion gegeben.

Die Existenz Greenscher Funktionen lässt sich für sehr allgemeine Gebiete  $\Omega$  nachweisen, auch für die anderen RWA. Das Konzept der „klassischen“ Lösung ist in vielen Anwendungsfällen zu restriktiv, z. B. wenn die rechte Seite  $f$  nicht regulär genug ist, um eine  $C^2$ -Lösung zuzulassen. Die Greensche Funktion selbst ist ein Extremfall in dieser Hinsicht. Als nächstes werden wir eine Abschwächung dieser Anforderungen kennenlernen, welche mehr Flexibilität bietet und für die man vergleichsweise leicht die Existenz von Lösungen garantieren kann.

ii) *Funktionalanalytische Methode:*

Wir haben bereits in der Einleitung diskutiert, dass eine enge Beziehung zwischen der Poisson-Gleichung und der Minimierung des zugehörigen Energiefunktional bestehen. Dies kann man zum Nachweis der Existenz von Lösungen ausnutzen. Wir betrachten das Funktional

$$E(v) := \frac{1}{2} \int_{\Omega} |\nabla v|^2 dx - \int_{\Omega} f v dx$$

auf dem Vektorraum  $\tilde{V}_0$  der „zulässigen“ Funktionen:

$$\tilde{V}_0 := \{v : \bar{\Omega} \rightarrow \mathbb{R} : v \in C^1(\Omega) \cap C(\bar{\Omega}), v|_{\partial\Omega} = 0, |\nabla v| \in L^2(\Omega)\}.$$

Dieser Raum wird mit der natürlichen sog. „Energie-Norm“

$$\|v\|_E := \|\nabla v\|_{\Omega}, \quad v \in \tilde{V}_0,$$

versehen. Dass dies wirklich eine Norm ist, folgt aus den entsprechenden Eigenschaften der  $L^2$ -Norm  $\|\cdot\| = \|\cdot\|_{\Omega}$ . Aus der Poincaréschen Ungleichung

$$\|v\|_{\Omega} \leq d_{\Omega} \|\nabla v\|_{\Omega}, \quad v \in \tilde{V}_0,$$

mit  $d_{\Omega} := \text{diam}(\Omega)$  folgt weiter, dass diese Norm stärker ist als die  $L^2$ -Norm. In kompakter Schreibweise ist  $E(v) = \frac{1}{2} \|\nabla v\|^2 - (f, v)$ . Wir verwenden jetzt eine Argumentation aus der Variationsrechnung, die dort als die „direkte Methode“ bekannt ist.

i) Wir zeigen zunächst, dass  $E(\cdot)$  nach unten beschränkt ist. Für  $v \in \tilde{V}_0$  folgt mit Hilfe der Hölderschen und der Poincaréschen Ungleichung

$$E(v) \geq \frac{1}{2} \|\nabla v\|^2 - \|f\| \|v\| \geq \frac{1}{2} \|\nabla v\|^2 - d_{\Omega} \|f\| \|\nabla v\|.$$

Anwendung der Ungleichung  $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$  liefert weiter

$$d_{\Omega} \|f\| \|\nabla v\| \leq \frac{1}{2} \|\nabla v\|^2 + \frac{1}{2} d_{\Omega}^2 \|f\|^2,$$

und folglich

$$E(v) \geq -\frac{1}{2} d_{\Omega}^2 \|f\|^2 > -\infty, \quad v \in \tilde{V}_0.$$

ii) Sei nun  $(u_k)_{k \in \mathbb{N}} \subset \tilde{V}_0$  eine „Minimalfolge“ des Funktionalen  $E(\cdot)$ , d. h.:

$$E(u_k) \rightarrow \inf_{v \in \tilde{V}} E(v) =: d > -\infty.$$

Wir wollen zeigen, dass  $(u_k)_{k \in \mathbb{N}}$  eine Cauchy-Folge bzgl. der Energie-Norm ist. Wichtiges Hilfsmittel dazu ist die sog. „Parallelogrammidentität“

$$\|v - w\|_E^2 + \|v + w\|_E^2 = 2\|v\|_E^2 + 2\|w\|_E^2,$$

die man durch direktes Nachrechnen verifiziert. Für beliebige Indizes  $n, m \in \mathbb{N}$  gilt folglich

$$\begin{aligned}\|u_n - u_m\|_E^2 &= 2\|u_n\|_E^2 + 2\|u_m\|_E^2 - 4\|\frac{1}{2}(u_n + u_m)\|_E^2 \\ &= 4E(u_n) + 4(f, u_n) + 4E(u_m) + 4(f, u_m) - 8E(\frac{1}{2}(u_n + u_m)) \\ &\quad - 8(f, \frac{1}{2}(u_n + u_m)) \\ &= 4E(u_n) + 4E(u_m) - 8E(\frac{1}{2}(u_n + u_m)).\end{aligned}$$

Wegen

$$\lim_{n,m \rightarrow \infty} \{E(u_n) + E(u_m)\} = 2d, \quad E(\frac{1}{2}(u_n + u_m)) \geq d,$$

folgt damit

$$\limsup_{n,m \rightarrow \infty} \|u_n - u_m\|_E^2 \leq 0,$$

d. h.:  $(u_n)_{n \in \mathbb{N}}$  ist wie behauptet eine Cauchy-Folge.

Die Cauchy-Folge  $(u_n)_{n \in \mathbb{N}}$  besitzt i. Allg. keinen Limes im normierten (unvollständigen) Raum  $\tilde{V}_0$ . Durch Vervollständigung von  $\tilde{V}_0$  erhält man den sog. „Sobolew<sup>11</sup>-Raum“  $V_0 = H_0^1(\Omega)$ . Die Elemente von  $H_0^1(\Omega)$  sind zunächst als Äquivalenzklassen von Cauchy-Folgen (analog wie bei der Konstruktion der reellen Zahlen aus den rationalen) definiert; sie lassen sich aber wieder als Funktionen interpretieren. Sie sind  $L^2$ -Funktionen, deren erste Ableitungen (im Distributionssinne) wieder in  $L^2$  liegen und die in einem abgeschwächten Sinn auf  $\partial\Omega$  verschwinden (siehe die angegebene Literatur zur Theorie partieller Differentialgleichungen). Wir werden unten in Abschnitt 1.3 die Eigenschaften dieser „Sobolew-Räume“ genauer diskutieren. Der Limes  $u \in H_0^1(\Omega)$  der Folge  $(u_n)_{n \in \mathbb{N}}$  wird als die „schwache“ oder auch „variationelle“ Lösung der 1. RWA des Laplace-Operators bezeichnet. Als Minimalpunkt des Funktionalen  $E(\cdot)$  genügt sie, wie wir schon früher gesehen haben, notwendig der Beziehung

$$(\nabla u, \nabla \varphi) = (f, \varphi) \quad \forall \varphi \in H_0^1(\Omega). \quad (1.2.11)$$

Umgekehrt gilt für eine Funktion  $u \in H_0^1(\Omega)$ , welcher dieser Variationsgleichung genügt, mit jeder anderen Funktion  $v \in H_0^1(\Omega)$

$$\begin{aligned}E(v) - E(u) &= \frac{1}{2}\|\nabla v\|^2 - (f, v) - \frac{1}{2}\|\nabla u\|^2 + (f, u) \\ &= \frac{1}{2}\|\nabla v\|^2 - (\nabla u, \nabla v) - \frac{1}{2}\|\nabla u\|^2 + (\nabla u, \nabla u) \\ &= \frac{1}{2}\|\nabla v\|^2 - (\nabla u, \nabla v) + \frac{1}{2}\|\nabla u\|^2 = \frac{1}{2}\|\nabla(v - u)\|^2 \geq 0.\end{aligned}$$

---

<sup>11</sup>Sergei Lvovich Sobolev (1908–1989): Russischer Mathematiker; wirkte zunächst in Leningrad (St. Petersburg) und dann am berühmten Steklov-Institut für Mathematik der Akademie der Wissenschaften in Moskau; fundamentale Beiträge zur Theorie der partiellen Differentialgleichungen, Konzept der verallgemeinerten (distributionellen) Lösung, Sobolew-Räume; beschäftigte sich auch mit numerischen Methoden, numerische Quadratur.

Folglich ist  $u$  automatisch auch Minimum des Energiefunktionalen und somit schwache Lösung.

Wenn die schwache Lösung  $u$  regulärer ist, etwa sogar die Regularität einer klassischen Lösung besitzt, so kann partiell integriert werden, und wir finden

$$(-\Delta u, \varphi) + (\partial_n u, \varphi)_{\partial\Omega} = (f, \varphi) \quad \forall \varphi \in H_0^1(\Omega)$$

bzw. wegen der Randbedingung  $\varphi|_{\partial\Omega} = 0$

$$(-\Delta u - f, \varphi) = 0 \quad \forall \varphi \in H_0^1(\Omega).$$

Hieraus folgt mit den üblichen Argumenten, dass  $-\Delta u = f$ , d. h.:  $u$  ist sogar klassische Lösung der RWA. Umgekehrt erfüllt natürlich jede klassische Lösung  $u \in C^2(\bar{\Omega}) \cap C(\bar{\Omega})$ ,  $|\nabla u| \in L^2(\Omega)$ , die Variationsgleichung

$$(\nabla u, \nabla \varphi) - (f, \varphi) = (-\Delta u - f, \varphi) + (\partial_n u, \varphi)_{\partial\Omega} = 0 \quad \forall \varphi \in H_0^1(\Omega).$$

Damit ist der „schwache“ Lösungsbegriff verträglich mit dem ursprünglichen „klassischen“. Der Nachweis höherer Regularität der schwachen Lösung  $u \in H_0^1(\Omega)$  ist allerdings schwierig und kann im Rahmen dieser Vorlesung nur diskutiert werden (siehe wieder die empfohlene Literatur).

Wir wollen noch kurz diskutieren, wie das obige Argument verwendet werden kann, um die Existenz von schwachen Lösungen der 1. RWA auch im Fall inhomogener Randdaten  $u|_{\partial\Omega} = g$  zu sichern. Dazu nehmen wir an, dass die Randfunktion  $g$  als „Spur“ einer auf ganz  $\Omega$  definierten Funktion  $\hat{g} \in H^1(\Omega)$  gegeben ist;  $g = \hat{g}|_{\partial\Omega}$ . Dann wäre die Funktion  $v := u - \hat{g}$  formale Lösung der RWA

$$-\Delta v = f - \Delta \hat{g} \quad \text{in } \Omega, \quad v|_{\partial\Omega} = 0.$$

Hierfür garantiert nun die variationelle Methode die Existenz einer (eindeutigen) schwachen Lösung  $v \in H_0^1(\Omega)$  mit der Eigenschaft

$$(\nabla v, \nabla \varphi) = (f, \varphi) + (\nabla \hat{g}, \nabla \varphi) \quad \forall \varphi \in H_0^1(\Omega).$$

Die schwache Lösung der ursprünglichen RWA ergibt sich dann als  $u := v + \hat{g}$ .

### 1.2.2 Eindeutigkeit von Lösungen

Die Eindeutigkeitsforderung an Lösungen dieser RWAn ist leicht zu gewährleisten. Wir diskutieren hier wieder nur die 1. RWA. Die entsprechenden Argumente für die 2. und die 3. RWA seien als Übung gestellt.

(i) Besonders einfach ist der Beweis für die schwachen Lösungen. Seien also  $u^{(1)}, u^{(2)} \in H_0^1(\Omega)$  zwei schwache Lösungen der 1. RWA, d. h.:

$$(\nabla u^{(i)}, \nabla \varphi) = (f, \varphi) \quad \forall \varphi \in H_0^1(\Omega).$$

Dann gilt für die Differenz  $w := u^{(1)} - u^{(2)}$

$$(\nabla w, \nabla \varphi) = 0 \quad \forall \varphi \in H_0^1(\Omega).$$

Mit  $\varphi := w$  folgt  $\|\nabla w\| = 0$  und folglich  $w \equiv \text{konst.}$  bzw.  $w \equiv 0$  wegen der Randbedingung.

(ii) Die Eindeutigkeit von klassischen Lösungen folgt unmittelbar aus der gerade bewiesenen Eindeutigkeit von schwachen Lösungen, da jede klassische Lösung ja auch schwache Lösung ist. Eine direktere Argumentation ergibt sich mit Hilfe des „Maximumprinzips“.

**Hilfssatz 1.1 (Maximumprinzip):** *Für den elliptischen Operator*

$$Lu := -\Delta u + au$$

mit  $a \geq 0$  auf einem Gebiet  $\Omega \subset \mathbb{R}^d$  gilt das sog. „Maximumprinzip“, d. h.: Eine Funktion  $u \in C^2(\Omega) \cap C(\bar{\Omega})$  mit der Eigenschaft  $Lu \leq 0$  hat in  $\Omega$  kein positives Maximum. Dies bedeutet, dass entweder  $u \leq 0$  auf ganz  $\Omega$  ist, oder

$$\max_{\bar{\Omega}} u \leq \max_{\partial\Omega} u. \quad (1.2.12)$$

**Beweis:** Wir führen den Beweis nur für den Fall, dass  $a > 0$  auf  $\Omega$ . Der allgemeine Fall  $a \geq 0$  erfordert eine aufwendigere Argumentation (siehe z. B. das Buch von Hellwig [5]). Ferner sei  $d = 2$ . Angenommen, die Funktion  $u$  habe im Fall  $u \not\leq 0$  in einem Punkt  $z \in \Omega$  ein positives Maximum,  $u(z) > 0$ . Dann ist notwendig

$$\nabla u(z) = 0, \quad \partial_x^2 u(z) \leq 0, \quad \partial_y^2 u(z) \leq 0.$$

Damit folgt  $0 \geq Lu(z) = -\Delta u(z) + au(z) \geq au(z)$ , was wegen  $a > 0$  den Widerspruch  $u(z) \leq 0$  erzwingt. Q.E.D.

Das Maximumprinzip für elliptische Operatoren 2. Ordnung ist die natürliche Verallgemeinerung der Tatsache, dass in einer Raumdimension aus  $u''(x) \geq 0$  die Konvexität von  $u$  folgt.

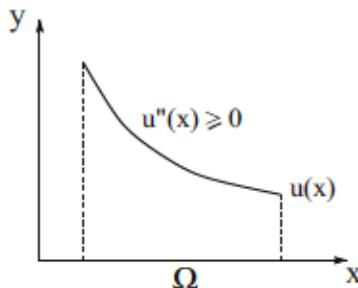


Abbildung 1.3: Maximumprinzip in einer Dimension

Aussagen vom Typ des obigen Maximumprinzips lassen sich für sehr allgemeine (auch nichtlineare) elliptische Operatoren 2. Ordnung herleiten. Wir betonen, dass das Maximumprinzip i. Allg. für elliptische Operatoren höherer Ordnung (z. B. den „biharmonischen Operator“  $\Delta^2 u$ ) und für elliptische Systeme (z. B. die Gleichungen der linearen Elastizitätstheorie) nicht mehr gilt.

Als erste, einfache Anwendung des Maximumprinzips erhalten wir einen alternativen Beweis für die Eindeutigkeit (klassischer) Lösungen der 1. RWA des Laplace-Operators. Sind  $u^{(1)}, u^{(2)}$  zwei Lösungen, so gilt für die Differenz  $w := u^{(1)} - u^{(2)}$  wieder

$$-\Delta w = 0 \quad \text{in } \Omega, \quad w = 0 \quad \text{auf } \partial\Omega.$$

Anwendung des Maximumprinzips auf  $w$  sowie  $-w$  impliziert dann, dass notwendig  $w \leq 0$  sowie  $-w \leq 0$ , d. h.:  $w \equiv 0$ .

### 1.2.3 Stetige Abhängigkeit der Lösungen von den Daten

(i) Die Frage nach der stetigen Abhängigkeit der Lösungen der 1. RWA von den Daten wollen wir wieder sowohl mit Hilfe des klassischen Ansatzes als auch mit der variationellen Methode angehen. Seien zunächst  $u^{(1)}, u^{(2)}$  zwei Lösungen (klassisch oder variationell) der 1. RWA des Laplace-Operators zu unterschiedlichen rechten Seiten  $f^{(1)}, f^{(2)}$ . Für die Differenz  $w = u^{(1)} - u^{(2)}$  folgt dann

$$\|\nabla w\|^2 = (f^{(1)} - f^{(2)}, w) \leq \|f^{(1)} - f^{(2)}\|_{\Omega} \|w\|.$$

Unter Ausnutzung der Poincaréschen Ungleichung folgt daraus

$$\|\nabla w\| \leq d_{\Omega} \|f^{(1)} - f^{(2)}\|,$$

d. h. die Stetigkeit der Lösung (in der Energie-Norm) gegenüber Störungen der rechten Seite.

(ii) Als nächstes betrachten wir Störungen der Randdaten. Dazu verwenden wir wieder das Maximumprinzip. Seien dazu  $u^{(1)}, u^{(2)}$  zwei Lösungen zu den Randdaten  $g^{(1)}, g^{(2)}$ . Für die Differenz  $w := u^{(1)} - u^{(2)}$  gilt dann

$$-\Delta w = 0 \quad \text{in } \Omega, \quad w = g := g^{(1)} - g^{(2)} \quad \text{auf } \partial\Omega.$$

Mit dem Maximumprinzip erschließen wir hieraus, dass  $w \equiv 0$  (was natürlich i. Allg. nicht eintritt) oder

$$\max_{\Omega} w \leq \max_{\partial\Omega} g, \quad \max_{\Omega} -w \leq \max_{\partial\Omega} -g.$$

Dies impliziert  $\max_{\Omega} |w| \leq \max_{\partial\Omega} |g|$ .

Schließlich ergibt sich mit Hilfe des Maximumprinzips noch, dass eine Lösung der 1. RWA des Laplace-Operators zu nichtnegativer rechter Seite und ebensolchen Randdaten,

$$-\Delta u \geq 0 \quad \text{in } \Omega, \quad u \geq 0 \quad \text{auf } \partial\Omega,$$

notwendig überall nicht-negativ ist:  $u \geq 0$ . Dies gilt dann z. B. auch für die zugehörige Greensche Funktion:  $g(\cdot, \cdot) \geq 0$ . Durch schärfere Argumente kann man darüber hinaus zeigen, dass die Greensche Funktion im Innern des Definitionsbereichs  $\Omega$  *positiv* ist. Dies bedeutet u. a., dass bei einem elliptischen Problem lokale Störungen in den Daten die Lösung im gesamten Lösungsgebiet verändern. Es liegt also gewissermaßen eine „unendliche Ausbreitungsgeschwindigkeit“ von Information vor. Dies ist charakteristisch für elliptische Randwertaufgaben.

### 1.2.4 Regularität von Lösungen

Auf glatt berandeten Gebieten  $\Omega$  besteht, ähnlich wie bei gewöhnlichen Differentialgleichungen, für (lineare) partielle Differentialgleichungen vom elliptischen Typ die Regel, dass sich die Regularität der Daten (rechte Seite und Randwerte) auf natürliche Weise auf die Lösung überträgt. Der Rand  $\partial\Omega$  sei aus der Klasse  $C^2$  (2-mal stetig differenzierbar parametrisierbar). Dann besitzt die schwache Lösung  $u \in H_0^1(\Omega)$  im Falle  $f \in L^2(\Omega)$  zweite Ableitungen mit der Regularität  $|\nabla^2 u| \in L^2(\Omega)$ , und es gilt die *a priori* Abschätzung

$$\left( \sum_{k=0}^2 \|\nabla^k u\|^2 \right)^{1/2} \leq c \|f\|. \quad (1.2.13)$$

Diese Aussage bleibt gültig, wenn  $\Omega$  ein konvexes Polygonalgebiet im  $\mathbb{R}^2$  oder ein konvexer Polyeder im  $\mathbb{R}^3$  ist. Ist darüber hinaus  $f$  Hölder-stetig, so ist die schwache Lösung  $u \in H_0^1(\Omega)$  sogar klassische Lösung. Höhere Regularitätseigenschaften von  $\partial\Omega$  und  $f$  übertragen sich entsprechend auf  $u$ .

Im Fall von Gebieten mit Ecken, insbesondere „einspringenden“ Ecken (Innenwinkel  $\omega > \pi$ ) treten allerdings dort Irregularitäten in der Lösung auf; z. B. ist die in ebenen Polarkoordinaten  $(r, \theta)$  ausgedrückte Funktion  $u(r, \theta) = r^{\frac{2}{3}} \sin(\frac{2}{3}\theta)$  auf der gelochten Ebene  $\mathbb{R}^2 \setminus \{0\}$  harmonisch, d. h.  $\Delta u \equiv 0$ . Auf dem „Tortenstück“

$$\Omega := \{(x, y) \in \mathbb{R}^2 \mid 0 < r < 1, 0 < \theta < \frac{3}{2}\pi\}$$

mit einer rechtwinkligen einspringenden Ecke ist  $\nabla u \in L^2(\Omega)^2$ , und  $u$  ist daher (klassische) Lösung der Poisson-Gleichung  $\Delta u = 0$  zu den Randbedingungen

$$u(r, \theta) = 0 \quad \text{für } \theta \in \{0, \frac{3}{2}\pi\}, \quad u(r, \theta) = \sin(\frac{2}{3}\pi) \quad \text{für } r = 1.$$

Wir sehen an diesem Beispiel, dass „klassische“ Lösungen elliptischer Gleichungen nicht unbedingt regulär bis zum Rand  $\partial\Omega$  zu sein brauchen. Für allgemeinen Innenwinkel  $\omega \in (0, 2\pi]$  (Der Fall  $\omega = 2\pi$  entspricht einem sog. „Schlitzgebiet“.) erhält man analoge klassische Lösungen in der Form  $u(r, \theta) = r^{\frac{2}{\omega}} \sin(\frac{\pi}{\omega}\theta)$ . Für  $\omega > \pi$ , d. h. für eine „einspringende“ Ecke hat die Lösung bei  $r = 0$  singuläre erste Ableitungen. Diese sog. „Eckensingularitäten“ sind gut analysiert und abschätzbar. Sie haben einen signifikanten, negativen Einfluss auf die Approximationsgüte von Diskretisierungen und erfordern besondere Vorkehrungen.

## 1.3 Hilfsmittel aus der Theorie von Funktionenräumen

In diesem Abschnitt stellen wir einige Aussagen über Räume verallgemeinert differenzierbarer Funktionen, sog. „Sobolew-Räume“, zusammen, soweit sie später bei der Analyse von Diskretisierungsverfahren benötigt werden.

### 1.3.1 Sobolew-Räume

Sei  $\Omega$  ein Gebiet im  $\mathbb{R}^d$  ( $d = 2, 3$ ) mit Rand  $\partial\Omega$ . Der Rand wird als „ausreichend“ glatt angenommen, was von der betrachteten Situation abhängt; diesbezügliche Einschränkungen werden von Fall zu Fall angegeben. Wir nehmen generell an, dass  $\partial\Omega$  eine Lipschitzstetige Parametrisierung besitzt und überall bis auf endlich viele Punkte oder Kanten eine wohl-definierte äußere Normale  $n$  besitzt.

Auf einem solchen Gebiet  $\Omega$  definieren wir zunächst für Funktionen aus  $C(\overline{\Omega})$  (Vektorraum der stetigen Funktionen auf dem Abschluss  $\overline{\Omega}$ ) das sog.  $L^2$ -Skalarprodukt und die zugehörige Norm

$$(u, v)_\Omega := \int_{\Omega} u(x)v(x) dx, \quad \|u\|_{0,\Omega} := \left( \int_{\Omega} |u(x)|^2 dx \right)^{1/2}.$$

Wenn Verwechslungen ausgeschlossen sind, wird auch die kürzere Notation  $(\cdot, \cdot) = (\cdot, \cdot)_\Omega$  sowie  $\|\cdot\| = \|\cdot\|_0 = \|\cdot\|_{0,\Omega}$  verwendet. Die Vervollständigung von  $C(\overline{\Omega})$  bzgl. der Norm  $\|\cdot\|_{0,\Omega}$  liefert den „Lebesgueschen<sup>12</sup> Hilbert-Raum“  $L^2(\Omega)$  der auf  $\Omega$  im Lebesgueschen Sinne messbaren und quadratintegrablen Funktionen. Eine „Funktion“  $v \in L^2(\Omega)$  ist dann dadurch charakterisiert, dass es eine Folge glatter Funktionen  $(v_k)_{k \in \mathbb{N}} \subset C(\overline{\Omega})$  gibt, welche bzgl. der  $L^2$ -Norm Cauchy-Folge ist und „fast überall“ (im Lebesgueschen Sinne) gegen  $v$  konvergiert. Mit einem einfachen Approximationssargument lässt sich zeigen, dass sich  $L^2(\Omega)$  auch als Vervollständigung des Raumes der „Testfunktionen“ (Begriff aus der Distributionen-Theorie)

$$C_0^\infty(\Omega) := \{v \in C^\infty(\Omega) : \text{Trg}(v) := \overline{\{x \in \Omega, v(x) \neq 0\}} \subset \Omega \text{ kompakt}\}$$

gewinnen lässt. Auf analoge Weise gewinnt man für  $1 \leq p < \infty$  die sog. „ $L^p$ -Räume“  $L^p(\Omega)$  als Vervollständigung von  $C(\overline{\Omega})$  bzgl. der Norm

$$\|v\|_{L^p(\Omega)} := \left( \int_{\Omega} |v(x)|^p dx \right)^{1/p}.$$

Der Fall  $p = \infty$  bedarf einer gesonderten Betrachtung. Der Lebesgue-Raum  $L^\infty(\Omega)$  besteht aus allen auf  $\Omega$  definierten, im Lebesgueschen Sinne messbaren und „wesentlich

---

<sup>12</sup>Henri Léon Lebesgue (1875–1941): Französischer Mathematiker, Prof. am Collège de France in Paris, lieferte grundlegende Beiträge zur modernen Integrationstheorie („Lebesgue-Integral“)

beschränkten“ Funktionen; seine Norm ist

$$\|v\|_{L^\infty(\Omega)} := \text{ess sup}_{x \in \Omega} |v(x)|.$$

Man beachte, dass sich  $L^\infty(\Omega)$  nicht als Vervollständigung von  $C(\bar{\Omega})$  bzgl. der Norm  $\|\cdot\|_\infty; \Omega$  gewinnen lässt, denn dies ergibt wieder  $C(\bar{\Omega})$ . Wenn Missverständnisse ausgeschlossen sind, werden die  $L^p$ -Normen auch kurz mit  $\|\cdot\|_p = \|\cdot\|_{L^p} = \|\cdot\|_{L^p(\Omega)}$  bezeichnet.

Die funktionalanalytische Methode zum Nachweis der Existenz von „schwachen“ Lösungen der Poisson-Gleichung

$$-\Delta u = f \quad \text{in } \Omega, \quad u|_{\partial\Omega} = 0, \quad (1.3.14)$$

bedient sich als natürlichen „Lösungsraum“ des Sobolew-Raums  $H_0^1(\Omega)$ . Dieser ist Teilraum des Sobolew-Raums  $H^1(\Omega)$ , welchen man erhält z. B. durch Vervollständigung des Vektorraums  $C^1(\bar{\Omega})$  bzgl. der sog.  $H^1$ -Norm

$$\|v\|_1 := (\|v\|_0^2 + \|\nabla v\|_0^2)^{1/2}.$$

Entsprechend ist  $H_0^1(\Omega)$  definiert als der Abschluss von  $C_0^\infty(\Omega)$  in  $H^1(\Omega)$ .

Die Definition des Sobolew-Raums  $H^1(\Omega)$  als Vervollständigung von  $C^1(\bar{\Omega})$  besagt zunächst, dass seine Elemente als Äquivalenzklassen von Cauchy-Folgen  $(v_k)_{k \in \mathbb{N}} \subset C^1(\bar{\Omega})$  bzgl. der  $H^1$ -Norm definiert sind. Dies ist ein sehr unhandliches Konzept, mit dem man schlecht arbeiten kann. Daher werden diesen Äquivalenzklassen Funktionen auf  $\Omega$  zugeordnet durch folgende Konstruktion:

$$\{(v_k)_{k \in \mathbb{N}}\} \mapsto v \in H^1(\Omega) : \quad v := \lim_{k \rightarrow \infty} v_k, \quad \nabla v := \lim_{k \rightarrow \infty} \nabla v_k,$$

wobei die Konvergenz jeweils im  $L^2$ -Sinne zu verstehen ist. Umgekehrt existiert dann für jede solche Funktion  $v \in H^1(\Omega)$  eine approximierende Folge „glatter“ Funktionen  $(v_k)_{k \in \mathbb{N}}$  mit der Eigenschaft  $\|v - v_k\|_1 \rightarrow 0$  ( $k \rightarrow \infty$ ). Auf diesem Wege, d. h. durch Konstruktion einer solchen approximierenden Folge wird auch für eine gegebene Funktion  $v \in L^2(\Omega)$  gegebenenfalls  $v \in H^1(\Omega)$  gezeigt. Die Limiten  $\partial_i v := \lim_{k \rightarrow \infty} \partial_i v_k$  werden „verallgemeinerte“ (oder auch „schwache“) Ableitungen von  $v$  genannt. Sie sind i. Allg. nicht stetig oder beschränkt und existieren nur im  $L^2$ -Sinne, d. h. im Lebesgueschen Sinn „fast überall“. Zum Nachweis, dass eine in fast allen Punkten  $x \in \Omega$  definierte Funktion  $v$  in  $H^1(\Omega)$  liegt, geht man üblicherweise wie folgt vor: Zunächst wird aus Kenntnis der Struktur von  $v$  eine Folge approximierender „glatter“ Funktionen  $(v_k)_{k \in \mathbb{N}} \subset C^1(\bar{\Omega})$  konstruiert, welche für  $k \rightarrow \infty$  samt ihrer ersten Ableitungen (fast überall) punktweise gegen  $v$  konvergieren. Kann dann noch gezeigt werden, dass

$$\overline{\lim} \|v_k\|_{1,\Omega} < \infty,$$

so folgt (nach dem Satz von der „dominierten Konvergenz“ der Maß-Theorie), dass  $(v_k)_{k \in \mathbb{N}}$  eine Cauchy-Folge bzgl. der  $H^1$ -Norm ist und den Limes  $v$  besitzt; d. h.  $v \in H^1(\Omega)$ .

**Beispiel 1.2:** i)  $L^2$ -Funktionen: Sei  $\Omega$  zunächst die gepunktete Kreisscheibe

$$\Omega_0 = \{x \in \mathbb{R}^2 : 0 < |x| < 1\}.$$

Auf  $\Omega_0$  ist die Funktion  $u(x) = \ln(|x|)$  stetig, aber unbeschränkt. Auf der vollen Kreisscheibe  $\Omega = \{x \in \mathbb{R}^2 : |x| < 1\}$  ist  $u(x) = \ln(|x|)$  aber als Funktion in  $L^2(\Omega)$  erklärt. Dies wird klar bei Betrachtung der approximierenden, stetigen Funktionen

$$u_k(x) := \begin{cases} \ln(|x|) & \text{für } k^{-1} < |x| < 1, \\ \ln(k^{-1}) & \text{für } 0 \leq |x| \leq k^{-1}. \end{cases}$$

Man rechnet leicht nach, dass  $(u_k)_{k \in \mathbb{N}}$  eine Cauchy-Folge bzgl. der  $L^2$ -Norm auf  $\Omega$  ist und in allen Punkten  $x \in \Omega$  (bis auf  $x = 0$ )  $u_k(x) \rightarrow u(x)$  ( $k \rightarrow \infty$ ) konvergiert.

Wir betrachten nun die Funktion  $u(x) := |x|^{-1}$ , welche ebenfalls auf  $\Omega_0$  stetig und unbeschränkt ist. In diesem Fall bilden die approximierenden Funktionen

$$u_k(x) := \begin{cases} |x|^{-1} & \text{für } k^{-1} < |x| < 1, \\ k & \text{für } 0 \leq |x| \leq k^{-1}, \end{cases}$$

wegen

$$\|u_k\|_\Omega^2 = 2\pi \int_0^{k^{-1}} k^2 r dr + 2\pi \int_{k^{-1}}^1 r^{-1} dr = 2\pi \left( \frac{1}{2} + \ln(k) \right) \rightarrow \infty \quad (k \rightarrow \infty)$$

keine Cauchy-Folge bzgl. der  $L^2$ -Norm, d. h.: Dieses  $u$  ist zu singulär, um als  $L^2$ -Funktion auf  $\Omega$  erklärt zu sein. Bei der analogen Betrachtung auf der Einheitskugel im  $\mathbb{R}^3$  ergibt sich dagegen, dass  $u(x) = |x|^{-1}$  in diesem Fall sehr wohl in  $L^2(\Omega)$  liegt. Die Zugehörigkeit von Funktionen mit lokalen Singularitäten zum Lebesgue-Raum  $L^2(\Omega)$  hängt also von der jeweiligen Raumdimension ab. Wir werden denselben Effekt auch beim Sobolew-Raum  $H^1(\Omega)$  finden.

ii)  $H^1$ -Funktionen: Wir betrachten wieder die Funktion  $u(x) = \ln(|x|)$  auf der punktierten Kreisscheibe  $\Omega_0 \subset \mathbb{R}^2$ . Ihr Gradient  $\nabla u(x) = |x|^{-2}x$  verhält sich bei Annäherung an  $x = 0$  wie  $|\nabla u(x)| \approx |x|^{-1}$ . Im Hinblick auf das eben diskutierte Beispiel ist  $\nabla u$  also nicht zu einer  $L^2$ -Funktion auf die volle Kreisscheibe  $\Omega$  fortsetzbar. Folglich ist  $u$  auch nicht in  $H^1(\Omega)$ . Wir sehen, dass insbesondere die Greensche Funktion zum Laplace-Operator in zwei Raumdimensionen *nicht* im „Energie-Raum“  $H^1(\Omega)$  liegt. Man beachte, dass  $v(x) = \ln(x)$  in *drei* Raumdimensionen aber sehr wohl in  $H^1(\Omega)$  liegt; der kritische Grenzfall ist hier die stärker singuläre Funktion  $u(x) = |x|^{-1}$ .

Als zweites Beispiel zeigen wir, dass  $H^1$ -Funktionen in mehr als einer Dimension nicht beschränkt sein müssen. Auf  $\Omega_0$  sei die Funktion

$$u(x) = \ln(\ln(|x|^{-1}) + 1)$$

betrachtet. Da  $|\ln \ln(r^{-1})|$  für  $r \rightarrow 0$  langsamer wächst als  $|\ln(r)|$ , ist  $u$  sicherlich zu

einer  $L^2$ -Funktion auf  $\Omega$  fortsetzbar. Wir berechnen nun den Gradienten

$$\nabla u(x) = -\frac{x}{|x|^2(\ln(|x|^{-1}) + 1)}.$$

Die zugehörigen „abgeschnittenen“ Funktionen

$$u_k(x) := \begin{cases} \ln(\ln(|x|^{-1}) + 1) & \text{für } k^{-1} < |x| < 1, \\ \ln(\ln(k) + 1) & \text{für } 0 \leq |x| \leq k^{-1}, \end{cases}$$

haben die „stückweise“ definierten Gradienten

$$\nabla u_k(x) := \begin{cases} \frac{x}{|x|^2(\ln(|x|^{-1}) + 1)} & \text{für } k^{-1} < |x| < 1, \\ 0 & \text{für } 0 \leq |x| \leq k^{-1}, \end{cases}$$

welche überall (bis auf  $x = 0$ ) gegen  $\nabla u$  konvergieren. Aus der Abschätzung

$$\|\nabla u_k\|^2 = 2\pi \int_{k^{-1}}^1 \left| \frac{r}{r^2(\ln(r^{-1}) + 1)} \right|^2 r dr = 2\pi \int_{k^{-1}}^1 \frac{1}{r(\ln(r^{-1}) + 1)^2} dr \quad (1.3.15)$$

$$= \frac{2\pi}{\ln(r^{-1}) + 1} \Big|_{k^{-1}}^1 = 2\pi - \frac{2\pi}{\ln(k) + 1} \leq 2\pi \quad (1.3.16)$$

ersehen wir ferner, dass  $(\nabla u_k)_{k \in \mathbb{N}}$  bzgl. der  $L^2$ -Norm eine Cauchy-Folge ist. Folglich ist  $u$  zu einer Funktion im Sobolew-Raum  $H^1(\Omega)$  fortsetzbar. Mit der eben verwendeten „Abschneidetechnik“ erhalten wir approximierende Funktionen  $u_k$ , welche i. Allg. nur stückweise stetig differenzierbar sind, d. h. nicht im strengen Sinne in  $C^1(\bar{\Omega})$  liegen und somit nicht direkt in das oben formulierte Approximationskonzept für  $H^1$ -Funktionen passen. Dieser Mangel kann behoben werden, in dem man statt abzuschneiden regularisiert, z. B. gemäss

$$u_k(x) = \ln(\ln((|x| + k)^{-1}) + 1).$$

Diese  $u_k \in C^1(\bar{\Omega})$  bilden dann ebenfalls eine approximierende Folge von  $u$  bzgl. der  $H^1$ -Norm. Eine ähnliche Modifikation (schon bei der Definition der Sobolew-Räume) muss auch vorgenommen werden, um spezielle Gebiete  $\Omega$  mit „schlitz-artigen“ Rand-einsprüngen einbeziehen zu können. Solche „Schlitz-Gebiete“ spielen eine wichtige Rolle z. B. in der Baumechanik, wenn die Ausbreitung von Rissen in Bauteilen beschrieben werden soll.

Analog zu dem Sobolew-Raum  $H^1(\Omega)$  „erster Ordnung“ quadrat-integrablen Funktionen kann man auch Sobolew-Räume  $H^{m,p}(\Omega)$  höherer Ordnung  $m \in \mathbb{N}$ , bestehend aus  $p$ -integrablen Funktionen ( $1 \leq p < \infty$ ), definieren. Diese erhält man durch Vervollständigung des Raumes  $C^m(\bar{\Omega})$  bzgl. der Norm

$$\|v\|_{H^{m,p}(\Omega)} := \left( \sum_{k=0}^m \|\nabla^k v\|_{L^p(\Omega)}^p \right)^{1/p}.$$

Der Fall  $p = \infty$  bedarf wieder einer gesonderten Betrachtung. Die Räume  $H^{m,\infty}(\Omega)$

werden über die Gleichsetzung  $H^{m,\infty}(\Omega) := W^{m,\infty}(\Omega)$  als Räume sog. „verallgemeinert differenzierbarer“ Funktionen mit distributionellen Ableitungen in  $L^\infty(\Omega)$  definiert. Wir wollen auf diese Begriffsbildungen nicht weiter eingehen und verweisen statt dessen auf die einschlägige Literatur über Sobolew-Räume (z. B.: Wloka [10]).

### 1.3.2 Eigenschaften von Lebesgue- und Sobolew-Räumen

Wir wollen im Folgenden einige wichtige Eigenschaften von Lebesgue- und Sobolew-Räumen zusammenstellen, welche später bei der Analyse numerischer Verfahren benötigt werden.

Sei  $1 < p < \infty$  und  $q := p/(p - 1)$ . Dann gilt für Funktionen  $u \in L^p(\Omega)$  und  $v \in L^q(\Omega)$  die allgemeine „Höldersche Ungleichung“

$$\left| \int_{\Omega} u(x)v(x) dx \right| \leq \left( \int_{\Omega} |u(x)|^p dx \right)^{1/p} \left( \int_{\Omega} |v(x)|^q dx \right)^{1/q}, \quad (1.3.17)$$

bzw. in Kurzform  $|(u, v)_\Omega| \leq \|u\|_{L^p(\Omega)} \|v\|_{L^q(\Omega)}$ . Für die Grenzfälle  $p = \infty$  bzw.  $q = 1$  gilt

$$|(u, v)_\Omega| \leq \|u\|_{L^\infty(\Omega)} \|v\|_{L^1(\Omega)}. \quad (1.3.18)$$

Die variationelle Methode zum Nachweis von „schwachen“ Lösungen der Poisson-Gleichung basierte auf der Eigenschaft der bilinearen Form  $(\nabla u, \nabla v)_\Omega$ , auf dem Teilraum  $H_0^1(\Omega) \subset H^1(\Omega)$  ein Skalarprodukt zu sein.

**Hilfssatz 1.2 (Poincarésche Ungleichung):** *Für Funktionen  $v \in H_0^1(\Omega)$  gilt die sog. „Poincarésche Ungleichung“*

$$\|v\|_\Omega \leq d_\Omega \|\nabla v\|_\Omega, \quad (1.3.19)$$

mit dem Durchmesser  $d_\Omega := \text{diam}(\Omega)$  des Gebiets  $\Omega$ .

**Beweis:** Wir geben den Beweis nur in zwei Raumdimensionen. In höheren Dimensionen verläuft die Argumentation ganz analog. Sei  $Q$  eine Quadrat der Kantenlänge  $L = d_\Omega$ , in welchem das Gebiet  $\Omega$  enthalten ist. O.B.d.A. sei das Koordinatensystem so verschoben und gedreht, dass  $Q = (0, L) \times (0, L)$ . Für irgendein  $v \in H_0^1(\Omega)$  sei  $(v_k)_{k \in \mathbb{N}} \subset C_0^\infty(\Omega)$  eine approximierende Folge. Mit  $\hat{v}_k$  bezeichnen wir die trivialen Fortsetzungen der  $v_k$  auf  $Q$ :

$$\hat{v}_k(x) := \begin{cases} v_k(x) & \text{für } x \in \Omega, \\ 0 & \text{für } x \in Q \setminus \Omega. \end{cases}$$

Diese sind dann ebenfalls in  $C_0^\infty(Q)$ . Wir setzen nun  $w := \hat{v}_k$ . Zunächst gilt in Punkten  $(x, y) \in Q$ :

$$w(x, y) = w(0, y) + \int_0^x \partial_\xi w(\xi, y) d\xi,$$

und folglich, bei Beachtung von  $w(0, y) = 0$ ,

$$|w(x, y)|^2 \leq L \int_0^L |\partial_\xi w(\xi, y)|^2 d\xi$$

Integration zunächst über  $y \in (0, L)$  und danach über  $x \in (0, L)$  ergibt

$$\int_0^L \int_0^L |w(x, y)|^2 dy dx \leq L^2 \int_0^L \int_0^L |\partial_\xi w(\xi, y)|^2 d\xi dy.$$

Dies bedeutet

$$\|w\|_Q \leq L \|\nabla w\|_Q.$$

und wegen  $w = \hat{v}_k \equiv 0$  auf  $Q \setminus \Omega$ :

$$\|v_k\|_\Omega \leq L \|\nabla v_k\|_\Omega.$$

Für  $k \rightarrow \infty$  überträgt sich diese Beziehung durch Stetigkeit auf  $v \in H_0^1(\Omega)$ . Q.E.D.

Die variationelle Methode liefert auch die Existenz von Lösungen der 1. RWA des Laplace-Operators, wenn die Randwertvorgaben inhomogen sind. Im allgemeinen Fall findet man eine schwache Lösung  $v \in H^1(\Omega)$ . Wir haben gesehen, dass  $H^1$ -Funktionen Singularitäten haben können. Es stellt sich also die Frage, in welchem Sinne die Randwerte von der „schwachen“ Lösung überhaupt angenommen werden.

**Hilfssatz 1.3 (Spur-Lemma):** Für Funktionen  $v \in C^1(\bar{\Omega})$  gilt die sog. „Spur-Abschätzung“

$$\|v\|_{L^2(\partial\Omega)} \leq c(\Omega) \|v\|_{H^1(\Omega)} \quad (1.3.20)$$

mit einer  $\Omega$ -abhängigen Konstante  $c(\Omega)$ .

**Beweis:** Zuerst betrachten wir den Spezialfall des Einheitsquadrats  $\Omega = (0, 1) \times (0, 1)$ . Sei  $v \in C^1(\bar{\Omega})$ . Für Punkte  $(0, y) \in \partial\Omega$  gilt

$$v(0, y) = - \int_0^x \partial_\xi v(\xi, y) d\xi + v(x, y), \quad x \in [0, 1].$$

und folglich

$$|v(0, y)|^2 \leq \left( \int_0^1 |\partial_\xi v(\xi, y)| d\xi + |v(x, y)| \right)^2 \leq 2 \int_0^1 |\partial_\xi v(\xi, y)|^2 d\xi + 2|v(x, y)|^2.$$

Integration zunächst über  $y \in (0, 1)$  und dann über  $x \in (0, 1)$  liefert

$$\int_0^1 |v(0, y)|^2 dy \leq 2 \int_0^1 \int_0^1 |\partial_\xi v(\xi, y)|^2 d\xi dy + 2 \int_0^1 \int_0^1 |v(x, y)|^2 dy dx.$$

Dieselbe Argumentation kann auch für die drei anderen Randkomponenten von  $\Omega$  ange-

wendet werden. Zusammenfassung der sich ergebenden Abschätzungen ergibt dann

$$\|v\|_{L^2(\partial\Omega)}^2 \leq 8\|u\|_{H^1(\Omega)}^2.$$

Die gezeigte Argumentation für das Einheitsquadrat lässt sich ohne Probleme für allgemeine Polygone gebiete modifizieren. Im allgemeineren Fall eines krumm berandeten Gebiets  $\Omega$  erhält man dasselbe Resultat mit Hilfe lokaler Transformationen, welche krumme Randstücke lokal gerade transformieren, so dass wieder das obige Argument angewendet werden kann. Q.E.D.

Mit Hilfe des Spurlemmas können wir Funktionen  $v \in H^1(\Omega)$  eine „Spur“  $v|_{\partial\Omega} \in L^2(\partial\Omega)$  zuordnen, was die Frage nach der Annahme von Randwerten durch die schwache Lösung der 1. RWA beantwortet. Sei  $v \in H^1(\Omega)$  und  $(v_k)_{k \in \mathbb{N}} \subset C^1(\bar{\Omega})$  eine approximierende Folge. Aufgrund der Spurabschätzung gilt dann

$$\|v_k - v_l\|_{L^2(\partial\Omega)} \leq c(\Omega) \|v_k - v_l\|_{H^1(\Omega)}, \quad k, l \in \mathbb{N}.$$

Da die Norm auf der rechten Seite für  $k, l \rightarrow \infty$  gegen Null konvergiert, folgt, dass die Spuren  $v_k|_{\partial\Omega}$  ( $k \in \mathbb{N}$ ) auf  $\partial\Omega$  eine Cauchy-Folge im Lebesgue-Raum  $L^2(\partial\Omega)$  bilden. Deren Limes  $v|_{\partial\Omega} \in L^2(\partial\Omega)$  wird dann als die „Spur“ der Funktion  $v \in H^1(\Omega)$  bezeichnet. In diesem Sinne nehmen schwache  $H^1$ -Lösungen vorgegebene Randwerte an.

Das Variationsargument liefert zunächst nur die Existenz einer „schwachen“ Lösung  $u \in H^1(\Omega)$  der Poisson-Gleichung. Um zu sehen, dass diese im Fall „glatter“ Daten auch „klassische“ Lösung ist, zeigt man (mit einem Aufwand)  $u \in H^m(\Omega)$  für sukzessive ansteigendes  $m \geq 2$ . Hieraus kann dann geschlossen werden, dass  $u$  auch bis zur gewünschten Stufe klassisch differenzierbar ist. Dazu bedient man sich einer sog. „Sobolewschen Ungleichung“. Zur Motivation sei zunächst der eindimensionale Fall betrachtet.

**Beispiel 1.3:** Wir betrachten das Intervall  $\Omega = (0, 1) \subset \mathbb{R}^1$ . Für eine Funktion  $u \in C^1(\bar{\Omega})$  impliziert der Fundamentalsatz der Differential- und Integralrechnung, dass

$$u(x) = u(y) + \int_y^x u'(\xi) d\xi, \quad x, y \in \Omega.$$

Daraus folgt nach Integration über  $y \in [0, 1]$ :

$$\sup_{x \in \Omega} |u(x)| \leq \int_0^1 |u'| d\xi + \int_0^1 |u| d\xi = \|u\|_{H^{1,1}(\Omega)},$$

mit der Norm des Sobolew-Raums  $H^{1,1}(\Omega)$ . Dies bedeutet, dass in einer Raumdimension eine  $H^{1,1}$ -Funktion beschränkt ist. Darüber hinaus ist sie sogar stetig (genauer im  $L^2$ -Sinne äquivalent zu einer stetigen Funktion), was man mit Hilfe des üblichen Approximationsarguments erschließt. Hierfür reicht hier schon die  $L^1$ -Integrabilität der ersten Ableitung aus.

In höheren Dimensionen ist die Situation komplizierter. Wir haben schon am Beispiel der Funktion  $v(x) = \ln(\ln(|x|^{-1}) + 1)$  gesehen, dass Funktionen in  $H^1(\Omega)$  (im  $\mathbb{R}^2$ ) i. Allg.

unbeschränkt sein können und man ihnen folglich auch nicht überall Punktwerte zuordnen kann. Dies ist aber möglich für Funktionen in Sobolew-Räumen höherer Ordnung. Wir präsentieren hier als Beispiel die folgende „Sobolewsche Ungleichung“.

**Hilfssatz 1.4 (Sobolewsche Ungleichung):** Für Funktionen  $v \in C^2(\bar{\Omega})$  gilt in zwei Raumdimensionen die Abschätzung

$$\sup_{x \in \Omega} |v(x)| \leq c(\Omega) \|v\|_{H^2(\Omega)} \quad (1.3.21)$$

mit einer  $\Omega$ -abhängigen Konstante  $c(\Omega)$ .

**Beweis:** Für den nicht trivialen Beweis verweisen wir auf die einschlägige Literatur über Sobolew-Räume (z. B.: Wloka [10]). Q.E.D.

Analog wie schon vorher bei der Spurabschätzung dient die Sobolewsche Ungleichung (1.3.21) zur Definition von Punktwerten von Funktionen in Sobolew-Räumen. Für ein  $v \in H^2(\Omega)$  sei wieder  $(v_k)_{k \in \mathbb{N}} \subset C^2(\bar{\Omega})$  eine approximierende Folge. Mit der Sobolewschen Ungleichung (1.3.21) erschließen wir, dass  $(v_k)_{k \in \mathbb{N}}$  auch Cauchy-Folge bzgl. der Maximumsnorm ist. Folglich können für  $v \in H^2(\Omega)$  Punktwerte definiert werden durch

$$v(x) := \lim_{k \rightarrow \infty} v_k(x), \quad x \in \bar{\Omega}.$$

In diesem Sinne besteht also eine („stetige“) Einbettung  $H^2(\Omega) \hookrightarrow C(\bar{\Omega})$ . Die Abschätzung (1.3.21) lässt sich (im  $\mathbb{R}^2$ ) übertragen auf die Normen der Sobolew-Räume  $H^{2,1}(\Omega)$  und  $H^{1,p}(\Omega)$  für  $p > 2$ . Folglich bestehen die stetigen Einbettungen

$$H^{2,1}(\Omega) \cup H^{1,p}(\Omega) \hookrightarrow C(\bar{\Omega}), \quad p > 2, \quad \text{im } \mathbb{R}^2. \quad (1.3.22)$$

Weitere Sobolewsche Ungleichungen führen auf die stetigen Einbettungen

$$H^1(\Omega) \hookrightarrow L^p(\Omega) \quad (1 \leq p < \infty) \quad \text{im } \mathbb{R}^2, \quad \dots \quad h.1(\Omega) \hookrightarrow L^6(\Omega) \quad \text{im } \mathbb{R}^3.$$

Von fundamentaler Bedeutung ist der sog. „Rellichsche<sup>13</sup> Auswahlsatz“. Wir formulieren hier nur eine einfache Variante, welche weiter unten benötigt wird.

**Hilfssatz 1.5 (Rellichscher Auswahlsatz):** Die natürliche Einbettung des Sobolew-Raumes  $H^1(\Omega)$  in  $L^2(\Omega)$  ist kompakt, d. h.: Aus jeder bzgl. der  $H^1$ -Norm beschränkten Folge  $(v_k)_{k \in \mathbb{N}} \subset H^1(\Omega) \subset L^2(\Omega)$  lässt sich eine Teilfolge auswählen, welche in  $L^2(\Omega)$  gegen einen Limes  $v$  konvergiert. Dieser Limes ist dann auch wieder in  $H^1(\Omega)$ .

---

<sup>13</sup>Franz Rellich (1906–1955): Deutscher Mathematiker; Promotion 1929 in Göttingen bei R. Courant; Prof. in Dresden und ab 1946 Institutsleiter in Göttingen; wichtige Beiträge zur Mathematischen Physik und Theorie partieller Differentialgleichungen.

**Beweis:** Für den Beweis wird auf die Literatur verwiesen; z. B. Wloka [10]. Q.E.D.

Der Rellichsche Auswahlssatz ist das Sobolew-Raum-Analogon des Auswahlssatzes von Arzelà-Ascoli für „gleichgradig stetige“ Folgen stetiger Funktionen (siehe den Band „Numerik 1“ dieser Reihe, Rannacher [2]).

### 1.3.3 Elemente der Spektraltheorie elliptischer Operatoren

Eine der wichtigsten Anwendungen des Rellichschen Auswahlssatzes findet sich in der „Spektral-Theorie“ elliptischer Differentialoperatoren, speziell des Laplace-Operators. Wir wollen deren Elemente hier kurz entwickeln. Dabei haben wir vor allem deren Anwendung in der Lösungstheorie für parabolische Anfangs-Randwert-Aufgaben im Auge. Ferner werden Resultate der Spektraltheorie bei der Analyse der Finite-Elemente-Approximation von Eigenwertaufgaben des Laplace-Operators benötigt.

Wir haben gesehen, dass für jede rechte Seite  $f \in L^2(\Omega)$  eine eindeutige „schwache“ Lösung  $u \in H_0^1(\Omega)$  der 1. RWA des Laplace-Operators existiert:

$$-\Delta u = f \quad \text{in } \Omega, \quad u|_{\partial\Omega} = 0. \quad (1.3.23)$$

Diese ist bestimmt durch die variationelle Beziehung

$$(\nabla u, \nabla \varphi) = (f, \varphi) \quad \forall \varphi \in H_0^1(\Omega). \quad (1.3.24)$$

Die Zuordnung  $f \mapsto Sf := u$  definiert dann einen linearen Operator  $S : L^2(\Omega) \mapsto L^2(\Omega)$ , der in diesem Sinne als die „ $L^2$ -Inverse“ des Laplace-Operators auf  $\Omega$  unter Dirichlet-Randbedingungen bezeichnet werden kann. Der Beziehung

$$\|\nabla u\|^2 = (f, u) \leq \|f\| \|u\| \leq d_\Omega \|f\| \|\nabla u\|$$

entnehmen wir, dass

$$\|Sf\| \leq d_\Omega \|\nabla Sf\| \leq d_\Omega^2 \|f\|,$$

d. h.: Der Lösungsoperator  $S : L^2(\Omega) \mapsto L^2(\Omega)$  ist beschränkt und wegen der kompakten Einbettung  $H^1(\Omega) \hookrightarrow L^2(\Omega)$  sogar kompakt. Wir haben schon gesehen, dass der Laplace-Operator als Operator im Hilbert-Raum  $L^2(\Omega)$  symmetrisch und positiv-definit ist:

$$(-\Delta u, v) = (\nabla u, \nabla v) = (u, -\Delta v), \quad (-\Delta u, u) = \|\nabla u\|^2 \geq d_\Omega^{-2} \|u\|^2.$$

Dies gilt für Funktionen  $u, v \in D(\Delta)$  im Definitionsbereich der „ $L^2$ -Realisierung“ des Laplace-Operators, welcher definiert ist durch

$$D(\Delta) := \{v \in H_0^1(\Omega) : |(\nabla v, \nabla \varphi)| \leq c(v) \|\varphi\|, \varphi \in H_0^1(\Omega)\} \subset L^2(\Omega).$$

Für glatt berandete Gebiete oder konvexe Polygonegebiete  $\Omega$  kann man zeigen, dass  $D(\Delta) = H_0^1(\Omega) \cap H^2(\Omega)$ . Im Fall einspringender Ecken ist dagegen  $D(\Delta) \not\subset H^2(\Omega)$ . Wir haben damit eine Realisierung des Laplace-Operators im Hilbert-Raum  $L^2(\Omega)$  konstruiert, welche auf ihrem Definitionsbereich  $D(\Delta)$  symmetrisch ist, diesen ein-eindeutig

auf ganz  $L^2(\Omega)$  abbildet, also bijektiv ist und deren Inverse  $S = (-\Delta)^{-1}$  kompakt ist. Damit ist der Rahmen für die Anwendung abstrakter Resultate der Funktionalanalysis kompakter Operatoren geschaffen.

Für das Eigenwertproblem des Laplace-Operators

$$-\Delta w = \lambda w \quad \text{in } \Omega, \quad w|_{\partial\Omega} = 0, \quad (1.3.25)$$

mit Eigenfunktion  $w \in H_0^1(\Omega)$  und Eigenwert  $\lambda \in \mathbb{R}$  gelten die folgenden Aussagen:

- Das Spektrum (Menge der „singulären“ Werte) besteht aus reellen, positiven Eigenwerten  $0 < \lambda_1 \leq \dots \leq \lambda_i \leq \dots$ , welche sich im Endlichen nicht häufen können. Die zugehörigen Eigenräume  $E(\lambda_i)$  sind endlich dimensional.
- Es existiert ein vollständiges Orthonormalsystem von Eigenfunktionen  $\{w_i\}_{i \in \mathbb{N}} \subset L^2(\Omega)$ , d. h.: Für jedes  $u \in L^2(\Omega)$  gilt die  $L^2$ -konvergente „Fourier<sup>14</sup>-Entwicklung“

$$u = \sum_{i=1}^{\infty} (u, w_i) w_i. \quad (1.3.26)$$

- Mit Hilfe der Eigenwerte  $\lambda_i$  (ihrer Vielfachheiten entsprechend oft gezählt) und zugehörigen (orthonormierten) Eigenfunktionen  $w_i$  lassen sich allgemeine Funktionen des Laplace-Operators definieren. Sei  $\Phi(z)$  eine meromorphe Funktion, so dass die Eigenwerte  $\lambda_i$  keine Pole sind. Dann wird durch

$$\Phi(-\Delta)u := \sum_{i=1}^{\infty} \Phi(\lambda_i)(u, w_i) w_i \quad (1.3.27)$$

ein linearer Operator in  $L^2(\Omega)$  erklärt. Wenn  $\Phi(z)$  beschränkt ist, wird auch  $\Phi(-\Delta)$  beschränkt und ist auf ganz  $L^2(\Omega)$  erklärt.

Diese Aussagen zeigen die starke Parallelität zwischen kompakten Operatoren bzw. von (dicht definierten) Operatoren mit kompakter Inverser im Hilbert-Raum und durch Matrizen dargestellte linearen Abbildungen im  $\mathbb{R}^n$ .

## 1.4 Parabolische Probleme

Die sog. „eindimensionale Wärmeleitungsgleichung“

$$\partial_t u = \partial_x^2 u, \quad (1.4.28)$$

---

<sup>14</sup>Jean-Baptiste Baron de Fourier (1768–1830): Französischer Mathematiker und Physiker; Mitglied der Pariser Akademie lehrte an der École Polytechnique; begleitete Napoleon auf seinem Feldzug nach Ägypten; zählt zu den bedeutendsten Mathematikern des 19. Jahrhunderts; fand bei seinen Arbeiten zur Theorie der Wärmeleitung die Darstellbarkeit periodischer Funktionen durch trigonometrische Reihen.

oder allgemeiner in höheren Ortsdimensionen

$$\partial_t u - \Delta u = f, \quad (1.4.29)$$

wird üblicherweise auf Zylindern  $Q_T := \Omega \times I$  des Orts/Zeit-Raumes betrachtet. Dabei sind  $\Omega \subset \mathbb{R}^d$  ein Ortsgebiet und  $I := (0, T]$  ein Zeitintervall. Im örtlich eindimensionalen Fall ( $d = 1$ ) ist die natürliche Anfangskurve  $\Gamma = \{(x, t) \in \mathbb{R}^2 : t = 0\}$  gerade Charakteristik, so dass das zugehörige Cauchysche Anfangswertproblem i. Allg. nicht lösbar ist. Entlang  $\Gamma$  dürfen, wie wir noch sehen werden, nur Anfangsbedingungen an  $u$  selbst gestellt werden:  $u|_{t=0} = u^0(x)$ . Entlang eines nicht-charakteristischen „örtlichen“ Randes  $\{(x, t) \in \mathbb{R}^{d+1} : x \in \partial\Omega, t > 0\}$  gilt dagegen dasselbe wie im elliptischen Fall, d. h.: Die zugehörige Anfangswertaufgabe ist lösbar, doch dürfen nur  $u$  oder  $\partial_n u$  vorgeschrieben werden, wenn man stetige Abhängigkeit von den Randdaten gewährleisten will.

Analog zum elliptischen Fall bieten sich drei verschiedene Typen von Randbedingungen entlang des örtlichen Randes  $\partial\Omega \times I$  für die Anfangs-Randwert-Aufgabe (kurz „ARWA“) der Wärmeleitungsgleichung an. Zusätzlich zu der Anfangsbedingung

$$u|_{t=0} = u^0 \quad (1.4.30)$$

wird gefordert:

- a) *Dirichletsche Randbedingungen* („1. ARWA“):  $u = g$  auf  $\partial\Omega \times I$ ;
- b) *Neumannsche Randbedingungen* („2. ARWA“):  $\partial_n u = g$  auf  $\partial\Omega \times I$ ;
- c) *Robinsche Randbedingungen* („3. ARWA“):  $\partial_n u + \alpha u = g$  auf  $\partial\Omega \times I$ .

Die Randfunktionen  $g$  werden i. Allg. als glatt und  $\alpha \geq 0$  angenommen. Alle diese ARWA sind, wie wir zum Teil zeigen werden, unter geeigneten Zusatzbedingungen an die Daten ebenfalls wohl gestellt. Unter einer „klassischen Lösung“ verstehen wir eine Funktion  $u \in C(\bar{Q}_T) \cap C^2(Q_T)$ , welche der Differentialgleichung sowie den Anfangs- und Randbedingungen genügt. Ähnlich wie bei elliptischen Problemen gibt es auch im parabolischen Fall den Begriff der „schwachen Lösung“, den wir hier aber wegen seiner Kompliziertheit nicht definieren wollen.

Zunächst diskutieren wir die Eindeutigkeitsfrage. Seien  $u^{(1)}, u^{(2)}$  wieder zwei klassische (analog zum elliptischen Fall definiert) Lösungen der 1. ARWA des Wärmeleitungoperators, für die  $\|\nabla u^{(i)}(t)\|_\Omega$  existiert und beschränkt ist. Für die Differenz  $w := u^{(1)} - u^{(2)}$  gilt dann

$$\partial_t w - \Delta w = 0 \quad \text{in } \Omega \times I, \quad w|_{t=0} = 0, \quad w|_{\partial\Omega} = 0.$$

Multiplikation mit  $w$ , Integration über  $\Omega$  und anschließende partielle Integration im Ort ergeben analog zum elliptischen Fall

$$0 = (\partial_t w, w) - (\Delta w, w) = \frac{1}{2} d_t \|w\|^2 + \|\nabla w\|^2.$$

Dies impliziert, dass  $\|w(t)\| \leq \|w(0)\| = 0$  für  $t \geq 0$ , und somit die Eindeutigkeit der Lösung und deren stetige (genauer sogar L-stetige) Abhängigkeit von den Anfangsdaten.

Die Existenzfrage lässt sich im Prinzip mit ähnlichen Methoden behandeln wie im elliptischen Fall. Wir wollen das hier aus Zeitgründen nicht weiter verfolgen. Im örtlich eindimensionalen Spezialfall  $\Omega = (-\infty, \infty)$  und  $f \equiv 0$  lässt sich die Lösung der Anfangswertaufgabe explizit angeben:

$$u(x, t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi t}} e^{-(x-s)^2/4t} u^0(s) ds. \quad (1.4.31)$$

Dies wird durch Nachrechnen verifiziert, wobei speziell auf die Existenz der auftretenden Integralterme zu achten ist. Man beachte, dass durch den Ansatz

$$s(x, t) = \frac{1}{\sqrt{4\pi t}} e^{-x^2/4t}$$

eine spezielle Lösung (sog. „Fundamentallösung“) der Wärmeleitungsgleichung gegeben ist.

Im Fall allgemeinerer, *beschränkter* Ortsgebiete  $\Omega \subset \mathbb{R}^d$  gewinnt man eine zu (1.4.31) korrespondierende Lösungsdarstellung mit Hilfe der „Methode der Variablenseparation“. Einsetzen des Lösungsansatzes  $u(x, t) = v(x)\psi(t)$  in die Wärmeleitungsgleichung ergibt

$$\psi'(t)v(x) = \psi(t)\Delta v(x) \quad \Rightarrow \quad \frac{\psi'(t)}{\psi(t)} = \frac{\Delta v(x)}{v(x)} \equiv \text{konst.},$$

für alle Argumente  $(x, t) \in Q_T$ . Die Separationsfaktoren  $v(\cdot) \in C(\bar{\Omega}) \cap C^2(\Omega)$ ,  $v|_{\partial\Omega} = 0$ , und  $\psi(\cdot) \in C(I)$  sind also notwendig Lösungen der Eigenwertprobleme

$$-\Delta v(x) = \lambda v(x), \quad x \in \Omega, \quad -\psi'(t) = \lambda \psi(t), \quad t \geq 0,$$

unter den Nebenbedingungen  $v|_{\partial\Omega} = 0$  bzw.  $\psi(0) = 1$ , mit Parametern  $\lambda \in \mathbb{R}$ . Die Eigenwertaufgabe für  $v(x)$  besitzt, wie schon oben diskutiert, eine abzählbare Folge von Lösungen  $\lambda_j > 0$  und  $v_j$ :

$$-\Delta v_j = \lambda_j v_j \quad (j = 1, 2, 3, \dots).$$

Die Eigenfunktionen  $(v_j)_{j \in \mathbb{N}}$  bilden ein vollständiges Orthonormal-System im Hilbert-Raum  $L^2(\Omega)$  der auf  $\Omega$  Lebesgue-messbaren und quadratintegrablen Funktionen.

Die zugehörigen Lösungen für  $\psi(t)$  sind  $\psi_j(t) = e^{-\lambda_j t}$ . Die Anfangsfunktion besitzt die (verallgemeinerte) Fourier-Entwicklung:

$$u^0(x) = \sum_{j=0}^{\infty} u_j^0 v_j(x), \quad u_j^0 = \int_I u^0(x) v_j(x) dx.$$

Durch Superposition der Einzellösungen für  $j \in \mathbb{N}$ ,

$$u(x, t) := \sum_{j=1}^{\infty} u_j^0 v_j(x) e^{-\lambda_j t}, \quad (1.4.32)$$

erhalten wir folglich eine Lösung der Wärmeleitungsgleichung, welche den Randbedingungen und insbesondere den Anfangsbedingungen genügt. (Zum Nachweis überprüfe man die Konvergenz der Reihen der jeweils nach  $x$  sowie  $t$  abgeleiteten Einzellösungen.) Im eindimensionalen Spezialfall  $\Omega = (0, 1) \subset \mathbb{R}^1$  ist gerade

$$v_j(x) = \alpha_j \sin(j\pi x), \quad \lambda_j = j^2\pi^2, \quad \alpha_j = \left( \int_I \sin^2(j\pi x) dx \right)^{-1/2} \quad (j \in \mathbb{N}),$$

und die Lösungsdarstellung erhält die explizite Form

$$u(x, t) = \sum_{j=1}^{\infty} u_j^0 \alpha_j \sin(j\pi x) e^{-aj^2\pi^2 t} \quad (1.4.33)$$

Anhand dieser Lösungsdarstellungen lassen sich einige wichtige Eigenschaften der ARWA der Wärmeleitungsgleichung ablesen. Wie bei den gewöhnlicher Differentialgleichungen entwickelt sich die Lösung ausgehend vom Anfangswert in der Zeit. Der adäquate numerische Ansatz wird also wieder ein Teilschrittverfahren in der Zeit sein. Im Ort pflanzen sich Störungen wie im elliptischen Fall „unendlich schnell“ fort. Irregularitäten in den Anfangs- oder Randdaten werden sofort ausgeglättet, d. h.: im Innern des Zylindergebiets  $Q_T := \Omega \times (0, T]$  ist die Lösung (im Falle glatter rechter Seite  $f$ ) stets glatt.

Im Folgenden wollen wir einige qualitative Eigenschaften von Lösungen der Wärmeleitungsgleichung diskutieren. Die Wärmeleitungsgleichung wird u. a. verwendet, um (ihrem Namen entsprechend) Wärmeausbreitungs- bzw. allgemein instationäre Diffusionsvorgänge zu beschreiben. Es ist wichtig, garantieren zu können, dass ihre Lösungen bei kompatiblen Daten auch stets positiv sind. Dies wird durch ein (dem elliptischen Fall ähnliches) Maximumprinzip geleistet.

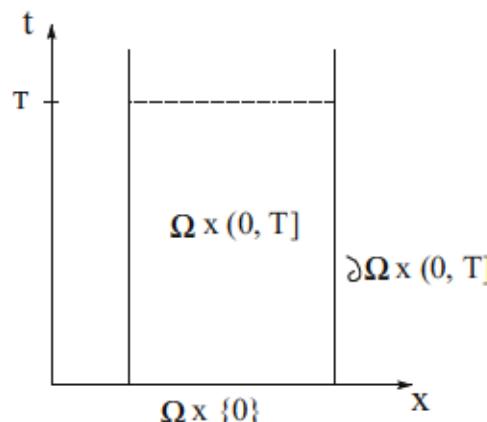


Abbildung 1.4: Parabolisches Raum-Zeit-Gebiet

**Satz 1.1 (Parabolisches Maximumprinzip):** Für jede klassische Lösung der Wärmeleitungs-Ungleichung

$$\partial_t u - \Delta u \leq 0 \quad \text{in } \Omega, \quad (1.4.34)$$

gilt das sog. „Maximumprinzip“, d. h.: Sie nimmt im (halboffenen) Zylinder  $Q_T := \Omega \times (0, T]$  kein striktes Maximum an.

**Beweis:** Wir geben den Beweis nur für eine Raumdimension. Die Verallgemeinerung für höhere Dimensionen ist dann evident. Für eine Lösung  $u$  der Wärmeleitungs-Ungleichung setzen wir  $v_\varepsilon := u - \varepsilon t$ , mit beliebigem  $\varepsilon > 0$ . Da  $v_\varepsilon$  stetig auf  $\bar{Q}_T$  ist, nimmt es in einem Punkt  $(x_0, t_0) \in \bar{Q}_T$  sein Maximum an. Angenommen,  $(x_0, t_0) \in Q_T$ . Dann gilt  $\partial_x^2 v_\varepsilon(x_0, t_0) \leq 0$ , und folglich

$$\partial_t v_\varepsilon(x_0, t_0) \leq \partial_t v_\varepsilon(x_0, t_0) - \partial_x^2 v_\varepsilon(x_0, t_0) = \partial_t u(x_0, t_0) - \varepsilon - \partial_x^2 u(x_0, t_0) \leq -\varepsilon.$$

Aus Stetigkeitsgründen ist dann auch  $\partial_t v_\varepsilon(x_0, t) \leq -\frac{1}{2}\varepsilon$  für  $t_0 - h \leq t \leq t_0$ , mit einem geeigneten  $h > 0$ . Hiermit folgern wir, dass

$$v_\varepsilon(x_0, t_0) - v_\varepsilon(x_0, t_0 - h) = \int_{t_0-h}^{t_0} \partial_t v_\varepsilon(x_0, t) dt \leq -\frac{1}{2}\varepsilon h < 0.$$

Dies führt auf den Widerspruch  $v_\varepsilon(x_0, t_0) < v_\varepsilon(x_0, t_0 - h)$ . Also nimmt  $v_\varepsilon$  notwendig sein Maximum für  $t = 0$  an. Da  $\varepsilon > 0$  beliebig klein gewählt werden darf, gilt diese Aussage auch für den (stetigen) Grenzfall  $\varepsilon = 0$ , d. h. für die Lösung  $u$ . Q.E.D.

Als Konsequenz des „parabolischen“ Maximumprinzips sehen wir insbesondere, dass eine Lösung der (homogenen) Wärmeleitungsgleichung

$$\partial_t u - \Delta u = 0 \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega, \quad (1.4.35)$$

zu nicht-negativen Anfangsdaten  $u^0 \geq 0$  ( $u|_{\partial\Omega} = 0$ ) nicht-negativ bleibt für alle  $t \geq 0$ :

$$u^0 \geq 0 \quad \Rightarrow \quad 0 \leq u(x, t) \leq \max_{\bar{\Omega}} u^0, \quad (x, t) \in Q_T. \quad (1.4.36)$$

Ferner sind wieder „klassische“ Lösungen  $u \in C(\bar{Q}_T) \cap C^2(Q_T)$  eindeutig bestimmt.

**Satz 1.2 (Globale Beschränktheit):** Für jede Lösung der inhomogenen Wärmeleitungsgleichung (1.4.29) gilt die a priori Abschätzung

$$\|u(t)\| \leq e^{-\lambda t} \|u^0\| + \lambda^{-1} \sup_{[0,t]} \|f\|, \quad (1.4.37)$$

mit dem kleinsten Eigenwert  $\lambda > 0$  des elliptischen Operators  $-\Delta$  auf  $\Omega$  zu homogenen Dirichlet-Randbedingungen.

**Beweis:** Wir betrachten die beiden Hilfsprobleme

$$\partial_t v - \Delta v = 0 \text{ in } Q_T, \quad v|_{\partial\Omega} = 0, \quad v|_{t=0} = u^0, \quad (1.4.38)$$

$$\partial_t w - \Delta w = f \text{ in } Q_T, \quad w|_{\partial\Omega} = 0, \quad w|_{t=0} = 0. \quad (1.4.39)$$

Offenbar ist dann  $u = v + w$  wegen der Linearität des Wärmeleitungsoperators (Superpositionsprinzip). Wir schätzen nun die beiden Lösungsanteile  $v$  und  $w$  separat ab.

i) Multiplikation von (1.4.38) mit  $v$  und Integration im Ort ergibt

$$\frac{1}{2}d_t\|v\|^2 + \|\nabla v\|^2 = 0.$$

Wir multiplizieren dies mit  $e^{2\lambda t}$  und finden

$$\frac{1}{2}d_t(e^{2\lambda t}\|v\|^2) + e^{2\lambda t}\|\nabla v\|^2 - \lambda e^{2\lambda t}\|v\|^2 = 0.$$

Wegen  $\lambda\|v\|^2 \leq \|\nabla v\|^2$  impliziert dies  $d_t(e^{2\lambda t}\|v\|^2) \leq 0$ , und Integration bzgl.  $t$  ergibt

$$e^{2\lambda t}\|v(t)\|^2 \leq \|u^0\|^2$$

bzw. wieder

$$\|v(t)\| \leq e^{-\lambda t}\|u^0\|.$$

ii) Multiplikation von (1.4.39) mit  $w$  und Integration im Ort ergibt

$$\frac{1}{2}d_t\|w\|^2 + \|\nabla w\|^2 = (f, w) \leq \frac{1}{2}\lambda\|w\|^2 + \frac{1}{2}\lambda^{-1}\|f\|^2.$$

Mit Hilfe von  $\lambda\|w\|^2 \leq \|\nabla w\|^2$  folgern wir

$$d_t\|w\|^2 + \|\nabla w\|^2 \leq \lambda^{-1}\|f\|^2.$$

Wir multiplizieren diese Ungleichung nun mit  $e^{\lambda t}$  und finden

$$d_t(e^{\lambda t}\|w\|^2) + e^{\lambda t}\|\nabla w\|^2 - \lambda e^{\lambda t}\|w\|^2 \leq \lambda^{-1}e^{\lambda t}\|f\|^2,$$

bzw.

$$d_t(e^{\lambda t}\|w\|^2) \leq \lambda^{-1}e^{\lambda t}\|f\|^2.$$

Integration bzgl.  $t$  ergibt

$$e^{\lambda t}\|w(t)\|^2 \leq \lambda^{-1} \int_0^t e^{\lambda s}\|f\|^2 ds,$$

$$\|w(t)\|^2 \leq \lambda^{-1}e^{-\lambda t} \int_0^t e^{\lambda s}\|f\|^2 ds.$$

Die Abschätzung

$$e^{-\lambda t} \int_0^t e^{\lambda s} ds \leq \lambda^{-1}.$$

impliziert dann  $\|w(t)\| \leq \lambda^{-1} \max_{[0,t]} \|f\|$ . Kombination der Resultate für  $v$  und  $w$  liefert die behauptete Abschätzung. Q.E.D.

Als Folgerung aus diesem Satz ersehen wir insbesondere, dass bei einem parabolischen Problem der Einfluss der Anfangsdaten exponentiell mit der Zeit abklingt. Weiter interessiert das Lösungsverhalten für Anfangsdaten  $u^0$  mit minimaler Regularität.

**Satz 1.3 (Glättungseigenschaft):** *Für jede Lösung der homogenen Wärmeleitungsgleichung (1.4.29) mit  $f \equiv 0$  gelten die a priori Abschätzungen*

$$\|\partial_t u(t)\| + \|\Delta u(t)\| \leq \|\Delta u^0\|, \quad t \geq 0, \quad (1.4.40)$$

$$\|\partial_t u(t)\| + \|\Delta u(t)\| \leq t^{-1} \|u^0\|, \quad t > 0, \quad (1.4.41)$$

vorausgesetzt, der Anfangswert  $u^0$  besitzt die erforderliche Regularität.

**Beweis:** Wir bedienen uns zum Beweis der sog. „Spektral-Methode“, welche aber auf symmetrische und autonome (d. h. nicht explizit von der Zeit abhängige) Operatoren beschränkt ist. Alternative Zugänge sind die „Halbgruppen-Methode“, welche nicht die Symmetrie des Operators erfordert, sowie die „Energie-Methode“, welche im allgemeinen Fall und sogar für nichtlineare Probleme anwendbar ist. Aus der Lösungsdarstellung (1.4.32) mit dem Orthonormalsystem von Eigenfunktionen  $\{v_j\}_{j \in \mathbb{N}}$  des Laplace-Operators,

$$u(x, t) := \sum_{j=1}^{\infty} u_j^0 v_j(x) e^{-\lambda_j t},$$

folgern wir

$$\partial_t u(x, t) = \Delta u(x, t) := - \sum_{j=1}^{\infty} u_j^0 \lambda_j v_j(x) e^{-\lambda_j t}.$$

Aufgrund der (verallgemeinerten) Parsevalschen<sup>15</sup> Identität gilt demnach

$$\|\partial_t u\|^2 = \|\Delta u\|^2 = \sum_{j=1}^{\infty} (u_j^0)^2 \lambda_j^2 e^{-2\lambda_j t}.$$

Als erstes Resultat entnehmen wir dieser Beziehung, dass

$$\|\partial_t u\|^2 = \|\Delta u\|^2 \leq \sum_{j=1}^{\infty} (u_j^0)^2 \lambda_j^2 = \|\Delta u^0\|^2.$$

Hieraus folgt wegen  $xe^{-x} \leq 1$ ,  $x \geq 0$ :

$$\|\partial_t u\|^2 = \|\Delta u\|^2 = t^{-2} \sum_{j=1}^{\infty} (u_j^0)^2 (\lambda_j t)^2 e^{-2\lambda_j t} \leq t^{-2} \sum_{j=1}^{\infty} (u_j^0)^2 = t^{-2} \|u^0\|^2,$$

was den Beweis vervollständigt. Q.E.D.

<sup>15</sup>Mare-Antoine Parseval des Chênes (1755–1836): Französischer Mathematiker; Privatgelehrter; Beiträge zu Reihen, insbesondere Fourier-Reihen.

Als Folgerung aus diesem Satz finden wir nochmal bestätigt, dass die Wärmeleitungsgleichung die „Glättungseigenschaft“ besitzt, d. h.: Irregularitäten in den Anfangsdaten werden für  $t > 0$  ausgeglättet. Durch Weiterführung der Argumentation im Beweis von Satz 1.3 lassen sich analog beliebig hohe Ableitungen der Lösung abschätzen:

$$\|\partial_t^p u(t)\| + \|\nabla^{2p} u(t)\| \leq c(p) \|\Delta^p u^0\|, \quad t \geq 0, \quad (1.4.42)$$

sowie

$$\|\partial_t^p u(t)\| + \|\nabla^{2p} u(t)\| \leq c(p) t^{-p} \|u^0\|, \quad t > 0, \quad p \in \mathbb{N}. \quad (1.4.43)$$

Später werden wir uns mit der Frage beschäftigen, ob und unter welchen Bedingungen auch numerische Verfahren zur Approximation der Wärmeleitungsgleichung ein solches „Glättungsverhalten“ in der Zeit aufweisen, d. h.: ob auch bei irregulären Anfangsdaten dennoch bei *festem*  $t > 0$  die volle Approximationsordnung garantiert ist.

## 1.5 Hyperbolische Probleme

Die Wellengleichung

$$\partial_t^2 u - \Delta u = 0 \quad (1.5.44)$$

wird in der Regel wieder auf einem Zylindergebiet  $Q_T := \Omega \times I$  mit einem (meist beschränkten) Gebiet  $\Omega \subset \mathbb{R}^d$  und einem Intervall  $I = (0, T]$  betrachtet. Die Frage nach der Wohlgestelltheit zugehöriger Anfangs-Randwertaufgaben wollen wir nur für den örtlich eindimensionalen Fall diskutieren. Die charakteristischen Steigungen der (örtlich) *eindimensionalen* Wellengleichung

$$\partial_t^2 u = \partial_x^2 u \quad (1.5.45)$$

sind gerade gegeben durch  $dt/dx = \pm 1$ , d. h.: die Charakteristiken sind alle Geraden in der  $(x, t)$ -Ebene mit der Steigung  $\pm 1$ . Die natürliche Anfangskurve  $\Gamma := \{(x, t) : x \in \Omega, t = 0\}$  ist also keine Charakteristik, so dass gemäß der Theorie die zugehörige Cauchysche Anfangswertaufgabe bei Vorgabe von Werten  $u(x, 0) = u^0(x)$  und  $\partial_t u(x, 0) = u^1(x)$  lösbar ist. Diese Lösung lässt sich im Fall einer Raumdimension leicht angeben. Wir betrachten den Sonderfall  $\Omega = \mathbb{R}^1$ . Die Koordinatentransformation  $\xi = x + t$ ,  $\eta = x - t$  überführt die Wellengleichung in die Form

$$\partial_\xi \partial_\eta u = 0.$$

Diese hat die Lösungen der Form

$$u(\xi, \eta) = F(\xi) + G(\eta)$$

mit beliebigen, hinreichend glatten Funktionen  $F(\cdot)$  und  $G(\cdot)$ . Die allgemeine Lösung der Wellengleichung lautet demnach

$$u(x, t) = F(x + t) + G(x - t).$$

Zur Erfüllung der Anfangsvorgaben auf  $\Gamma$  muss nun gelten:

$$F(x) + G(x) = u^0(x), \quad F'(x) - G'(x) = u^1(x).$$

Hieraus entnehmen wir, dass

$$F(x + t) + G(x + t) + F(x - t) + G(x - t) = u^0(x + t) + u^0(x - t), \quad (1.5.46)$$

$$F(x + t) - G(x + t) - F(x - t) + G(x - t) = \int_{x-t}^{x+t} u^1(s) ds, \quad (1.5.47)$$

und folglich,

$$u(x, t) = \frac{1}{2} \left\{ u^0(x + t) + u^0(x - t) + \int_{x-t}^{x+t} u^1(s) ds \right\}.$$

Dies ist die (eindeutige) klassische Lösung der Wellengleichung zu den vorgegebenen Anfangsdaten  $u^0(x)$ ,  $u^1(x)$ . Eine analoge Konstruktion ist auch in höheren Raumdimensionen möglich.

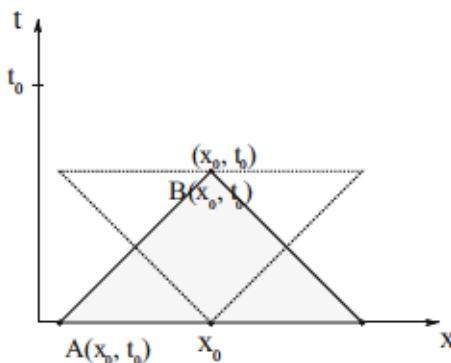


Abbildung 1.5: Hyperbolisches Raum-Zeit-Gebiet

Die Form der Lösung  $u(x, t)$  zeigt, dass bei einem hyperbolischen Problem die Ausbreitungsgeschwindigkeit von Information endlich ist. Lokale Störungen pflanzen sich entlang der Charakteristiken (Geraden mit Steigung  $\pm 1$ ) fort. Insbesondere erzeugen unstetige Anfangsdaten notwendig auch unstetige Lösungen. Dies erfordert im Falle irregulärer Anfangs- oder Randdaten einen neuartigen Lösungsbegriff, der auch Unstetigkeiten zulässt. Für jeden Punkt  $(x_0, t_0) \in Q_T$  gibt es demnach einen „Abhängigkeitsbereich“  $A(x_0, t_0)$  sowie einen „Bestimmtheitsbereich“  $B(x_0, t_0)$ , innerhalb deren sich das

Anfangswertproblem unabhängig vom restlichen Bereich lösen lässt:

$$A(x_0, t_0) := \{x \in \mathbb{R} : |x - x_0| \leq t_0\}, \quad B(x_0, t_0) := \{(x, t) \in \mathbb{R} \times \mathbb{R}_+ : |x - x_0| \leq t_0\}.$$

Die Eindeutigkeit von Lösungen der Wellengleichung erschließt man wieder am leichtesten mit „Hilbertraum-Argumenten“. Sei  $u(x, t)$  eine klassische Lösung der ARWA

$$\partial_t^2 u = \Delta u \quad \text{in } \Omega, \quad u|_{t=0} = u^0, \quad \partial_t u|_{t=0} = u^1, \quad u|_{\partial\Omega} = 0, \quad (1.5.48)$$

mit endlicher „Energie“ (kinetische + potentielle Energie)

$$E(t) := \|\partial_t u(t)\|_\Omega^2 + \|\nabla u(t)\|_\Omega^2 < \infty.$$

Multiplikation der Differentialgleichung mit  $\partial_t u$ , Integration über  $\Omega$  und anschließende partielle Integration ergibt

$$0 = (\partial_t^2 u - \Delta u, \partial_t u) = \frac{1}{2} d_t (\|\partial_t u\|^2 + \|\nabla u\|^2).$$

Dies impliziert, dass

$$\|\partial_t u(t)\|^2 + \|\nabla u(t)\|^2 = \|u^1\|^2 + \|\nabla u^0\|^2, \quad (1.5.49)$$

d. h. die Lösung ist eindeutig und hängt bzgl. der natürlichen Energie-Norm stetig von den Anfangsdaten ab. Ferner bleibt die Gesamtenergie  $E(t)$  im System in der Zeit erhalten. Dies entspricht der Vorstellung, dass bei einem Schwingungsprozess, etwa der Schwingung eines elastischen Körpers oder einer Schallwelle, bei Vernachlässigung von Dämpfung im Verlaufe der Zeit keine Energie verloren geht. Ein „gutes“ Diskretisierungsverfahren für die Wellengleichung sollte diese kritische Eigenschaft möglichst gut nachbilden.

## 1.6 Übungen

**Übung 1.1:** Gegeben sei die Anfangswertaufgabe (AWA) einer skalaren ODE

$$u'(t) = f(t, u(t)), \quad t \geq t_0, \quad u(t_0) = u^0,$$

mit einer analytischen Funktion  $f(t, x)$  mit gleichmäßig beschränkten Ableitungen

$$\sup_{j, k \geq 0} \|\partial_t^j \partial_x^k f(t, \cdot)\|_\infty \leq K < \infty, \quad t \geq t_0,$$

(z. B. die Funktion  $f(t, x) := \sin(x)$ ). Man zeige, dass man durch den Taylor-Ansatz

$$u(t) = u^0 + \sum_{k=1}^{\infty} \frac{f^{(k-1)}(t_0, u^0)}{k!} (t - t_0)^k$$

mit  $f^{(r)}(t_0, u^0) := (d/dt)^r f(t, u(t))|_{t=t_0}$  eine globale (d. h. für alle  $t \geq t_0$  existierende) Lösung der AWA erhält.

**Übung 1.2:** Im Text wurde die Typeneinteilung von linearen Differentialoperatoren 2. Ordnung mit der Aufgabe motiviert, aus gegebenen Werten  $u(x_0, y_0)$  und  $\partial_n u(x_0, y_0)$  entlang einer Kurve  $\Gamma$  die Lösung  $u(x, y)$  über einen Taylor-Reihenansatz zu bestimmen. Diese Konstruktion wurde allerdings nur bis zu den drei zweiten Ableitungen  $\partial_x^2 u(x_0, y_0)$ ,  $\partial_x \partial_y u(x_0, y_0)$  und  $\partial_y^2 u(x_0, y_0)$  durchgeführt und hing von der Regularität einer gewissen Matrix  $A$  ab. Man zeige, dass nach Bestimmung der zweiten Ableitungen die Konstruktion der vier dritten Ableitungen  $\partial_x^3 u(x_0, y_0)$ ,  $\partial_x^2 \partial_y u(x_0, y_0)$ ,  $\partial_x \partial_y^2 u(x_0, y_0)$  und  $\partial_y^3 u(x_0, y_0)$  auf dieselbe Matrix  $A$  führt. Diese Aussage gilt auch für die weiteren, höheren Ableitungen. Die Vorgenommene Klassifizierung des Differentialoperators als „elliptisch“, „parabolisch“ oder „hyperbolisch“ basierend auf der Konstruierbarkeit der Lösung aus den Randdaten ist also sinnvoll.

**Übung 1.3:** Man bestimme den Typ der Differentialgleichungen

- a)  $\partial_x \partial_y u - \partial_x u = 0,$
- b)  $\partial_x^2 u + \partial_x \partial_y u + y \partial_y^2 u + 4u = 0,$
- c)  $2(\partial_x + \partial_y)^2 u + \partial_y u = 0.$

(Hinweis: Das im Text angegebene Kriterium für den Typ einer Gleichung kann auch bei variablen Koeffizienten separat in jedem einzelnen Ortspunkt verwendet werden.)

**Übung 1.4:** Eine (skalare) lineare partielle Differentialgleichung (PDE) 2. Ordnung der Form

$$a_{11} \partial_1^2 u + a_{12} \partial_1 \partial_2 u + a_{21} \partial_2 \partial_1 u + a_{22} \partial_2^2 u = f$$

lässt sich durch Setzung  $u_1 := \partial_1 u$ ,  $u_2 := \partial_2 u$  in ein System von PDE 1. Ordnung umformen:

$$\begin{aligned} \partial_2 u_1 - \partial_1 u_2 &= 0, \\ a_{11} \partial_1 u_1 + a_{12} \partial_1 u_2 + a_{21} \partial_2 u_1 + a_{22} \partial_2 u_2 &= f. \end{aligned}$$

Man zeige, dass sich die in der Vorlesung für (skalare) lineare PDE 2. Ordnung durchgeführte Typeneinteilung analog auch für Systeme 1. Ordnung der allgemeinen Form

$$\begin{aligned} b_{11}^1 \partial_1 u_1 + b_{12}^1 \partial_1 u_2 + b_{21}^1 \partial_2 u_1 + b_{22}^1 \partial_2 u_2 &= f_1, \\ b_{11}^2 \partial_1 u_1 + b_{12}^2 \partial_1 u_2 + b_{21}^2 \partial_2 u_1 + b_{22}^2 \partial_2 u_2 &= f_2, \end{aligned}$$

vornehmen lässt (mit analogen Resultaten). Ziel ist dabei die Bestimmung von Kurven  $\Gamma \subset \mathbb{R}^2$ , bei denen die Vorgabe von Anfangswerten für  $u_1$  und  $u_2$  entlang  $\Gamma$  die Konstruktion aller Ableitungen von  $u_1$  und  $u_2$ , beginnend mit den zweiten Ableitungen  $\partial_1^2 u_1$ ,  $\partial_1 \partial_2 u_1$ ,  $\partial_2^2 u_1$ ,  $\partial_1^2 u_2$ ,  $\partial_1 \partial_2 u_2$ ,  $\partial_2^2 u_2$  und damit einen Taylor-Reihenansatz für die Lösung erlaubt. (Bem.: Wenn die Konstruktion von  $u_1$  und  $u_2$  auf diesem Wege möglich ist, erhält man für die gegebene PDE 2. Ordnung dann durch weitere Vorgabe von  $u$  entlang  $\Gamma$  aus der Kenntnis von  $\partial_1 u = u_1$  und  $\partial_2 u = u_2$  im ganzen Lösungsgebiet dort auch eine Lösung  $u$  durch einfaches Aufintegrieren.)

**Übung 1.5:** Im Text wurde die Poincarésche Ungleichung

$$\int_G |u(x)|^2 dx \leq d_G^2 \int_G \|\nabla u(x)\|^2 dx, \quad d_G := \text{diam}(G),$$

nur für Funktionen  $u \in V_0(G)$  formuliert, d. h. welche auf dem ganzen Rand  $\partial G$  null sind. Der Beweis funktioniert aber auch für Funktionen, die nur entlang eines Teils  $\Gamma \subset \partial G$  des Randes mit Länge  $|\Gamma| \neq 0$  null sind, d. h. auf dem Raum

$$V_0(\Gamma; G) := \{v \in C^1(G) \cap C(\bar{G}) : \nabla v \in L^2(G)^n, v|_{\Gamma} = 0\}.$$

- (i) Man führe den Beweis dieser Verallgemeinerung der Poincaréschen Ungleichung für das Einheitsquadrat  $Q = (0, 1)^2 \subset \mathbb{R}^2$  und den Randteil  $\Gamma := \{x = (x_1, 0) : 0 \leq x_1 \leq 1\}$ .
- (ii) Kann die Poincarésche Ungleichung gültig bleiben, wenn der Randteil  $\Gamma \subset \partial G$  trivial ist, etwa nur aus einem Punkt besteht? Man untersuche diese Frage anhand der in (i) gegebenen Situation mit  $\Gamma := \{(0, 0)\}$ . Welche Konsequenzen hat die Antwort auf diese Frage für die 1. RWA des Laplace-Operators? (Hinweis: Man betrachte die Folge der Funktionen  $u_k(r, \theta) = r^{1/k}$ .)

**Übung 1.6:** Auf einem beschränkten Gebiet  $\Omega \subset \mathbb{R}^n$  mit glattem Rand  $\partial\Omega$  werden die folgende (a) zweite und (b) dritte Randwertaufgabe betrachtet:

$$\begin{aligned} (a) \quad & -\Delta u + au = f \quad \text{in } \Omega, \quad \partial_n u = g \quad \text{auf } \partial\Omega, \\ (b) \quad & -\Delta u + au = f \quad \text{in } \Omega, \quad \partial_n u + \alpha u = g \quad \text{auf } \partial\Omega, \end{aligned}$$

mit Konstanten  $a > 0$  und  $\alpha \geq 0$ . Man zeige, dass diese RWAn jeweils höchstens eine „klassische“ Lösung  $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$  haben können. Welches Problem ergibt sich im Fall  $a = 0$ , d. h. für den reinen Laplace-Operator?

**Übung 1.7:** Für die klassische Lösung der Randwertaufgabe

$$-\Delta u = 1 \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega,$$

auf einem glatt berandetem Gebiet  $\Omega \subset Q_1 := \{(x, y) \in \mathbb{R}^2 \mid 0 < x, y < 1\}$  zeige man mit Hilfe des Maximumprinzips die Einschließung  $0 \leq u(x) \leq \frac{1}{8}$ . (Hinweis: Man vergleiche  $u$  mit der quadratischen Funktion  $v = \frac{1}{4}x(1-x) + \frac{1}{4}y(1-y)$ .)

**Übung 1.8:** Der Laplace-Operator  $\Delta = \operatorname{div} \operatorname{grad}$  hat für Funktionen  $u = u(r, \theta)$  in Polarkoordinaten  $(r, \theta) \in [0, \infty) \times [0, 2\pi]$  die folgende Form:

$$\Delta u = (\partial_r^2 + r^{-1}\partial_r + r^{-2}\partial_{\theta}^2)u.$$

- (i) Für ein  $\omega \in (0, 2\pi]$  sei  $S_{\omega} := \{(r, \theta) : r > 0, \theta \in (0, \omega)\}$  der zugehörige Sektor der  $(x, y)$ -Ebene. Man zeige, dass die auf dem Gebiet  $G := S_{\omega} \cap K_1(0)$  definierte Funktion

$$s_{\omega}(r, \theta) := r^{\pi/\omega} \sin(\theta\pi/\omega)$$

harmonisch ist, d. h.  $\Delta s_\omega \equiv 0$ , und den Randbedingungen  $s_\omega(r, 0) = s_\omega(r, \omega) = 0$  sowie  $s_\omega(1, \theta) = \sin(\theta\pi/\omega)$  genügt.

(ii) Man zeige, dass im Fall  $\pi < \omega \leq 2\pi$ , d. h. im Fall eines *stumpfen* Innenwinkels, die ersten Ableitungen dieser Funktion zwar unbeschränkt aber noch (uneigentlich) quadrat-integrabel sind, dass ihre zweiten Ableitungen aber nicht mehr quadrat-integrabel sind. Wie sieht das bei *spitzen* Innenwinkeln, d. h.  $0 < \omega < \pi$ , aus?

Dieses Beispiel zeigt, dass klassische Lösungen von elliptischen RWA auch zu glatten Daten am Gebietsrand nicht regulär zu sein brauchen.

**Übung 1.9:** Man untersuche, ob die folgenden Funktionen auf dem Einheitsquadrat  $\Omega = \{(x, y) \in \mathbb{R}^2 | 0 < x, y < 1\}$  im Sobolew-Raum  $H^1(\Omega)$  liegen:

$$a) \quad u(x, y) = |x - y|^{1/2}, \quad b) \quad u(x, y) = \sin(\ln(1/r)), \quad r = (x^2 + y^2)^{1/2}.$$

(Hinweis: Man untersuche die “uneigentliche” Riemann-Integrabilität der Ableitungen.)

**Übung 1.10:** Man zeige, dass für Funktionen  $u \in H^1(\Omega)$  unter der Mittelwertbedingung

$$\int_{\Omega} u(x) dx = 0$$

ebenfalls die Poincarésche Ungleichung gilt (mit einer Konstante  $c_\Omega$ ):

$$\|u\|_{\Omega} \leq c_\Omega \|\nabla u\|_{\Omega}.$$

(Hinweis: Zum Beweis gibt es zwei alternative Wege: Modifikation des direkten Beweises aus dem Text unter Verwendung von „Randwerten“  $u_{|\Gamma} = 0$ , oder Widerspruchssargument unter Verwendung des Rellichschen Auswahlsatzes.)

**Übung 1.11:** Man zeige mit Hilfe der Poincaréschen Ungleichung aus Aufgabe 3.2 die eindeutige Lösbarkeit der „Neumannschen Randwertaufgabe“ (2. RWA)

$$-\Delta u = f \text{ in } \Omega, \quad \partial_n u = 0 \text{ auf } \partial\Omega,$$

im Quotientenraum  $H^1(\Omega)/\mathbb{R}$  für rechte Seiten  $f \in L^2(\Omega)$  mit Mittelwert Null:

$$\int_{\Omega} f(x) dx = 0.$$

(Hinweis: Man modifiziere den Existenzbeweis für schwache Lösungen der 1. RWA des Laplace-Operators im Sobolew-Raum  $H_0^1(\Omega)$  für die 2. RWA im Sobolewschen Quotientenraum  $H^1(\Omega)/\mathbb{R}$ . Dabei bedeutet „Eindeutigkeit in  $H^1(\Omega)/\mathbb{R}$ “, dass eine „schwache“ Lösung in  $H^1(\Omega)$  existiert und bis auf eine additive Konstante eindeutig bestimmt ist.)

**Übung 1.12:** Auf welchem der folgenden Gebiete ist die RWA

$$-\Delta u = 0 \quad \text{in } \Omega, \quad u|_{\partial\Omega} = 0,$$

wohl gestellt, d. h. besitzt eine eindeutige schwache Lösung  $u \in H_0^1(\Omega)$  (mit Begründung)?

a) gepunktete Kreisscheibe

$$\Omega = \{x \in \mathbb{R}^2 \mid 0 < |x| < 1\} \subset \mathbb{R}^2;$$

b) geschlitzte Kreisscheibe

$$\Omega = \{x \in \mathbb{R}^2 \mid |x| < 1\} \setminus \Gamma, \quad \Gamma := \{x \in \mathbb{R}^2 \mid -\frac{1}{2} \leq x_1 \leq \frac{1}{2}, x_2 = 0\} \subset \mathbb{R}^2.$$

**Übung 1.13:** Man gebe eine variationelle Formulierung der folgenden Randwertaufgabe an:

$$-\Delta u + u = f \quad \text{in } \Omega, \quad \partial_n u + u|_{\partial\Omega} = g,$$

und begründe, dass deren Lösung im Falle ausreichender Glattheit die RWA löst.

**Übung 1.14:** Welche von den folgenden Sobolewschen Ungleichungen sind richtig?

- a)  $\|u\|_{L^\infty(\Omega)} \leq c \|u\|_{H^2(\Omega)}, \quad u \in H^2(\Omega), \quad \Omega \subset \mathbb{R}^3;$
- b)  $\|u\|_{L^\infty(\Omega)} \leq c \|u\|_{H^{1,1}(\Omega)}, \quad u \in H^{1,1}(\Omega), \quad \Omega \subset \mathbb{R}^1;$
- c)  $\|u\|_{L^\infty(\Omega)} \leq c \|u\|_{H^1(\Omega)}, \quad u \in H^1(\Omega), \quad \Omega \subset \mathbb{R}^2;$
- d)  $\|u\|_{L^1(\partial\Omega)} \leq c \|u\|_{H^{1,1}(\Omega)}, \quad u \in H^{1,1}(\Omega), \quad \Omega \subset \mathbb{R}^2.$

Man erkläre die Bedeutung der verwendeten Funktionenräume und Normen.

**Übung 1.15:** Für die (klassische) Lösung der Wärmeleitungsgleichung

$$\partial_t u(x, t) - \Delta u(x, t) = 0 \quad \text{in } Q_T := \Omega \times (0, T], \quad u|_{t=0} = u^0(x), \quad u|_{\partial\Omega} = 0,$$

zeige man mit Hilfe der „Spektraltechnik“ für  $u^0 \in H_0^1(\Omega)$  die a priori Abschätzung

$$\|\partial_t u(\cdot, t)\| + \|\Delta u(\cdot, t)\| \leq t^{-1/2} \|\nabla u^0\|, \quad t > 0.$$

**Übung 1.16:** Man konstruiere mit Hilfe der Methode der Variablenseparation eine Lösung der sog. „1. Anfangs-Randwert-Aufgabe (1. ARWA)“ der Wellengleichung

$$\partial_t^2 u(x, t) - \Delta u(x, t) = 0, \quad u|_{t=0} = u^0, \quad \partial_t u|_{t=0} = u^1, \quad u|_{\partial\Omega} = 0.$$

Welche Regularität muss dabei für die Anfangswerte  $u^0$  und  $u^1$  gefordert werden? (Hinweis: Man orientiere sich am entsprechenden Beweis für die Wärmeleitungsgleichung unter Verwendung des ONS von Eigenfunktionen des Laplace-Operators.)



## 2 Differenzen-Verfahren für elliptische Probleme

In diesem Kapitel werden wir zunächst die klassischen Differenzenapproximationen zur Lösung elliptischer Randwertaufgaben (RWAn) diskutieren. Der Übersichtlichkeit halber beschränken wir uns dabei auf das Modellproblem der Poisson-Gleichung in zwei Raumdimensionen mit Dirichletschen Randbedingungen, d. h. auf die 1. RWA:

$$Lu := -\Delta u = f \quad \text{in } \Omega, \quad u = g \quad \text{auf } \partial\Omega. \quad (2.0.1)$$

Das Definitionsgebiet  $\Omega \subset \mathbb{R}^2$  wird zunächst wieder als glatt berandet oder als konkaves Polygongebiet vorausgesetzt. Die Problemdaten  $f, g$  sind ebenfalls glatt, so dass die im vorigen Kapitel beschriebenen Resultate anwendbar sind. Erweiterungen für Probleme mit variablen Koeffizienten oder anderen Randbedingungen sowie auf den dreidimensionalen Fall werden gegebenenfalls in Bemerkungen diskutiert.

### 2.1 Allgemeine Differenzenapproximationen

Zur Definition einer sog. „Differenzenapproximation“ der RWA wird das Lösungsgebiet  $\Omega$  durch ein endliches (nicht notwendig äquidistantes oder kartesisches) Punktgitter überdeckt. Wir definieren disjunkte Punktmengen

$$\Omega_h := \{\text{„innere“ Gitterpunkte in } \Omega\}, \quad \partial\Omega_h := \{\text{„Rand-Gitterpunkte“ nahe bei } \partial\Omega\},$$

und setzen  $\bar{\Omega}_h := \Omega_h \cup \partial\Omega_h$ . Welche Punkte zu  $\partial\Omega_h$  gehören, hängt von der Eigenart der gewählten Differenzenapproximation ab; im folgenden werden verschiedene Beispiele betrachtet. Der Parameter  $h > 0$  beschreibt wie üblich die „Feinheit“ des Gitters  $\bar{\Omega}_h$ , d. h. so etwas wie den maximalen (horizontalen oder vertikalen) Abstand benachbarter Gitterpunkte. Die Gitterpunkte in  $\partial\Omega_h$  seien ähnlich dicht verteilt wie die in  $\Omega_h$ . Die Differentialgleichung wird nun in den Punkten in  $\Omega_h$  betrachtet und jede Ableitung durch einen Differenzenquotient ersetzt. Diese Differenzenapproximation greift dabei auch auf Randpunkte in  $\partial\Omega_h$  zurück. So ergibt sich ein Differenzenschema zur Bestimmung einer diskreten Lösung  $u_h : \bar{\Omega}_h \rightarrow \mathbb{R}$  der allgemeinen Gestalt

$$L_h u_h(P) = f_h(P) \quad \text{für } P \in \Omega_h, \quad u_h(P) = g_h(P) \quad \text{für } P \in \partial\Omega_h, \quad (2.1.2)$$

wobei  $f_h$  und  $g_h$  geeignete Approximationen der rechten Seite  $f$  bzw. der Randwerte  $g$  sind (im einfachsten Fall etwa  $f_h(P) = f(P)$  und  $g_h(P) = g(P)$ ). Auf natürliche Weise verstehen wir die Anwendung des Operators  $L_h$  auch auf die kontinuierliche Lösung  $u \in C(\bar{\Omega})$ . Zur Konstruktion konkreter Differenzenoperatoren definieren wir zu jedem Punkt eine Umgebung von (verschiedenen) Punkten

$$N(P) := \{Q_i, i = 0, \dots, r_P\} \quad (\text{Konvention: } Q_0 := P),$$

auf denen eine Differenzenapproximation des Laplace-Operators definiert ist. Der Differenzenoperator hat dann die folgende Form:

$$L_h u_h(P) = \sum_{Q \in N(P)} \sigma(P, Q) u_h(Q), \quad (2.1.3)$$

mit gewissen Koeffizienten  $\sigma(P, Q)$ . Diese sind so zu bestimmen, dass die folgenden Forderungen erfüllt sind:

i) *Konsistenz*: Das Schema ist „konsistent“, d. h.: Für den „Abschneidefehler“

$$\tau_h(P) := L_h u(P) - f_h(P), \quad P \in \Omega_h,$$

gilt

$$\max_{P \in \Omega_h} |\tau_h(P)| \rightarrow 0 \quad (h \rightarrow 0). \quad (2.1.4)$$

Wünschenswert wäre eine möglichst hohe „Konsistenzordnung“  $m \geq 1$ , d. h.:

$$\max_{P \in \Omega_h} |\tau_h(P)| = O(h^m). \quad (2.1.5)$$

Wir werden uns im Folgenden üblicherweise meist mit  $m = 2$  begnügen. Aus Ökonomiegründen wird man  $r_P$  von moderater Größe wählen ( $r_P \approx 4 - 24$  in zwei Raumdimensionen).

ii) *Stabilität*: Das Differenzenschema ist wohl-gestellt, d. h.: Es bestimmt eindeutige diskrete Lösungen  $(u_h(P))_{P \in \bar{\Omega}_h}$ , welche gleichmäßig bzgl.  $h$  stetig von den Daten des Problems abhängen. Dazu wäre z. B. eine Stabilitätsabschätzung der folgenden Form zu beweisen:

$$\max_{P \in \bar{\Omega}_h} |u_h(P)| \leq c_{\text{stab}} \left\{ \max_{P \in \Omega_h} |L_h u_h(P)| + \max_{P \in \partial\Omega_h} |u_h(P)| \right\}, \quad (2.1.6)$$

mit einer von  $h$  unabhängigen Konstante  $c_{\text{stab}} > 0$ .

iii) *Verträglichkeit*: Der Differenzenoperator  $L_h$  soll analoge charakteristische Eigenschaften wie der kontinuierliche  $L$  besitzen; z. B.: Symmetrie, Definitheit, Maximumsprinzip etc..

Das Hauptziel der Konvergenzanalyse von Differenzenschemata ist der Nachweis, dass eine Konsistenzordnung  $m$  auch eine Konvergenzordnung  $m$  des Fehlers  $e_h := u - u_h$  impliziert:

$$\max_{P \in \bar{\Omega}_h} |e_h(P)| = O(h^m), \quad (2.1.7)$$

vorausgesetzt, die Lösung  $u$  ist ausreichend regulär.

### 2.1.1 Konsistenz

Für den Fall  $f_h(P) = f(P)$  betrachten wir o.B.d.A. den Punkt  $P = Q_0 = 0$  und die Punktumgebung  $N(P) = \{Q_i = (x_i, y_i), i = 0, \dots, r_P\}$ . Taylorentwicklung ergibt:

$$u(Q_i) = u(P) + x_i \partial_x u(P) + y_i \partial_y u(P) + \frac{1}{2} x_i^2 \partial_x^2 u(P) + x_i y_i \partial_x \partial_y u(P) + \frac{1}{2} y_i^2 \partial_y^2 u(P) + \dots.$$

Wir wollen aus diesem Ansatz die Koeffizienten  $\sigma_i := \sigma(P, Q_i)$  im Differenzenoperator so bestimmen, dass eine vorgegebene Konsistenzordnung garantiert ist. Mit dem Ansatz  $\tau_h(P) = O(h^m)$  erhält man durch Koeffizientenvergleich als notwendige und hinreichende Bedingung für Konsistenz:

(B0) *Konsistenz*: Für die Koeffizienten des Differenzenschemas gilt:

$$\sum_{i=0}^{r_P} \sigma_i = \sum_{i=0}^{r_P} x_i \sigma_i = \sum_{i=0}^{r_P} y_i \sigma_i = \sum_{i=0}^{r_P} x_i y_i \sigma_i = 0, \quad \frac{1}{2} \sum_{i=0}^{r_P} x_i^2 \sigma_i = \frac{1}{2} \sum_{i=0}^{r_P} y_i^2 \sigma_i = 1. \quad (2.1.8)$$

Aus diesen Beziehungen sind die Koeffizienten  $\sigma_i, i = 0, \dots, r_P$ , im Differenzenoperator  $L_h$  zu bestimmen. Für Konsistenz sind also im allgemeinen Fall eines völlig unstrukturierten Gitters mindestens 6 Punkte in  $N(P)$  erforderlich. Die Konsistenz des Differenzenschemas ist offenbar äquivalent dazu, daß der Differenzenoperator „exakt“ ist für quadratische Polynome:

$$u \in P_2 : \quad L_h u(P) = Lu(P), \quad P \in \Omega_h. \quad (2.1.9)$$

Da man in der Regel Diskretisierungen auf Folgen von Gittern mit Gitterweiten  $h \rightarrow 0$  betrachtet, führt man skalierte Parameter  $\xi_i = h^{-1} x_i, \eta_i = h^{-1} y_i$  ein, um sich von der  $h$ -Abhängigkeit in den Koeffizienten zu befreien. Wegen

$$\frac{1}{2} h^2 \sum_{i=0}^{r_P} \xi_i^2 \sigma_i = \frac{1}{2} h^2 \sum_{i=0}^{r_P} \eta_i^2 \sigma_i = 1,$$

ist  $\sigma_i = 0$  oder  $\sigma_i \sim h^{-2}$ . Wenn also eine Lösung für  $(\sigma_0, \dots, \sigma_{r_P})$  existiert, dann ergibt sich für den Konsistenzfehler

$$\tau_h(P) = (L_h u - Lu)(P) = O(h^3 \xi_i^3 \sigma_i + \dots) = O(h). \quad (2.1.10)$$

Auch bei ganz irregulärer Gitterpunktsanordnung erhält man somit schon mindestens eine Konsistenzordnung  $m = 1$ . Es gibt mehrere Möglichkeiten, die Ordnung zu erhöhen:

- Man nimmt mehr Gitterpunkte in die Menge  $N(P)$  auf.
- Man ordnet das Gitter regulär an, um in der Taylor-Entwicklung des Abschneidefehlers Weghebeffekte zu erzielen.

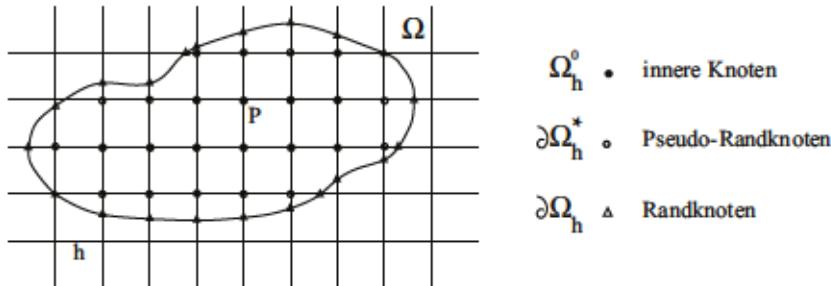


Abbildung 2.1: Gitter für Differenzenapproximation

Wir werden uns nun im Folgenden mit regulären, speziell äquidistanten, kartesischen Gittern beschäftigen. Der Einfachheit halber sollen die Gitterpunkte äquidistant entlang von Parallelen zu den Koordinatenachsen angeordnet sein. Dabei bezeichnet  $\mathbb{R}_h^2$  das gesamte den  $\mathbb{R}^2$  überdeckende Punktgitter. Die Schnittpunkte der Gitterlinien mit dem Rand  $\partial\Omega$  bilden dabei natürliche Stützpunkte für die Approximation der Randwerte. Die „Gitterweite“  $h$  hat auf solchen Gittern eine natürliche Bedeutung.

Die einfachste Approximation  $\Delta_h \approx \Delta$  des Laplace-Operators verwendet zentrale Differenzenquotienten 2. Ordnung in jeder der Koordinatenrichtungen. Man erhält den sog. „5-Punkte-Operator“:

$$\Delta_h^{(5)} u_h(x, y) := \frac{1}{h^2} \left\{ u(x \pm h, y) + u(x, y \pm h) - 4u(x, y) \right\}$$

für innere Gitterpunkte  $P = (x, y) \in \Omega_h$ . Wir verwenden hier die Konvention, dass „ $\pm h$ “ abkürzend für die Summe der beiden Terme „ $+h$ “ und „ $-h$ “ steht. Die Konsistenzordnung dieser Differenzenapproximation ist wegen der Äquidistanz des Gitters und der symmetrischen Plazierung der Stützpunkte  $m = 2$ :

$$|\Delta_h^{(5)} u(P) - \Delta u(P)| \leq \frac{1}{6} M_4(u) h^2, \quad (2.1.11)$$

wobei  $M_4(u) := \max_{\Omega} \{ |\partial_x^i \partial_y^j u|, i+j=4 \}$ . Diese Aussage bleibt sinngemäß gültig, wenn das Gitter nur in jeder einzelnen der Koordinatenrichtungen äquidistant ist. Dagegen ginge die Konsistenzordnung auf  $m = 1$  zurück, wenn das Gitter zwar kartesisch, aber innerhalb einer Koordinatenrichtung nicht äquidistant wäre. Höhere Approximationsordnungen lassen sich z. B. auf einem gleichförmigen Gitter durch Hinzunahme von weiteren Punkten in der Umgebung des Auswertungspunkts  $(x, y)$  erzielen:

- i) Approximation der zweiten Ableitungen im Laplace-Operator durch zentrale Differenzenquotienten auf jeweils 5 Punkten ergibt den „gestreckten“ 9-Punkte-Operator:

$$\Delta_h^{(9)} u_h(x, y) = \frac{1}{12h^2} \left\{ -u(x \pm 2h, y) + 16u(x \pm h, y) - u(x, y \pm 2h) + 16u(x, y \pm h) - 60u(x, y) \right\}.$$

Dieser Differenzenoperator hat offensichtlich die Konsistenzordnung  $m = 4$ , doch hat die zugehörige Koeffizientenmatrix wegen des Vorzeichenwechsels ungünstige Eigenschaften.

ii) Der „kompakte“ 9-Punkte-Operator verwendet neben dem Auswertungspunkt  $(x, y)$  die 8 direkten Nachbarpunkte  $(x \pm h, y), (x, y \pm h), (x, y \pm h)$ :

$$\bar{\Delta}_h^{(9)} u_h(x, y) = \frac{1}{6h^2} \left\{ 4u(x \pm h, y) + 4u(x, y \pm h) + u(x \pm h, y \pm h) - 20u(x, y) \right\}.$$

Dieses Schema hat zunächst auch nur die Ordnung  $m = 2$ , doch kann man es durch eine Modifikation bei der Auswertung der rechten Seite auf die Ordnung  $m = 4$  bringen (Übungsaufgabe):

$$f(x, y) \rightarrow f_h(x, y) := f(x, y) + \frac{1}{12} h^2 \Delta_h^{(5)} f(x, y). \quad (2.1.12)$$

**Approximation entlang des Randes:** Bei der Approximation am (möglicherweise gekrümmten) Rand des Gebiets gibt es verschiedene Möglichkeiten, die, wie wir später sehen werden, durchaus auf unterschiedliche Approximationsordnungen führen. Wir betrachten wieder den 5-Punkte-Operator und definieren zunächst die folgenden Gitterpunktmengen:

$$\Omega_h := \{P \in \mathbb{R}_h^2 \mid N(P) \subset \overline{\Omega}\}, \quad \partial\Omega_h := \bigcup_{P \in \Omega_h} N(P) \setminus \Omega_h.$$

i) *Konstante Randwertextrapolation:* In Punkten von  $\Omega_h$  wird der 5-Punkte-Operator angesetzt, und in Punkten  $P \in \partial\Omega_h$  werden die Randwerte  $u_h(P) = g(P^0)$  verwendet. Dabei ist  $P^0$  der Punkt  $P$  entlang einer der Koordinatenachsen am nächsten gelegene Punkt auf  $\partial\Omega$ . Dies ergibt entlang des Randes nur eine Approximationsordnung  $m = 1$ .

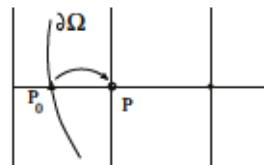


Abbildung 2.2: Schema der konstanten Randapproximation

ii) *Lineare Randwertinterpolation:* Jeder Punkt  $P \in \partial\Omega_h$  liegt in  $x$ - und/oder  $y$ -Richtung zwischen zwei Punkten  $P_0 \in \partial\Omega$  und  $P_1 \in \Omega_h$  mit Abständen  $0 \leq \alpha h < h$  zu  $P_0$  und  $h$  zu  $P_1$ . Man setzt dann (lineare Interpolation):

$$u_h(P) := \frac{1}{1+\alpha} \{g(P_0) + \alpha u_h(P_1)\}. \quad (2.1.13)$$

Damit wird eine implizite Kopplung der Randwerte an die „inneren“ Lösungswerte bewirkt. Diese Randwertapproximation hat die Ordnung  $m = 2$ .

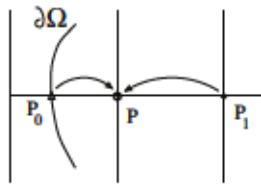


Abbildung 2.3: Schema der linearen Randapproximation

iii) *Shortley<sup>1</sup>-Weller<sup>2</sup>-Approximation:* Wir definieren die folgenden Punktmengen:

$$\Omega_h^0 := \{P \in \mathbb{R}_h^2 \mid N(P) \subset \Omega\}, \quad \partial\Omega_h^* := \bigcup_{P \in \Omega_h^0} N(P) \setminus \Omega_h,$$

$$\Omega_h := \Omega_h^0 \cup \partial\Omega_h^*, \quad \partial\Omega_h := \{\text{Schnittpunkte der Gitterlinien mit } \partial\Omega\}.$$

In Punkten  $P \in \Omega_h$  wird wieder der normale 5-Punkte-Operator und in den „fiktiven“ Randpunkten  $P \in \partial\Omega_h^*$  der modifizierte 5-Punkte-Operator verwendet (siehe die schematische Darstellung):

$$\begin{aligned} -\Delta_h^* u_h(x, y) := h^{-2} & \left\{ \left( \frac{2}{\alpha} + \frac{2}{\beta} \right) u_h(x, y) - \frac{2}{1+\alpha} u_h(x+h, y) - \frac{2}{\alpha(1+\alpha)} u_h(x-\alpha h, y) \right. \\ & \left. - \frac{2}{1+\beta} u_h(x, y+h) - \frac{2}{\beta(1+\beta)} u_h(x, y-\beta h) \right\} = f(x, y), \end{aligned}$$

gemäß

$$-\Delta_h^* u_h = f \quad \text{auf } \partial\Omega_h^*, \tag{2.1.14}$$

mit den Werten  $u_h(P) := g(P)$  auf  $\partial\Omega_h$ . Für diesen Differenzenoperator gilt

$$|\Delta_h^* u(P) - \Delta u(P)| \leq \frac{2}{3} M_3(u) h, \tag{2.1.15}$$

wobei  $M_3(u) := \max_{\overline{\Omega}} \{ |\partial_x^i \partial_y^j u|, i+j=3 \}$ .

---

<sup>1</sup>George H. Shortley (1910–????): US-Amerikanischer Astro-Physiker; wirkte 1935–2011 als Prof. an der Johns Hopkins University.

<sup>2</sup>Royal Weller (????–????): US-Amerikanischer Physiker und Ingenieur; die nach ihm und G. H. Shortley benannte Methoden findet sich in „The numerical solution of Laplace's equation“, J. Appl. Physics 9, 334 (1938); Mitherausgeber zweier Lehrbücher „Modern Physics for the Engineer“ (1954) und „Modern Mathematics for the Engineer“ (2013).

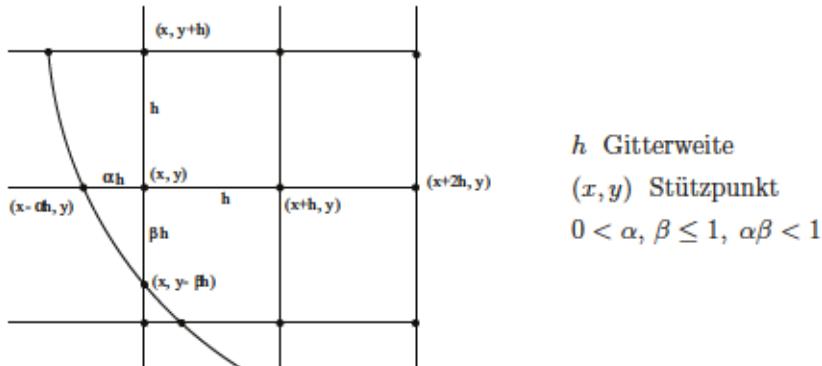


Abbildung 2.4: Schema der Shortley-Weller-Approximation

Die Verwendung dieser Differenzenapproximationen führt auf Schemata der Form

$$L_h u_h = f_h \quad \text{in } \Omega_h, \quad R_h u_h = g_h \quad \text{auf } \partial\Omega_h^*, \quad (2.1.16)$$

wobei der „Randoperator“  $R_h$  die jeweilige Randwertapproximation beschreibt. Dabei werden die echten Randwerte von  $u_h$  auf  $\partial\Omega_h$  unter Umständen implizit mit Werten im Innern verkoppelt. In diesem Fall sind alle Werte  $u_h(P)$ ,  $P \in \bar{\Omega}_h$ , zu bestimmen. Der Differenzenoperator  $L_h$  steht für den normalen 5-Punkte-Operator in  $\Omega_h$  und gegebenenfalls den modifizierten Shortley-Weller-Operator auf  $\partial\Omega_h^*$ . Im einfachsten Fall der trivialen Randwertapproximation  $R_h u_h = u_h$  können die Randwerte  $u_h(P) = g_h(P)$ ,  $P \in \partial\Omega_h$ , direkt eliminiert werden. Wir werden im folgenden nur diesen Fall weiter diskutieren.

**Bemerkung:** Die obigen Differenzenschemata haben natürliche Analoga in drei Raumdimensionen. Man spricht dann aus naheliegendem Grund vom „7-Punkte-Operator“. Dessen Abschneidefehler genügt der Abschätzung

$$|\Delta_h^{(7)} u(P) - \Delta u(P)| \leq \frac{1}{4} M_4(u) h^2. \quad (2.1.17)$$

## 2.2 Eigenschaften der Differenzengleichungen

Ausgangspunkt der folgenden Untersuchungen ist das Differenzenschema der Gestalt

$$L_h u_h(P) := \sum_{Q \in N(P)} \sigma(P, Q) u_h(Q) = f_h(P), \quad P \in \Omega_h, \quad (2.2.18)$$

$$u_h(P) = g_h(P), \quad P \in \partial\Omega_h, \quad (2.2.19)$$

mit geeigneten Approximationen  $f_h(\cdot)$  zu  $f$  und  $g_h(\cdot)$  zu  $g$ . Wir definieren die Koeffizienten  $\sigma(P, Q) := 0$  für Punkte  $Q \notin N(P)$ . Für  $P \in \Omega_h$  gilt dann

$$\sum_{Q \in \Omega_h} \sigma(P, Q) u_h(Q) = f_h(P) - \sum_{Q \in \partial\Omega_h} \sigma(P, Q) g_h(Q). \quad (2.2.20)$$

Bei (beliebiger) Numerierung der Gitterpunkte etwa gemäß  $\Omega_h = \{P_n, n = 1, \dots, N\}$ ,  $\partial\Omega_h = \{P_n, n = N+1, \dots, N+M\}$  ergibt sich ein quadratisches Gleichungssystem für den Vektor der (inneren) Knotenwerte  $U = (U_n)_{n=1}^N$ ,  $U_n := u_h(P_n)$ :

$$AU = F, \quad (2.2.21)$$

mit  $A = (A_{nm})_{n,m=1}^N$ ,  $F = (F_n)_{n=1}^N$ , wobei

$$A_{nm} := \sigma(P_n, P_m), \quad F_n := f_h(P_n) - \sum_{m=N+1}^{N+M} \sigma(P_n, P_m) g_h(P_m).$$

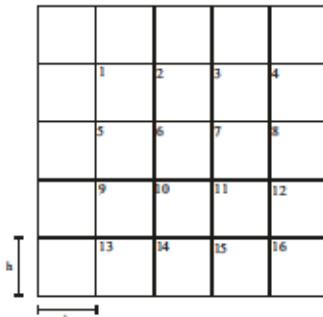


Abbildung 2.5: Differenzengitter

$$h = \frac{1}{m+1} \quad \text{Gitterweite}$$

$$N = m^2 \quad \text{"innere" Gitterpunkte}$$

**Beispiel („Modellproblem“):** Im Fall des Einheitsquadrats  $\Omega = (0, 1)^2$  ergibt sich für den 5-Punkte-Operator bei zeilenweiser Durchnumerierung des Gitters  $\overline{\Omega}_h = \{P_{ij}\}_{i,j=0}^{m+1}$  die folgende dünn-besetzte Matrix der Dimension  $N = m^2$ :

$$A = \frac{1}{h^2} \left. \begin{bmatrix} B_m & -I_m & & & \\ -I_m & B_m & -I_m & & \\ & -I_m & B_m & \ddots & \\ & & \ddots & \ddots & \end{bmatrix} \right\} N \quad B_m = \left. \begin{bmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & -1 & 4 & \ddots & \\ & & \ddots & \ddots & \end{bmatrix} \right\} m$$

mit der  $m \times m$ -Einheitsmatrix  $I_m$ . Die rechte Seite ist  $F := (f(x_{11}), \dots, f(x_{mm}))^T$ . Die Matrix  $A$  ist eine dünn besetzte Bandmatrix mit der Halbbandbreite  $m$ , symmetrisch und (irreduzibel) diagonal-dominant. Damit ist sie auch regulär und positiv definit.

**Bemerkung:** In drei Raumdimensionen hat die entsprechende Matrix die Dimension  $N = m^3$  und die Halbbandbreite  $m^2$ . Ansonsten hat sie dieselben Eigenschaften wie in zwei Dimensionen.

Im allgemeinen Fall ist für moderates  $r_P$  die Matrix  $A$  dünn besetzt, aber häufig wegen der Randwertapproximation *nicht* symmetrisch. Dies wäre ein schwerwiegender Nachteil, etwa bei der Approximation von Eigenwertaufgaben zum Laplace-Operator. Um die Regularität von  $A$  bzw. die Lösbarkeit der Differenzengleichung garantieren zu können, formulieren wir die folgenden Strukturbedingungen in numerierungs-unabhängiger Form:

(B1) *Erweiterte Diagonaldominanz*: Für Punkte  $P \in \Omega_h$  gilt:

$$(i) \quad \sum_{Q \in \bar{\Omega}_h, Q \neq P} |\sigma(P, Q)| \leq |\sigma(P, P)|, \quad (2.2.22)$$

und für Punkte  $P \in \Omega_h^* := \{Q \in \Omega_h : N(Q) \cap \partial\Omega_h \neq \emptyset\}$ :

$$(ii) \quad \sum_{Q \in \Omega_h, Q \neq P} |\sigma(P, Q)| < |\sigma(P, P)|. \quad (2.2.23)$$

(B2) *Nicht-negativer Typ*: Für Punkte  $P \in \Omega_h$  und  $Q \in \bar{\Omega}_h \setminus \{P\}$  gilt:

$$\sigma(P, P) > 0, \quad \sigma(P, Q) \leq 0. \quad (2.2.24)$$

(B3) *Zusammenhang*: Es ist  $\partial\Omega_h \neq \emptyset$ , und mit  $N(P) := \{Q \in \bar{\Omega}_h, \sigma(P, Q) \neq 0\}$  gilt für jede echte Teilmenge  $S_h \subset \Omega_h$ :

$$\left( \bigcup_{P \in S_h} N(P) \right) \cap (\Omega_h \setminus S_h) \neq \emptyset. \quad (2.2.25)$$

Die ersten beiden Bedingungen (B1) und (B2) werden von vielen Differenzenschemata, insbesondere solchen höherer als zweiter Ordnung (z. B.: der „gestreckte“ 9-Punkte-Operator), nicht erfüllt. Die hier betrachteten 5-Punkte-Schemata mit Randapproximation genügen ihnen aber. Wir zeigen im Folgenden, wie bei den einzelnen Randapproximationen die Bedingung (B1ii) erfüllt ist.

i) Konstante Randwertextrapolation: Für  $P \in \Omega_h^*$  gilt

$$\sum_{Q \in \Omega_h, Q \neq P} |\sigma(P, Q)| \leq 3h^{-2} = \frac{3}{4} |\sigma(P, P)|,$$

ii) Lineare Randwertinterpolation: Für  $P \in \Omega_h^*$  gilt

$$\sum_{Q \in \Omega_h, Q \neq P} |\sigma(P, Q)| \leq \left(4 - \frac{\alpha}{1+\alpha}\right) h^{-2} \leq 4h^{-2} = |\sigma(P, P)|,$$

iii) Shortley-Weller-Randwertapproximation: Für  $P \in \Omega_h^*$  gilt

$$\sum_{Q \in \Omega_h, Q \neq P} |\sigma(P, Q)| \leq \left(\frac{2}{1+\alpha} + \frac{2}{1+\beta}\right) h^{-2} \leq \frac{1}{2} \left(\frac{2}{\alpha} + \frac{2}{\beta}\right) h^{-2} = \frac{1}{2} |\sigma(P, P)|.$$

Die dritte Bedingung (B3) sichert die Kopplung eines jeden inneren Gitterpunktes  $P \in \bar{\Omega}_h$  mit allen anderen, insbesondere mit den Randpunkten. Die Eigenschaft des kontinuierlichen Problems, dass sich eine kleine Störung in einem Punkt global bemerkbar macht, spiegelt sich hier wider.

Die obigen Eigenschaften des Differenzenschemas korrespondieren zu den bereits bekannten der zugehörigen Koeffizientenmatrix  $A$ . unabhängig von der gewählten Numerierung der Gitterpunkte implizieren die Bedingungen (B1) und (B3) zunächst die einfache Diagonaldominanz von  $A$ , darüber hinaus aber auch die stärkere „irreduzible Diagonaldominanz“ (analog dem „schwachen Zeilensummenkriterium“ für die Konvergenz des bekannten Jacobi- oder des Gauß-Seidel-Verfahrens). Die Bedingung (B2) entspricht einer analogen für  $A$ .

Um Existenz und Eindeutigkeit einer Lösung für das allgemeine Differenzenschema garantieren zu können, geht man ähnlich wie im Kontinuierlichen vor und nutzt ein diskretes Analogon des „Maximumsprinzips“.

**Satz 2.1 (Diskretes Maximumsprinzip):** Unter den Voraussetzungen (B1i), (B2) und (B3) genügt das zugehörige Differenzenschema einem „diskreten Maximumsprinzip“, d. h.: Gitterfunktionen  $u_h$  mit der Eigenschaft

$$L_h u_h(P) \leq 0, \quad P \in \Omega_h, \quad (2.2.26)$$

haben in  $\Omega_h$  kein positives Maximum. Genauer gilt  $u_h \leq 0$  auf  $\bar{\Omega}_h$  oder

$$\max_{\bar{\Omega}_h} u_h \leq \max_{\partial\Omega_h} u_h. \quad (2.2.27)$$

**Beweis:** Wir nehmen an, dass (2.2.26) für ein  $u_h$  erfüllt ist, und führen den Beweis indirekt. Es gebe also einen Punkt  $P_0 \in \Omega_h$ , so dass

$$M := u_h(P_0) = \max_{\bar{\Omega}_h} u_h > 0, \quad \max_{\partial\Omega_h} u_h < M.$$

Ausgehend von  $L_h u_h(P) \leq 0$ , folgt mit Bedingung (B2):

$$\begin{aligned} u_h(P_0) &\leq - \sum_{Q \neq P_0} \frac{\sigma(P_0, Q)}{\sigma(P_0, P_0)} u_h(Q) = \sum_{Q \neq P_0} \left| \frac{\sigma(P_0, Q)}{\sigma(P_0, P_0)} \right| u_h(Q) \\ &= \sum_{Q \neq P_0} \left| \frac{\sigma(P_0, Q)}{\sigma(P_0, P_0)} \right| u_h(P_0) + \sum_{Q \neq P_0} \left| \frac{\sigma(P_0, Q)}{\sigma(P_0, P_0)} \right| \{u_h(Q) - u_h(P_0)\}. \end{aligned}$$

Weiter gilt wegen Bedingung (B1i)

$$\sum_{Q \neq P_0} \left| \frac{\sigma(P_0, Q)}{\sigma(P_0, P_0)} \right| \leq 1,$$

und folglich ( $u_h(P_0) = M$ )

$$M \leq M + \sum_{Q \in N(P_0) \setminus \{P_0\}} \left| \frac{\sigma(P_0, Q)}{\sigma(P_0, P_0)} \right| \{u_h(Q) - M\}.$$

Da nach Voraussetzung  $u_h(Q) \leq M$  ist, muss also  $u_h(Q) = M$  in allen Punkten  $Q \in N(P_0)$  (d. h. solchen mit  $\sigma(P_0, Q) \neq 0$ ) sein, da sonst ein Widerspruch entsteht. Anwendung derselben Schlussweise für alle Punkte in  $N(P_0)$  liefert

$$u_h(Q) = M, \quad Q \in \bigcup_{P \in N(P_0)} N(P).$$

Mit Hilfe der Bedingung (B3) erschließen wir dann durch sukzessive Fortsetzung dieses Arguments, dass  $u_h \equiv M$  auf  $\Omega_h$ . Nach Voraussetzung ist  $\partial\Omega_h \neq \emptyset$ , so dass es wegen (B3) einen Punkt  $Q \in \partial\Omega_h \cap N(P)$  mit einem „inneren“ Punkt  $P \in \Omega_h$  geben muss. Für diesen folgt dann  $u_h(Q) = M$ , was im Widerspruch zur Annahme  $\max_{\partial\Omega_h} u_h < M$  steht. Q.E.D.

Das diskrete Maximumprinzip hat eine Reihe von wichtigen Folgerungen für die obigen Differenzenschemata.

**Korollar 2.1 (Eindeutigkeit):** *Unter den Voraussetzungen (B1), (B2) und (B3) besitzt das Differenzenschema (2.2.18), (2.2.19) genau eine Lösung  $u_h$ . Im Falle  $f_h \geq 0$  folgt aus  $g_h \geq 0$  auch  $u_h \geq 0$ .*

**Beweis:** i) Wegen der Äquivalenz von (2.2.18), (2.2.19) zu dem Gleichungssystem  $A_h U = F_h$  genügt es, die Eindeutigkeit zu zeigen. Seien also  $u_h^{(1)}$  und  $u_h^{(2)}$  zwei Lösungen:

$$L_h u_h^{(i)} = f_h \quad \text{in } \Omega_h, \quad u_h^{(i)} = g_h \quad \text{auf } \partial\Omega_h.$$

Für die Differenz  $w_h := u_h^{(1)} - u_h^{(2)}$  gilt dann

$$L_h w_h = 0 \quad \text{in } \Omega_h, \quad w_h = 0 \quad \text{auf } \partial\Omega_h.$$

Nun wird das diskrete Maximumprinzip auf  $L_h w_h \leq 0$  sowie auf  $L_h(-w_h) \leq 0$  angewendet. Ersteres ergibt  $w_h \leq 0$  und letzteres  $w_h \geq 0$  und folglich  $w_h \equiv 0$ .

ii) Sei nun  $f_h \geq 0$  und  $g_h \geq 0$ . Dann impliziert das diskrete Maximumprinzip angewendet für  $-u_h$ , dass entweder  $u_h \geq 0$  oder  $\min_{\Omega_h} u_h = -\max_{\Omega_h} (-u_h) \geq -\max_{\partial\Omega_h} (-g_h) = \min_{\partial\Omega_h} g_h$ , was zu beweisen war. Q.E.D.

Im nächsten Abschnitt werden wir das diskrete Maximumprinzip verwenden, um die stetige Abhängigkeit der Lösungen von den Problemdaten zu zeigen. Dies wird sich als Nebenprodukt einer sehr viel stärkeren Stabilitätsungleichung ergeben.

**Korollar 2.2 (Inverse Monotonie):** Unter den Voraussetzungen (B1), (B2) und (B3) ist die zum Differenzenoperator  $L_h$  bei beliebiger Numerierung der Gitterpunkte gehörende Koeffizientenmatrix  $A$  eine sog. „M-Matrix“ (inverse-monotone Matrix), d. h.: Ihre Inverse  $A^{-1}$  ist elementweise nicht negativ:

$$A^{-1} \geq 0. \quad (2.2.28)$$

**Beweis:** Die Matrix  $A$  wird wie üblich geschrieben als Summe  $A = L + D + R$  einer linken unteren Dreiecksmatrix  $L$ , einer Diagonalmatrix  $D$  und einer rechten oberen Dreiecksmatrix  $R$ . Die Bedingungen (B1), (B2) und (B3) implizieren, wie oben bereits bemerkt, die (irreduzible) Diagonaldominanz von  $A$  und damit 1) die Regularität von  $A$  und  $D$  und 2) die Kontraktivität  $\text{spr}(J) < 1$  der Jacobi-Matrix  $J := -D^{-1}(L + R)$ . Dann existiert eine (natürliche) Matrizennorm  $\|\cdot\|$ , bzgl. derer  $\|J\| < 1$  ist. Hiermit folgt die Existenz der Reihe (im Sinne der Matrizenkonvergenz)

$$\sum_{k=0}^{\infty} J^k = (I - J)^{-1}.$$

Bedingung (B2) garantiert, dass (elementweise)  $J = -D^{-1}(L + R) \geq 0$  und folglich auch

$$A^{-1}D = (D^{-1}A)^{-1} = (I + D^{-1}(L + R))^{-1} = (I - J)^{-1} = \sum_{k=0}^{\infty} J^k \geq 0.$$

Wegen  $D^{-1} \geq 0$  impliziert dies  $A^{-1} \geq 0$ .

Q.E.D.

Die Matrix  $A^{-1}$  ordnet der rechten Seite  $f_h$  und den Randwerten  $g_h$  eindeutig den Lösungsvektor  $U$  der Differenzengleichung zu:

$$A^{-1} : \{f_h, g_h\} \rightarrow U = A^{-1}F(f_h, g_h).$$

Damit ist  $A^{-1}$  das algebraische Äquivalent der kontinuierlichen Greenschen Funktion  $G(\cdot, \cdot)$ :

$$G : \{f, g\} \rightarrow u(x) = \int_{\Omega} G(x, y)f(y) dy - \int_{\partial\Omega} \partial_n G(x, y)g(y) dy.$$

Als Folgerung des Maximumprinzips ist  $G(x, y) \geq 0$ ,  $x \neq y$ , was analog zur gerade gezeigten Eigenschaft  $A^{-1} \geq 0$  ist.

### 2.2.1 Das Konvergenzverhalten von Differenzenverfahren

Die Grundlage der Fehleranalyse für die oben eingeführten Differenzenschemata ist wie im Fall von Differenzenverfahren für gewöhnliche Differentialgleichungen eine asymptotische Stabilitätsabschätzung. Wir formulieren diese im folgenden Satz für ein allgemeines Differenzenschema

$$L_h u_h = f_h \quad \text{in } \Omega_h, \quad u_h = g_h \quad \text{auf } \partial\Omega_h. \quad (2.2.29)$$

Wir setzen im folgenden stets voraus, dass dieses Schema konsistent mit der RWA ist.

**Satz 2.2 (Stabilität):** Das Differenzenschema (2.2.29) sei konsistent und genüge den Bedingungen (B1), (B2) und (B3). Dann gilt für jede Gitterfunktion  $u_h$  die Stabilitätsabschätzung

$$\max_{P \in \Omega_h} |u_h(P)| \leq \frac{1}{4} d_\Omega^2 \max_{\Omega_h} |L_h u_h(P)| + \max_{P \in \partial\Omega_h} |u_h(P)|, \quad (2.2.30)$$

wobei  $d_\Omega := \text{diam}(\Omega)$ .

**Beweis:** Wir argumentieren im folgenden ähnlich wie auf der kontinuierlichen Ebene. Durch die folgende Vorschrift für beliebiges, festes  $Q \in \bar{\Omega}_h$

$$L_h G_h(\cdot, Q) = h^{-2} \delta(\cdot, Q) \quad \text{in } \Omega_h, \quad G_h(\cdot, Q) = \delta(\cdot, Q) \quad \text{auf } \partial\Omega_h, \quad (2.2.31)$$

wird ein diskretes Analogon  $G_h(P, Q) : \bar{\Omega}_h \times \bar{\Omega}_h \rightarrow \mathbb{R}$  zur kontinuierlichen Greenschen Funktion (des Laplace-Operators) definiert. Dabei ist  $\delta(P, Q)$  das übliche „Kronecker<sup>3</sup>-Symbol“. Aus dem diskreten Maximumsprinzip folgt, dass  $G_h(P, Q) \geq 0$ . Für Gitterfunktionen  $v_h$  gilt dann die diskrete „Greensche Identität“

$$v_h(P) = h^2 \sum_{Q \in \Omega_h} G_h(P, Q) L_h v_h(Q) + \sum_{Q \in \partial\Omega_h} G_h(P, Q) v_h(Q), \quad P \in \bar{\Omega}_h. \quad (2.2.32)$$

Um dies zu sehen, bezeichnen wir die rechte Seite von (2.2.32) mit  $w_h$  und erhalten mit den Eigenschaften der Greenschen Funktion

$$L_h w_h = L_h v_h \quad \text{in } \Omega_h, \quad w_h = v_h \quad \text{auf } \partial\Omega_h.$$

Die Eindeutigkeit der Lösungen des Differenzenschemas impliziert dann  $w_h \equiv v_h$ . Aus der diskreten Greenschen Identität folgt für jede Gitterfunktion  $v_h$

$$|v_h(P)| \leq h^2 \sum_{Q \in \Omega_h} G_h(P, Q) \max_{\Omega_h} |L_h v_h| + \sum_{Q \in \partial\Omega_h} G_h(P, Q) \max_{\partial\Omega_h} |v_h|, \quad P \in \bar{\Omega}_h. \quad (2.2.33)$$

i) Wegen der Konsistenz des Schemas gilt für die Gitterfunktion  $w_h \equiv 1$  notwendig  $L_h w_h \equiv 0$ , so dass aus der Greenschen Identität folgt:

$$1 = h^2 \sum_{Q \in \Omega_h} G_h(P, Q) L_h w_h(Q) + \sum_{Q \in \partial\Omega_h} G_h(P, Q) w_h(Q) = \sum_{Q \in \partial\Omega_h} G_h(P, Q). \quad (2.2.34)$$

Dies impliziert eine Schranke für die zweite Summe in (2.2.33).

<sup>3</sup>Leopold Kronecker (1823–1891): Deutscher Mathematiker; wirkte in Berlin als „Privatgelehrter“; betrieb die Arithmetisierung der Mathematik; wichtiger Vertreter des „Konstruktivismus“, welcher die generelle Verwendung des Widerspruchsbeweises und des „aktual Unendlichen“ in Form z. B. der allgemeinen reellen Zahlen ablehnt.

ii) Jeder Punkt  $P_0 \in \bar{\Omega}_h$  ist Mittelpunkt eines  $\bar{\Omega}$  enthaltenden Kreises mit Radius  $d_\Omega$ . O.b.d.A. sei hier  $P_0 = 0$ . Für die Funktion  $w(P) := \frac{1}{4}|P|^2$  gilt in Punkten  $P \in \Omega_h$ :

$$L_h w(P) = L_h w(P) + \Delta w(P) - \Delta w(P) = M_3(w)O(h) - \Delta w(P) = -1,$$

da  $M_3(w) = 0$ . Wir definieren nun die Gitterfunktion

$$v_h(P) := h^2 \sum_{Q \in \Omega_h} G_h(P, Q).$$

Durch elementares Nachrechnen verifiziert man, dass

$$\begin{aligned} L_h v_h &= 1 \quad \text{in } \Omega_h, \quad v_h = 0 \quad \text{auf } \partial\Omega_h, \\ L_h(v_h + w) &= 0 \quad \text{in } \Omega_h, \quad v_h + w \leq \frac{1}{4}d_\Omega^2 \quad \text{auf } \partial\Omega_h. \end{aligned}$$

Aus dem diskreten Maximumprinzip folgt dann wegen  $w \geq 0$  notwendig

$$\max_{\bar{\Omega}_h} v_h \leq \max_{\bar{\Omega}_h} (v_h + w) \leq \max_{\partial\Omega_h} (v_h + w) = \max_{\partial\Omega_h} w \leq \frac{1}{4}d_\Omega^2,$$

und somit

$$h^2 \sum_{Q \in \Omega_h} G_h(P, Q) \leq \frac{1}{4}d_\Omega^2, \quad P \in \bar{\Omega}_h. \quad (2.2.35)$$

Die Abschätzungen (2.2.34) und (2.2.35) ergeben zusammen mit (2.2.33) die Behauptung.  
Q.E.D.

Als unmittelbare Folgerung aus Satz 2.2 erhalten wir (in Analogie zum kontinuierlichen Fall) die stetige Abhängigkeit der Lösung von den Daten. Das wichtigste Resultat ist eine a priori Konvergenzabschätzung für das Differenzenschema (2.2.29).

**Korollar 2.3 (Allgemeines Konvergenzresultat):** *Unter den Voraussetzungen von Satz 2.2 gilt für das Differenzenschema (2.2.29) die a priori Konvergenzabschätzung*

$$\max_{P \in \bar{\Omega}_h} |e_h(P)| \leq \frac{1}{4}d_\Omega^2 \max_{P \in \Omega_h} |\tau_h(P)| + \max_{P \in \partial\Omega_h} |e_h(P)|, \quad (2.2.36)$$

mit dem Fehler  $e_h = u - u_h$  und dem Abschneidefehler  $\tau_h(P) := L_h u(P) + \Delta u(P)$ .

**Beweis:** Für den Fehler  $e_h$  gilt die Differenzengleichung

$$L_h e_h(P) = \tau_h(P) \quad P \in \Omega_h,$$

so dass die Stabilitätsungleichung (2.2.30) unmittelbar die Behauptung liefert. Q.E.D.

Die Abschätzung (2.2.36) garantiert wieder, dass für ein „gutes“ Differenzenschema der globale Diskretisierungsfehler mit derselben Ordnung wie der lokale Abschneidefehler konvergiert. Wir wollen diese allgemeinen Resultate für die obigen speziellen Diskretisierungen

anwenden. Dabei wird wieder die folgende Bezeichnung für Normschränken der exakten Lösung verwendet:

$$M_m(u) := \max_{\Omega} \{ |\partial_x^i \partial_y^j u|, i + j = m \}.$$

**Korollar 2.4 (Konvergenz des 5-Punkte-Operators):** *Unter den Voraussetzungen von Satz 2.2 gilt für den Fehler den 5-Punkte-Operator mit konstanter Randwertextrapolation die a priori Fehlerabschätzung*

$$\max_{P \in \bar{\Omega}_h} |e_h(P)| \leq \frac{1}{24} d_\Omega^2 M_4(u) h^2 + M_1(u) h. \quad (2.2.37)$$

**Beweis:** Für den Abschneidefehler gilt

$$\max_{\Omega_h} |\tau_h| \leq \frac{1}{6} M_4(u) h^2,$$

und in Randgitterpunkten

$$\max_{\partial\Omega_h} |e_h| = \max_{\partial\Omega_h} |u(P) - u(P^*)| \leq M_1(u) h.$$

Die Stabilitätsabschätzung (2.2.30) impliziert damit die Behauptung.

Q.E.D.

Der ordnungsreduzierende Effekt der mangelhaften Randwertapproximation kann durch die oben beschriebene „lineare Randwertinterpolation“ behoben werden. Für das so modifizierte 5-Punkte-Schema erhält man ähnlich wie eben die a priori Fehlerabschätzung

$$\max_{P \in \bar{\Omega}_h} |e_h(P)| \leq \frac{1}{12} d_\Omega^2 M_4(u) h^2 + \frac{1}{2} M_2(u) h^2. \quad (2.2.38)$$

**Korollar 2.5 (Konvergenz des Shortley-Weller-Operators):** *Unter den Voraussetzungen von Satz 2.2 gilt für den modifizierten 5-Punkte-Operator nach Shortley-Weller die a priori Fehlerabschätzung*

$$\max_{P \in \bar{\Omega}_h} |e_h(P)| \leq \frac{1}{24} d_\Omega^2 M_4(u) h^2 + \frac{1}{3} M_3(u) h^3. \quad (2.2.39)$$

**Beweis:** Für den Fehler  $e_h$  gelten die Beziehungen

$$-\Delta_h e_h = \tau_h \text{ in } \Omega_h, \quad -\Delta_h^* e_h = \tau_h^* \text{ auf } \partial\Omega_h^*, \quad e_h = 0 \text{ auf } \partial\Omega_h.$$

Für die Abschneidefehler gilt

$$\max_{\Omega_h} |\tau_h| \leq \frac{1}{6} M_4(u) h^2, \quad \max_{\partial\Omega_h^*} |\tau_h^*| \leq \frac{2}{3} M_3(u) h.$$

In diesem Fall können wir nicht direkt die Stabilitätsungleichung (2.2.30) anwenden, da sie nur auf eine reduzierte Konvergenzordnung  $O(h)$  führen würde. Statt dessen modifizieren wir in geeigneter Weise den Beweis dieser Abschätzung. Mit den Bezeichnungen

des Beweises von Satz 2.2 ergibt die Greensche Identität ( $e_h = 0$  auf  $\partial\Omega_h$ )

$$e_h(P) = h^2 \sum_{Q \in \Omega_h^0} G_h(P, Q) \tau_h(Q) + h^2 \sum_{Q \in \partial\Omega_h^*} G_h(P, Q) \tau_h^*(Q)$$

und folglich

$$|e_h(P)| \leq h^2 \sum_{Q \in \Omega_h^0} G_h(P, Q) \max_{\Omega_h^0} |\tau_h| + h^2 \sum_{Q \in \partial\Omega_h^*} G_h(P, Q) \max_{\partial\Omega_h^*} |\tau_h^*|. \quad (2.2.40)$$

Wie im Beweis von Satz 2.2 folgt

$$h^2 \sum_{Q \in \Omega_h^0} G_h(P, Q) \leq h^2 \sum_{Q \in \Omega_h} G_h(P, Q) \leq \frac{1}{4} d_\Omega^2. \quad (2.2.41)$$

Wir wollen jetzt zeigen, dass

$$\sum_{Q \in \partial\Omega_h^*} G_h(P, Q) \leq \frac{1}{2}. \quad (2.2.42)$$

Dazu definieren wir die Gitterfunktion  $w_h$  durch

$$w_h = 1 \text{ in } \Omega_h, \quad w_h = 0 \text{ auf } \partial\Omega_h.$$

Für diese verifiziert man durch Nachrechnen

$$-\Delta_h w_h = 0 \text{ in } \Omega_h^0, \quad -\Delta_h^* w_h \geq 2h^{-2} \text{ auf } \partial\Omega_h^*.$$

Mit Hilfe der Greenschen Identität folgt damit

$$1 = h^2 \sum_{Q \in \partial\Omega_h^*} G_h(P, Q) (-\Delta_h^*) w_h(Q) \geq 2 \sum_{Q \in \partial\Omega_h^*} G_h(P, Q),$$

was (2.2.42) impliziert. Dies vervollständigt den Beweis. Q.E.D.

**Bemerkung:** In drei Raumdimensionen gelten Satz 2.2 und Korollar 2.3 mit der Konstante  $\frac{1}{6}d_\Omega^2$ . Hiermit und der zugehörigen Abschätzung (2.1.17) für den Konsistenzfehler ergeben sich dann auch die a priori Fehlerabschätzungen der Korollare 2.4 und 2.4 mit der führenden Konstante  $\frac{1}{24}d_\Omega^2$ .

**Bemerkung:** Wir betonen, dass die Konvergenz der betrachteten Differenzenschemata eine vergleichsweise hohe Regularität der zu approximierenden Lösung erfordert. Das Shortley-Weller-Schema erfordert z. B. für seine „maximale“ Konvergenzordnung  $m = 2$ , dass  $u$  beschränkte vierte Ableitungen auf ganz  $\bar{\Omega}$  besitzt ( $M_4(u) < \infty$ ). Dies ist eine sehr einschneidende Forderung, wie wir im vorigen Kapitel gesehen haben. Sie kann i. Allg. nur für glatt berandete Gebiete sowie unter gewissen Zusatzbedingungen für spezielle Geometrien wie z. B. Rechtecke garantiert werden. Weiter erfordert sie auch hohe Glattheit der Daten  $f$  und  $g$ . Auf Gebieten mit *einspringenden* Ecken oder sonstwie

reduzierter Regularität von  $u$  können diese Sätze nicht verwendet werden. In solchen realitätsnäheren Situationen werden sich die im nächsten Abschnitt behandelten „Finite-Elemente-Verfahren“ als flexibler erweisen.

## 2.3 Lösungsaspekte

Wir diskutieren nun die Lösung der durch eine Differenzendiskretisierung entstehenden algebraischen Gleichungssysteme. Zugrunde gelegt wird wieder die Situation des Modellproblems der 1. RWA des Laplace-Operators

$$Lu := -\Delta u = f \quad \text{in } \Omega, \quad u = g \quad \text{auf } \partial\Omega, \quad (2.3.43)$$

auf einem Gebiet  $\Omega \subset \mathbf{R}^2$ . Erweiterungen für Probleme mit variablen Koeffizienten, anderen Randbedingungen, Unsymmetrien sowie auf drei Raumdimensionen werden wieder in Bemerkungen berücksichtigt. Die zugehörigen algebraischen Systeme haben die Gestalt

$$Ax = b, \quad (2.3.44)$$

Die Matrix  $A = (a_{nm})_{n,m=1}^N$  und der Vektor  $b = (b_n)_{n=1}^N$  haben die Dimension  $N := \text{Anzahl der inneren Gitterpunkte bzw. Knotenfreiheitsgrade}$ . In der Praxis ist meist  $N \gg 1000$ , so dass neben dem Rechenaufwand auch der Speicherbedarf ein wichtiger Aspekt ist. Die Matrix ist extrem dünn besetzt und besitzt abhängig von der gewählten Numerierung der Gitterpunkte bzw. Knoten eine Bandstruktur. Für die Wahl eines geeigneten (d. h.: möglichst sparsamen) Lösungsverfahrens ist die zu erwartende Dimension  $N$  entscheidend.

Wir orientieren die folgende Diskussion an der Modellsituation der Poisson-Gleichung auf dem Einheitsquadrat  $\Omega = (0, 1)^2$  und der Diskretisierung mit dem üblichen 5-Punkteschema auf einem äquidistanten, kartesischen Gitter  $\bar{\Omega}_h = \Omega_h \cup \partial\Omega_h$  der Gitterweite  $h$ . Das Gebiet  $\Omega = (0, 1)^2$  hat den Durchmesser  $d_\Omega = \sqrt{2}$ . Wir wählen die Funktion  $u(x, y) = \sin(\pi x) \sin(\pi y)$  als Lösung der Randwertaufgabe

$$-\Delta u = 2\pi^2 u =: f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega. \quad (2.3.45)$$

Es ist  $M_4(u) = \pi^4$ . Die a priori Fehlerschranke (2.2.39) für die Shortley-Weller-Approximation liefert damit die (pessimistische) Fehlerschätzung

$$\max_{\bar{\Omega}_h} |u - u_h| \approx \frac{1}{24} d_\Omega^2 \pi^4 h^2 \approx 8h^2. \quad (2.3.46)$$

Zur Erreichung garantierter 3-stelliger (relativer) Genauigkeit,  $\text{TOL} = 10^{-3}$ , ist in diesem Fall also eine Gitterweite  $h \approx 10^{-2}$  erforderlich. Dies führt auf ein (symmetrisches) Gleichungssystem der Dimension  $N \approx 10^4$ .

**Bemerkung 2.1:** Die Fehlerabschätzung (2.3.46) ist in der Tat sehr pessimistisch. Der wirkliche Fehler ist ungefähr um einen Faktor  $10^{-1}$  kleiner. Dies zeigt, dass unsere a

*a priori* Fehleranalyse selbst für ein so einfaches Modellproblem zu unscharf und für praktische Zwecke nur bedingt brauchbar ist.

**Bemerkung 2.2:** In drei Raumdimensionen ist  $d_\Omega = \sqrt{3}$ , und die a priori Fehlerabschätzung (2.3.46) liefert den Wert  $\approx 12h^2$ , so dass hier mindestens auch  $h = 10^{-2}$  und damit  $N \approx 10^6$  erforderlich wäre.

Die Struktur der Matrix  $A$  hängt von der gewählten Nummerierung der Gitterpunkte ab. Die gängigen Alternativen sind:

1) *Zeilenweise Nummerierung*: Die lexikographische Anordnung der Gitterpunkte,

$$(x_i, y_j) \leq (x_p, y_q) \quad \text{wenn} \quad j \leq q \quad \text{oder} \quad j = q, i \leq p$$

führt auf eine Bandmatrix mit Bandbreite  $2m + 1 \approx h^{-1}$ .

21	22	23	24	25	
16	17	18	19	20	
11	12	13	14	15	
6	7	8	9	10	
1	2	3	4	5	

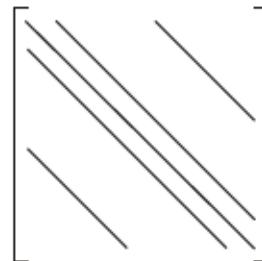


Abbildung 2.6: Lexikographische Nummerierung

2) *Diagonale Nummerierung*: Die sukzessive Numerierung diagonal zu den Koordinatenrichtungen führt auf eine Bandmatrix mit geringem „Bandinhalt“ ( $\Rightarrow$  Speicherersparnis beim Gaußschen Eliminationsverfahren).

11	16	20	23	25	
7	12	17	21	24	
4	8	13	18	22	
2	5	9	14	19	
1	3	6	10	15	

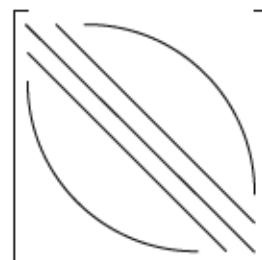


Abbildung 2.7: Diagonale Nummerierung

3) *Schachbrett-Nummerierung:* Die versetzte zeilenweise- sowie spaltenweise Nummerierung führt auf eine  $2 \times 2$ -Blockmatrix mit diagonalen Hauptblöcken.

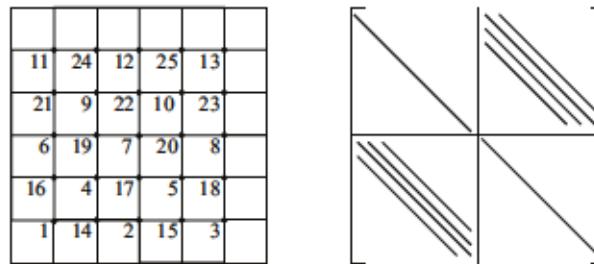


Abbildung 2.8: Schachbrett-Nummerierung

a) **Direkte Lösung:** Prinzipiell wäre das klassische Gaußsche Eliminationsverfahren zur Lösung des Systems (2.3.44) geeignet. Zu seiner Durchführung benötigt man bei zeilenweiser Nummerierung und Ausnutzung der Symmetrie etwa  $mN$  Speicherplätze (im Folgenden abgekürzt als „SP“) und  $m^2N$  arithmetische Operationen (im Folgenden abgekürzt als „OP“). Im Falle  $N \approx 10^4$  sind dies etwa  $10^6$  SP und  $10^8$  OP. Die kurze Überschlagsrechnung zeigt, dass direkte Lösungsmethoden für das Gleichungssystem (2.3.44) nur in speziellen Situationen in Frage kommen. Sie spielen eine Rolle bei geringer Problemgröße ( $N \leq 10^3$ ) oder im Falle sehr uniformer Matrixeinträge (bei der Approximation von Differentialoperatoren mit konstanten Koeffizienten) ( $N \leq 10^5$ ). Speziell für die 5-Punkte-Approximation des Laplace-Operators auf Rechtecken gibt es sehr effiziente direkte Lösungstechniken auf Grundlage der sog. „schnellen Fourier-Transformation (FFT)“ ( $N \leq 10^7$ ). Diese sind aber bei allgemeineren Problemstellungen meist nicht anwendbar und werden daher hier nicht weiter diskutiert.

**Bemerkung 2.3:** Zur Illustration machen wir dieselbe Überschlagsrechnung auch für die entsprechende dreidimensionale Situation  $\Omega = (0, 1)^3$ . Hier hat die Matrix  $A$  die Dimension  $N = m^3 \approx 10^6$  und die Bandbreite  $2m^2 + 1 \approx 10^4$ . Folglich betrüge der benötigte Lösungsaufwand  $\approx 10^{10}$  SP und  $\approx 10^{14}$  OP, was noch jenseits der Kapazität von Arbeitsplatzrechnern liegt.

b) **Iterative Lösung:** Wie bereits erwähnt, übertragen sich die wichtigen Eigenschaften (B1), (B2) und (B3) des Differenzenschemas unabangig von der gewählten Numerierung auf die jeweilige Matrix  $A$ . Diese ist (irreduzibel) diagonaldominant und von nichtnegativem Typ und folglich eine M-Matrix, d. h.: Es gilt elementweise  $A^{-1} \geq 0$ . Zusätzlich ist  $A$  oft auch symmetrisch und positiv definit. In diesem Falle konvergieren die meisten gängigen iterativen Verfahren. Ausgehend von einem Defektkorrekturansatz

$$d^t = F - Ax^t, \quad Cv^t = d^t, \quad x^{t+1} = x^t + v^t, \quad (2.3.47)$$

mit einer regulären Matrix  $C$  („Vorkonditionierer“) wird die Iteration zunächst als Fixpunktiteration geschrieben

$$Cx^{t+1} = Cx^t + b - Ax^t \Leftrightarrow x^{t+1} = (I - C^{-1}A)x^t + C^{-1}b. \quad (2.3.48)$$

Für den Fehler  $e^{(t)} := x^{(t)} - x$  gilt

$$\begin{aligned} e^{(t)} &= (I - C^{-1}A)x^{(t-1)} + C^{-1}b - x \\ &= (I - C^{-1}A)x^{(t-1)} + C^{-1}b - (I - C^{-1}A)x - C^{-1}b \\ &= (I - C^{-1}A)e^{(t-1)} = \dots = (I - C^{-1}A)^t e^{(0)}. \end{aligned} \quad (2.3.49)$$

Dies zeigt, dass die Iteration konvergiert, wenn die „Iterationsmatrix“  $B := I - C^{-1}A$  eine Kontraktion ist, d. h.:  $\rho(B) := \max\{|\lambda|, \lambda \text{ Eigenwert von } B\} < 1$ . Die „Konvergenzrate“ der Iteration ist dann gegeben durch

$$\rho := \sup_{x^{(0)} \in \mathbf{R}^N} \lim_{t \rightarrow \infty} \left( \frac{\|x^{(t)} - x\|}{\|x^{(0)} - x\|} \right)^{1/t} = \rho(B). \quad (2.3.50)$$

Es ist dann näherungsweise

$$\|x^{(t)} - x\| \leq \rho^t \|x^{(0)} - x\|. \quad (2.3.51)$$

Die Anzahl der zur Erreichung einer Reduktion des Anfangsfehlers um  $\varepsilon$  erforderlichen Iterationsschritte ergibt sich damit zu

$$t_\varepsilon \approx \frac{\ln(\varepsilon)}{\ln(\rho)}. \quad (2.3.52)$$

Der „Vorkonditionierer“  $C$  sollte folgende Bedingungen erfüllen:

- einfache Invertierung (mit  $O(N)$  OP und Speicherbedarf  $O(N)$  SP);
- möglichst kleines  $\rho(I - C^{-1}A)$ .

Dies sind leider gegenläufige Zielsetzungen. Die einfachsten Verfahren dieses Typs verwenden ausgehend von der natürlichen Aufspaltung  $A = L + D + R$  die Vorkonditionierer:

1. (Gedämpftes) Richardson<sup>4</sup>-Verfahren:  $0 < \theta \leq 2\lambda_{\max}(A)^{-1}$

$$C = \theta I, \quad B = I - \theta A. \quad (2.3.53)$$

---

<sup>4</sup>Lewis Fry Richardson (1881–1953): Englischer Mathematiker und Physiker; wirkte an verschiedenen Institutionen in England und Schottland; typischer „angewandter Mathematiker“; leistete Pionierbeiträge zur Modellierung und Numerik in der Wettervorhersage.

2. Jacobi<sup>5</sup>-Verfahren (Gesamtschrittverfahren):

$$C = D, \quad B = -D^{-1}(L + R). \quad (2.3.54)$$

3. Gauß-Seidel<sup>6</sup>-Verfahren (Einzelschrittverfahren):

$$C = D + L, \quad B = -(D + L)^{-1}R. \quad (2.3.55)$$

4. SOR-Verfahren („Successive Over-Relaxation“):  $\omega = \omega_{opt} \in (0, 2)$

$$C = D + \omega L, \quad B = (D + \omega L)^{-1}\{(1 - \omega)D - \omega R\}. \quad (2.3.56)$$

5. ILU-Verfahren („Incomplete LU Decomposition“):

$$C = \tilde{L}\tilde{R}, \quad B = I - \tilde{R}^{-1}\tilde{L}^{-1}A. \quad (2.3.57)$$

Für eine symmetrische, positiv definite Matrix wird das ILU- zum  $ILL^T$ -Verfahren („Incomplete Cholesky<sup>7</sup> Decomposition“). Die ILU-Zerlegung erhält man mit Hilfe des üblichen, rekursiven Prozesses zur Bestimmung der LU-Zerlegung aus der Gleichung  $LU = A$  durch Nullsetzung aller Matrixeinträge zu Indexpaaren  $\{n, m\}$  mit  $a_{nm} = 0$ :

$$\begin{aligned} n = 1, \dots, N : \quad \tilde{u}_{nm} &= a_{nm} - \sum_{i=1}^{n-1} \tilde{l}_{ni} \tilde{u}_{im} \quad (m = 1, \dots, N) \\ \tilde{l}_{nn} &= 1, \quad \tilde{l}_{in} = \tilde{u}_{nn}^{-1} \left\{ a_{in} - \sum_{j=1}^{n-1} \tilde{l}_{ij} \tilde{u}_{jn} \right\} \quad (i = n+1, \dots, N) \\ \tilde{l}_{nm} &= 0, \quad \tilde{u}_{nm} = 0, \text{ wenn } a_{nm} = 0. \end{aligned}$$

6. ADI-Verfahren („Alternating-Direction Implicit Iteration“):

$$\begin{aligned} C &= (A_x + \omega I)(A_y + \omega I), \\ B &= (A_y + \omega I)^{-1}(\omega I - A_x)(A_x + \omega I)^{-1}(\omega I - A_y). \end{aligned} \quad (2.3.58)$$

Das ADI-Verfahren ist für beliebige Wahl des Parameters  $\omega > 0$  konvergent. Wenn die Struktur der Matrix seine Anwendung zulässt, ist es bei optimaler Wahl von  $\omega$  mindestens so schnell wie das optimale SOR-Verfahren.

<sup>5</sup>Carl Gustav Jakob Jacobi (1804–1851): Deutscher Mathematiker; schon als Kind h.ehbegabt; wirkte in Königsberg und Berlin; Beiträge zu vielen Bereichen der Mathematik: zur Zahlentheorie, zu elliptischer Funktionen, zu partiellen Differentialgleichungen, zu Funktionaldeterminanten und zur theoretischen Mechanik.

<sup>6</sup>Philipp Ludwig von Seidel (1821–1896): Deutscher Mathematiker; Prof. in München; Beiträge zur Analysis (u. a. Methode der kleinsten Fehlerquadrate) sowie Himmelsmechanik und Astronomie.

<sup>7</sup>André Louis Cholesky (1857–1918): Französischer Mathematiker; Militärkarriere; Beiträge zur Numerischen Linearen Algebra.

Die Konvergenzraten dieser einfachen Iterationsverfahren verhalten sich in Abhängigkeit von der (gleichförmigen) Gitterweite wie

$$\rho = \rho(I - C^{-1}A) = 1 - O(h^r),$$

in Abhängigkeit von der Gitterweite  $h$  der Diskretisierung (bei fester Problemkonfiguration) mit einem geeigneten  $r \geq 0$ . Die Anzahl  $T$  der zur Gewinnung einer Dezimalstelle Genauigkeit erforderlichen Iterationsschritte ist also ungefähr bestimmt durch

$$\rho^T \approx 10^{-1} \quad \Rightarrow \quad T \approx -\frac{\ln(10)}{\ln(\rho)} \approx h^{-r}. \quad (2.3.59)$$

Hierzu beachte man, dass  $\ln(1 - ch^r) = -ch^r + O(h^{2r})$ . Da die Durchführung eines Iterationsschritts approximativ  $N \approx h^{-2}$  OP (in zwei Raumdimensionen) kostet, ergibt sich ein Gesamtaufwand pro Dezimalstelle an Genauigkeit von  $\approx h^{-2-r}$  OP. Der für die Durchführung dieser Iterationsverfahren benötigte Speicherplatz entspricht etwa dem zur Speicherung der wesentlichen (d. h. von Null verschiedenen) Elemente der Matrix  $A$  erforderlichen. Für das Jacobi- und Gauß-Seidel-Verfahren ist  $r = 2$  und für das (optimale) SOR-Verfahren ist  $r = 1$ . Das ILU- und das ADI-Verfahren liegen bei speziellen Konfigu. a.ionen etwa gleich auf zum SOR-Verfahren. Damit ergibt sich ein Gesamtlösungsaufwand von jeweils  $O(N^2)$  bzw.  $O(N^{3/2})$  OP für die einzelnen Verfahren. Ein wirklich „effizientes“ Verfahren sollte ein möglichst kleines  $r$  aufweisen; optimal wäre  $r = 0$ . Dies lässt sich durch den Einsatz von sog. „Multi-Level-Techniken“ („Mehrgitterverfahren“) erreichen, welche später im Zusammenhang mit den Finite-Elemente-Diskretisierungen diskutiert werden.

### 2.3.1 Aufwandsanalyse: ein Beispiel

Wir wollen das Konvergenzverhalten der bisher betrachteten Iterationsverfahren und deren sich daraus ergebende Effizienz anhand des obigen Modellproblems eingehender diskutieren. Dabei soll insbesondere die Bedeutung der Formel (2.3.59) für die Konvergenzrate illustriert werden.

Für die obige Modellsituation (5-Punkte-Differenzenoperator auf einem äquidistanten, kartesischen Gitter des Einheitsquadrats) lassen sich die Eigenwerte und zugehörigen Eigenvektoren der Systemmatrix  $A$  explizit angeben. Für  $k, l = 1, \dots, m$  ergibt sich mit der Bezeichnung  $Aw^{kl} = \lambda_{kl}w^{kl}$ :

$$\begin{aligned} \lambda_{kl} &= h^{-2}\{4 - 2(\cos(kh\pi) + \cos(lh\pi))\}, \quad k, l = 1, \dots, m, \\ w^{kl} &= (\sin(ih\pi) \ \sin(jh\pi))_{i,j=1,\dots,m} \quad (h = 1/(m+1)). \end{aligned}$$

Also ist (für  $h \ll 1$ )

$$\begin{aligned} \lambda_{\max} &= h^{-2}\{4 - 4 \cos(1-h)\pi\} \approx 8h^{-2}, \\ \lambda_{\min} &= h^{-2}\{4 - 4 \cos(h\pi)\} = h^{-2}\{4 - 4(1 - \frac{1}{2}\pi^2 h^2 + O(h^4))\} \approx 2\pi^2. \end{aligned}$$

und somit

$$\kappa := \text{cond}_2(A) \approx \frac{4}{\pi^2 h^2}. \quad (2.3.60)$$

Die Eigenwerte der Jacobi-Matrix  $J = -D^{-1}(L + R)$  sind

$$\mu_{kl} = \frac{1}{2} (\cos(kh\pi) + \cos(lh\pi)) \quad (k, l = 1, \dots, m)$$

Folglich wird

$$\rho_J = \mu_{\max} = \cos(h\pi) = 1 - \frac{\pi^2}{2} h^2 + O(h^4). \quad (2.3.61)$$

Für die Iterationsmatrizen des Gauß-Seidel- und des (optimalen) SOR-Verfahrens folgt:

$$\begin{aligned} \rho_{GS} &= \rho_{GS}^2 = 1 - \pi^2 h^2 + O(h^4), \\ \rho_{SOR} &= \frac{1 - \sqrt{1 - \rho_J^2}}{1 + \sqrt{1 - \rho_J^2}} = \frac{1 - \pi h + O(h^2)}{1 + \pi h + O(h^2)} = 1 - 2\pi h + O(h^2). \end{aligned}$$

Für die Iterationszahlen (pro Dezimalstelle Fehlerreduktion) ergibt sich also asymptotisch:

$$\begin{aligned} T_J &\approx -\frac{\ln(10)}{\ln(1 - \frac{\pi^2}{2} h^2)} \approx \frac{4,6}{\pi^2 h^2} \approx \frac{1}{2} N, \\ T_{GS} &\approx -\frac{\ln(10)}{\ln(1 - \pi^2 h^2)} \approx \frac{2,3}{\pi^2 h^2} \approx \frac{1}{4} N, \\ T_{SOR} &\approx -\frac{\ln(10)}{\ln(1 - 2\pi h)} \approx \frac{2,3}{2\pi h} \approx \frac{1}{3} \sqrt{N}. \end{aligned}$$

Der Vollständigkeit halber geben wir hier auch die entsprechenden Werte für das CG-Verfahren („Verfahren der konjugierten Richtungen“):

$$T_{CG} = \frac{1}{2} \sqrt{\kappa} \ln(20) \approx \frac{3}{\pi h} \approx \sqrt{N}.$$

Das CG-Verfahren ist zwar langsamer als das „optimale“ SOR-Verfahren, erfordert aber nicht die Bestimmung eines Iterationsparameters.

Zum Vergleich der Effizienz der Iterationsverfahren muss natürlich auch der Aufwand pro Iterationsschritt berücksichtigt werden. Für die Anzahl OP der arithmetischen Operationen pro Iterationsschritt gilt (optimistische Schätzung)

$$OP_J, OP_{GS}, OP_{SOR} \approx 6N, \quad OP_{CG} \approx 10N.$$

Als Endresultat finden wir, dass zur Bestimmung der Lösung des diskretisierten Modellproblems das Jacobi-Verfahren, das Gauß-Seidel-Verfahren und das Gradientenverfahren  $O(N^2)$  OP benötigen. Das direkte Cholesky-Verfahren würde für die Berechnung der „exakten“ Lösung (auf Rundungsfehlergenauigkeit) ebenfalls  $O(m^2N) = O(N^2)$  OP benötigen.

gen, erscheint also dem Gauß-Seidel-Verfahren überlegen zu sein. Es ist jedoch zu berücksichtigen, dass letzteres nur  $O(N)$  SP benötigt, im Gegensatz zu den  $O(mN) = O(N^{3/2})$  SP für das Cholesky-Verfahren. Die schnelleren, auf Mehrgitterkonzepten basierenden Iterationsverfahren sind dagegen dem einfachen direkten Löser asymptotisch klar überlegen.

Für das Beispiel mit  $N = 10^4$  ergibt sich überschlagsmäßig der folgende Gesamtaufwand „GA“ zur sicheren Lösung des Systems (2.3.44) unter die Diskretisierungsgenauigkeit ( $\text{TOL} = 10^{-3}$ ,  $h = 10^{-2}$ ,  $N = 10^4$ ):

$$\begin{aligned}\text{GA}_J(\text{TOL}) &\approx 4 \cdot 3N^2 \approx 1,2 \cdot 10^9 \text{ OP}, \\ \text{GA}_{GS}(\text{TOL}) &\approx 4 \cdot 1,5N^2 \approx 6 \cdot 10^8 \text{ OP}, \\ \text{GA}_{SOR}(\text{TOL}) &\approx 4 \cdot 2N^{3/2} \approx 8 \cdot 10^6 \text{ OP}, \\ \text{GA}_{CG}(\text{TOL}) &\approx 4 \cdot 10N^{3/2} \approx 4 \cdot 10^7 \text{ OP}.\end{aligned}$$

Für ein „optimales“ Verfahren wie z. B. das Mehrgitterverfahren „MG“ würde man hier wesentlich bessere Werte erwarten:  $\text{GA}_{MG}(\text{TOL}) \approx 4 \cdot 25N \approx 10^6 \text{ OP}$ .

**Bemerkung:** In drei Raumdimensionen erhalten wir näherungsweise

$$\lambda_{\max} \approx 12h^{-2}, \quad \lambda_{\min} \approx 3\pi^2, \quad \kappa \approx \frac{8}{3\pi^2 h^2},$$

und folglich im wesentlichen dieselben Abschätzungen für  $\rho_J$ ,  $\rho_{GS}$  und  $\rho_{SOR}$  sowie für die Iterationszahlen  $T_J$ ,  $T_{GS}$ ,  $T_{SOR}$  und  $T_{CG}$  wie im zweidimensionalen Fall. Der Aufwand für die Durchführung eines Iterationsschritts ist hier  $\text{OP}_J, \text{OP}_{GS}, \text{OP}_{SOR} \approx 8N$  bzw.  $\text{OP}_{CG} \approx 12N$ . Für den Gesamtlösungsaufwand ergibt dies dann:

$$\begin{aligned}\text{GA}_J(\text{TOL}) &\approx 4 \cdot 4N^2 \approx 1,6 \cdot 10^{13} \text{ OP}, \\ \text{GA}_{GS}(\text{TOL}) &\approx 4 \cdot 2N^2 \approx 8 \cdot 10^{12} \text{ OP}, \\ \text{GA}_{SOR}(\text{TOL}) &\approx 4 \cdot 3N^{3/2} \approx 1,2 \cdot 10^{10} \text{ OP}, \\ \text{GA}_{CG}(\text{TOL}) &\approx 4 \cdot 12N^{3/2} \approx 4,8 \cdot 10^{10} \text{ OP}.\end{aligned}$$

Der Aufwand des Mehrgitterverfahrens erhöht sich aber nur vergleichsweise unwesentlich auf  $\text{GA}_{MG}(\varepsilon) \approx 4 \cdot 50N \approx 2 \cdot 10^8 \text{ OP}$ . Zur Bewertung dieser Komplexitätsschätzungen muss man sich die Leistung verfügbarer Rechner vergegenwärtigen. Ein Arbeitsplatzrechner leiste real etwa 200 MFlops (200 Millionen „Floating-Point“ Operationen pro Sekunde). Das SOR-Verfahren braucht dann auf einem solchen Rechner zur Lösung des Gleichungssystems (auf Diskretisierungsfehlergenauigkeit) etwa 1,5 Minuten, wogegen das Mehrgitterverfahren „nur“ 1 Sekunde benötigt.

## 2.4 Übungen

**Übung 2.1:** Man betrachte die Diskretisierung der 1. RWA des Laplace-Operators auf dem Einheitsquadrat des  $\mathbb{R}^2$  mit dem 9-Punkte-Differenzenschema (sog. „Mehrstellenfor-

mel“)

$$-\Delta_h^{(9)} u_h(x, y) = f(x, y), \quad (x, y) \in \Omega_h,$$

mit dem „gestreckten“ Differenzenoperator

$$\Delta_h^{(9)} u(x, y) := \frac{1}{12h^2} \left\{ -u(x \pm 2h, y) + 16u(x \pm h, y) - u(x, y \pm 2h) + 16u(x, y \pm h) - 60u(x, y) \right\}$$

und dem 9-Punkte-Differenzenschema

$$-\Delta_h^{(9)} u_h(x, y) = f(x, y) + \frac{1}{12} h^2 \Delta f(x, y), \quad (x, y) \in \Omega_h$$

mit dem „kompakten“ Differenzenoperator

$$\bar{\Delta}_h^{(9)} u(x, y) = \frac{1}{6h^2} \left\{ 4u(x \pm h, y) + 4u(x, y \pm h) + u(x \pm h, y \pm h) - 20u(x, y) \right\}.$$

Man zeige, dass dies Approximationen mit der Konsistenzordnung  $m = 4$  sind.

**Übung 2.2:** In vielen Fällen kann die asymptotische Konvergenzordnung eines Differenzenverfahrens nur experimentell bestimmt werden. Dazu werden bei bekannter exakter Lösung für zwei Schrittweiten  $h$  und  $h/2$  die Fehler  $e_h := u - u_h$  und  $e_{h/2} := u - u_{h/2}$  berechnet und dann die Ordnung  $\alpha$  über den formalen Ansatz  $\|u - u_h\|_h = h^\alpha$  mit einer geeigneten Gitternorm  $\|\cdot\|_h$  aus der folgenden Formel ermittelt:

$$\alpha = \frac{\log(\|e_h\|_h / \|e_{h/2}\|_h)}{\log(2)}.$$

a) Man rechtfertige diese Formel und überlege, wie man vorgehen kann, wenn keine exakte Lösung  $u$  bekannt ist.

b) Man bestimme die inhärente Konvergenzordnungen der folgenden Zahlenfolgen:

$h = 2^{-1}$	33.627	26.570
$h = 2^{-2}$	30.318	27.008
$h = 2^{-3}$	29.100	27.883
$h = 2^{-4}$	28.586	28.072
$h = 2^{-5}$	28.351	28.117

**Übung 2.3:** Man betrachte die Diskretisierung der 1. RWA des Laplace-Operators auf dem Einheitsquadrat des  $\mathbb{R}^2$  mit dem 9-Punkte-Differenzenschema

$$\Delta_h^{(9)} u_h(x, y) = f_h(x, y) := f(x, y) + \frac{1}{12} h^2 \Delta f(x, y), \quad (x, y) \in \Omega_h,$$

mit dem „kompakten“ Differenzenoperator

$$\Delta_h^{(9)} u(x, y) = \frac{1}{6h^2} \left\{ 4u(x \pm h, y) + 4u(x, y \pm h) + u(x \pm h, y \pm h) - 20u(x, y) \right\}.$$

Diese Approximationen hat nach Aufgabe 2.1 die Konsistenzordnung  $m = 4$ .

a) Man zeige mit den Mitteln der Vorlesung die Fehlerabschätzung

$$\max_{P \in \bar{\Omega}_h} |u(P) - u_h(P)| \leq c M_6(u) h^4.$$

b) (Zusatzaufgabe für Leser mit Ergeiz und Zeit) Im Falle eines allgemeinen, glattberandeten Gebiets  $\Omega \subset \mathbb{R}^2$  werde entlang der gekrümmten Randabschnitte die Shortley-Weller-Approximation betrachtet:

$$-\Delta_h^{(9)} u_h = f + \frac{1}{12} h^2 \Delta f \text{ in } \Omega_h, \quad -\Delta_h^* u_h = f \text{ in } \Omega_h^*, \quad u_h = g \text{ auf } \partial\Omega_h.$$

Man zeige hierfür mit den Mitteln des Textes die Fehlerabschätzung

$$\max_{P \in \bar{\Omega}_h} |u(P) - u_h(P)| \leq c \{M_6(u)h^4 + M_3(u)h^3\}.$$

**Übung 2.4:** Sei  $A_h$  die zum 5-Punkte-Operator auf dem Einheitsquadrat gehörende  $N \times N$ -Matrix (bei zeilenweiser Nummerierung der Gitterpunkte mit  $m$  Punkten in jeder Zeile). Die  $N = m^2$  Eigenvektoren  $w^{\nu\mu}$ ,  $\nu, \mu = 1, \dots, m$ , und die zugehörigen Eigenwerte  $\lambda_{\nu\mu}$  von  $A_h$  sind gegeben durch:

$$w^{\nu\mu}(x, y) = \sin(\nu\pi x) \sin(\mu\pi y), \quad (x, y) \in \Omega_h, \quad \lambda_{\nu\mu} = \frac{1}{h^2} (4 - 2(\cos(\nu h\pi) + \cos(\mu h\pi))).$$

Man zeige, dass für die Spektralkondition von  $A_h$  gilt:

$$\text{cond}_2(A_h) =: \frac{\lambda_{\max}(A_h)}{\lambda_{\min}(A_h)} = \frac{4}{\pi^2 h^2} + \mathcal{O}(1).$$

**Übung 2.5:** Eine Matrix  $A \in \mathbb{R}^{N \times N}$  heißt „M-Matrix“, wenn sie von nichtnegativem Typ und regulär ist und wenn ihre Inverse  $A^{-1} = (a_{ij}^{(-1)})_{i,j=1}^N$  elementweise nichtnegativ ist:  $a_{ij}^{(-1)} \geq 0$ . Das 5-Punkte-Differenzenschema (bzw. das Shortley-Weller-Schema) zur Approximation der 1. RWA des Laplace-Operators führt z. B. auf eine solche M-Matrix.

a) Man zeige, dass M-Matrizen „invers-monoton“ sind, d. h.: Für Vektoren  $v, w \in \mathbb{R}^N$  gilt komponentenweise:

$$Av \geq Aw \quad \Rightarrow \quad v \geq w.$$

b) Ist ferner  $A_h w \geq (1, \dots, 1)^T$  für einen Vektor  $w \in \mathbb{R}^N$ , so folgt bzgl. der Maximumsnorm bzw. Maximalen-Zeilensummen-Norm:

$$\|A_h^{-1}\|_\infty \leq \|w\|_\infty.$$

c) Man zeige mit Hilfe von (b), dass für die Systemmatrix  $A_h$  des 5-Punkte-Schemas auf dem Einheitsquadrat die Abschätzung

$$\|A_h^{-1}\|_\infty \leq 1/8$$

gilt, und folgere hiermit für die  $l_\infty$ -Kondition von  $A_h$ :

$$\text{cond}_\infty(A_h) := \|A_h\|_\infty \|A_h^{-1}\|_\infty \leq h^{-2}.$$

Die  $l_\infty$ -Kondition von  $A_h$  verhält sich also in Abhängigkeit von der Gitterweite  $h$  genauso wie die Spektralkondition. (Hinweis: Man versuche es mit der mit  $w(x, y) = x(1-x)/2 + y(1-y)/2$  gebildeten Gitterfunktion.)

**Übung 2.6:** Gegeben sei die 1. RWA des Laplace-Operators

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega,$$

auf dem Dreiecksgebiet  $\Omega = \{(x, y) \in \mathbb{R}^2 \mid x, y > 0, x+y < 1\}$ . Man stelle zu einem äquidistanten, kartesischen Gitter das Gleichungssystem der 5-Punkte-Differenzenapproximation auf und vergleiche (a) die zeilenweise, (b) die diagonale und (c) die schachbrettartige Gitterpunktnumerierung in Bezug auf Matrixstruktur, Speicherplatzbedarf und Rechenaufwand bei der Lösung mit der LR-Zerlegung unter Ausnutzung der Bandstruktur.

**Übung 2.7:** Zur Auffrischung der Kenntnise über iterative Lösungsverfahren: Eine Vereinfachung des Jacobi-Verfahrens zur Lösung eines linearen  $N \times N$ -Gleichungssystems  $Ax = b$  ist das sog. „Richardson-Verfahren“. Dabei wird ausgehend von einem beliebigen Startvektor  $x^0 \in \mathbb{R}^N$  mit einem Dämpfungsparameter  $\theta \in \mathbb{R}$  wie folgt iteriert:

$$x^{t+1} = x^t - \theta(Ax^t - b), \quad t = 0, 1, 2, \dots$$

a) Im Falle, dass  $A$  nur reelle Eigenwerte  $\lambda_{\min} \leq \dots \leq \lambda \leq \dots \leq \lambda_{\max}$  besitzt, zeige man für den Spektralradius  $\rho(B_\theta)$  der zugehörigen Iterationsmatrix  $B_\theta = I - \theta A$  die Gleichung

$$\rho(B_\theta) = \max \{ |1 - \theta\lambda_{\min}|, |1 - \theta\lambda_{\max}| \}.$$

b) Im Falle, dass zusätzlich alle Eigenwerte positiv sind, zeige man

$$\rho(B_\theta) < 1 \iff 0 < \theta < \frac{2}{\lambda_{\max}}.$$

c) Für welchen Wert von  $\theta$  wird  $\rho(B_\theta)$  in diesem Falle minimal?

**Übung 2.8:** Betrachtet werde wieder das Modellproblem

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega,$$

auf dem Einheitsquadrat  $\Omega \subset \mathbb{R}^2$ . Die Systemmatrix des 5-Punkte-Differenzenoperators auf einem äquidistanten, kartesischen Punktgitter lässt sich schreiben als Summe der Anteile der beiden Differenzenquotienten in x- und y-Richtung:  $A = A_x + A_y$ . Bei lexikographischer Numerierung sind dabei  $A_x$  und  $A_y$  reguläre Tridiagonalmatrizen mit den Einträgen  $2h^{-2}$  auf den Hauptdiagonalen und  $-h^{-2}$  verteilt auf den Nebendiagonalen.

Das zugehörige Gleichungssystems  $AU = F$  besitzt damit die äquivalenten Formen

$$(\sigma I + A_x)U = (\sigma I - A_y)U + F, \quad (\sigma I + A_y)U = (\sigma I - A_x)U + F$$

mit einem beliebigen Parameterwert  $\sigma > 0$ . Dies legt das folgende zweistufige Iterationsverfahren (sog. „ADI-Verfahren“ = „Alternating Direction Implicit Iteration“) nahe:

$$(\sigma I + A_x)U^{t+1/2} = (\sigma I - A_y)U^t + F, \quad (\sigma I + A_y)U^{t+1} = (\sigma I - A_x)U^{t+1/2} + F.$$

Man zeige die Konvergenz dieses Verfahrens. Für welche Wahl von  $\sigma$  wird die Konvergenz am schnellsten? (Hinweis: Man überlege sich, dass die Zerlegungsmatrizen  $A_x$  und  $A_y$  ein gemeinsames System von Eigenvektoren besitzen und folglich vertauschbar sind.)

### 3 Finite-Elemente-Verfahren für elliptische Probleme

In diesem Kapitel werden wir die modernen Finite-Elemente-(Galerkin)-Methoden zur Lösung elliptischer Randwertaufgaben (RWA) diskutieren. Der Übersichtlichkeit halber werden wir uns dabei auf das Modellproblem der Poisson-Gleichung mit Dirichletschen Randbedingungen, d. h. auf die 1. RWA, beschränken:

$$Lu := -\Delta u = f \quad \text{in } \Omega, \quad u = g \quad \text{auf } \partial\Omega. \quad (3.0.1)$$

Das Definitionsbereich  $\Omega \in \mathbb{R}^2$  wird zunächst wieder als glatt berandet oder als konkaves Polygongebiet vorausgesetzt. Die Problemdaten  $f, g$  sind ebenfalls glatt, so dass die im vorigen Kapitel beschriebenen Resultate anwendbar sind. Erweiterungen für Probleme mit variablen Koeffizienten oder anderen Randbedingungen sowie auf drei Raumdimensionen werden wieder in Bemerkungen berücksichtigt.

#### 3.1 Allgemeine Projektionsverfahren

Ausgangspunkt ist die variationelle Formulierung der RWA. Wir erinnern an den oben diskutierten Ansatz zu einer allgemeinen Lösungstheorie. Eine „schwache“ bzw. „verallgemeinerte“ Lösung der 1. RWA des Laplace-Operators (zu den Randdaten  $g \equiv 0$ ) ist definiert als das (eindeutige) Minimum auf dem Sobolew-Raum  $H_0^1(\Omega)$  des Energiefunktionalen

$$E(v) := \frac{1}{2} \|\nabla v\|^2 - (f, v) \rightarrow \min.$$

Wir verwenden hier und im folgenden wieder die Bezeichnungen

$$(v, w) := \int_{\Omega} v(x)w(x) dx, \quad \|v\| := \left( \int_{\Omega} |v(x)|^2 dx \right)^{1/2}, \quad \|\nabla v\| := \left( \int_{\Omega} |\nabla v(x)|^2 dx \right)^{1/2}.$$

Über den Variationsansatz

$$\frac{d}{d\varepsilon} E(u + \varepsilon\varphi)_{|\varepsilon=0} = 0 \quad \forall \varphi \in H_0^1(\Omega),$$

erhalten wir die äquivalente Variationsgleichung (Stationaritätsbedingung)

$$u \in V : \quad (\nabla u, \nabla \varphi) = (f, \varphi) \quad \forall \varphi \in H_0^1(\Omega). \quad (3.1.2)$$

Wir erinnern daran, dass das natürliche Skalarprodukt des Raumes  $H_0^1(\Omega)$  gerade durch das sogen. „Dirichlet-Produkt“  $(\nabla v, \nabla w)$  gegeben ist. Zum Nachweis der Definitheit dieses Ausdrucks haben wir die Poincarésche Ungleichung verwendet:

$$\|v\| \leq d_{\Omega} \|\nabla v\|, \quad v \in H_0^1(\Omega). \quad (3.1.3)$$

**Bemerkung 3.1:** Im Falle *inhomogener Randbedingungen*  $u|_{\partial\Omega} = g$  geht man wie folgt vor. Wir nehmen an, dass die Randwerte als Spur einer Funktion  $\bar{g} \in H^1(\Omega)$  gegeben

sind:  $g = \bar{g}|_{\partial\Omega}$ . Für die Funktion  $v := u - \bar{g} \in H_0^1(\Omega)$  gilt dann im Falle  $\Delta\bar{g} \in L^2(\Omega)$ :

$$(\nabla v, \nabla \varphi) = (f, \varphi) - (\Delta\bar{g}, \varphi) =: (\tilde{f}, \varphi) \quad \forall \varphi \in H_0^1(\Omega),$$

d. h.: Die Funktion  $v$  genügt einer Variationsgleichung der Art (3.1.2). Im folgenden können wir also o.B.d.A. stets homogene Dirichlet-Randbedingungen annehmen.

**Bemerkung 3.2:** Im Fall von *Neumannschen Randbedingungen*  $\partial_n u|_{\partial\Omega} = g$  wird der Sobolew-Raum  $H^1(\Omega)$  (ohne Vorgabe von Randwerten) verwendet und die zugehörige variationelle Formulierung lautet

$$u \in H^1(\Omega) : \quad (\nabla u, \nabla \varphi) = (f, \varphi) + (g, \varphi)_{\partial\Omega} \quad \forall \varphi \in H^1(\Omega). \quad (3.1.4)$$

Um die eindeutige Lösbarkeit zu sichern, muss in diesem Fall noch eine Zusatzbedingung gestellt werden, um *konstante* Lösungen auszuschließen, z. B.: die Normierungsbedingung  $(u, 1)_\Omega = 0$ . Ferner muss die Verträglichkeitsbedingung  $(f, 1) + (g, 1)_{\partial\Omega} = 0$  erfüllt sein. Jede hinreichend glatte Lösung von (3.1.4) erfüllt dann neben der Differentialgleichung  $-\Delta u = f$  auch notwendig die in der variationellen Formulierung implizit enthaltene *natürliche* Randbedingung  $\partial_n u|_{\partial\Omega} = g$  (Beweis ähnlich wie im Fall der 1. RWA durch partielle Integration und Variation der Testfunktion). Eine ähnliche Konstruktion liefert auch die variationelle Formulierung im Fall der 3. RWA, d. h. für Robinsche Randbedingungen.

**Bemerkung 3.3:** Nicht alle elliptischen RWAn lassen sich nicht über den Energieminimierungsansatz behandeln. Ein typisches Beispiel ist die Diffusions-Transport-Gleichung

$$-\Delta u + \partial_1 u = f \quad \text{in } \Omega, \quad u = g \quad \text{auf } \partial\Omega. \quad (3.1.5)$$

Ihre variationelle Formulierung lautet

$$(\nabla u, \nabla \varphi) + (\partial_1 u, \varphi) = (f, \varphi) \quad \forall \varphi \in H_0^1(\Omega). \quad (3.1.6)$$

Diese besitzt ebenfalls eine eindeutige Lösung  $u \in H_0^1(\Omega)$ , was sich weiter unten als Folgerung eines allgemeineren Resultats ergeben wird.

Die folgende Diskussion wird in einem etwas abstrakteren Rahmen durchgeführt, welche an den obigen Beispielen orientiert ist und diese als Sonderfälle beinhaltet. Seien  $V$  ein allgemeiner Hilbert-Raum mit Skalarprodukt  $(\cdot, \cdot)_V$  und zugehöriger Norm  $\|\cdot\|_V := (\cdot, \cdot)_V^{1/2}$  und  $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$  eine *beschränkte* Bilinearform sowie  $l(\cdot) : V \rightarrow \mathbb{R}$  eine *beschränkte* Linearform:

$$|a(v, w)| \leq \alpha \|v\|_V \|w\|_V, \quad |l(v)| \leq \gamma \|v\|_V, \quad v, w \in V. \quad (3.1.7)$$

Mit diesen Bezeichnungen betrachten wir die folgende allgemeine Variationsgleichung: Bestimme  $u \in V$ , so dass

$$a(u, \varphi) = l(\varphi) \quad \forall \varphi \in V. \quad (3.1.8)$$

Zum Nachweis, dass diese Aufgabe auch eine Lösung besitzt, postulieren wir, dass die Bilinearform  $a(\cdot, \cdot)$  „stark  $V$ -elliptisch“ ist, d. h.:

$$a(v, v) \geq \kappa \|v\|_V^2, \quad v \in V, \quad (3.1.9)$$

mit einer Konstante  $\kappa > 0$ . Allgemeiner wird die Bilinearform  $a(\cdot, \cdot)$  „koerzitiv“ (oder „regulär“) genannt, wenn gilt:

$$\sup_{\varphi \in V} \frac{a(v, \varphi)}{\|\varphi\|_V} \geq \gamma \|v\|_V, \quad \sup_{\varphi \in V} \frac{a(\varphi, v)}{\|\varphi\|_V} \geq \gamma \|v\|_V, \quad v \in V, \quad (3.1.10)$$

mit einer Konstante  $\gamma > 0$ .

**Hilfssatz 3.1: (Lax<sup>1</sup>-Milgram<sup>2</sup>-Lemma)** Unter den obigen Voraussetzungen besitzt die Gleichung (3.1.8) eine eindeutige Lösung  $u \in V$ , für welche die a priori Abschätzung gilt:

$$\|u\|_V \leq \frac{1}{\kappa} \|l\|_{V^*}, \quad (3.1.11)$$

mit der „Dualnorm“  $\|l\|_{V^*} := \sup_{\{\varphi \in V, \|\varphi\|_V=1\}} |l(\varphi)|$ .

**Beweis:** Für jedes feste  $v \in V$  definiert  $a(v, \cdot)$  ein lineares, stetiges Funktional auf  $V$ . Nach dem Rieszschen<sup>3</sup> Darstellungssatz existieren Elemente  $Av \in V$  und  $f \in V$ , so dass

$$a(v, \varphi) = (Av, \varphi)_V, \quad l(\varphi) = (f, \varphi)_V, \quad \varphi \in V.$$

Die Zuordnung  $v \mapsto Av$  definiert eine lineare Abbildung mit der Eigenschaft

$$\|Av\|_{V^*} \leq \alpha \|v\|_V,$$

d. h.:  $A$  ist beschränkt. Die Aufgabe (3.1.8) ist offenbar äquivalent zu der Gleichung

$$Au = f. \quad (3.1.12)$$

Wir wollen zeigen, dass die Abbildung

$$v \in V \mapsto T_\delta v := v - \delta(Av - f) \in V$$

---

<sup>1</sup>Peter David Lax (1926–): US-Amerikanischer Mathematiker ungarischer Abstammung; Prof. an der New York University und am Courant-Institut; wichtige Beiträge zur Analysis, insbesondere zu den partiellen Differentialgleichungen der Math. Physik, und zur Numerik.

<sup>2</sup>Arthur Norton Milgram (1912–1961): US-Amerikanischer Mathematiker; Prof. an der Univ. of Minnesota, Minneapolis, USA; Beiträge u. a. zur Funktionalanalysis und ihren Anwendungen in der Theorie partieller Differentialgleichungen; am besten bekannt durch das sog. „Lax-Milgram-Lemma“ zus. mit P. Lax (1954).

<sup>3</sup>Frigyes Riesz (1880–1956): Ungarischer Mathematiker; Prof. in Szeged und Budapest; fundamentale Beiträge zur Funktionalanalysis, insbesondere der Fourier-Analyse im Hilbert-Raum als theoretische Grundlage der frühen Quantenmechanik.

für einen geeigneten Wert  $\delta > 0$  eine Kontraktion auf ganz  $V$  ist. Dann besitzt die Fixpunktgleichung

$$T_\delta v = v$$

eine eindeutige Lösung  $u \in V$ , welche wegen  $0 = v - T_\delta v = \delta(Av - f)$  dann auch (eindeutige) Lösung von (3.1.12) bzw. (3.1.8) ist. Die Kontraktions-eigenschaft ergibt sich aus der Beziehung

$$\begin{aligned}\|v - \delta Av\|_V^2 &= \|v\|_V^2 - 2\delta a(v, v) + \delta^2 \|Av\|_V^2 \\ &\leq (1 - 2\delta\kappa + \delta^2\alpha^2)\|v\|_V^2,\end{aligned}$$

für  $0 < \delta < 2\kappa/\alpha^2$ . Die a priori Abschätzung (3.1.11) ergibt sich dann direkt durch Testen mit  $\varphi := u$  in der Variationsgleichung (3.1.8). Q.E.D.

Die beiden obigen Beispiele zur 1. RWA passen in diesen Rahmen mit den natürlichen Setzungen  $V := H_0^1(\Omega)$ ,  $l(v) := (f, \varphi)$  und

$$a(v, w) := (\nabla v, \nabla w), \quad a(v, w) := (\nabla v, \nabla w) + (\partial_1 v, w).$$

Die Beschränktheit dieser Formen ergibt sich direkt mit Hilfe der Hölderschen und der Poincaréschen Ungleichung. Ihre  $V$ -Elliptizität ergibt sich unmittelbar:

$$a(v, v) = \|\nabla v\|^2 = \|v\|_V^2,$$

bzw. unter Beachtung von  $v|_{\partial\Omega} = 0$ :

$$\begin{aligned}a(v, v) &= \|\nabla v\|^2 + (\partial_1 v, v) = \|\nabla v\|^2 + \frac{1}{2}(\partial_1 v^2, 1) \\ &= \|\nabla v\|^2 + \frac{1}{2}(n_1 v^2, 1)|_{\partial\Omega} = \|\nabla v\|^2 = \|v\|_V^2.\end{aligned}$$

Durch geeignete Setzungen lassen sich auch die variationellen Formulierungen der 2. und 3. RWA in diesen abstrakten Rahmen einordnen.

Zur Approximation der Variationsgleichung (3.1.2) werden endlich dimensionale Teilmäume

$$V_h \subset V \quad (0 < h \leq h_0)$$

ausgewählt, deren *Feinheit* durch einen Diskretisierungsparameter  $h$  (z. B.: Gitterweite) charakterisiert ist.

i) Im Fall einer symmetrischen Bilinearform  $a(\cdot, \cdot)$  bestimmt das klassische „Ritzsche<sup>4</sup> Projektions-Verfahren“ Näherungslösungen  $u_h \in V_h$  durch die Vorschrift

$$E(u_h) = \min_{v_h \in V} E(v_h) \tag{3.1.13}$$

---

<sup>4</sup>Walter Ritz (1878–1909): Schweizer Physiker; Prof. in Zürich und Göttingen; Beiträge zu Spektraltheorie in der Kernphysik und Elektro-Magnetismus.

oder äquivalent durch die diskrete Variationsgleichung

$$a(u_h, \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h. \quad (3.1.14)$$

Die (eindeutige) Existenz der diskreten Lösung  $u_h \in V_h$  folgt mit demselben Argument wie beim kontinuierlichen Problem. Diese Analogie der Schlüsselelemente von kontinuierlicher und diskreter (endlich dimensionaler) Situation ist die charakteristische Stärke der Projektionsmethoden im Gegensatz zu den Differenzenverfahren. Die Bezeichnung Projektionsverfahren ist motiviert durch die Beziehung

$$a(u - u_h, \varphi_h) = 0 \quad \forall \varphi_h \in V_h, \quad (3.1.15)$$

welche man durch Subtraktion der Gleichungen (3.1.2) und (3.1.14) erhält. Sie kann geometrisch dahingehend interpretiert werden, dass der Fehler  $e_h := u - u_h$  bzgl. des Skalarprodukts  $a(\cdot, \cdot)$  senkrecht auf dem Ansatzraum  $V_h$  steht. Dies impliziert auch die sog. „Bestapproximationseigenschaft“ für den Approximationsfehler  $e_h$  bzgl. der natürlichen „Energie-Norm“  $\|\cdot\|_a := a(\cdot, \cdot)^{1/2}$ , denn mit beliebigem  $\varphi_h \in V_h$  gilt:

$$\|e_h\|_a^2 = a(e_h, e_h) = a(e_h, u - \varphi_h) + a(e_h, \varphi_h - u_h) \leq \|e_h\|_a \|u - \varphi_h\|_a$$

bzw.

$$\|e_h\|_a \leq \inf_{\varphi_h \in V_h} \|u - \varphi_h\|_a. \quad (3.1.16)$$

Da die Normen  $\|\cdot\|_a$  und  $\|\cdot\|_V$  auf  $V$  äquivalent sind, ist die Frage nach der Konvergenz des Projektionsverfahrens,

$$\|e_h\|_V \rightarrow 0 \quad (h \rightarrow 0), \quad (3.1.17)$$

damit zurückgeführt auf die Frage der Approximierbarkeit von Funktionen  $u \in V$  durch Ansatzfunktionen  $\varphi_h \in V_h$ :

$$\inf_{\varphi_h \in V_h} \|u - \varphi_h\|_V \rightarrow 0 \quad (h \rightarrow 0). \quad (3.1.18)$$

ii) Wenn die Bilinearform  $a(\cdot, \cdot)$  nicht symmetrisch ist, wie beim obigen Diffusions-Transport-Problem, kann die zugehörige RWA nicht mehr durch ein Minimierungsproblem charakterisiert werden. Das allgemeine „Galerkinsche<sup>5</sup> (Projektions)-Verfahren“ geht direkt von der Variationsgleichung (3.1.2) aus und bestimmt Näherungen durch die Beziehung

$$a(u_h, \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h. \quad (3.1.19)$$

Wegen der  $V$ -Elliptizität der Bilinearform  $a(\cdot, \cdot)$  auf dem endlich dimensionalen Teilraum  $V_h$  folgt unmittelbar die Existenz der (eindeutigen) Lösung  $u_h \in V_h$ . Die Ortho-

---

<sup>5</sup>Boris Grigorjewich Galerkin (1871–1945): Russischer Bauingenieur und Mathematiker; Prof. in St. Petersburg; Beiträge zur Struktur-Mechanik, insbesondere zur Plattentheorie.

gonalitätsbeziehung (3.1.15) bleibt dabei gültig. Damit erschließen wir die Quasi-Best-Approximationseigenschaft (Übungsaufgabe)

$$\|e_h\|_V \leq \frac{\alpha}{\kappa} \min_{\varphi \in V_h} \|u - \varphi_h\|_V. \quad (3.1.20)$$

iii) Eine noch allgemeinere Variante, bei der Ansatzraum  $V_h^{\text{ansatz}}$  und Testraum  $V_h^{\text{test}}$  unterschiedlich gewählt werden,

$$u \in V_h^{\text{ansatz}} : \quad a(u_h, \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h^{\text{test}}, \quad (3.1.21)$$

ist das sog. „Petrow<sup>6</sup>-Galerkin-Verfahren“. Wir werden später Beispiele für diesen unkonventionellen Ansatz kennenlernen.

Zur praktischen Realisierung des Projektionsverfahrens muss die zunächst abstrakte Variationsgleichung (3.1.19) im Funktionenraum algebraisiert werden, d. h.: in ein äquivalentes algebraisches Gleichungssystem umgewandelt werden. Dazu wählen wir zunächst eine Basis  $\{\varphi_h^{(i)}, i = 1, \dots, N\}$ ,  $N := \dim V_h$ , von  $V_h$  aus und machen für die zu bestimmende diskrete Lösung den Ansatz  $u_h = \sum_{j=1}^N \xi_j \varphi_h^{(j)}$ . Wird dies in (3.1.19) eingesetzt und lässt man die Testfunktionen  $\varphi_h \in V_h$  alle Basisfunktionen durchlaufen, ergibt sich ein lineares (algebraisches)  $N \times N$ -Gleichungssystem

$$\sum_{j=1}^N \xi_j a(\varphi_h^{(j)}, \varphi_h^{(i)}) = (f, \varphi_h^{(i)}), \quad i = 1, \dots, N,$$

für den Vektor  $\xi = (\xi_j)_{j=1}^N$  der Entwicklungskoeffizienten, bzw. in kompakter Schreibweise

$$A_h \xi = b_h. \quad (3.1.22)$$

Dabei sind die Koeffizientenmatrix  $A_h = (a_{ij})_{i,j=1}^N$  sowie die rechte Seite  $b_h = (b_i)_{i=1}^N$  durch die spezielle Wahl der Basis bestimmt:

$$a_{ij} = a(\varphi_h^{(j)}, \varphi_h^{(i)}), \quad b_j = (f, \varphi_h^{(j)}).$$

Die Entwicklungskoeffizienten  $\xi_j$  können sehr unterschiedliche Bedeutung haben; z. B.: Monom-Koeffizienten einer Polynomdarstellung, Fourier-Koeffizienten einer trigonometrischen Entwicklung, Knotenwerte einer stückweise polynomialen Funktion, u.s.w.. Die Eigenschaften der Bilinearform  $a(\cdot, \cdot)$  übertragen sich direkt auf die zugehörige Matrix  $A_h$ . Ist  $a(\cdot, \cdot)$  symmetrisch, so auch  $A_h$ ,

$$a_{ij} = a(\varphi_h^{(j)}, \varphi_h^{(i)}) = a(\varphi_h^{(i)}, \varphi_h^{(j)}) = a_{ji},$$

---

<sup>6</sup>Georgi Iwanowitsch Petrow (1912–1987): Russischer Ingenieur; 1965–1973 Direktor des Instituts für Raumfahrtforschung; Publ.: „Application of the Galerkin method and the problem of flow stability of a viscous liquid“ (russ.), Prikl. Mat. Mekh. 4, 36–47 (1947)

und die  $V$ -Elliptizität von  $a(\cdot, \cdot)$  impliziert die Definitheit von  $A_h$ , denn für  $x \in \mathbb{R}^n \setminus \{0\}$  gilt:

$$\begin{aligned} (A_h x, x) &= \sum_{i,j=1}^N a_{ij} x_i x_j = \sum_{i,j=1}^N a(\varphi_h^{(j)}, \varphi_h^{(i)}) x_i x_j \\ &= a\left(\sum_{j=1}^N x_j \varphi_h^{(j)}, \sum_{i=1}^N x_i \varphi_h^{(i)}\right) \geq \kappa \left\| \sum_{i=1}^N x_i \varphi_h^{(i)} \right\|_V^2 > 0. \end{aligned}$$

### 3.1.1 Beispiele von Galerkin-Ansatzräumen

Wir wollen einige konkrete Realisierungen für den beschriebenen abstrakten Rahmen diskutieren. Bei der Wahl der Ansatzräume  $V_h \subset V = H_0^1(\Omega)$  sowie der Basen zur Aufstellung der Gleichungssysteme (3.1.22) sind einige Bedingungen zu beachten:

- Die Berechnung der Matrixelemente  $a_{ij} = a(\varphi_h^{(j)}, \varphi_h^{(i)})$  sowie die der rechten Seite  $(f, \varphi_h^{(i)})$  sollte „billig“ sein.
- Aus Genauigkeitsgründen wird die Problemdimension in der Regel sehr groß sein:  $N \gg 100$ . Die Matrix  $A_h$  sollte daher möglichst dünn besetzt sein, d. h. möglichst viele Nullen enthalten.
- Die Matrix  $A_h$  sollte nicht zu schlecht konditioniert sein; akzeptabel sind z. B. beim vorliegenden Problem  $\text{cond}_2(A_h) \approx O(N) - O(N^2)$ , wogegen  $\text{cond}_2(A_h) \approx O(N^4) - O(e^N)$  nicht praktikabel wären.

Beispiele von solchen Ansätzen sind:

1) *Globaler Polynomansatz*: Auf einem Quadrat  $\Omega = (0, 1)^2$  wird der Tensor-Produkt-Ansatz gemacht

$$V_h := Q_m(\Omega) := \left\{ p(x, y) = \sum_{i,j=0}^m c_{ij} x^i y^j \right\}, \quad h := 1/m, \quad N = (m+1)^2.$$

Als Basen kommen dabei in Frage:

a) Monombasis  $1, x, y, x^2, xy, y^2, \dots$ ; die zugehörige Matrix  $A$  mit den Elementen

$$a_{ij} = \int_0^1 \int_0^1 \nabla x^i \cdot \nabla y^j \, dx \, dy, \quad 0 \leq i, j \leq m,$$

verhält sich dann wie die bekannte Hilbert-Matrix mit exponentiell mit  $N$  wachsender Kondition  $\text{cond}_2(A_h) = O(e^N)$ . Dieser Ansatz ist also praktisch unbrauchbar.

b) Tensorprodukt-Basen  $L^2$ -orthogonaler Polynome wie z. B. Legendre-Polynome oder Tschebyscheff-Polynome; die zugehörige Matrix  $A_h$  mit den Elementen

$$a_{ij} = \int_0^1 \int_0^1 \nabla L_i^{(m)} \cdot \nabla L_j^{(m)} \, dx \, dy, \quad 0 \leq i, j \leq m,$$

ist dann zwar voll besetzt, hat aber eine wesentlich günstigere Kondition,  $\text{cond}_2(A_h) = O(N)$ . Dieser Ansatz führt auf die sog. „Spektral-Galerkin-Verfahren“, welche bei Problemen auf geometrisch einfachen (rechteckigen) Gebieten sehr leistungsfähig sind. Die Bezeichnung „Spektralverfahren“ röhrt daher, dass man die orthogonalen Polynome auch als Eigenfunktionen gewisser Differentialoperatoren 2. Ordnung charakterisieren kann. Wegen ihrer konzeptionellen Beschränkung auf einfache Geometrien wollen wir derartige Methoden hier nicht weiter diskutieren. Stichworte für Entwicklungen in Richtung auf eine Überwindung dieser Restriktion sind z. B. „Spektral-Elemente-Methoden“ und „ $h/p$ -Finite-Elemente-Methode“.

2) *Globaler trigonometrischer Ansatz („echte Spektralverfahren“):* Wieder auf einem Quadrat  $\Omega = (0, 1)^2$  wird der Tensor-Produkt-Ansatz gemacht

$$V_h := T_m(\Omega) := \left\{ t(x, y) = \sum_{i,j=0}^m c_{ij} \sin(i\pi x) \sin(j\pi y) \right\}, \quad h := 1/m, N = (m+1)^2.$$

Als Basen verwendet man dabei die trigonometrische Basis  $\{1, \sin(n\pi x) \sin(m\pi y), \dots\}$ . Die zugehörige Matrix  $A_h$  ist dann vergleichsweise gut konditioniert,  $\text{cond}_2(A_h) = O(N)$ . In diesem Fall gibt es mit der schnellen Fourier-Transformation („FFT“) einen fast optimalen Algorithmus mit der Komplexität  $O(N \log(N))$  zur Lösung des Gleichungssystems (3.1.22). Der Nachteil dieses in Spezialfällen sehr leistungsfähigen Ansatzes ist wieder seine Beschränktheit auf einfache Rechteckgeometrien und sog. „separable“ Differentialoperatoren mit konstanten Koeffizienten.

3) *Stückweise polynomieller Ansatz („Finite Elemente“):* Um das Problem der Approximation allgemeiner Gebiete zu lösen, werden Ansatzfunktionen (auch „Formfunktionen“ genannt) verwendet, welche bzgl. einer Zerlegung von  $\bar{\Omega}$  in einfache Teilgebiete  $T$ , sog. „Zellen“, stückweise polynomial sind. Gängige Beispiele von Zellen sind Dreiecke oder (konvexe) Vierecke in zwei bzw. Tetraeder oder (konvexe) Hexaeder in drei Dimensionen. Der Parameter  $h$  ist in diesem Fall etwa der maximale Zelldurchmesser.

Wir illustrieren diesen Finite-Elemente-Ansatz anhand eines einfachen Beispiels. Die 1. RWA (3.1.2) sei auf einem (konvexen) polygonalen Gebiet  $\Omega \subset \mathbb{R}^2$  mit homogenen Randwerten  $u|_{\partial\Omega} = 0$  und rechter Seite  $f \in L^2(\Omega)$  gestellt. Die zugehörige Lösung  $u \in H_0^1(\Omega)$  ist dann auch im Sobolew-Raum  $H^2(\Omega)$  und genügt der *a priori* Abschätzung

$$\|\nabla^2 u\| \leq c_S \|f\|, \tag{3.1.23}$$

wobei  $c_S = 1$  im Falle eines konvexen Gebiets.

Weiter sei eine Folge von Zerlegungen  $T_h = \{T\}$  des Gebiets  $\bar{\Omega}$  in abgeschlossene Dreiecke  $T$  („Triangulierung“) gegeben mit  $h := \max_T \text{diam}(T) \rightarrow 0$ . Wir stellen die folgenden Regularitätsbedingungen an diese Triangulierung:

i) *Strukturregularität:* Je zwei Dreiecke der Zerlegung  $\bar{\Omega} = \bigcup\{T \in T_h\}$  überlappen sich höchstens in gemeinsamen Eckpunkten oder in ganzen Seiten, d. h.: Sog. „hängende“ Knoten auf Dreiecksseiten sind hier nicht erlaubt.

ii) *Formregularität*: Alle Dreiecke der Triangulierungen  $T \in \mathbb{T}_h$  sind von ähnlicher Gestalt, d.h.: Für den Inkreisradius  $\rho_T$  und Umkreisradius  $h_T$  eines jeden Dreiecks  $T$  gilt gleichmäßig für  $h \rightarrow 0$ :

$$\max_{T \in \mathbb{T}_h} \frac{h_T}{\rho_T} \leq c_1 \quad (3.1.24)$$

iii) *Größenregularität*: Alle Dreiecke einer Triangulierung  $\mathbb{T}_h$  sind von gleicher Größenordnung, d.h.: Es gilt gleichmäßig für  $h \rightarrow 0$ :

$$\max_{T \in \mathbb{T}_h} h_T \leq c_2 \min_{T \in \mathbb{T}_h} h_T. \quad (3.1.25)$$

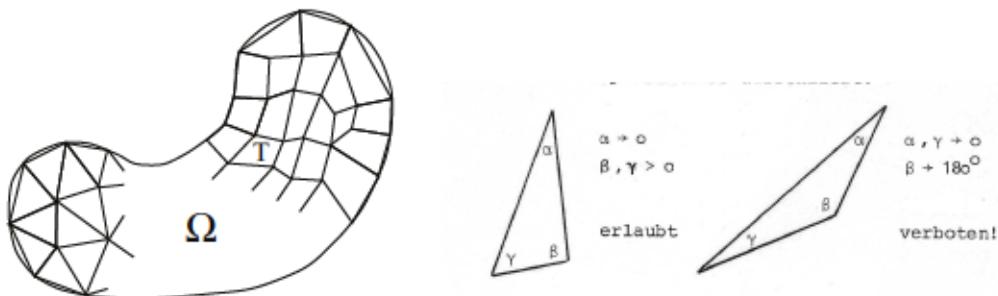


Abbildung 3.1: Finite-Elemente: Dreiecks- bzw. Vierecksgitter

Auf den Triangulierungen  $\mathbb{T}_h$  definieren wir Ansatzräume stückweise linearer Funktionen („lineare finite Elemente“: Verallgemeinerung des Konzepts eines Polygonzugs auf höhere Raumdimensionen):

$$V_h^{(1)} := \{v_h \in C(\bar{\Omega}) \mid v_{h|T} \in P_1(T), T \in \mathbb{T}_h, v_{h|\partial\Omega} = 0\}.$$

Dabei bezeichnet allgemein  $P_r(T)$  den Vektorraum der Polynome bis zum Grad  $r \geq 0$  über  $T$ . Man überlegt sich leicht, dass dadurch tatsächlich Teilräume  $V_h^{(1)} \subset H_0^1(\Omega)$  erklärt sind. Sei  $N$  die Anzahl der „inneren“ Knoten (Dreieckseckpunkte) der Triangulierung. Jedes  $v_h \in V_h^{(1)}$  ist als stückweise lineare Funktion eindeutig durch Vorgabe ihrer Funktionswerte („Knotenwerte“) in den „inneren“ Dreieckseckpunkten („Knoten“) festgelegt. In den Eckpunkten auf dem Gebietsrand  $\partial\Omega$  ist  $v_h = 0$  wegen der Dirichlet-Randbedingung. In  $V_h^{(1)}$  gibt es daher eine natürliche Basis, die sog. „Knotenbasis“ in Analogie zur „Lagrange“<sup>7</sup>-Basis bei der eindimensionalen Lagrange-Interpolation. Jedem Knoten  $a_i$  wird durch die Bedingung

$$\varphi_h^i(a_j) = \delta_{ij}, \quad j = 1, \dots, N,$$

<sup>7</sup>Joseph Louis de Lagrange (1736–1813): Französischer Mathematiker; 1766–87 Direktor der mathem. Klasse der Berliner Akademie, dann Prof. in Paris; bahnbrechende Arbeiten zur Variationsrechnung, zur komplexen Funktionentheorie sowie zur theor. Mechanik und Himmelsmechanik.

eindeutig eine Funktion  $\varphi_h^i \in V_h^{(1)}$  zugeordnet. Damit gilt dann für jedes  $v_h \in V_h^{(1)}$  die Darstellung

$$v_h = \sum_{i=1}^N v_h(a_i) \varphi_h^{(i)}.$$

Daraus folgt, dass  $\{\varphi_h^{(i)}, i = 1, \dots, N\}$  tatsächlich eine Basis von  $V_h^{(1)}$  ist. Umgekehrt lässt sich jeder kontinuierlichen Funktion  $v \in C(\bar{\Omega})$  durch die Vorschrift

$$I_h v := \sum_{i=1}^N v(a_i) \varphi_h^{(i)}$$

eindeutig eine (stückweise lineare) „Interpolierende“  $I_h v \in V_h^{(1)}$  zuordnen. Offenbar ist  $I_h v_h \equiv v_h$  für  $v_h \in V_h^{(1)}$ .

Dieser Diskretisierungsansatz erfüllt offensichtlich die oben formulierten Anforderungen an ein brauchbares Galerkin-Verfahren: Die resultierende Systemmatrix  $A_h$  ist dünn besetzt (wegen der geringen Überlappung der Träger der Basisfunktionen), und ihre Elemente sind sehr leicht zu berechnen. Wir werden die praktische Berechnung von  $A_h$  im Zusammenhang mit allgemeineren Finite-Elemente-Ansätzen dieser Art noch eingehender diskutieren.

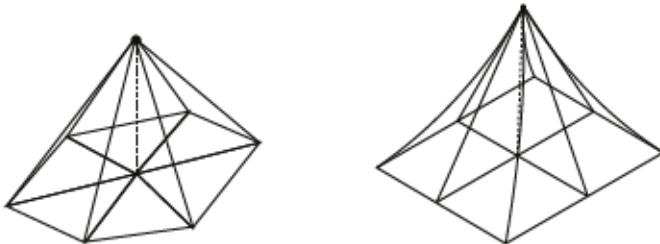


Abbildung 3.2: Knotenbasisfunktionen: Lineare und bilineare FE-Ansätze

Für die Kondition der Matrix  $A_h$  werden wir später in zwei Dimensionen  $\text{cond}_2(A_h) = O(h^{-2}) = O(N)$  zeigen (ähnlich wie bei der 5-Punkte-Differenzendiskretisierung). Wir rekapitulieren für das Galerkin-Verfahren mit dem Finite-Elemente-Raum  $V_h^{(1)}$  die asymptotische Abschätzung für den Fehler  $e_h := u - u_h$ :

$$\|\nabla e_h\| = \min_{\varphi_h \in V_h^{(1)}} \|\nabla(u - \varphi_h)\|. \quad (3.1.26)$$

Die Frage ist also, ob es  $\varphi_h \in V_h^{(1)}$  gibt, so dass  $\|\nabla(u - \varphi_h)\| \rightarrow 0$  ( $h \rightarrow 0$ ). Wenn man nur weiß, dass  $u \in H_0^1(\Omega)$  ist, dann kann man nur qualitative Konvergenz zeigen. Viel interessanter wäre es, wieder die Konvergenzgeschwindigkeit in Potenzen der Gitterweite  $h$  zu kennen. Hierzu ist aber natürlich mehr Regularität der Lösung  $u$  erforderlich. Für den stückweise linearen Ansatz werden wir später im Rahmen einer allgemeinen Theorie

die folgende Interpolationsabschätzung zeigen:

$$\|\nabla(v - I_h v)\| \leq c_I h \|\nabla^2 v\|, \quad v \in H_0^1(\Omega) \cap H^2(\Omega). \quad (3.1.27)$$

(Man vergleiche die analoge Abschätzung, welche in einer Dimension bei der Finite-Elemente-Approximation eindimensionaler Sturm-Liouville-Probleme verwendet wird.) Unter den bisher formulierten Voraussetzungen lässt sich eine erste quantitative Konvergenzaussage für das Finite-Elemente-Verfahren ableiten.

**Satz 3.1 (Konvergenzsatz):** Für die Galerkin-Approximation von Problem (3.1.2) mit „linearen“ finiten Elementen gelten unter den obigen Voraussetzungen die Fehlerabschätzungen

$$\|\nabla e_h\| \leq c_I c_S h \|f\|, \quad (3.1.28)$$

$$\|e_h\| \leq c_I^2 c_S^2 h^2 \|f\|, \quad (3.1.29)$$

mit den Konstanten  $c_I, c_S$  aus den Ungleichungen (3.1.27) und (3.1.23).

**Beweis:** (i) Die Abschätzung des sog. „Energienorm-Fehlers“ ergibt sich unmittelbar aus der Best-Approximationsbeziehung (3.1.26), der Interpolationsabschätzung (3.1.27) und der Regularitätsabschätzung (3.1.23):

$$\|\nabla e_h\| \leq \|\nabla(u - I_h u)\| \leq c_I h \|\nabla^2 u\| \leq c_I c_S h \|f\|.$$

(ii) Zum Beweis der Fehlerabschätzung in der  $L^2$ -Norm verwenden wir ein sog. „Dualitätsargument“ („Aubin<sup>8</sup> -Nitsche<sup>9</sup> -Trick“). Sei  $z \in H_0^1(\Omega)$  die (schwache) Lösung des Hilfsproblems

$$-\Delta z = \|e_h\|^{-1} e_h \text{ in } \Omega, \quad z|_{\partial\Omega} = 0.$$

Diese ist dann auch in  $H^2(\Omega)$ , und es gilt die *a priori* Abschätzung

$$\|\nabla^2 z\| \leq c_S \|\Delta z\| = c_S,$$

wobei wieder  $c_S = 1$  auf konvexem Gebiet  $\Omega$ . Nach Konstruktion folgt mit Hilfe der Galerkin-Orthogonalität:

$$\begin{aligned} \|e_h\| &= (\nabla e_h, \nabla z) = (\nabla e_h, \nabla(z - I_h z)) \\ &\leq \|\nabla e_h\| \|\nabla(z - I_h z)\| \leq c_I h \|\nabla e_h\| \|\nabla^2 z\| \leq c_I c_S h \|\nabla e_h\|. \end{aligned}$$

Mit dem Ergebnis (i) ergibt sich damit die gewünschte Abschätzung.

Q.E.D.

Die im Beweis von Satz 3.1 verwendete Schlussweise über ein Dualitätsargument ist „das“ zentrale Hilfsmittel bei der Konvergenzanalyse von Finite-Elemente-Verfahren. Die-

<sup>8</sup>Jean-Pierre Aubin (1939–): Französischer Mathematiker; Prof. an der Univ. Paris-Dauphine (2004) emeritiert); Beiträge zur Theorie partieller Differentialgleichungen und ihrer Numerik.

<sup>9</sup>Joachim A. Nitsche (1926–1996): Deutscher Mathematiker; Prof. in Freiburg; fundamentale Beiträge zur Theorie der Finite-Elemente-Methode (u. a.  $L^\infty$ -Fehlerabschätzungen).

ses abstrakte Argument entspricht der allgemeinen Regel, dass sich die Analyse der Projektionsverfahren eng an die abstrakten Hilbertraum-Methoden zur Behandlung des kontinuierlichen Problems anlehnt. Das zentrale Hilfsmittel bei der Untersuchung von Differenzenverfahren war dagegen das „(diskrete) Maximumprinzip“, welches sich mehr an den klassischen Techniken für partielle Differentialgleichungen orientiert. Wir wollen diesen Vergleich „Finite-Elemente (FEM) - Finite-Differenzen (FDM)“ anhand des Resultats von Satz 3.1 noch etwas weiterführen.

Die a priori Fehlerabschätzung (3.1.29) für das Finite-Elemente-Verfahren ist zu vergleichen mit der Abschätzung (2.2.37) für das Differenzenverfahren (5-Punkte-Diskretisierung mit Shortley-Weller-Randapproximation auf polygonalen Gebieten):

$$\max_{\Omega} |e_h| \leq \frac{1}{24} d_{\Omega}^2 M_4(u) h^2 + \mathcal{O}(h^3), \quad (3.1.30)$$

mit der Schranke  $M_4(u)$  für die vierten Ableitungen von  $u$ . Beide Abschätzungen zeigen dieselbe asymptotische Konvergenzordnung  $\mathcal{O}(h^2)$ , was aufgrund der verwendeten Diskretisierungsansätze auch zu erwarten ist. Die Unterschiede liegen zum einen in der Art der Norm, in der der Fehler gemessen wird, und zum anderen in der benötigten Regularität der approximierten Lösung. Beim Differenzenverfahren erhält man wegen der Verwendung des Maximumprinzips punktweise Abschätzungen, wie sie auch der Anwender gern hat. (Der Ingenieur ist z. B. an der maximalen Auslenkung einer belasteten Brückenkonstruktion interessiert.) Dagegen liefert die Hilbert-Raum-Theorie für das Finite-Elemente-Verfahren zunächst nur Abschätzungen im quadratischen Mittel, was etwa lokale „Ausreißer“ an kritischen Stellen nicht ausschließt. (Dem Brückenbauer genügt so etwas nicht, wenn Fehlerspitzen etwa in kritischen Lagerungspunkten der Brücke auftreten können.) Wir werden später die Frage diskutieren, ob und wie man auch für das Finite-Elemente-Verfahren Fehlerabschätzungen in der Maximumnorm herleiten kann. Die in der Abschätzung (3.1.30) geforderte hohe Regularität der Lösung ist ein sehr viel schwerwiegender Nachteil unserer Analyse des Differenzenverfahrens, da diese Regularitätsstufe i. Allg. auf Polygonegebieten und unter realistischen Annahmen an die Problemdaten nicht erwartet werden kann. Wir bemerken, dass man für das Finite-Elemente-Verfahren mit wesentlich mehr technischem Aufwand „optimale“ Maximumnorm-Fehlerabschätzungen der Form ( $h \geq h_0 > 1$ )

$$\max_{\Omega} |e_h| \leq c M_2(u) h^2 |\log h| \quad (3.1.31)$$

beweisen kann. Allerdings ist auch die abgeschwächte Annahme  $M_2(u) < \infty$  i. Allg. noch zu restriktiv. Für das Finite-Elemente-Verfahren ist auch noch unter der Minimalvoraussetzung  $u \in H_0^1(\Omega)$  wenigstens qualitative Konvergenz gesichert. Seinen eigentlichen Vorteil, nämlich die große Flexibilität bei der Approximation von komplizierten Geometrien auf unstrukturierten Gittern werden wir später im Zusammenhang mit der Frage nach adaptiver Gittersteuerung und Fehlerkontrolle erkennen.

### 3.1.2 Diskretes Maximumprinzip für Finite-Elemente-Approximationen

Als nächstes wollen wir zeigen, dass Galerkin-Verfahren mit finiten Elementen als Ansatzfunktionen tatsächlich eng verwandt mit Differenzenverfahren sind. Dazu werden die Elemente der „Steifigkeitsmatrix“  $A_h$  des Finite-Elemente-Verfahrens für den Fall „linearer“ Ansatzfunktionen auf einem Dreiecksgitter explizit bestimmt.

Dazu wird zunächst ein einzelnes Dreieck  $T$  mit den Eckpunkten  $P_i$  ( $i = 1, 2, 3$ ) betrachtet (siehe Abb. 3.3). Die dem Eckpunkt  $P_i$  gegenüberliegende Seite sei mit  $S_i$  und die zugehörige Höhe mit  $H_i$  bezeichnet. Die Seiten werden dabei als im Gegenuhrzeigersinn und die Höhen gegen den Eckpunkt orientierte Vektoren aufgefasst. Weiter bezeichne  $\psi_i$  die (stückweise lineare) Knotenbasisfunktion zum Punkt  $P_i$ , welche auf  $T$  definiert ist durch  $\psi_i(P_k) = \delta_{ik}$ ,  $i, k = 1, 2, 3$ .

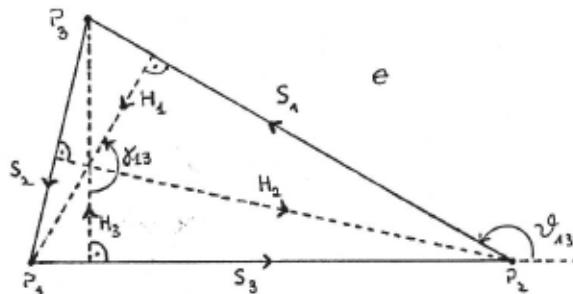


Abbildung 3.3: Dreiecksschema

Es gilt  $\nabla \psi_i \equiv \text{konst.}$  und wegen  $\psi_i(P_j) = \psi_i(P_k) = 0$ ,  $i \neq j, k$ , hat  $\nabla \psi_i$  in Richtung  $S_i$  die Komponente Null. Folglich zeigt  $\nabla \psi_i$  in Richtung  $H_i$  und hat wegen  $\psi_i(P_i) = 1$  den Betrag  $|H_i|^{-1}$ :

$$\nabla \psi_i = \frac{H_i}{|H_i|^2}.$$

Wir erhalten also

$$(\nabla \psi_i, \nabla \psi_j)_T = |T| \frac{(H_i, H_j)}{|H_i|^2 |H_j|^2}.$$

Für den Winkel  $\gamma_{ij}$  zwischen den Höhen  $H_i$  und  $H_j$  gilt

$$\cos(\gamma_{ij}) = \frac{(H_i, H_j)}{|H_i| |H_j|}$$

und, da  $\gamma_{ij}$  gleich dem Winkel  $\theta_{ij}$  zwischen den Seitenvektoren  $S_i$  und  $S_j$  ist, auch (Man beachte die Orientierung von  $S_i$  und  $S_j$ .):

$$\cos(\gamma_{ij}) = \frac{(S_i, S_j)}{|S_i| |S_j|}$$

Damit folgt bei Beachtung von  $2|T| = |H_i| |S_i|$  die Beziehung

$$(\nabla \psi_i, \nabla \psi_j)_T = |T| \frac{(S_i, S_j)}{|S_i| |S_j| |H_i| |H_j|} = \frac{(S_i, S_j)}{4|T|}.$$

Hieraus lesen wir ab, dass

$$(\nabla \psi_i, \nabla \psi_i)_T > 0, \quad (\nabla \psi_i, \nabla \psi_j)_T \leq 0, \quad i \neq j, \quad (3.1.32)$$

falls alle Winkel im Dreieck  $T$  kleiner oder gleich  $\pi/2$  sind. Weiter ist nach Konstruktion

$$\sum_{j=1}^3 \psi_j \equiv 1, \quad (\nabla \psi_i, \sum_{j=1}^3 \nabla \psi_j)_T = 0,$$

bzw.

$$\sum_{j=1, j \neq i}^N |(\nabla \psi_i, \nabla \psi_j)_T| \leq (\nabla \psi_i, \nabla \psi_i)_T, \quad i = 1, 2, 3.$$

Für die Elemente  $a_{ij} = \sum_{T \in T_h} (\nabla \psi_j, \nabla \psi_i)_T$  der Matrix  $A_h$  erhalten wir somit, dass

$$a_{ii} > 0, \quad a_{ij} \leq 0 \quad (i \neq j),$$

wenn alle Dreiecksinnenwinkel in der Triangulierung kleiner oder gleich  $\pi/2$  sind. Darüber hinaus gilt dann

$$\sum_{j \neq i} |a_{ij}| \leq a_{ii}, \quad \sum_{j \neq i_0} |a_{i_0 j}| < a_{i_0 i_0} \quad \text{für ein } i_0. \quad (3.1.33)$$

Die Steifigkeitsmatrix  $A_h$  ist in diesem Fall also „(irreduzibel) diagonal-dominant“, „von nicht-negativem Typ“ und eine „ $M$ -Matrix“. Wir fassen die sich daraus ergebenden Konsequenzen in einem Satz zusammen.

**Satz 3.2 (Maximumprinzip für finite Elemente):** *Wenn alle Innenwinkel der Triangulierung  $T_h$  kleiner oder gleich  $\pi/2$  sind, genügt das Finite-Elemente-Schema mit stückweise linearen Ansatzfunktionen einem diskreten Maximumprinzip, d. h.:*

$$(\nabla v_h, \nabla \varphi_h^{(n)}) \leq 0 \quad (n = 1, \dots, N) \quad \Rightarrow \quad \max_{\Omega} v_h \leq \max\{0, \max_{\partial\Omega} v_h\}. \quad (3.1.34)$$

Ferner ist die Steifigkeitsmatrix  $A_h$  eine  $M$ -Matrix, d. h.: Es gilt  $A_h^{-1} \geq 0$  sowie

$$x \in \mathbb{R}^N, \quad A_h x \geq 0 \quad \Rightarrow \quad x \geq 0. \quad (3.1.35)$$

Dieses Resultat kann so interpretiert werden, dass unter den gegebenen Voraussetzungen auch die Finite-Elemente-Diskretisierung ein „diskretes Maximumprinzip“ erfüllt. Leider gilt die kritische Eigenschaft (3.1.32) praktisch nur in der oben beschriebenen Situation. Insbesondere Finite-Elemente-Ansätze höherer Ordnung erfüllen dies nicht (z. B. quadratische Ansätze nur auf „gleichseitigen“ Triangulierungen).

**Bemerkung 3.4:** Im Spezialfall einer gleichförmigen, kartesischen Triangulierung (mit Kantenlänge  $h$ ) des Einheitsquadrats erhalten wir aus der obigen expliziten Darstellung für die Matrixelemente  $a_{ij} = (\nabla \psi_i, \nabla \psi_j)$  die Beziehung:

$$a_{ii} = 4, \quad a_{i,i\pm 1} = -1, \quad a_{i,i\pm m} = -1;$$

alle anderen Elemente  $a_{ij}$  sind Null. In diesem Fall stimmt die Steifigkeitsmatrix  $A_h^{FEM}$  also bis auf den Faktor  $h^{-2}$  mit der Matrix  $A_h^{FDM}$  des „5-Punkte-Operators“ überein:

$$A_h^{FEM} = h^2 A_h^{FDM}. \quad (3.1.36)$$

Für die Elemente des zugehörigen „Lastvektors“ gilt entsprechend:

$$b_i^{FEM} = \int_{\Omega} f \psi_i \, dx \approx h^2 f(P_i) + \mathcal{O}(h^4) = h^2 b_i^{FDM} + \mathcal{O}(h^4). \quad (3.1.37)$$

Dies zeigt, dass aus algebraischer Sicht FEM und FDM eng verwandt sind. Für eine stückweise „bi-linearen“ Ansatz auf einer gleichförmigen Quadratzerlegung erhält man ein Analogon zu einem „kompakten 9-Punkte-Operator“ (siehe Abb. 3.4):

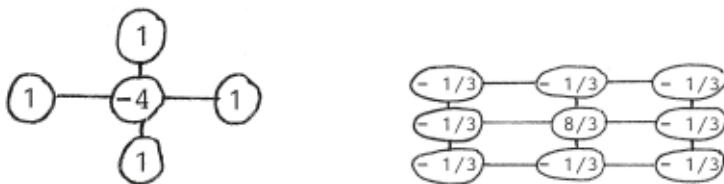


Abbildung 3.4: Differenzensterne: „5-Punkte-Stern“ (links) zur Approximation von  $\Delta$  und „9-Punkte-Stern“ (rechts) zur Approximation von  $-\Delta$ .

### 3.1.3 Approximation krummer Ränder

Zum Abschluss dieser einführenden Diskussion wollen wir noch darstellen, wie in der FEM krumme Ränder approximiert werden. Dazu sei angenommen, dass der Rand  $\partial\Omega$  regulär genug ist, dass die schwache Lösung  $u \in V := H_0^1(\Omega)$  auch in  $H^2(\Omega)$  ist und der a priori Abschätzung

$$\|u\|_{H^2} \leq c_S \|f\| \quad (3.1.38)$$

genügt.

i) **Der „konvexe Fall“:** Sei  $\Omega \subset \mathbb{R}^2$  ein glatt berandetes, *konvexes* Gebiet. Dieses sei überdeckt durch eine reguläre Triangulierung  $T_h = \{T\}$ , so dass alle Eckpunkte des Polygongebiets

$$\Omega_h := \bigcup \{T \in T_h\} \subset \bar{\Omega},$$

auf dem Rand  $\partial\Omega$  liegen. Die Länge der Polygonkanten von  $\partial\Omega$  ist dann durch die Gitterweite  $h$  der Triangulierung  $T_h$  beschränkt (siehe Abb. 3.5).

Auf  $\bar{\Omega}$  wird nun zunächst der einfachste Finite-Elemente-Ansatz (mit linearen Formfunktionen) wie folgt definiert:

$$V_h^{(1)} := \{v_h \in C(\bar{\Omega}) \mid v_{h|T} \in P_1(T), T \in T_h, v_{h|\Omega \setminus \Omega_h} \equiv 0\} \subset V = H_0^1(\Omega).$$

Die zugehörigen Galerkin-Approximationen  $u_h \in V_h^{(1)}$  sind durch die Variationsgleichung

$$(\nabla u_h, \nabla \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h \quad (3.1.39)$$

bestimmt. Wegen der Teilraumbeziehung  $V_h \subset V$  gilt dann wieder für den Fehler  $e_h := u - u_h$  die Bestapproximationsbeziehung (3.1.26).

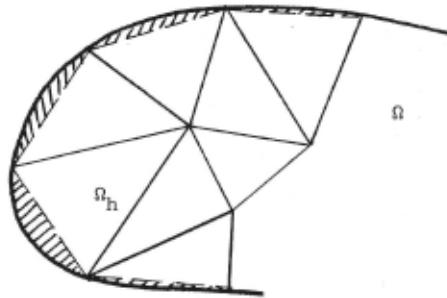


Abbildung 3.5: Polygonale Approximation eines krumm berandeten Gebiets; Randstreifen  $S_h = \cup\{S_T\}$  schraffiert.

**Satz 3.3 (FEM auf konvexem Gebiet):** Für das FE-Schema (3.1.39) auf einem glatt berandeten, konvexen Gebiet  $\Omega$  gelten die a priori Konvergenzabschätzungen

$$\|\nabla e_h\| \leq (c_I + c_\Omega) c_S h \|f\|, \quad (3.1.40)$$

$$\|e_h\| \leq (c_I + c_\Omega)^2 c_S^2 h^2 \|f\|, \quad (3.1.41)$$

mit Stabilitäts- und Interpolationskonstanten  $c_S$ ,  $c_I$  und einer generischen Konstante  $c_\Omega$ , die nur vom Gebiet  $\Omega$  abhängt.

**Beweis:** i) Sei  $I_h u \in V_h^{(1)}$  die natürliche Knoteninterpolierende von  $u$ , welche auf dem Streifen  $S_h = \Omega \setminus \Omega_h$  zu Null gesetzt wird. Für diese gilt wieder auf jeder Zelle  $T \in T_h$  (wird später gezeigt werden)

$$\|\nabla(u - I_h u)\|_T \leq c_I h_T \|\nabla^2 u\|_T, \quad (3.1.42)$$

und folglich

$$\|\nabla(u - I_h u)\|_{\Omega_h} \leq c_I h \|\nabla^2 u\|_{\Omega_h}.$$

Mit Hilfe der Approximationsbeziehung (3.1.26) ergibt sich somit

$$\|\nabla e_h\|_{\Omega}^2 \leq c_I^2 h^2 \|\nabla^2 u\|_{\Omega_h}^2 + \|\nabla u\|_{S_h}^2. \quad (3.1.43)$$

Es bleibt, das Integral über den Randstreifen  $S_h$  zu behandeln.

ii) Für ein glatt berandetes Gebiet (d. h.:  $\partial\Omega$  ist  $C^2$ -parametrisiert.) ist nun  $|S_h| = \mathcal{O}(h^2)$ . Um dies zu sehen, nehmen wir an, dass der Randabschnitt  $\partial\Omega_T$ , welcher durch  $\Gamma$  von  $\partial\Omega$  abgetrennt wird, als Graph einer Funktion  $\psi(s)$  der Bogenlänge über  $\Gamma$  aufgefasst werden kann. Diese nehme ihr Maximum  $\psi_0$  für  $s = s_0$  an, so dass

$$(\psi - \psi_0)(s_0) = 0, \quad (\psi - \psi_0)'(s_0) = 0.$$

Durch Taylor-Entwicklung von  $\psi - \psi_0$  um  $s_0$  ergibt sich dann

$$\max_{\Gamma} |\psi(s)| = \max_{\Gamma} |\psi(s) - \psi_0| \leq \delta := \frac{1}{2} \max_{\Gamma} |\psi''| h_T^2.$$

Folglich ist  $|S_h| \leq ch^2$ . Damit ist bewiesen, dass

$$\|\nabla e_h\|_{\Omega} \leq ch \|u\|_{H^{2,p}}. \quad (3.1.44)$$

iii) Zur Abschätzung des Integrals über  $S_h$  gehen wir ähnlich vor wie beim Beweis der Poincaréschen Ungleichung. Sei  $T \in \mathbb{T}_h$  ein Randdreieck und  $S_T$  der zugehörige Teilstreifen des Randstreifens  $S_h$ , welcher von der Kante  $\Gamma$  von  $T$  begrenzt ist. O.B.d.A. sei angenommen, dass ein Rechteck  $Q_T$  mit  $\Gamma$  als kurzer Seite und Länge  $L > 0$  (unabhängig von  $h$ ) ganz in  $\bar{\Omega}$  enthalten ist (s. Abb. 3.6).

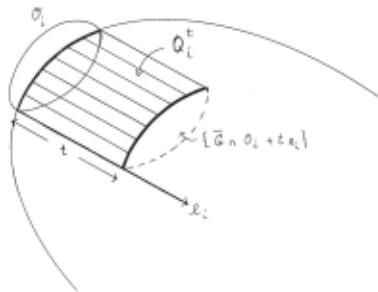


Abbildung 3.6: Schema der Randapproximation

Sei nun weiter  $v \in C^1(\bar{\Omega})$  beliebig. Es bezeichne  $n_{\Gamma}(x)$  den bzgl.  $S_h$  nach innen gerichteten Normaleneinheitsvektor zu  $\Gamma$  im Punkt  $x \in \Gamma$  mit Parameterwert  $s$ . Damit gilt für  $0 \leq t \leq \delta$ :

$$v(x + tn_{\Gamma}(x)) = v(x) + \int_0^t \partial_r v(x + rn_{\Gamma}(x)) dr,$$

bzw.

$$|v(x + tn_\Gamma(x))|^2 \leq 2|v(x)|^2 + 2\delta \int_0^{\psi(s)} |\nabla v(x + rn_\Gamma(x))|^2 dr.$$

Wir integrieren dies zunächst über  $0 \leq t \leq \psi(s) \leq \delta$  (Man beachte, dass für  $x \in \Gamma$  gilt:  $x + \psi(s)n_\Gamma(x) \in \partial\Omega_T$ ),

$$\int_0^{\psi(s)} |v(x + tn_\Gamma(x))|^2 dt \leq 2\delta|v(x)|^2 + 2\delta^2 \int_0^{\psi(s)} |\nabla v(x + rn_\Gamma(x))|^2 dr,$$

und dann über  $x \in \Gamma$  und erhalten

$$\int_{S_T} |v(x)|^2 dx \leq 2\delta \int_\Gamma |v(x)|^2 ds + 2\delta^2 \int_{S_T} |\nabla v(x)|^2 dx.$$

Für das Randintegral rechts erhalten wir mit Hilfe der Spurabschätzung

$$\int_\Gamma |v(x)|^2 ds \leq c_\Omega \|v\|_{H^1(Q_T)}^2$$

mit einer von  $\max_\Gamma |\psi'|$  abhängigen Konstante  $c_\Omega$ . Damit gewinnen wir schließlich

$$\int_{S_h} |v(x)|^2 dx \leq c_\Omega h^2 \|v\|_{H^1(\Omega)}^2, \quad (3.1.45)$$

mit einer generischen Konstante  $c_\Omega$ . Durch das übliche Stetigkeitsargument überträgt sich diese Abschätzung auf alle Funktionen  $v \in H^1(\Omega)$ . Wir wenden die Abschätzung (3.1.45) nun für die Funktion  $|\nabla u| \in H^1(\Omega)$  an und erhalten

$$\|\nabla u\|_{S_h}^2 \leq c_\Omega h^2 \|u\|_{H^2}^2,$$

so dass sich mit (3.1.43) schließlich das erste gewünschte Resultat ergibt:

$$\|\nabla e\| \leq c_I h \|\nabla^2 u\| + c_\Omega h \|u\|_{H^2} \leq (c_I + c_\Omega) c_S h \|f\|. \quad (3.1.46)$$

iv) Zur Abschätzung des  $L^2$ -Fehlers wird wieder ein Dualitätsargument verwendet. Sei  $z \in H_0^1(\Omega) \cap H^2(\Omega)$  die Lösung des Hilfsproblems

$$-\Delta z = \|e\|^{-1} e \quad \text{in } \Omega, \quad z|_{\partial\Omega} = 0.$$

Wie im Fall eines Polygongebiets argumentieren wir nun wie folgt:

$$\|e\| = (\nabla e, \nabla z) = (\nabla e, \nabla(z - I_h z)) \leq \|\nabla e\| \|\nabla(z - I_h z)\|.$$

Mit Hilfe der Interpolationsabschätzung (3.1.42) sowie der Abschätzung (3.1.45) für  $v := |\nabla z|$  folgt weiter

$$\|e\| \leq \|\nabla e\| \{c_I h \|\nabla^2 z\| + c_\Omega h \|z\|_{H^2}\} \leq (c_I + c_\Omega) h \|\nabla e\| \|z\|_{H^2}.$$

Hiermit folgt dann unter Verwendung des ersten Resultats (3.1.46) sowie der a priori Schranke  $\|z\|_{H^2} \leq c_S$  auch die zweite gewünschte Ungleichung

$$\|e\| \leq (c_I + c_\Omega)^2 c_S^2 h^2 \|f\|. \quad (3.1.47)$$

Dies vervollständigt den Beweis.

Q.E.D.

**ii) Der „nicht-konvexe Fall“:** Sei  $\Omega \subset \mathbb{R}^2$  ein glatt berandetes, aber nicht notwendig *konvexes* Gebiet. Dieses sei wieder überdeckt durch eine reguläre Triangulierung  $T_h = \{T\}$ , so dass alle Eckpunkte des Polygongebiets  $\Omega_h := \bigcup\{T \in T_h\}$  auf dem Rand  $\partial\Omega$  liegen (s. Abb. 3.7). Die Länge der Polygonkanten von  $\partial\Omega_h$  ist dann durch die Gitterweite  $h$  der Triangulierung  $T_h$  beschränkt. Ist  $\Omega$  nicht konvex, so ist  $\Omega_h \not\subset \bar{\Omega}$ . Der auf  $\Omega_h = \bigcup\{T \in T_h\}$  definierte Finite-Elemente-Raum  $V_h^{(1)}$  ist dann auch nicht in  $V$  enthalten, und die Approximation wird „nicht-konform“ (bzgl.  $V = H_0^1(\Omega)$ ) genannt. Die Analyse dieser Approximation gestaltet sich technisch etwas schwieriger als im konformen Fall. Doch auch hierfür kann man Konvergenz mit der optimalen Ordnung beweisen.

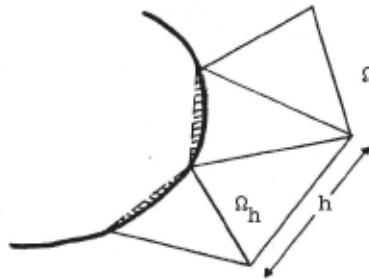


Abbildung 3.7: Approximation nicht-konvexer Randteile

**Satz 3.4 (FEM auf nicht-konvexem Gebiet):** Für das FE-Schema (3.1.39) auf einem glatt berandeten, nicht-notwendig konvexen Gebiet  $\Omega$  gelten die a priori Konvergenzabschätzungen

$$\|\nabla e_h\|_\Omega \leq (c_I + c_\Omega)c_S h\|f\|_\Omega \quad (3.1.48)$$

$$\|e_h\|_\Omega \leq (c_I + c_\Omega)^2 c_S^2 h^2 \|f\|_\Omega, \quad (3.1.49)$$

mit Stabilitäts- und Interpolationskonstanten  $c_S$ ,  $c_I$  und einer generischen Konstante  $c_\Omega$ , die nur vom Gebiet  $\Omega$  abhängt.

**Beweis:** i) Das Problem beim Beweis des Satzes besteht in der „Nicht-Konformität“ der Diskretisierung, d. h.:  $V_h^{(1)} \not\subset V$ . Zunächst gilt auch im Fall eines nichtkonvexen Gebiets

$$\|\nabla(u - I_h u)\|_\Omega \leq c_I h \|\nabla^2 u\|_\Omega \leq c_I c_S h \|f\|_\Omega. \quad (3.1.50)$$

Weiter haben wir

$$\begin{aligned}\|\nabla e_h\|_\Omega^2 &= (\nabla e_h, \nabla(u - I_h u))_\Omega + (\nabla e_h, \nabla(I_h u - u_h))_\Omega \\ &= (\nabla e_h, \nabla(u - I_h u))_\Omega + (\nabla u, \nabla(I_h u - u_h))_\Omega - (\nabla u, \nabla(I_h u - u_h))_\Omega \\ &= (\nabla e_h, \nabla(u - I_h u))_\Omega + (\nabla u, \nabla(I_h u - u_h))_\Omega - (f, I_h u - u_h)_\Omega \\ &\leq (\nabla e_h, \nabla(u - I_h u))_\Omega + N_h(u) \|\nabla(I_h u - u_h)\|_\Omega,\end{aligned}$$

mit

$$N_h(u) := \sup_{\psi_h \in V_h^{(1)}} \frac{|(\nabla u, \nabla \psi_h)_\Omega - (f, \psi_h)_\Omega|}{\|\nabla \psi_h\|_\Omega}.$$

Hieraus ergibt sich

$$\|\nabla e_h\|_\Omega^2 \leq 3\|\nabla(u - I_h u)\|_\Omega^2 + 2N_h(u)^2.$$

ii) Zur Abschätzung des Nichtkonformitätsterms schätzen wir wie folgt ab:

$$\begin{aligned}(\nabla u, \nabla \psi_h)_\Omega - (f, \psi_h)_\Omega &= (-\Delta u - f, \psi_h)_\Omega + (\partial_n u, \psi_h)_{\partial\Omega} \\ &= (\partial_n u, \psi_h)_{\partial\Omega} \leq \|\partial_n u\|_{\partial\Omega} \|\psi_h\|_{\partial\Omega}.\end{aligned}$$

Zunächst gilt aufgrund einer bekannten Spurabschätzung und der üblichen Regularitätsabschätzung auf glatt-berandeten Gebieten

$$\|\partial_n u\|_{\partial\Omega} \leq c_\Omega \|u\|_{H^2} \leq c_\Omega c_S \|f\|_\Omega. \quad (3.1.51)$$

Zur Behandlung des Terms  $\|\psi_h\|_{\partial\Omega}$  benötigen wir etwas zusätzliche Notation. Da wir nur lineare oder (isoparametrische) bilineare Ansatzfunktionen betrachten, gibt es auf  $\partial\Omega$  ein stetiges (nach außen orientiertes) Richtungsvektorfeld  $r(x)$ , so dass jedes  $\psi_h \in V_h^{(1)}$  entlang der von  $r(x)$  aufgespannten Gerade  $\{x + sr(x), s \in \mathbb{R}\}$  linear ist. Für den Schnittpunkt  $x_h = x + d(x)r(x)$  dieser Gerade ausgehend von  $x \in \partial\Omega$  mit dem Rand  $\partial\Omega_h$  gilt wieder  $|x_h - x| = d(x) \leq c_\Omega h^2$ . Durch Taylor-Entwicklung folgt dann

$$0 = \psi_h(x_h) = \psi_h(x) + \partial_r \psi_h(x) d(x),$$

und bei Integration über  $\partial\Omega$ :

$$\|\psi_h\|_{\partial\Omega} \leq c_\Omega h^2 \|\partial_r \psi_h\|_{\partial\Omega}. \quad (3.1.52)$$

Dies impliziert insbesondere, dass

$$\|\psi_h\|_{\partial\Omega} \leq c_\Omega h \|\nabla \psi_h\|_\Omega,$$

und damit

$$N_h(u) \leq c_\Omega c_S h \|f\|_\Omega.$$

Dies in Verbindung mit der Interpolationsfehlerabschätzung (3.1.50) beweist die Energienormfehlerabschätzung.

iii) Zum Beweis der  $L^2$ -Normfehlerabschätzung verwenden wir wieder ein Dualitätsragu-

ment. Sei  $z \in H_0^1(\Omega) \cap H^2(\Omega)$  die (eindeutig bestimmte) Lösung des Hilfsproblems

$$-\Delta z = e_h \|e_h\|^{-1} \text{ in } \Omega, \quad z|_{\partial\Omega} = 0. \quad (3.1.53)$$

Auf dem Gebiet  $\Omega$  gilt die a priori Abschätzung  $\|z\|_{H^2} \leq c_\Omega \|\Delta z\|_\Omega \leq c_\Omega$ . Damit gilt

$$\begin{aligned} \|e_h\|_\Omega &= (e_h, -\Delta z)_\Omega = (\nabla e_h, \nabla z)_\Omega - (e_h, \partial_n z)_{\partial\Omega} \\ &= (\nabla e_h, \nabla(z - I_h z))_\Omega + (\nabla e_h, \nabla I_h z)_\Omega - (e_h, \partial_n z)_{\partial\Omega} \\ &= (\nabla e_h, \nabla(z - I_h z))_\Omega + (\nabla u, \nabla I_h z)_\Omega - (f, I_h z)_\Omega - (e_h, \partial_n z)_{\partial\Omega} \\ &= (\nabla e_h, \nabla(z - I_h z))_\Omega + (\partial_n u, I_h z)_{\partial\Omega} + (u_h, \partial_n z)_{\partial\Omega}. \end{aligned}$$

Wie oben unter (i) und (ii) schätzen wir ab

$$|(\nabla e_h, \nabla(z - I_h z))_\Omega| \leq \|\nabla e_h\|_\Omega \|\nabla(z - I_h z)\|_\Omega \leq c_I h \|\nabla e_h\|_\Omega \|\nabla^2 z\|_\Omega \leq c_I c_S h \|\nabla e_h\|_\Omega.$$

Weiter folgt mit Hilfe der Abschätzung (3.1.52) und den schon oben verwendeten Interpolations-, Spur- und Regularitätsabschätzungen:

$$\begin{aligned} |(\partial_n u, I_h z)_{\partial\Omega}| &\leq \|\partial_n u\|_{\partial\Omega} \|I_h z\|_{\partial\Omega} \leq c_\Omega c_S \|f\|_\Omega h^2 \|\partial_r I_h z\|_{\partial\Omega} \\ &\leq c_\Omega c_S \|f\|_\Omega h^2 \{\|\partial_r(I_h z - z)\|_{\partial\Omega} + \|\partial_r z\|_{\partial\Omega}\} \\ &\leq c_\Omega c_S \|f\|_\Omega h^2 \{c_I h^{1/2} \|z\|_{H^2} + c_\Omega \|z\|_{H^2}\} \leq c_\Omega c_I c_S h^2 \|f\|_\Omega. \end{aligned}$$

Analog erschließen wir

$$\begin{aligned} |(u_h, \partial_n z)_{\partial\Omega}| &\leq \|u_h\|_{\partial\Omega} \|\partial_n z\|_{\partial\Omega} \leq c_\Omega c_S h^2 \|\partial_r u_h\|_{\partial\Omega} \|\Delta z\|_\Omega \\ &\leq c_\Omega c_S h^2 \{\|\partial_r(u_h - I_h u)\|_{\partial\Omega} + \|\partial_r(I_h u - u)\|_{\partial\Omega} + \|\partial_r u\|_{\partial\Omega}\} \\ &\leq c_\Omega c_S h^2 \{h^{-1/2} \|\nabla(u_h - I_h u)\|_\Omega + c_I h^{1/2} \|u\|_{H^2} + c_\Omega \|u\|_{H^2}\} \\ &\leq c_\Omega c_S h^2 \{h^{-1/2} \|\nabla e_h\|_\Omega + h^{-1/2} \|\nabla(u - I_h u)\|_\Omega + \|\partial_r(I_h u - u)\|_{\partial\Omega} + \|\partial_r u\|_{\partial\Omega}\} \\ &\leq c_\Omega c_S h^2 \{(c_I + c_\Omega) c_S h^{1/2} \|f\|_\Omega + c_I c_\Omega c_S h^{1/2} \|f\|_\Omega + c_\Omega c_S \|f\|_\Omega\} \\ &\leq (c_I + c_\Omega)^2 c_S^2 h^2 \|f\|_\Omega. \end{aligned}$$

Combination der obigen Abschätzungen liefert schließlich

$$\|e_h\|_\Omega \leq c_I c_S h \|\nabla e_h\|_\Omega + c_\Omega c_I c_S h^2 \|f\|_\Omega + (c_I + c_\Omega)^2 c_S^2 h^2 \|f\|_\Omega.$$

Dies zusammen mit der schon bewiesenen Energienormfehlerabschätzung (3.1.48) ergibt die behauptete  $L^2$ -Fehlerabschätzung. Q.E.D.

## 3.2 Allgemeine Finite-Elemente-Ansätze

Wir wollen nun Finite-Elemente-Ansatzräume allgemeineren Typs konstruieren und Fragen der praktischen Realisierung der Methode diskutieren. Zunächst wird  $\Omega$  als ein Polygonegebiet (Polyeder in 3-D) angenommen. Seien  $T_h$  Zerlegungen von  $\overline{\Omega}$  in Dreiecke oder

Vierecke (Tetraeder oder Hexaeder in 3-D), welche den im vorigen Abschnitt formulierten Bedingungen genügen. Für die folgenden Konstruktionen von Finite-Elemente-Ansätzen verwenden wir die Bezeichnungen

$$P_r := \left\{ p(x) = \sum_{0 \leq i+j \leq r} c_{ij} x_1^i x_2^j \right\}, \quad Q_r := \left\{ q(x) = \sum_{0 \leq i,j \leq r} c_{ij} x_1^i x_2^j \right\},$$

für Polynom-Vektorräume im  $\mathbb{R}^2$  (analog für solche im  $\mathbb{R}^3$ ). Ein Finite-Elemente-Ansatz ist definiert durch Vorgabe eines Polynomraumes  $P(T) \subset P_r(T)$  oder  $P(T) \subset Q_r(T)$  auf  $T \in \mathbb{T}_h$  sowie eines Satzes von „Knotenwerten“ (gegeben durch lineare „Knotenfunktionale“), z. B.:

$$\{v_h(a), v_h(m), \partial_n v_h(m), \nabla v_h(a), (v_h, 1)_\Gamma, (v_h, 1)_T, \dots\},$$

welcher Polynome aus  $P(T)$  eindeutig bestimmt.

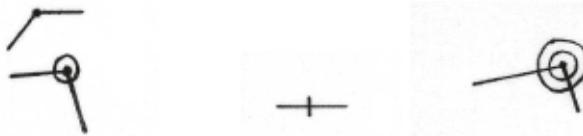


Abbildung 3.8: *Funktionswerte  $v_h(a)$  sowie  $v_h(a), \nabla v_h(a)$  (links), Normalableitung  $\partial_n v_h(m)$  in Seitenmittelpunkten (Mitte), Funktionswerte  $v_h(a), \nabla v_h(a), \nabla^2 v_h(a)$  (rechts).*

Wir verwenden die Bezeichnungen

$$\begin{aligned} \mathbb{T}_h &= \{T\} \quad \text{Zerlegung von } \bar{\Omega}, \\ \partial \mathbb{T}_h &= \{\Gamma\} \quad \text{Menge aller Kanten (bzw. Flächen)}, \\ \partial^2 \mathbb{T}_h &= \{a\} \quad \text{Menge aller Eckpunkte („Knoten“)}, \end{aligned}$$

sowie die in Abb. 3.8 skizzierte Symbolik für einige typische Knotenwertvorgaben.

**Definition 3.1 (Unisolvenz):** Ein Polynomraum  $P(T)$  und ein zugehöriger Satz von linearen „Knotenfunktionalen“  $K(T)$  heißen „unisolvent“, wenn jedes  $p \in P(T)$  eindeutig durch die Vorgabe von  $\chi(p)$  für alle  $\chi \in K(T)$  bestimmt ist.

**Definition 3.2 (Lagrange- und Hermite-Ansatz):** Man spricht bei einem FE-Ansatz  $P(T)$  mit zugehörigem Satz von Knotenfunktionalen  $K(T) = \{\chi_r, r = 1, \dots, R\}$  von „Lagrange-Elementen“, wenn die Knotenfunktionale nur auf Funktionswerte zurückgreifen; werden auch Ableitungswerte verwendet, spricht man von „Hermite<sup>10</sup>-Elementen“.

<sup>10</sup>Charles Hermite (1822–1901): Französischer Mathematiker; Prof. an der École Polytechnique und der Sorbonne in Paris; Beiträge zur Zahlentheorie und zur Theorie elliptischer Funktionen; Beweis der Transzendenz von  $e$ .

Notwendig für Unisolvenz ist offenbar  $\dim P(T) = \#K(T)$  und hinreichend, dass für ein  $p \in P(T)$  aus  $\chi(p) = 0$  für alle  $\chi \in K(T)$  notwendig  $p \equiv 0$  folgt. Dies wird in der Regel zum Nachweis von Unisolvenz verwendet.

**Definition 3.3 (Interpolation):** Für jede Zelle  $T \in \mathbb{T}_h$  sei ein Polynomraum  $P(T)$  mit Dimension  $R$  und ein Satz  $K_T = \{\chi_r, r = 1, \dots, R\}$  von Knotenfunktionalen

$$\chi_r : H^m(T) \rightarrow \mathbb{R} \quad (r = 1, \dots, R),$$

spezifiziert, welche „unisolvent“ sind. Durch die Vorgabe

$$\chi_r(I_h v) = \chi_r(v), \quad r = 1, \dots, R,$$

ist dann eindeutig eine „Finite-Elemente-Interpolierende“  $I_h v \in P(T)$  definiert.

Durch Zusammensetzen der zunächst zellweise definierten Formfunktionen  $v_T \in P(T)$  erhält man global definierte Funktionen

$$v_h : \overline{\Omega} \rightarrow \mathbb{R}, \quad v_{h|T} := v_T, \quad T \in \mathbb{T}_h,$$

mit denen der Finite-Elemente-Ansatzraum  $V_h$  gebildet wird. Durch Gleichsetzen geeigneter Knotenwerte auf dem gemeinsamen Rand  $\Gamma = T \cap T'$  jeweils benachbarter Zellen wird Stetigkeit, Differenzierbarkeit und auf analogem Wege auch die Randbedingung  $v_h|_{\partial\Omega} = 0$  implementiert. Die Dimension des endlich dimensionalen Teilraumes  $V_h \subset V$  ist dann gleich der Anzahl der Knotenfunktionalwerte zur eindeutigen Festlegung einer Funktion  $v_h \in V_h$ .

**Definition 3.4 (Konformität):** Ein Finite-Elemente-Ansatzraum  $V_h$  dieser Art heißt „ $H_0^1$ -konform“, wenn  $V_h \subset H_0^1(\Omega)$ , und andernfalls „nicht-konform“.

**A) Dreieckselemente im  $\mathbb{R}^2$ :** Als erstes betrachten wir Finite-Elemente-Ansätze auf Triangulierungen von Polygonegebieten.

1) Konstanter Ansatz:  $P(T) = P_0$ ,  $\dim P(T) = 1$ .

Mit konstanten Polynomansätzen kann offensichtlich für die Variationsgleichung (3.1.2) keine konforme Diskretisierung gewonnen werden. Diese spielen aber bei anderen Typen von variationellen Formulierungen, den sog. „dual-gemischten“, eine Rolle

2a) Linearer Ansatz:  $P(T) = P_1$ ,  $\dim P(T) = 3$ .

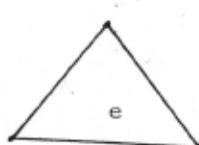


Abbildung 3.9: Knotenpunkte des (konformen) linearen Ansatzes ( $e = T$ )

Der Polynomraum  $P_1(T)$  und der Satz von Knotenfunktionalen  $\{\chi_i(p) = p(a_i), i = 1, 2, 3\}$  sind unisolvant, denn für  $p \in P_1(T)$  mit  $p(a_i) = 0$  gilt notwendig  $p|_{\partial T} \equiv 0$  und damit auch  $p \equiv 0$ . Mit dem Ansatz

$$V_h^{(1)} = \{v_h : \bar{\Omega} \rightarrow \mathbb{R} \mid v_h|_T \in P_1, v_h \text{ stetig in Eckpunkten, } v_h = 0 \text{ in Eckpunkten auf } \partial\Omega\}.$$

erhält man einen  $H_0^1$ -konformen Finite-Elemente-Raum. Denn der Sprung von  $[v_h]$  über eine gemeinsame Kante  $\Gamma$  zweier Dreiecke  $T_1$  und  $T_2$  ist in  $P_1(\Gamma)$ . Folglich ist  $v_h$  stetig, da seine Restriktionen auf  $T_1$  und  $T_2$  gemeinsame Werte in den beiden Endpunkten von  $\Gamma$  haben.

2b) Einen *nicht-konformen* linearen Ansatz erhält man durch

$$\tilde{V}_h^{(1)} = \{v_h : \bar{\Omega} \rightarrow \mathbb{R} \mid v_h|_T \in P_1, v_h \text{ stetig in Kantenmitten, } v_h = 0 \text{ in Kantenmitten auf } \partial\Omega\}.$$

Die Unisolvanz folgt analog wie die des entsprechenden konformen Ansatzes. Dieses nicht-konforme „lineare“ Element spielt z. B. eine Rolle bei der Diskretisierung der Navier-Stokes-Gleichungen in der Strömungsmechanik.

3) *Quadratischer Ansatz*:  $P(T) = P_2$ ,  $\dim P(T) = 6$ .

Der Polynomraum  $P_2(T)$  und der Satz von Knotenfunktionalen  $\{\chi_i(p) = p(a_i), \psi_i(p) = p(m_i), i = 1, 2, 3\}$  sind unisolvant. Für  $p \in P_2(T)$  mit  $p(a_i) = p(m_i) = 0$  gilt notwendig  $p|_{\partial T} \equiv 0$ . Dies impliziert, dass  $\nabla p(a_i) = 0$ , woraus wegen  $\partial_i p \in P_1(T)$  wiederum  $\nabla p \equiv 0$  und schließlich  $p \equiv 0$  folgt. Mit dem Ansatz

$$V_h^{(2)} = \{v_h : \bar{\Omega} \rightarrow \mathbb{R} \mid v_h|_T \in P_2, v_h \text{ stetig in Eckpunkten und Kantenmitten, } v_h = 0 \text{ in solchen Punkten auf } \partial\Omega\}.$$

erhält man einen  $H_0^1$ -konformen Finite-Elemente-Raum. Denn der Sprung von  $[v_h]$  über eine gemeinsame Kante  $\Gamma$  zweier Dreiecke  $T_1$  und  $T_2$  ist in  $P_2(\Gamma)$ . Folglich ist  $v_h$  stetig, da seine Restriktionen auf  $T_1$  und  $T_2$  gemeinsame Werte in den beiden Endpunkten und dem Mittelpunkt von  $\Gamma$  haben. Alternativ zu den Werten in den Kantenmitten  $m$  kann man auch die Mittelwerte  $|\Gamma|^{-1} \int_{\Gamma} v_h ds$  über Kanten  $\Gamma \in \partial T_h$  als Knotenwerte verwenden.

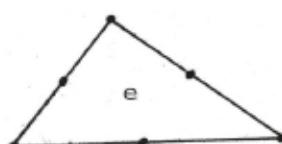


Abbildung 3.10: Knotenpunkte des (konformen) quadratischen Ansatzes ( $e = T$ )

Der naheliegende *nicht-konforme* Ansatz

$$\tilde{V}_h^{(2)} = \{v_h : \bar{\Omega} \rightarrow \mathbb{R} \mid v_h|_T \in P_2, v_h \text{ stetig in jeweils zwei Gauß-Punkten auf Kanten}, \\ v_h = 0 \text{ in solchen Punkten auf } \partial\Omega\}$$

ist *nicht* unisolvent, denn es existieren nicht-triviale, stückweise quadratische Funktionen, welche in allen Knotenpunkten verschwinden. Man betrachte dazu auf jeder Kante die Legendre-Polynome  $L_2$  zweiten Grades, deren Nullstellen gerade die beiden Gauß-Punkte  $m_1, m_2$  sind. Sie können wegen ihrer Symmetrie zum Kantenmittelpunkt so normiert werden, dass sie in den Eckpunkten  $a$  gleiche Werte  $L_2(a) = 1$  haben. Dann lässt sich eine Funktion  $L \in P_2(T)$  finden, durch konforme Interpolation in Eckpunkten und Seitenmitten:  $L(a) = 1, L(m) = L_2(m)$ . Auf jeder Kante ist dann  $L \equiv L_2$ , so dass sich ein Widerspruch zur Unisolvenz des Ansatzes ergibt.

Eine weitere *nicht-konforme* Variante des quadratischen Elements (das sog. „Morley<sup>11</sup>-Plattenelement“) erhält man bei Wahl der Knotenwerte  $\{v_h(a), \partial_n v_h(m)\}$ . Dieser Ansatz ist wieder unisolvent. Für  $p \in P_2(T)$  folgt aus  $p(a_i) = \partial_n p(m_i) = 0, i = 1, 2, 3$ , notwendig  $\partial_\tau p(m_i) = 0$  und somit  $\nabla p(m_i) = 0$ . Wegen  $\partial_i p \in P_1(T)$  folgt  $\nabla p \equiv 0$  und damit auch  $p \equiv 0$ .

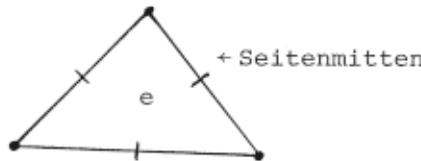


Abbildung 3.11: Knotenpunkte des (nicht-konformen) quadratischen „Morley-Elements“ ( $e = T$ )

Das Morley-Element ist zwar nicht-konform bzgl. der  $H^1$ - wie auch der  $H^2$ -Norm, trotzdem kann es bei geeigneter Modifikation der variationellen Formulierungen,

$$u_h \in V_h^M : \quad \sum_{T \in \mathbb{T}_h} (\nabla^2 u_h, \nabla^2 \varphi_h)_T = (f, \varphi_h) \quad \forall \varphi_h \in V_h^M,$$

sogar zur Approximation des biharmonischen Operators  $\Delta^2 u = f$  verwendet werden.

5) *Kubischer Ansatz* (sog. „kubisches Membran-Element“:  $P(T) = P_3$ ,  $\dim P(T) = 10$ ). Mit dem Ansatz

$$V_h^{(3)} = \{v_h : \bar{\Omega} \rightarrow \mathbb{R} \mid v_h|_T \in P_3, v_h \text{ stetig in Eckpunkten und in je} \\ \text{zwei Gauß-Punkten auf Kanten}, v_h = 0 \text{ in solchen Punkten auf } \partial\mathbb{T}_h\}.$$

<sup>11</sup>Leslie Sydney Dennis Morley (1924–2011): Englischer Ingenieur; wirkte an der Brunel University in Uxbridge, England; Beiträge zur FEM für nichtlineare Schalenmodelle.

ist  $H_0^1$ -konform. Denn der Sprung von  $[v_h]$  über eine gemeinsame Kante  $\Gamma$  zweier Dreiecke  $T_1$  und  $T_2$  ist in  $P_3(\Gamma)$ . Folglich ist  $v_h$  stetig, da seine Restriktionen auf  $T_1$  und  $T_2$  gemeinsame Werte in den beiden Endpunkten und zwei Gauß-Punkten auf  $\Gamma$  haben. Alternativ zu den Gauß-Punkten auf den Kanten kann man auch in jedem Knoten die beiden partiellen Ableitungen, d.h. den Gradienten  $\nabla v_h(a)$ , als Knotenwerte verwenden. Auch damit erhält man einen  $H_0^1$ -konformen Ansatzraum  $\hat{V}_h^{(3)}$ ; dieser ist offenbar echt kleiner als  $V_h^{(3)}$  (Übungsaufgabe).



Abbildung 3.12: Knotenpunkte der (konformen) kubischen Ansätze ( $e = T$ )

Zur Diskretisierung der biharmonischen Gleichung kann man wieder eine *nicht-konforme* Variante mit den Knotenwerten  $\{v_h(a), \partial_n v_h(m_1), \partial_n v_h(m_2), v_h(z)\}$  ( $m_1, m_2$  zwei Gauß-Punkte auf  $\Gamma$  und  $z$  der Mittelpunkt von  $T$ ) verwenden. Ein  $H^2$ -konformes Platten-Element erhält man durch den sog. „Clough-Tocher-Ansatz“ als „zusammengesetztes“ Element ( $v_h \in C^1(T)$  stückweise kubisch).

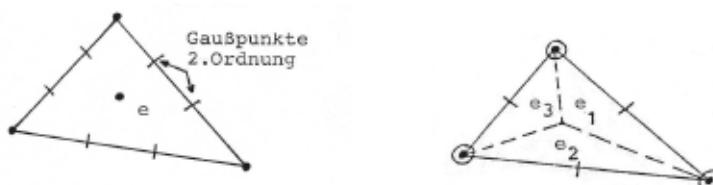


Abbildung 3.13: Knotenpunkte des (nicht-konformen) kubischen „Platten-Elemente“ ( $e_i = T_i$ )

6) *Quartischer Ansatz:*  $P(T) = P_4(T)$ ,  $\dim P(T) = 15$ .

Mit dem Satz von Knotenwerten  $\{v_h(a), \nabla v_h(a), v_h(m_1), v_h(m_2)\}$  ( $m_1, m_2$  die beiden Gauß-Punkte auf jeder Kante  $\Gamma \in \partial T_h$ ) erhält man hier einen  $H^1$ -konformen Ansatz (Übungsaufgabe).

7) *Quintischer Ansatz (sog. „Argyris<sup>12</sup> -Plattenelement“):*  $P(T) = P_5(T)$ ,  $\dim P(T) = 21$ . Mit dem Satz von Knotenwerten  $\{v_h(a), \nabla v_h(a), \nabla^2 v_h(a), \partial_n v_h(m)\}$  erhält man hier sogar einen  $H^2$ -konformen Ansatz (Übungsaufgabe). Dieses finite Element ist ein Beispiel für einen konformen Ansatz zur Lösung der biharmonischen Gleichung  $\Delta^2 u = f$ , für welche stetig differenzierbare Übergänge von Zelle zu Zelle erforderlich sind.

<sup>12</sup>John Hadji Argyris (1913–2004): Griechischer Bauingenieur; Prof. in Stuttgart; einer der „Erfinder“ der Finite-Elemente-Methode.

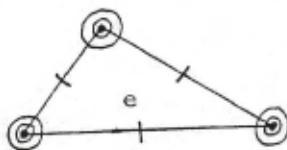


Abbildung 3.14: Knotenpunkte des (konformen) quintischen „Argyris-Elements“ ( $e = T$ )

**B) Viereckselemente:** Als nächstes betrachten wir Finite-Elemente-Ansätze auf (kartesischen) Rechteckszerlegungen.

1) *Bi-linearer Ansatz:*  $P(T) = Q_1 = \text{span}\{1, x_1, x_2, x_1x_2\}$ ,  $\dim P(T) = 4$ .

Der Polynomraum  $Q_1(T)$  und der Satz von Knotenfunktionalewn  $\{\chi_i(p) = p(a_i), i = 1, \dots, 4\}$  sind unisolvent. Ein  $p \in Q_1(T)$  ist entlang der Kanten von  $T$  linear. Aus  $p(a_i) = 0$  folgt also  $p|_{\partial T} \equiv 0$ , und weiter  $\nabla p(a_i) = 0$ . Wegen  $\partial_i p \in P_1(T)$  impliziert dies  $\nabla p \equiv 0$  und schließlich  $p \equiv 0$ . Der Ansatz

$$V_h^{(1)} = \{v_h : \bar{\Omega} \rightarrow \mathbb{R} \mid v_h|_T \in Q_1, v_h \text{ stetig in Eckpunkten, } v_h = 0 \text{ in Eckpunkten auf } \partial\Omega\}$$

ist  $H_0^1$ -konform, da die Sprünge von  $v_h$  entlang von Kanten linear sind.

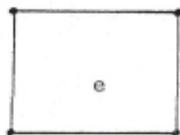


Abbildung 3.15: Knotenpunkte des (konformen) bi-linearen Ansatzes ( $e = T$ )

Der naheliegende *nicht-konforme* Ansatz

$$\tilde{V}_h^{(1)} = \{v_h : \bar{\Omega} \rightarrow \mathbb{R} \mid v_h|_T \in Q_1, v_h \text{ stetig in Kantenmitten, } v_h = 0 \text{ in diesen auf } \partial\Omega\}$$

ist aber i. Allg. *nicht* unisolvent, da z. B. die Funktion  $v_h(x_1, x_2) = x_1x_2$  in den Kantenmitten des Quadrats  $T_1 = [-1, 1] \times [-1, 1]$  verschwindet. Auf  $T_1$  erhält man aber durch  $P(T) = \text{span}\{1, x_1, x_2, x_1^2 - x_2^2\}$  einen mit den Kantenmitten als Knotenfunktionale unisolventen Ansatz. Alternativ zu den Funktionswerten in den Seitenmitten kann man auch die Mittelwerte  $|\Gamma|^{-1} \int_{\Gamma} v_h ds$  über die Kanten als Knotenwerte verwenden. Dies ergibt aber einen von  $\tilde{V}_h^{(1)}$  verschiedenen Ansatzraum. Beide Ansätze sind offensichtlich *nicht-konform*.

2) *Bi-quadratischer Ansatz:*  $P(T) = Q_2 = \text{span}\{1, x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^2x_2, x_1x_2^2, x_1^2x_2^2\}$ ,  $\dim P(T) = 9$ . Die Konstruktion eines „zulässigen“ Satzes von Knotenwerten ist Übung.

3) Reduzierter bi-quadratischer Ansatz (sog. „Wilson<sup>13</sup>-Membran-Element“):

$P(T) = P_2(T) \oplus \text{span}\{x_1^2 x_2, x_1 x_2^2\}$ , dim  $P(T) = 8$ . Dieser Ansatz wird mit den Knotenwerten  $\{v_h(a), v_h(m)\}$   $H^1$ -konform.

4) Bi-kubischer Ansatz:  $P(T) = Q_3 = \text{span}\{1, x_1, x_2, x_1^2, \dots, x_1^3 x_2^3\}$ , dim  $P(T) = 16$ .

Die Konstruktion eines „zulässigen“ Satzes von Knotenwerten wird als Übung gestellt.

5) Reduzierter bi-kubischer Ansatz (sog. „Adini<sup>14</sup>-Platten-Element“):

$P(T) = P_3(T) \oplus \text{span}\{x_1^3 x_2, x_1 x_2^3\}$ , dim  $P(T) = 12$ . Dieser Ansatz wird mit den Knotenwerten  $\{v_h(a), \nabla v_h(a)\}$  unisolvent und  $H^1$ -konform, ist aber nicht  $H^2$ -konform.

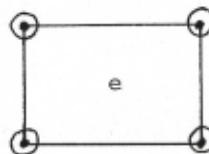


Abbildung 3.16: Knotenpunkte des (nicht-konformen) „Adini-Plattenelments“ ( $e = T$ )

Viele der aufgeführten zwei-dimensionalen Finite-Elemente-Ansätze haben natürliche Erweiterungen auf drei Dimensionen. Die gebräuchlichsten Beispiele sind:

1) Lineares Tetraederelement:  $P(T) = \text{span}\{1, x_1, x_2, x_3\}$ , dim  $P(T) = 4$ . Mit den Funktionswerten in den Eckpunkten als Knotenwerte ist dieser Ansatz unisolvent und  $H^1$ -konform.

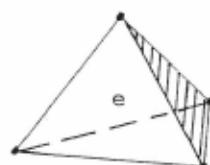


Abbildung 3.17: Knotenpunkte des (konformen) linearen Ansatzes ( $e = T$ )

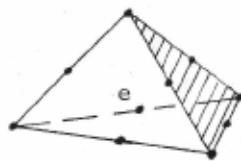
2) Nicht-konf., lineares Tetraederelement:  $P(T) = \text{span}\{1, x_1, x_2, x_3\}$ , dim  $P(T) = 4$ .

Mit den Funktionswerten in den Flächenmitteln als Knotenwerte ist dieser Ansatz unisolvent, aber *nicht-konform*.

3) Quadratisches Tetraederelement:  $P(T) = \text{span}\{1, x_1, x_2, x_3, x_1 x_2, x_1 x_3, x_2 x_3, x_1^2, x_2^2, x_3^2\}$ , dim  $P(T) = 10$ . Mit den Funktionswerten in den Ecken und den Kantenmitteln als Knotenwerte ist dieser Ansatz unisolvent und  $H^1$ -konform.

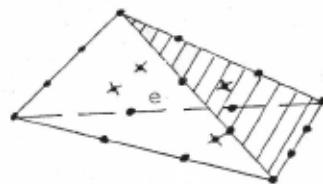
<sup>13</sup>Edward L. Wilson (1931–): US-Amerikanischer Ingenieur; Prof. für Ingenieurwissenschaften an der Univ. of California (Berkeley, USA); ein früher Pionier der (praktischen) Finite-Elemente-Methode; Koautor des Buches „Numerical Methods in Finite Element Analysis“ (zus. mit K. J. Bathe), 1976.

<sup>14</sup>Avner Adini (????–): Promotion 1961 an der Univ. of California (Berkeley, USA) als Bauingenieur; Beiträge u. a. zur Finite-Elemente-Methode in der Plattenstatik.

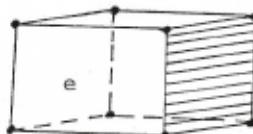
Abbildung 3.18: Knotenpunkte des (konformen) quadratischen Ansatzes ( $e = T$ )

4) *Kubisches Tetraederelement*:  $P(T) = P_3$ ,  $\dim P(T) = 20$ .

Mit den Funktionswerten in den Ecken, in zwei Gauß-Punkten auf den Kanten und in den Seitenmitteln als Knotenwerte ist dieser Ansatz unisolvant und  $H^1$ -konform.

Abbildung 3.19: Knotenpunkte des (konformen) Kubischen Tetraederelement ( $e = T$ )

5) *Tri-lineares Quaderelement*:  $P(T) = \text{span}\{1, x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1x_2x_3\}$ ,  $\dim P(T) = 8$ . Mit den Funktionswerten in den Eckpunkten als Knotenwerte ist dieser Ansatz unisolvant und  $H^1$ -konform.

Abbildung 3.20: Knotenpunkte des (konformen) tri-linearen Quaderelements ( $e = T$ )

Ein zugehöriges *nicht-konformes* Quaderelement erhält man durch den Ansatz  $P(T) = \text{span}\{1, x_1, x_2, x_3, x_1^2 - x_2^2, x_1^2 - x_3^2\}$ ,  $\dim P(T) = 6$ . Mit den Funktionswerten in den Flächenmitteln als Knotenwerte ist dieser Ansatz unisolvant.

6) *Tri-quadratisches Quaderelement*:  $P(T) = \text{span}\{1, x_1, x_2, x_3, \dots, x_1^2 x_2^2 x_3^2\}$ ,  $\dim P(T) = 27$ . Mit den Funktionswerten in den Eckpunkten, den Kantenmitteln, den Seitenmitteln und dem Mittelpunkt als Knotenwerte ist dieser Ansatz unisolvant und  $H^1$ -konform.

In allgemeinen Situationen können die Zellen bzgl. des Koordinatensystems gedreht oder gezerrt werden. Daraus folgt, dass es Fälle gibt, in denen man für zwei Zellen der

Zerlegung  $T_h$  nicht denselben Ansatz nehmen kann (z. B. das bi-lineare Viereckselement). Deshalb müssen wir uns bei der Definition von Finite-Elemente-Ansätzen von dem festen Koordinatensystem befreien. Dies induziert die Idee des „Referenzelements“. Wir verwenden als Referenzelement ein natürliches Einheitselement  $\hat{T}$  (Einheitsdreieck, Einheitsviereck, ...) und definieren zunächst einen Polynomansatz  $\hat{P}(\hat{T})$  auf diesem Referenzelement.

Sei  $\sigma_T$  eine (polynomiale) Transformation des Referenzelements auf das („physikalische“) Element mit der Inversen  $\sigma_T^{-1} : T \rightarrow \hat{T}$ . Der Ansatz auf der Zelle  $T$  ist dann gegeben durch

$$P(T) := \{v_h : T \rightarrow \mathbb{R} \mid v_h(\sigma_T(\cdot)) \in \hat{P}(\hat{T})\}. \quad (3.2.54)$$

Der Funktionenraum  $P(T)$  ist nicht notwendig ein Raum von Polynomen, auch wenn  $\hat{P}(\hat{T})$  ein solcher ist. Dies liegt daran, dass i. Allg. die inverse Abbildung  $\sigma_T^{-1}$  und damit die Funktion  $v_h(x) = v_h(\sigma(\sigma^{-1}(x)))$  nicht polynomial ist, z. B. im Fall (echt) bilinearer Abbildungen  $\sigma_T$ . Wenn man  $T$  aus  $\hat{T}$  durch eine Verschiebung, eine Rotation, eine Scherung und eine Skalierung gewinnen kann, so ist  $\sigma_T$  eine affin-lineare Transformation:

$$\sigma_T(\hat{x}) = B_T \hat{x} + b_T$$

mit einer Matrix  $B_T \in \mathbb{R}^{d \times d}$  und einem Verschiebungsvektor  $b_T \in \mathbb{R}^d$ . Dies ist möglich bei Dreiecken (in 2-D) bzw. Tetraedern (in 3-D) sowie bei Parallelogrammen (in 2-D) bzw. Parallelepipeden (in 3-D). Für allgemeine (konvexe) Vierecke (in 2-D) oder Hexaeder (in 3-D) benötigt man für die Transformation echt bi- bzw. tri-lineare Abbildungen.

Wir gehen nun zum allgemeinen  $d$ -dimensionalen Fall über und betrachten Zerlegungen  $\mathbb{T}_h = \{T\}$  des Abschlusses eines Gebiets  $\Omega \subset \mathbb{R}^d$  in  $d$ -Simplizes. Dabei ist ein (nicht degeneriertes) Simplex  $T \subset \mathbb{R}^d$  die konvexe lineare Hülle von  $d+1$  linear unabhängigen Punkten  $a^i \in \mathbb{R}^d$ ,  $i = 0, \dots, d$ :

$$T = \left\{ x \in \mathbb{R}^d \mid x = \sum_{i=0}^d \lambda_i a^i, \sum_{i=0}^d \lambda_i = 1, 0 \leq \lambda_i \leq 1 \right\}. \quad (3.2.55)$$

Das System  $\{a^i, i = 0, \dots, d\}$  heißt linear unabhängig, wenn die erzeugenden Vektoren  $\{w^i = a^i - a^0, i = 1, \dots, n\}$  eine Basis des  $\mathbb{R}^d$  bilden. Die einfachsten Beispiele sind wieder Dreiecke für  $d = 2$  und Tetraeder für  $d = 3$ . Sei  $\hat{T}$  das Einheitssimplex im  $\mathbb{R}^d$ , welches von den Punkten  $e^0 := 0$ ,  $e^i = (\delta_{1i}, \dots, \delta_{di})^T$ ,  $i = 1, \dots, d$ , aufgespannt wird.

**Hilfssatz 3.2 (Referenztransformation):** Jedes (nicht degenerierte) Simplex  $T \subset \mathbb{R}^d$  lässt sich mittels einer umkehrbaren affinen Abbildung

$$x = \sigma_T(\hat{x}) := B_T \hat{x} + b_T, \quad B_T \in \mathbb{R}^{d \times d}, \quad b_T \in \mathbb{R}^d, \quad (3.2.56)$$

aus dem Einheitssimplex  $\hat{T}$  gewinnen:  $T = B_T \hat{T} + b_T$ . Dabei ist die Umkehrabbildung gegeben durch  $\hat{x} = \sigma_T^{-1}(x) = B_T^{-1}x - B_T^{-1}b_T$ .

**Beweis:** Wir lassen im Folgenden den Zusatz  $T$  weg. Sei  $A \in \mathbb{R}^{d \times d}$  die reguläre Matrix, welche die Basis  $\{w^i = a^i - a^0, i = 1, \dots, d\}$  auf die kartesische Einheitsbasis  $\{e^i, i = 1, \dots, d\}$  abbildet:  $e^i = Aw^i, i = 1, \dots, d$ . Man gewinnt ihre Elemente  $a_{\nu\mu}$  als Lösungen der Gleichungssysteme

$$\sum_{\mu=1}^n a_{\nu\mu} w_\mu^i = e_\nu^i, \quad i = 1, \dots, d,$$

für  $\nu = 1, \dots, d$ . Die Koeffizientenmatrix  $(w_\mu^i)_{i,\mu=1}^d$  enthält die linear unabhängigen Vektoren  $w^i, i = 1, \dots, d$ , als Zeilenvektoren und ist folglich regulär. Die affine Abbildung  $\hat{x} = Ax - Aa^0$  ist dann umkehrbar und bildet das Simplex  $T$  auf das Einheitssimplex  $\hat{T}$  ab, denn für  $x = \sum_{i=0}^d \lambda_i a^i \in T$  ist (Man beachte  $\sum_{i=1}^d \lambda_i = 1$ .)

$$Ax - Aa^0 = \sum_{i=0}^d \lambda_i A(a^i - a^0) = \sum_{i=0}^d \lambda_i e^i \in \hat{T},$$

und umgekehrt für  $\hat{x} = \sum_{i=0}^d \lambda_i e^i \in \hat{T}$

$$A^{-1}\hat{x} + a^0 = \sum_{i=0}^d \lambda_i A^{-1}e^i + a^0 = \sum_{i=0}^d \lambda_i w^i + a^0 = \sum_{i=0}^d \lambda_i a^i + (1 - \sum_{i=1}^d \lambda_i)a^0 \in T.$$

Dies komplettiert den Beweis mit  $B_T := A^{-1}$  und  $b_T := a_0$ .

Q.E.D.

**Bemerkung 3.5:** Zu Hilfssatz 3.2 gibt es ein Analogon für Vierecks-Zerlegungen im  $\mathbb{R}^2$  sowie Hexaeder-Zerlegungen im  $\mathbb{R}^3$ . Zu jedem konvexen Viereck  $T \in \mathbb{R}^2$  oder Hexaeder  $T \in \mathbb{R}^3$  („6-Flächner“) existieren bi- bzw. tri-lineare Abbildungen  $\sigma_T : \hat{T} \rightarrow T$  des Einheitsquadrats bzw. Einheitswürfels  $\hat{T}$  auf  $T$ . Dabei werden die Eckpunkte von  $\hat{T}$  auf die Eckpunkte von  $T$  sowie die Kanten bzw. Seitenflächen von  $\hat{T}$  auf die Kanten bzw. Seitenflächen von  $T$  abgebildet (jeweils in derselben Orientierung).

Die Erzeugung von Finite-Elemente-Ansätzen über Transformation von einem Referenzelement hat auch den Zweck, auf allgemeinen Vierecks- oder Hexaeder-Zerlegungen *konforme* Ansätze zu gewinnen. Wir wollen das anhand des bi-linearen Ansatzes diskutieren:

$$V_h^{(1)} = \{v_h : \bar{\Omega} \rightarrow \mathbb{R} \mid v_h|_T \in Q_1, v_h \text{ stetig in Knoten, } v_h = 0 \text{ in Knoten auf } \partial\Omega\}.$$

Wir betrachten allgemeine Vierecke  $T$  und  $T'$ , die eine gemeinsame Kante  $\Gamma$  haben, und setzen  $P(T) = P_1 \oplus \text{span}\{x_1 x_2\}$ . Der Sprung von  $v_h$  ist zwar gleich Null in den Endpunkten von  $\Gamma$ , doch ist er i. Allg. nicht linear auf  $\Gamma$ , so dass die Stetigkeit über  $\Gamma$  nicht gesichert ist. Hierfür muss erreicht werden, dass die Restriktion von  $v_h \in V_h^{(1)}$  auf alle Kanten  $\Gamma \in \partial T_h$  linear ist. Dies kann durch Konstruktion von  $V_h^{(1)}$  mit Hilfe der bilinearen Transformationen  $\sigma_T : \hat{T} \rightarrow T$  erreicht werden.

Für eine bi-lineare Transformation  $\sigma_T : \hat{T} \rightarrow T$  ist die Inverse  $\sigma_T^{-1} : T \rightarrow \hat{T}$  i. Allg. nicht bi-linear. In diesem Fall ist der gemäß (3.2.54) erzeugte lokale Ansatzraum  $P(T)$

auch kein Polynomraum. Dennoch ist  $v|_{\Gamma}$  auf jeder Kante  $\Gamma \in \partial T_h$  linear, so dass Stetigkeit in allen Eckpunkten auch automatisch globale Stetigkeit auf  $\bar{\Omega}$  sowie  $v|_{\partial\Omega} = 0$  garantiert. Dieser Transformationsansatz löst also das Problem der globalen Stetigkeit.

**Definition 3.5 (Parametrischer Ansatz):** Der Ansatz  $P(T)$  wird „parametrisch“ genannt, wenn er durch Transformationen  $\sigma_T : \hat{T} \rightarrow T$  von einem Referenzelement  $\hat{T}$  erzeugt wird. Er heißt „isoparametrisch“, wenn die Transformation  $\sigma_T$  vom selben Polynomtyp wie die Ansatzfunktionen in  $\hat{P}(\hat{T})$  ist.

Der Begriff des „isoparametrischen“ Finite-Elemente-Ansatzes lässt sich auf höheren Polynomgrad  $r \geq 2$  übertragen. Dies wird wichtig bei der Approximation eines krummen Randes bei Verwendung von Ansätzen höherer Ordnung. (s. Abb. 3.21). Die einfache Polygonzugapproximation würde hier zu einer Ordnungsreduktion führen:

$$\|\nabla(u - u_h)\| \leq c\{h^r\|u\|_r + h\|u\|_2\},$$

im Gegensatz zur optimalen Abschätzung

$$\|\nabla(u - u_h)\| \leq ch^r\|u\|_r.$$

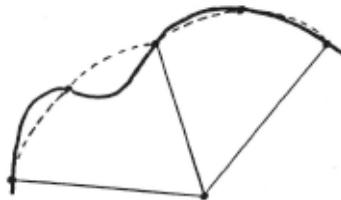


Abbildung 3.21: Parametrische Randapproximation

### 3.3 Interpolation mit finiten Elementen

Dieser Abschnitt ist dem grundlegenden Aspekt bei der mathematischen Analyse der Methode der finiten Elemente gewidmet: Wie gut lassen sich hinreichend glatte Funktionen durch stückweise polynomiale approximieren? Wir gehen von der im vorigen Abschnitt anhand von Beispielen beschriebenen abstrakten Situation aus. Sei  $T \subset \mathbb{R}^d$  eine Zelle eines Finite-Elemente-Gitters  $\mathbb{T}_h$ ; der Durchmesser von  $T$  wird wieder mit  $\text{diam}(T) = h_T$  und der Radius einer (maximalen) einbeschriebenen Kugel mit  $\rho_T$  bezeichnet.

Wir betrachten im Folgenden ausschließlich „parametrische“ Finite-Elemente-Ansätze. Jedes  $T \in \mathbb{T}_h$  sei Bild eines „Referenz-Einheitselements“  $\hat{T} \subset \mathbb{R}^d$  mit Durchmesser  $\text{diam}(\hat{T}) = \hat{h} \approx 1$  und Inkugelradius  $\hat{\rho} > 0$ . Die zugehörigen Abbildungen  $\sigma_T : \hat{T} \rightarrow T$

seien der Einfachheit halber als affin-linear angenommen:

$$x = \sigma_T(\hat{x}) = B_T \hat{x} + b_T, \quad B_T \in \mathbb{R}^{d \times d}, \quad b_T \in \mathbb{R}^d. \quad (3.3.57)$$

Der Fall allgemeiner Vierecke mit erzeugenden *bi-linearen* Transformationen wird gegebenenfalls in Bemerkungen berücksichtigt werden.

**Allgemeine Interpolationsaufgabe:** Auf einem beliebigen, aber festen Element  $T$  (z. B. dem Einheitselement  $\hat{T}$ ) seien ein Vektorraum  $P(T)$  von Polynomen über  $T$  mit  $\dim P(T) = R$  sowie ein System von linearen „Knotenfunktionalen“  $K_T = \{\chi_r, r = 1, \dots, R\}$  gegeben, so dass die folgenden Bedingungen erfüllt sind:

i) Der Ansatz ist unisolvent:

$$q \in P(T) : \chi_r(q) = 0 \quad (r = 1, \dots, R) \quad \Rightarrow \quad q = 0. \quad (3.3.58)$$

ii) Für ein  $m \geq 1$  gilt  $P_{m-1} \subset P(T)$ .

iii) Die Knotenfunktionale aus  $K_T$  sind stetig auf  $H^m(T)$ :

$$|\chi_r(v)| \leq c_b \|v\|_{m;T}, \quad v \in H^m(T), \quad r = 1, \dots, R. \quad (3.3.59)$$

Unter der Bedingung (i) ist die zugehörige Lagrangesche bzw. Hermitesche Interpolationsaufgabe eindeutig lösbar, d. h. Zu jeder Funktion  $v \in H^m(T)$  existiert ein eindeutig bestimmtes „Interpolationspolynom“  $I_T v \in P(T)$  mit den Eigenschaften

$$\chi_r(I_T v) = \chi_r(v), \quad r = 1, \dots, R.$$

Wenn die Knotenfunktionale zu „singulär“ sind, um für Funktionen aus  $H^m(\Omega)$  definiert zu sein (z. B. die Ableitung  $\partial_n^{m-1} v(m)$  in einer Seitenmitte  $m \in \partial T$ ), kann auch ein stärkerer Sobolew-Raum  $H^{m,p}(\Omega)$  mit  $p > d$  verwendet werden. Wir werden diesen Fall hier aber nicht weiter verfolgen.

**Notation:** Im folgenden verwenden wir eine gebräuchliche „Multiindex“-Schreibweise für mehrfach indizierte Größen. Für einen Indexvektor  $\alpha = (\alpha_1, \dots, \alpha_d)^T \in \mathbb{N}_0^d$  mit ganzzahligen, nichtnegativen Komponenten setzen wir

$$|\alpha| := \sum_{i=1}^d \alpha_i, \quad x^\alpha := \prod_{i=1}^d x_i^{\alpha_i}, \quad D^\alpha := \prod_{i=1}^d \partial_i^{\alpha_i}, \quad P_k = \left\{ q(x) = \sum_{|\alpha| \leq k} a_\alpha x^\alpha \right\}.$$

Mit dieser Notation schreiben sich z. B. die Sobolew-Normen bzw. -Halbnormen über  $T$  in der Form

$$\|v\|_{m;T} = \left( \sum_{0 \leq |\alpha| \leq m} \|D^\alpha v\|_T^2 \right)^{1/2}, \quad |v|_{m;T} = \left( \sum_{|\alpha|=m} \|D^\alpha v\|_T^2 \right)^{1/2}.$$

Wir leiten nun eine Reihe von technischen Hilfsätzen ab, die am Schluss zu den

gewünschten allgemeinen Abschätzungen für den Interpolationsfehler bei Finite-Elemente-Ansätze führen werden. Die dabei verwendete Schlussweise geht in diesem Zusammenhang auf Bramble<sup>15</sup> und Hilbert<sup>16</sup> (1971) zurück, weswegen die ganze Theorie auch „Bramble-Hilbert-Theorie“ und das Hauptresultat „Bramble-Hilbert-Lemma“ genannt werden.

**Hilfssatz 3.3 (Nullraum von Ableitungsoperatoren):** *Jede Funktion  $v \in H^m(T)$  mit der Eigenschaft*

$$D^\alpha v = 0, \quad |\alpha| = m, \quad (3.3.60)$$

*ist fast überall gleich einem Polynom aus  $P_{m-1}(T)$ .*

**Beweis:** Aus den Voraussetzungen folgt  $D^\beta D^\alpha v \equiv 0$  für beliebiges  $\beta$  und somit  $v \in \bigcap_{k=1}^{\infty} H^k(T)$ . Nach dem Sobolewschen Einbettungssatz ist damit  $v \in C^m(T)$ , so dass sich die Behauptung mit Hilfe „klassischer“ Argumente ergibt. Q.E.D.

**Hilfssatz 3.4 (Polynomprojektion):** *Zu jeder Funktion  $v \in H^m(T)$  existiert ein eindeutig bestimmtes Polynom  $q \in P_{m-1}(T)$  mit der Eigenschaft*

$$\int_T D^\alpha (v - q) dx = 0, \quad 0 \leq |\alpha| \leq m-1. \quad (3.3.61)$$

**Beweis:** Zur Lösung der Aufgabe machen wir den Ansatz

$$q(x) := \sum_{|\beta| \leq m-1} \xi^\beta x^\beta \in P_{m-1}(T)$$

mit unbekannten Koeffizienten  $\xi = (\xi^\beta)_{|\beta| \leq m-1}$  (bei lexikographischer Anordnung der Indexkomponenten). Dies führt auf das quadratische, lineare Gleichungssystem

$$\sum_{0 \leq |\beta| \leq m-1} \xi^\beta \int_T D^\alpha x^\beta dx = \int_T D^\alpha v dx, \quad 0 \leq |\alpha| \leq m-1.$$

Dessen Koeffizientenmatrix

$$M = \left( \int_T D^\alpha x^\beta dx \right)_{0 \leq |\alpha|, |\beta| \leq m-1}$$

ist regulär. Andernfalls gäbe es ein  $\xi = (\xi^\beta)_{|\beta| \leq m-1} \neq 0$  mit  $M\xi = 0$ . Das damit gebildete Polynom  $q(x) = \sum_{0 \leq |\beta| \leq m-1} \xi^\beta x^\beta \in P_{m-1}(T)$  hätte dann die Eigenschaft

$$\int_T D^\alpha q dx = 0, \quad 0 \leq |\alpha| \leq m-1, \quad (3.3.62)$$

---

<sup>15</sup>James H. Bramble (1932–): US-Amerikanischer Mathematiker: Prof. an der Cornell University und der Texas A&M University; fundamentale Beiträge zur Theorie der Finite-Elemente-Methode und von Iterationsverfahren, insbesondere Mehrgitterverfahrens.

<sup>16</sup>Stephen R. Hilbert (????–): US-amerikanischer Mathematiker: Prof. am Ithaca College, New York; Student von J. Bramble; bekannt durch das sog. „Bramble-Hilbert-Lemma“ (1970).

woraus offensichtlich  $q \equiv 0$  und damit der Widerspruch  $\xi \equiv 0$  folgte. Also existiert ein eindeutig bestimmtes Polynom mit den verlangten Eigenschaften. Q.E.D.

**Hilfssatz 3.5 (Verallg. Poincarésche Ungleichung):** *Für jede Funktion  $v \in H^m(T)$  mit der Eigenschaft*

$$\int_T D^\alpha v \, dx = 0, \quad 0 \leq |\alpha| \leq m-1, \quad (3.3.63)$$

*gilt mit einer Konstante  $c_0 = c(d, m, T)$*

$$\|v\|_{m,T} \leq c_0 |v|_{m,T}. \quad (3.3.64)$$

**Beweis:** Angenommen, die Behauptung ist falsch. Dann existiert eine Folge von Funktionen  $v_k \in H^m(T)$ ,  $k \in \mathbb{N}$  mit den Eigenschaften

$$1 = \|v_k\|_{m,T} \geq k |v_k|_{m,T}, \quad k \in \mathbb{N}. \quad (3.3.65)$$

Aufgrund der Kompaktheit der Einbettung von  $H^m(T)$  in  $H^{m-1}(T)$  konvergiert eine Teilfolge, welche wir wieder mit  $(v_k)_{k \in \mathbb{N}}$  bezeichnen, in  $H^{m-1}(T)$  gegen ein  $v \in H^{m-1}(T)$ :

$$\|v_k - v\|_{m-1;T} \rightarrow 0 \quad (k \rightarrow \infty). \quad (3.3.66)$$

Mit der Annahme folgt  $|v_k|_{m,T} \rightarrow 0$  ( $k \rightarrow \infty$ ). Also ist  $(v_k)_{k \in \mathbb{N}}$  Cauchy-Folge in  $H^m(\Omega)$  mit Limes  $\tilde{v} \in H^m(T)$ . Wegen  $v_k \rightarrow_{H^{m-1}} v$  muss  $\tilde{v} = v$  sein. Damit folgt  $|v|_{m,T} = 0$ . Nach Hilfssatz 3.3 ist also  $v \in P_{m-1}(T)$  und besitzt die Eigenschaft

$$\int_T D^\alpha v \, dx = \lim_{k \rightarrow \infty} \int_T D^\alpha v_k \, dx = 0, \quad 0 \leq |\alpha| \leq m-1. \quad (3.3.67)$$

Dies bedeutet aber wegen Hilfssatz 3.4 notwendig  $v \equiv 0$ , was im Widerspruch zur Annahme  $\|v\|_{m,T} = \lim_{k \rightarrow \infty} \|v_k\|_{m,T} = 1$  steht. Q.E.D.

Nach diesen Vorbereitungen können wir das zentrale Resultat dieses Abschnitts, das sog. „Bramble-Lemma“, beweisen.

**Satz 3.5 (Bramble-Hilbert-Lemma):** *Sei  $F(\cdot) : H^m(T) \rightarrow \mathbb{R}$  ein beschränktes, sublineares Funktional, welches auf  $P_{m-1}(T)$  verschwindet, d. h.:*

- i)  $|F(v)| \leq c_1 \|v\|_{m,T}$  (Beschränktheit),
- ii)  $|F(u+v)| \leq c_2 \{|F(u)| + |F(v)|\}$  (Sublinearität),
- iii)  $F(q) = 0, \quad q \in P_{m-1}(T)$  (Annulierungseigenschaft).

*Dann gilt mit der Konstante  $c_0$  aus Hilfssatz 3.5:*

$$|F(v)| \leq c_0 c_1 c_2 |v|_{m,T}, \quad v \in H^m(T). \quad (3.3.68)$$

**Beweis:** Für ein  $v \in H^m(T)$  gilt mit beliebigem  $q \in P_{m-1}(T)$ :

$$|F(v)| = |F(v - q + q)| \leq c_2\{|F(v - q)| + |F(q)|\} \leq c_1 c_2 \|v - q\|_{m;T}.$$

Wir wählen nun  $q \in P_{m-1}(T)$  als das gemäß Hilfssatz 3.4 zu  $v$  gehörende Polynom, so dass gemäß Hilfssatz 3.5 folgt:

$$\|v - q\|_{m;T} \leq c_0 |v - q|_{m;T} = c_0 |v|_{m;T}.$$

Dies impliziert dann

$$|F(v)| \leq c_3 |v|_{m;T},$$

mit  $c_3 := c_0 c_1 c_2$ , was zu beweisen war. Q.E.D.

**Korollar 3.1 (Allgemeiner Interpolationssatz):** Seien die obigen Voraussetzungen erfüllt. Für jede Funktion  $v \in H^m(T)$  und das zugehörige interpolierende Polynom  $I_T v \in P(T)$  gilt mit einer beliebigen beschränkten Halbnorm  $|\cdot|$  auf  $H^m(T)$ :

$$|v - I_T v| \leq c |v|_{m;T} \tag{3.3.69}$$

mit einer Konstante  $c = c(d, m, R, T, |\cdot|)$ .

**Beweis:** O.B.d.A. sei  $|v| \leq \|v\|_{m;T}$ ,  $v \in H^m(T)$ . Durch  $F(v) := |v - I_T v|$  wird auf  $H^m(T)$  ein sublineares Funktional definiert. Die Interpolierende  $I_T v$  besitzt die Darstellung

$$I_T v = \sum_{r=1}^R \chi_r(v) \varphi^{(r)}$$

mit der durch die Bedingung  $\chi_r(\varphi^{(s)}) = \delta_{rs}$ , ( $r, s = 1, \dots, R$ ) eindeutig bestimmten verallgemeinerten Lagrange-Basis  $\{\varphi^{(r)}, r = 1, \dots, R\}$  des Polynomraums  $P(T)$ . Wegen der Beschränktheit der Knotenfunktionale  $\chi_r$  folgt

$$|F(v)| \leq |v| + |I_T v| \leq |v| + \sum_{r=1}^R |\chi_r(v)| |\varphi^{(r)}| \leq (1 + R c_b \max_{r=1, \dots, R} |\varphi^{(r)}|) \|v\|_{m;T},$$

und damit die Beschränktheit von  $F(\cdot)$ . Wegen  $I_T q = q$  für  $q \in P(T)$  gilt weiter

$$F(q) = 0, \quad q \in P_{m-1}(T).$$

Aus Satz 3.5 folgt damit die behauptete Abschätzung. Q.E.D.

Beispiele von Halbnormen, für die das obige Resultat angewendet wird, sind etwa:

1.  $L^2$ -Norm über  $T$ :

$$|v - I_T v| := \left( \int_T |v - I_T v|^2 dx \right)^{1/2}.$$

2.  $L^2$ -Norm über den Rand  $\partial T$ :

$$|v - I_T v| := \left( \int_{\partial T} |v - I_T v|^2 dx \right)^{1/2}.$$

3. Mittelwert über eine Kante  $\Gamma \subset \partial T$ :

$$|v - I_T v| := \left| \int_{\Gamma} (v - I_T v) dx \right|.$$

4. Maximum-Norm über  $T$ :

$$|v - I_T v| := \max_T |v - I_T v|.$$

5. Wert in einem Punkt  $P \in \Omega$ :

$$|v - I_T v| := |(v - I_T v)(P)|.$$

Als nächstes greifen wir nun unser eigentliches Problem an, nämlich den Interpolationsfehler auf den einzelnen Zellen  $T$  der Zerlegung  $T_h$  abzuschätzen. Wir tun dies für den repräsentativen Spezialfall der klassischen Lagrange/Hermite-Interpolation, bei der die Knotenfunktionale  $\chi_r$  als Punktfunktionale für Funktionswerte sowie Ableitungswerte in gewissen Knotenpunkten  $\hat{a}_r \in \hat{T}$  ( $r=1, \dots, R$ ) gegeben sind.

Sei  $\hat{T}$  das Referenzelement der Größe  $\hat{h} := \text{diam}(\hat{T}) = 1$  und Inkreisradius  $\hat{\rho} > 0$ . Für eine einzelne Zelle  $T \in T_h$  bezeichnen wir mit  $a_r = B\hat{a}_r + b$  die aus den Stützstellen  $\hat{a}_r \in \hat{T}$  durch Anwendung der Transformation  $\sigma(\hat{x}) = B\hat{x} + b$  erzeugten Punkte  $a_r \in T$ . Entsprechend seien  $h_T := \text{diam}(T)$  sowie  $\rho_T > 0$  der Inkreisradius von  $T$ . Die Umkehrabbildung  $\sigma^{-1} : T \rightarrow \hat{T}$  hat die Darstellung  $\sigma^{-1}(x) = B^{-1}x - B^{-1}b$  mit der inversen Matrix  $B^{-1} = (b_{ij}^{(-1)})_{i,j=1}^d$ . Unter Verwendung dieser Abbildung  $x = \sigma(\hat{x})$  werden für Funktionen  $v : T \rightarrow \mathbb{R}$  und  $\hat{w} : \hat{T} \rightarrow \mathbb{R}$  zugehörige Funktionen  $\hat{v} : \hat{T} \rightarrow \mathbb{R}$  und  $w : T \rightarrow \mathbb{R}$  definiert durch

$$\hat{v}(\hat{x}) := v(x), \quad w(x) := \hat{w}(\hat{x}). \quad (3.3.70)$$

Entsprechend lassen sich die partiellen Ableitungen nach  $\hat{x}$  und  $x$  durch die jeweils anderen ausdrücken:

$$\hat{\partial}_i := \sum_{j=1}^d b_{ij} \partial_j, \quad \partial_i := \sum_{j=1}^d b_{ij}^{(-1)} \hat{\partial}_j.$$

Wir nehmen an, dass der Polynomansatz  $\hat{P}(\hat{T})$  unisolvant ist mit einem Satz von Knotenfunktionalen der Form  $\hat{D}_r \hat{v}(\hat{a}_r)$ ,  $r = 1, \dots, R$ , wobei die Punkte  $\hat{a}_r$  auch mehrfach auftreten können und die Ableitungsoperatoren die Gestalt  $\hat{D}_r = \hat{\partial}^{\alpha^r}$  mit geeigneten Multiindizes  $\alpha^r = (\alpha_i^r, \dots, \alpha_d^r)$  haben. Der Ansatzraum  $\hat{P}(\hat{T})$  habe die Lagrange-Basis  $\{\hat{\varphi}_r, r = 1, \dots, R\}$ , d.h.:

$$\hat{D}_r \hat{\varphi}_s(\hat{a}_r) = \delta_{rs}.$$

Wir nehmen weiter an, dass alle auftretenden Ableitungen  $\hat{D}_r$  auf  $H^m(\hat{T})$  wohl definiert und stetig sind:

$$|\hat{D}_r \hat{v}(\hat{a}_r)| \leq c \|\hat{v}\|_{m;\hat{T}}, \quad \hat{v} \in H^m(\hat{T}).$$

Die lokalen Polynomräume  $P(T)$  werden wieder erzeugt via Koordinatentransformation aus dem Ansatzraum  $\hat{P}(\hat{T})$  auf dem Referenzelement:

$$P(T) := \{q : T \rightarrow \mathbb{R} \mid q(\sigma(\cdot)) \in \hat{P}(\hat{T})\}.$$

Dasselbe gilt für die zugehörigen Basen  $\{\varphi^{(r)}, r = 1, \dots, R\}$  von  $P(T)$

$$\varphi_r(x) := \hat{\varphi}_r(\sigma^{-1}(x)), \quad x \in T.$$

Für Funktionen  $v \in H^m(\Omega)$  erhält man dann durch Setzung

$$I_T v := \sum_{s=1}^R \hat{D}_s \hat{v}(\hat{a}_s) \varphi_s \in P(T)$$

eine zellweise Lagrange/Hermite-Interpolierende  $I_T v \in P(T)$  mit den Eigenschaften (Man beachte  $\hat{D}_r \hat{\varphi}_s(\hat{a}_r) = \delta_{rs}$ ):

$$\begin{aligned} \hat{D}_r I_T v(a_r) &= \sum_{s=1}^R \hat{D}_s \hat{v}(\hat{a}_s) \hat{D}_r \varphi_s(a_r) = \sum_{s=1}^R \hat{D}_s \hat{v}(\hat{a}_s) \hat{D}_r \hat{\varphi}_s(\hat{a}_r) \\ &= \hat{D}_r \hat{v}(\hat{a}_r) = \hat{D}_r v(a_r), \quad r = 1, \dots, R. \end{aligned}$$

Je nach Art der Ableitungsoperatoren  $\hat{D}_r$  lässt sich dies gegebenenfalls auch als eine Interpolation mit lokal auf den Einzelzellen  $T$  definierten Ableitungsoperatoren  $D_r$  bzgl. der (physikalischen) Variablen  $x$  ausdrücken; z. B. in den einfachsten Fällen  $D_r v(a_r) := v(a_r)$  bzw.  $D_r v(a_r) \approx \nabla v(a_r)$ .

Wir beweisen nun den Hauptsatz dieses Abschnittes zur Polynominterpolation.

**Satz 3.6 (Spezieller Interpolationssatz):** Für jedes  $v \in H^m(T)$  und die zugehörige Interpolierende  $I_T v \in P(T)$  auf der Zelle  $T \in \mathbb{T}_h$  gilt:

$$|v - I_T v|_{k;T} \leq c_I \frac{h_T^m}{\rho_T^k} |v|_{m;T}, \quad 0 \leq k \leq m, \quad (3.3.71)$$

mit Durchmesser und Inkreisradius  $h_T$  bzw.  $\rho_T$  von  $T$  und einer Konstante  $c_I = c_I(d, m, \hat{T})$ .

**Beweis:** i) Für eine Funktion  $f \in L^1(T)$  bezeichnet  $\hat{f} \in L^1(\hat{T})$  die zugehörige Transformierte

$$\hat{f}(\hat{x}) := f(x), \quad \hat{x} = \sigma^{-1}(x) = B^{-1}x - B^{-1}b \in \hat{T}. \quad (3.3.72)$$

Die Transformation  $\sigma^{-1}$  hat die Funktionalmatrix  $\nabla\sigma^{-1} = B^{-1}$ , so dass gilt:

$$\int_{\hat{T}} \hat{f}(\hat{x}) d\hat{x} = |\det B^{-1}| \int_T \hat{f}(\sigma^{-1}(x)) dx = |\det B|^{-1} \int_T f(x) dx. \quad (3.3.73)$$

Wir schätzen nun die Elemente  $b_{ij}$  sowie  $b_{ij}^{(-1)}$  der Matrizen  $B$  und  $B^{-1}$  ab. Jedes  $x \in \mathbb{R}^d$  mit  $|x| = \rho$  gestattet mit zwei Punkten  $\xi, \eta \in T$  die Darstellung  $x = \xi - \eta$ , wobei  $\xi$  etwa als Mittelpunkt der Inkugel von  $T$  gewählt werden kann. Dann sind  $\hat{\xi} = B^{-1}\xi - B^{-1}b$ ,  $\hat{\eta} = B^{-1}\eta - B^{-1}b \in \hat{T}$ , und wir erhalten

$$|B^{-1}x| = |B^{-1}\xi - B^{-1}b - B^{-1}\eta + B^{-1}b| = |\hat{\xi} - \hat{\eta}| \leq \hat{h}. \quad (3.3.74)$$

Da alle Matrizennormen auf  $\mathbb{R}^{d \times d}$  äquivalent sind, gilt mit einer Konstante  $c = c(d)$

$$\max_{i,j=1,\dots,d} |b_{ij}^{(-1)}| \leq c \sup_{x \in \mathbb{R}^d} \frac{|B^{-1}x|}{|x|} = c \sup_{|x|=\rho} \frac{|B^{-1}x|}{|x|} \leq c \frac{\hat{h}}{\rho}. \quad (3.3.75)$$

Analog wird bewiesen:

$$\max_{i,j=1,\dots,d} |b_{ij}| \leq c \frac{h}{\hat{\rho}}. \quad (3.3.76)$$

ii) Es seien nun  $T \in \mathbb{T}_h$  sowie  $v \in H^m(T)$  beliebig gegeben. Es ist  $I_T v \in P(T)$  die Interpolierende von  $v$  auf  $T$  und

$$I_T \hat{v}(\cdot) = I_T v(\sigma(\cdot)) \in P(\hat{T}) \quad (3.3.77)$$

die Interpolierende der Transformierten  $\hat{v} \in H^m(\hat{T})$  auf  $\hat{T}$ . Nach Korollar 3.1 gilt auf  $\hat{T}$ :

$$|\hat{v} - I_T \hat{v}|_{k;\hat{T}} \leq \hat{c} |\hat{v}|_{m;\hat{T}}, \quad 0 \leq k \leq m, \quad (3.3.78)$$

mit einer festen Konstante  $\hat{c} = \hat{c}(d, m, \hat{T})$ . Durch Koordinatentransformation zwischen  $\hat{T}$  und  $T$  werden wir nun zeigen, dass

$$\begin{aligned} |v - I_T v|_{k;T} &\leq \frac{c}{\rho^k} |\det B|^{1/2} |\hat{v} - I_T \hat{v}|_{k;\hat{T}}, \\ |\hat{v}|_{m;\hat{T}} &\leq c h^m |\det B|^{-1/2} |v|_{m;T}, \end{aligned}$$

mit Konstanten  $c = c(d, m, \hat{T})$ . Diese beiden Beziehungen ergeben dann zusammen mit (3.3.78) die Behauptung.

iii) Für die Ableitungen von  $v$  bzw.  $\hat{v}$  gilt mit  $\hat{x} = \sigma^{-1}(x) = B^{-1}x - B^{-1}b$ :

$$\begin{aligned} \widehat{\partial_i v}(\hat{x}) &= \partial_i v(x) = \partial_i \hat{v}(\sigma^{-1}(x)) = \sum_{j=1}^d \hat{\partial}_j \hat{v}(\hat{x}) \partial_i \sigma_j^{-1}(x) = \sum_{j=1}^d \hat{\partial}_j \hat{v}(\hat{x}) b_{ji}^{(-1)}, \\ \hat{\partial}_i \hat{v}(\hat{x}) &= \hat{\partial}_i v(\sigma(\hat{x})) = \sum_{j=1}^d \partial_j v(x) \hat{\partial}_i \sigma_j(\hat{x}) = \sum_{j=1}^d \partial_j v(x) b_{ji} = \sum_{j=1}^d \widehat{\partial_j v}(\hat{x}) b_{ji}, \end{aligned}$$

und folglich

$$|\widehat{\partial_i v}(\hat{x})| \leq c \frac{\hat{h}}{\rho} \max_{j=1,\dots,d} |\widehat{\partial_j v}(\hat{x})| \quad |\widehat{\partial_i v}(\hat{x})| \leq c \frac{h}{\hat{\rho}} \max_{j=1,\dots,d} |\widehat{\partial_j v}(\hat{x})|.$$

Für allgemeine Ableitungen  $D^\alpha$  bzw.  $\hat{D}^\alpha$  der Ordnung  $k = |\alpha|$  gewinnen wir durch  $k$ -malige Anwendung dieser Beziehungen:

$$|\widehat{D^\alpha v}(\hat{x})| \leq c \max_{i,j=1,\dots,d} |b_{ij}^{(-1)}|^{|\alpha|} \max_{|\beta|=|\alpha|} |\widehat{D^\beta v}(\hat{x})| \leq c \left( \frac{\hat{h}}{\rho} \right)^{|\alpha|} \max_{|\beta|=|\alpha|} |\widehat{D^\beta v}(\hat{x})|,$$

$$|\widehat{D^\alpha v}(\hat{x})| \leq c \max_{i,j=1,\dots,d} |b_{ij}|^{|\alpha|} \max_{|\beta|=|\alpha|} |\widehat{D^\beta v}(\hat{x})| \leq c \left( \frac{h}{\hat{\rho}} \right)^{|\alpha|} \max_{|\beta|=|\alpha|} |\widehat{D^\beta v}(\hat{x})|.$$

Mit Hilfe dieser Abschätzungen und der Substitutionsformel (3.3.73) folgt nun

$$\begin{aligned} \int_T |D^\alpha v(x)|^2 dx &= |\det B| \int_{\hat{T}} |\widehat{D^\alpha v}(\hat{x})|^2 d\hat{x} \\ &\leq c |\det B| \left( \frac{\hat{h}}{\rho} \right)^{2|\alpha|} \max_{|\beta|=|\alpha|} \int_{\hat{T}} |\widehat{D^\beta v}(\hat{x})|^2 d\hat{x} \\ \int_{\hat{T}} |\widehat{D^\alpha v}(\hat{x})|^2 d\hat{x} &\leq c \left( \frac{h}{\hat{\rho}} \right)^{2|\alpha|} \max_{|\beta|=|\alpha|} \int_{\hat{T}} |\widehat{D^\beta v}(\hat{x})|^2 d\hat{x} \\ &\leq c |\det B|^{-1} \left( \frac{h}{\hat{\rho}} \right)^{2|\alpha|} \max_{|\beta|=|\alpha|} \int_T |D^\beta v(x)|^2 dx. \end{aligned}$$

Damit ist für  $0 \leq k \leq m$  bewiesen:

$$|v|_{k;T} \leq c |\det B|^{1/2} \left( \frac{\hat{h}}{\rho} \right)^k |\hat{v}|_{k;\hat{T}}, \quad |\hat{v}|_{k;\hat{T}} \leq c |\det B|^{-1/2} \left( \frac{h}{\hat{\rho}} \right)^k |v|_{k;T}.$$

Anwendung dieser Beziehungen für  $v$  und  $v - I_T v$  ergibt schließlich wegen  $0 < \rho \leq h \leq 1$  die behauptete Abschätzungen (3.3.71). Q.E.D.

**Bemerkung 3.6:** Die Fehlerabschätzung (3.3.71) für die Polynominterpolation hat natürliche Verallgemeinerungen auf andere Halbnormen  $|\cdot|_T$  sowie auf andere Regularitätsstufen  $v \in H^{m,p}(T)$ ,  $1 \leq p \leq \infty$ . Wir geben ohne Beweis das folgende allgemeine Resultat an:

$$|v - I_T v|_{k,q;T} \leq c_I \frac{h_T^{m-d/p}}{\rho_T^{k-d/q}} |v|_{m,p;T}, \quad (3.3.79)$$

für  $0 \leq k \leq m$ ,  $1 \leq p \leq q \leq \infty$ , mit einer Konstante  $c_I = c_I(d, k, m, p, q, \hat{T})$ . Für spätere Zwecke sind hiervon insbesondere die Fälle  $p = q = 1$  (für  $m \geq 2$ ),  $p = q = \infty$  sowie  $p = 2$ ,  $q = \infty$  (für  $k \leq m + 2$ ) von Interesse; z. B. gilt für  $d = 2$ ,  $m = 2$ ,  $q = \infty$ ,  $p = 2$ :

$$\max_T |v - I_T v| \leq c_I h_T |v|_{2,2;T}. \quad (3.3.80)$$

Auf ähnliche Art wie im Beweis von Satz 3.6 gewinnt man durch Transformation auf das Einheitselement  $\hat{T}$  die folgende sog. „inverse Beziehung“ für finite Elemente:

**Satz 3.7 (Inverse Beziehung):** Unter den obigen Voraussetzungen gilt auf jeder Zelle  $T \in \mathbb{T}_h$  für Finite-Elemente-Funktionen  $v \in P(T)$  mit einer Konstante  $c = c(d, \rho_T, k, s)$ :

$$|v|_{k,T} \leq c \frac{h_T^s}{\rho_T^k} |v|_{s,\hat{T}}, \quad 0 \leq s \leq k \leq m. \quad (3.3.81)$$

**Beweis:** Für Polynome  $\hat{q} \in P(\hat{T})$  gilt wegen der Äquivalenz von Normen auf dem (endlich dimensionalen) Quotientenraum  $P(T)/P_{k-1}(T)$  mit einer Konstante  $\hat{c} = \hat{c}(d, m, \hat{T})$ :

$$|\hat{q}|_{k,\hat{T}} \leq \hat{c} |\hat{q}|_{s,\hat{T}}, \quad 0 \leq s \leq k \leq m. \quad (3.3.82)$$

Die Behauptung ergibt sich nun wieder durch Transformation auf die Zelle  $T$ . Q.E.D.

**Bemerkung 3.7:** Analoge Abschätzungen wie die in Satz 3.6 und Satz 3.7 gelten auch für Tensorprodukt-Polynomansätze auf Vierecken bzw. Hexaedern.

Wir wollen diese Resultate auf die obigen konkreten Beispiele anwenden. In diesen sind alle formulierten Bedingungen erfüllt. Insbesondere gilt die „uniform shape“ Bedingung

$$\sup_{h>0} \max_{T \in \mathbb{T}_h} \frac{h_T}{\rho_T} \leq c.$$

Dann folgt aus Satz 3.6 für die FE-Ansatzräume  $V_h^{(m-1)} \subset H_0^1(\Omega)$  vom Polynomgrad  $m-1 \in \mathbb{N}_0$  die Interpolationsfehlerabschätzungen

$$|v - I_T v|_{k,T} \leq c h_T^{m-k} |v|_{m,T}, \quad k = 0, \dots, m, \quad T \in \mathbb{T}_h, \quad (3.3.83)$$

mit Konstanten  $c = c(d, m, \hat{T})$ . Manchmal möchte man den Interpolationsfehler auch über den Rand der Zelle oder punktweise abschätzen. Hierfür gilt

$$\|v - I_T v\|_{\partial T} \leq c h_T^{m-1/2} \|\nabla^2 v\|_{m,T}, \quad (3.3.84)$$

und z. B. in zwei Dimensionen

$$\max_{x \in T} |v - I_T v| \leq c h_T^{m-1} \|\nabla^2 v\|_{m,T}. \quad (3.3.85)$$

Als Folgerung aus diesem lokalen Abschätzungen erhalten wir die folgenden globale Approximationsabschätzungen

$$|v - I_h v|_k \leq c h^{m-k} |v|_m, \quad k = 0, \dots, m. \quad (3.3.86)$$

Für „lineare“ finite Elemente gilt also speziell

$$\|u - I_h u\| + h \|\nabla(u - I_h u)\| \leq c h^2 \|\nabla^2 u\|. \quad (3.3.87)$$

In diesem Fall ergibt die „inverse“ Beziehung (3.3.81):

$$\|\nabla v_h\| \leq h^{-1} \|v_h\|, \quad v_h \in V_h^{(1)}. \quad (3.3.88)$$

Entsprechend gilt für „quadratische“ finite Elemente

$$\|u - I_h u\| + h \|\nabla(u - I_h u)\| \leq ch^3 \|\nabla^3 u\|, \quad (3.3.89)$$

und

$$\|\nabla^2 v_h\| \leq h^{-2} \|v_h\|, \quad v_h \in V_h^{(2)}. \quad (3.3.90)$$

Damit ist jetzt die theoretische Grundlage für die *a priori* Fehlerabschätzungen für das Galerkin-Finite-Elemente-Verfahren aus Satz 3.1 geschaffen.

### 3.4 A priori Fehleranalyse

Die im Abschnitt 3.3 hergeleiteten Abschätzungen für den Fehler bei der Interpolation mit stückweise polynomialem Funktionen sind die Basis für die *a priori* Fehleranalyse des Finite-Elemente-Verfahrens. Wir formulieren die folgende Verallgemeinerung von Satz 3.1 für FE-Approximationen allgemeiner Ordnung  $m \geq 2$ ,

$$(\nabla u_h, \nabla \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h, \quad (3.4.91)$$

der Poisson-Gleichung. Dabei ist die sog. „Ordnung“ eines FE-Verfahrens im wesentlichen durch den Polynomgrad der verwendeten Ansatzfunktionen bestimmt, d. h. durch die Beziehung  $P_{m-1} \subset P(T)$  für alle  $T \in \mathbb{T}_h$ .

**Korollar 3.2 (Allgemeine FE-Konvergenz):** Für den Fehler  $e_h := u - u_h$  einer FE-Methode mit Ansatzräumen  $V_h \subset H_0^1(\Omega)$  der Ordnung  $m \geq 2$  zur Approximation der Poisson-Gleichung gilt die *a priori* Fehlerabschätzung:

$$\|e_h\| + h \|\nabla e_h\| \leq c h^m \|\nabla^m u\|. \quad (3.4.92)$$

Wir wollen eine Schwäche dieses Resultats nicht unerwähnt lassen. Im allgemeinen sind Lösungen  $u \in H_0^1(\Omega)$  elliptischer Gleichungen zweiter Ordnung über Gebieten mit Polygonrand nicht aus  $H^m(\Omega)$  für  $m \geq 3$ . An den Ecken von  $\partial\Omega$  treten starke Singularitäten der Ableitungen von  $u$  auf. Die obige Voraussetzung  $u \in H^m(\Omega)$  ist also für  $m > 2$  unrealistisch. Finite Elemente höherer Ordnung  $m > 2$  können jedoch bei Gebieten mit hinreichend glattem Rand  $\partial\Omega$  erfolgreich verwendet werden, denn in diesem Fall kann die Regularitätstheorie meist  $u \in H^m(\Omega)$  sichern. Dabei sind aber besondere Maßnahmen, wie z. B. die Verwendung isoparametrischer Ansätze, zur Approximation entlang des krummen Randes erforderlich.

Wir haben gesehen, dass der Fehler z. B. bei „linearen“ finiten Elementen gemessen in der  $L^2$ -Norm eine verbesserte Konvergenzordnung  $O(h^2)$  gegenüber der von  $O(h)$  in der

Energie-Norm zulässt. Es stellt sich die Frage, ob man durch weitere Abschwächung der Norm vielleicht noch höhere Konvergenzordnungen erzielen kann. Diese Hoffnung wird zunächst bestärkt durch die Beobachtung, dass dies für die  $L^2$ -Projektion durchaus der Fall ist. Die in Frage kommenden Normen werden illustrativ als „negative“ Sobolew-Normen (genauer als Sobolew-Normen mit „negativer Ordnung“) bezeichnet und sind in den einfachsten Fällen definiert durch

$$\|v\|_{-1} := \sup_{\varphi \in V} \frac{(v, \varphi)}{\|\varphi\|_1}, \quad \|v\|_{-2} := \sup_{\varphi \in V \cap H^2(\Omega)} \frac{(v, \varphi)}{\|\varphi\|_2}.$$

wobei wieder  $V := H_0^1(\Omega)$ .

**Hilfssatz 3.6 ( $L^2$ -Projektion):** Für die  $L^2$ -Projektion  $P_h : V \rightarrow V_h$  auf den Raum  $V_h^{(1)}$  der „linearen“ finiten Elementen gilt die Fehlerabschätzungen

$$\|u - P_h u\|_{-2} + h \|u - P_h u\|_{-1} + h^2 \|u - P_h u\| \leq ch^4 \|\nabla^2 u\|. \quad (3.4.93)$$

**Beweis:** Zunächst rekapitulieren wir die „Bestapproximationseigenschaft“ der  $L^2$ -Projektion:

$$\|u - P_h u\| = \min_{\varphi_h \in V_h^{(1)}} \|u - \varphi_h\|. \quad (3.4.94)$$

Daraus folgt mit der lokalen Interpolationsabschätzung (3.3.83) die Beziehung

$$\|u - P_h u\| \leq ch^2 \|\nabla^2 u\|. \quad (3.4.95)$$

Mit einem beliebigen  $\varphi \in H_0^1(\Omega)$  gilt entsprechend für  $k \in \{1, 2\}$ :

$$\begin{aligned} (u - P_h u, \varphi) &= (u - P_h u, \varphi - P_h \varphi) \\ &\leq \|u - P_h u\| \|\varphi - P_h \varphi\| \\ &\leq ch^{2+k} \|\nabla^2 u\| \|\nabla^k \varphi\|. \end{aligned}$$

Dies impliziert

$$\sup_{\varphi \in H_0^1(\Omega) \cap H^k(\Omega)} \frac{(u - P_h u, \varphi)}{\|\varphi\|_k} \leq ch^{2+k} \|\nabla^2 u\|,$$

was zu beweisen war. Q.E.D.

Der Beweis von Hilfssatz 3.6 zeigt, dass eine weitere Erhöhung jenseits  $\mathcal{O}(h^4)$  der Approximationsordnung der  $L^2$ -Projektion auf den Ansatzraum  $V_h^{(1)}$  auch in einer noch schwächeren Norm nicht mehr möglich ist. Für die „Ritz-Projektion“  $R_h : V \rightarrow V_h^{(1)}$  ist in diesem Fall sogar  $\mathcal{O}(h^2)$  die Obergrenze für die erreichbare Ordnung. Dies zeigt der folgende Satz.

**Satz 3.8 (Ritz-Projektion):** Für die Ritz-Projektion  $R_h : V \rightarrow V_h^{(1)}$  auf den Raum der „linearen“ finiten Elemente gilt die Abschätzung

$$\|u - R_h u\|_{-1} \geq c(f) \|\nabla(u - R_h u)\|^2, \quad (3.4.96)$$

mit einer positiven Konstante  $c(f) > 0$ .

**Beweis:** Sei  $f \in H_0^1(\Omega)$ . Für die Lösung  $u \in V$  der Gleichung  $-\Delta u = f$  gilt unter Ausnutzung der „Galerkin-Orthogonalität“:

$$(u - R_h u, f) = (\nabla(u - R_h u), \nabla u) = \|\nabla(u - R_h u)\|^2.$$

Wegen

$$\sup_{\varphi \in V} \frac{(u - R_h u, \varphi)}{\|\varphi\|_1} \geq \frac{(u - R_h u, f)}{\|f\|_1}$$

impliziert dies die Behauptung. Q.E.D.

Da die Energie-Fehlerabschätzung

$$\|\nabla(u - R_h u)\| \leq ch \|\nabla^2 u\|$$

bzgl. der Ordnung bestmöglich ist, folgt aus (3.4.96) die Unmöglichkeit einer Fehlerabschätzung mit einer Ordnung größer als zwei für „lineare“ finite Elemente.

### 3.4.1 Punktweise Fehlerabschätzung

In den vorangegangenen Abschnitten haben wir gesehen, dass die Methode der finiten Elemente als Projektionsmethode zunächst ganz natürlich zu a priori Abschätzungen in der „Energienorm“  $\|\nabla e\|$  und dann über ein Dualitätsargument auch zu verbesserten Abschätzungen in der  $L^2$ -Norm  $\|e\|$  führt. Diese Konvergenzaussagen im „quadratischen Mittel“ gestatten aber noch keinen unmittelbaren Schluss auf die punktweise Konvergenz des Verfahrens. Dieser Frage soll jetzt wieder anhand des einfachen Modellproblems „Poisson-Gleichung“,

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ auf } \partial\Omega, \quad (3.4.97)$$

auf einem (konvexen) Polygonegebiet  $\Omega \subset \mathbb{R}^2$  nachgegangen werden. Wir beschränken uns dabei auf die Approximation (3.4.91) mit stückweise linearen Ansätzen  $V_h^{(1)} \subset V$  auf regulären Triangulierungen. Wir rekapitulieren hierfür die a priori Fehlerabschätzung

$$\|e\| + h\|\nabla e\| \leq ch^2 \|\nabla^2 u\|, \quad (3.4.98)$$

wobei die Konstante  $c$  wesentlich durch die Konstante in der Interpolationsfehlerabschätzung

$$\|\nabla(u - I_h u)\|_T \leq c_I h \|\nabla^2 u\|_T, \quad T \in \mathbb{T}_h, \quad (3.4.99)$$

bestimmt ist. Wir erinnern daran, dass im vorliegenden Fall eines konvexen Grundgebiets jedes  $v \in V$  mit  $\Delta v \in L^2(\Omega)$  automatisch in  $H^2(\Omega)$  ist und der a priori Abschätzung genügt:

$$\|\nabla^2 v\| \leq \|\Delta v\|. \quad (3.4.100)$$

**Satz 3.9 (Sub-optimale  $L^\infty$ -Norm-Fehlerabschätzung):** *Unter den obigen Voraussetzungen konvergiert die Methode der finiten Elemente punktweise mit der Ordnung  $\mathcal{O}(h)$ :*

$$\max_{\Omega} |e| \leq ch \|\nabla^2 u\|. \quad (3.4.101)$$

**Beweis:** Sei  $T$  ein beliebiges Dreieck aus  $\mathbb{T}_h$ . Für eine Funktion  $v_h \in V_h^{(1)}$  gilt dann

$$\max_T |v_h| \leq c|T|^{-1} \int_T |v_h| dx. \quad (3.4.102)$$

Dies folgt leicht mit Hilfe der Transformation auf die Referenzzelle  $\hat{T}$  ( $dx \approx |T|d\hat{x}$ ) und der dort geltenden Beziehung (Äquivalenz von Normen)

$$\max_{\hat{T}} |\hat{v}_h| \leq \hat{c} \int_{\hat{T}} |\hat{v}_h| d\hat{x}.$$

Mit der Knoteninterpolierenden  $I_h u \in V_h^{(1)}$  zu  $u \in V \cap H^2(\Omega)$  gilt

$$\max_T |u - I_h u| \leq ch_T \|\nabla^2 u\|_T. \quad (3.4.103)$$

Damit erschließen wir dann:

$$\begin{aligned} \max_T |e| &\leq \max_T |u - I_h u| + \max_T |I_h e| \\ &\leq \max_T |u - I_h u| + c|T|^{-1} \int_T |I_h e| dx \\ &\leq \max_T |u - I_h u| + c|T|^{-1} \int_T |e| dx \leq ch_T \|\nabla^2 u\|_T + ch_T^{-1} \|e\|_T. \end{aligned}$$

Mit Hilfe der  $L^2$ -Fehlerabschätzung (3.4.98) folgt also die Behauptung. Q.E.D.

**Bemerkung 3.8:** Unter der bloßen Annahme  $u \in V \cap H^2(\Omega)$  ist die Fehlerabschätzung (3.4.101) bzgl. der  $h$ -Potenz optimal. Um dies einzusehen, betrachte man z. B. auf dem Einheitskreis  $B_1 = \{x \in \mathbb{R}^2 \mid |x| < 1\}$  die Funktionen

$$u^h(x) := (|x|^2 + h^4)^{1/2}$$

Offenbar ist  $u^h \in H^2(B_1)$ , und es gilt

$$\|\nabla^2 u^h\|_{B_1} \leq c \left( \int_{B_1} (|x|^2 + h^4)^{-1} dx \right)^{1/2} \leq c |\ln(h)|^{1/2}.$$

Wir nehmen nun an, dass unser Lösungsgebiet  $\Omega$  den Kreis  $B_1$  enthält und die Funktionen  $u^h$  geeignet zu Funktionen  $\bar{u}^h \in H_0^1(\Omega) \cap H^2(\Omega)$  fortgesetzt sind. Ferner gehöre zu jeder der Triangulierungen  $T_h$  ein Dreieck  $T_0$  mit Durchmesser  $h$  und dem Inkreismittelpunkt  $a_0 = 0$ . Dann ist in den Eckpunkten  $a_i$  und dem Mittelpunkt  $a_0$  von  $T_0$  stets

$$\bar{u}^h(a_i) \geq h, \quad \bar{u}^h(a_0) = h^2.$$

Würde nun für die Ritz-Projektion  $\bar{u}_h^h \in V_h$  zu  $\bar{u}^h$  mit einem  $\varepsilon > 0$  gelten

$$\max_{T_0} |\bar{u}^h - \bar{u}_h^h| \leq ch^{1+\varepsilon} \|\nabla^2 \bar{u}^h\|_\Omega,$$

so ergäbe sich für hinreichend kleines  $h$  im Widerspruch zur Linearität der  $\bar{u}_h^h$  auf  $T_0$ :

$$\begin{aligned} |\bar{u}_h^h(a_0)| &\leq |\bar{u}_h^h(a_0) - \bar{u}^h(a_0)| + |\bar{u}^h(a_0)| \leq ch^{1+\varepsilon_1}, \quad 0 < \varepsilon_1 < \varepsilon, \\ |\bar{u}_h^h(a_i)| &\geq |\bar{u}^h(a_i)| - |\bar{u}^h(a_i) - \bar{u}_h^h(a_i)| \geq ch. \end{aligned}$$

Numerische Experimente zeigen, dass im Falle höherer Regularität von  $u$  (etwa  $u \in C^2(\bar{\Omega})$ ) die optimale Ordnung  $O(h^2)$  des  $L^2$ -Fehlers auch für die punktweise Konvergenz vorliegt. Ein ähnliches Phänomen haben wir bereits bei der Differenzenapproximation mit dem 5-Punkte-Operator gesehen, bei dem sich auf gleichförmigen Gittern die Konvergenzordnung  $O(h)$  im Falle  $u \in C^3(\bar{\Omega})$  auf  $O(h^2)$  im Falle  $u \in C^4(\bar{\Omega})$  erhöht. Für die Methode der finiten Elemente beweisen wir nun das folgende optimale Resultat auf allgemeinen regulären Gittern:

**Satz 3.10 (Optimale  $L^\infty$ -Abschätzung):** Im Falle  $u \in V \cap C^2(\bar{\Omega})$  gilt die Konvergenzabschätzung

$$\sup_{\Omega} |e| \leq ch^2 L(h) \sup_{\Omega} |\nabla^2 u| \tag{3.4.104}$$

mit dem logarithmischen Faktor  $L(h) := |\log(h)| + 1$ .

**Beweis:** Wir notieren zunächst die Interpolationsfehlerabschätzung

$$\sup_{\Omega} |u - I_h u| + h \sup_{\Omega} |\nabla(u - I_h u)| \leq ch^2 \sup_{\Omega} |\nabla^2 u| \tag{3.4.105}$$

i) Für ein  $h > 0$  sei  $T_* \in T_h$  beliebig, aber fest gewählt. Mit der Knoteninterpolierenden  $I_h u \in V_h^{(1)}$  von  $u$  gilt wieder (s. Beweis von Satz 3.9):

$$\begin{aligned}\max_{T_*} |e| &\leq c \max_{T_*} |u - I_h u| + c |T_*|^{-1} \int_{T_*} |e| dx \\ &\leq ch_{T_*}^2 \max_{T_*} |\nabla^2 u| + c |T_*|^{-1} \int_{T_*} |e| dx.\end{aligned}$$

Damit ist die Abschätzung des  $L^\infty$ -Fehlers zurückgeführt auf eine lokale  $L^1$ -Fehlerabschätzung. Mit der durch

$$\delta^h := |T_*|^{-1} \operatorname{sign}(e) \text{ in } T_*, \quad \delta^h := 0 \text{ sonst,}$$

definierten Funktion  $\delta^h \in L^2(\Omega)$  („regularisierte“ Dirac<sup>17</sup>-Funktion) gilt weiter

$$|T_*|^{-1} \int_{T_*} |e| dx = (\delta^h, e).$$

ii) Jetzt wird wieder ein Dualitätsargument verwendet. Wir betrachten das Hilfsproblem

$$-\Delta g^h = \delta^h \text{ in } \Omega, \quad g^h = 0 \text{ auf } \partial\Omega. \quad (3.4.106)$$

Die Funktion  $g^h$  kann als „regularisierte“ Greensche Funktion angesehen werden. Damit gilt

$$|T_*|^{-1} \int_{T_*} |e| dx = (\nabla e, \nabla g^h). \quad (3.4.107)$$

Unter Verwendung der durch

$$(\nabla g_h^h, \nabla \varphi_h) = (\nabla g^h, \nabla \varphi_h) \quad \forall \varphi_h \in V_h^{(1)},$$

definierten „Ritz-Projektion“  $g_h^h \in V_h^{(1)}$  von  $g^h$  erhalten wir durch zweimalige Anwendung der Galerkin-Orthogonalität die Beziehung

$$|T_*|^{-1} \int_{T_*} |e| dx = (\nabla e, \nabla(g^h - g_h^h)) = (\nabla(u - I_h u), \nabla(g^h - g_h^h)). \quad (3.4.108)$$

Mit der Hölderschen Ungleichung folgt

$$|T_*|^{-1} \int_{T_*} |e| dx \leq \max_{\Omega} |\nabla(u - I_h u)| \int_{\Omega} |\nabla(g^h - g_h^h)| dx. \quad (3.4.109)$$

Unter Beachtung der Abschätzung (3.4.105) erhalten wir schließlich die Beziehung

$$\max_{T_*} |e| \leq ch \left\{ h + \int_{\Omega} |\nabla(g^h - g_h^h)| dx \right\} \sup_{\Omega} |\nabla^2 u|. \quad (3.4.110)$$

---

<sup>17</sup>Paul Adrien Maurice Dirac (1902–1984): Französischer Physiker und Mathematiker; Prof. in Cambridge; wichtige Beiträge zur Quanten Mechanik und Kosmologie, 1933 Nobel-Preis.

Die punktweise Abschätzung des Fehlers  $e$  ist also zurückgeführt auf eine globale  $L^1$ -Fehlerabschätzung für den Gradienten der „Greenschen Funktion“:  $\nabla(g^h - g_h^h)$ . Dieser wird unten weiter abgeschätzt. Dazu benötigen wir einige a priori Abschätzungen für  $g^h$ , die im folgenden bereitgestellt werden.

iii) Sei  $x_*$  der Inkreismittelpunkt von  $T_*$ . Wir definieren die Gewichtsfunktion

$$\sigma(x) := (|x - x_*|^2 + \kappa^2 h^2)^{1/2}.$$

Durch Nachrechnen verifiziert man leicht die Beziehungen

$$\kappa h \leq \sigma \leq c, \quad |\nabla \sigma| \leq c_*, \quad |\nabla^2 \sigma| \leq c\sigma^{-1}, \quad \|\sigma^{-1}\| \leq c L(h)^{1/2},$$

mit von  $h$  und  $\kappa$  unabhängigen Konstanten. Für die Größen  $\bar{\sigma}_T := \max_T \sigma$  und  $\underline{\sigma}_T := \min_T \sigma$  gilt daher

$$\bar{\sigma}_T \leq \underline{\sigma}_T + h \max_T |\nabla \sigma| \leq \underline{\sigma}_T + c_* h,$$

und bei Wahl von  $\kappa := 2c_*$  (unabhängig von  $h$ ):

$$\bar{\sigma}_T \leq \underline{\sigma}_T + \frac{1}{2} \bar{\sigma}_T,$$

und damit

$$\max_{T \in \mathbf{T}_h} \frac{\bar{\sigma}_T}{\underline{\sigma}_T} \leq 2. \tag{3.4.111}$$

**Hilfssatz 3.7 (Greensche Funktion):** Für die regularisierte „Greensche Funktion“  $g^h$  gelten die a priori Abschätzungen

$$\sup_{\Omega} |g^h| \leq c L(h), \tag{3.4.112}$$

$$\|\nabla g^h\| + \|\sigma \nabla^2 g^h\| \leq c L(h)^{1/2}, \tag{3.4.113}$$

$$\|\nabla^2 g^h\| \leq ch^{-1}, \tag{3.4.114}$$

mit von  $h$  unabhängigen Konstanten  $c$ .

**Beweis:** i) Die „richtige“ Greensche Funktion  $g(\cdot) = g(x_*, \cdot)$  zum Aufpunkt  $x_*$  erlaubt die Abschätzung (Beweis mit Hilfe des Maximumsprinzips)

$$|g(x)| \leq c \{ |\ln(|x - x_*|)| + 1 \}.$$

Konstruktionsgemäß folgt damit

$$|g^h(x)| = |(\nabla g^h, \nabla g)| = |(\delta^h, g)_\Omega| \leq ch^{-2} \int_{T_*} |g| dx \leq cL(h).$$

Dies impliziert die Abschätzung (3.4.112).

ii) Zum Beweis von (3.4.114) verwenden wir die übliche a priori  $L^2/H^2$ -Abschätzung

$$\|\nabla^2 g^h\| \leq c \|\delta^h\| \leq ch^{-1}.$$

iii) Als nächstes notieren wir die einfache a priori Abschätzung

$$\|\nabla^2 g^h\| \leq \|\Delta g^h\| \leq ch^{-1}. \quad (3.4.115)$$

Weiter gilt

$$\|\nabla g^h\|^2 = (\delta^h, g^h) \leq c \sup_{\Omega} |g^h| \leq cL(h).$$

Dies impliziert den ersten Teil der Abschätzung (3.4.113).

iii) Schließlich setzen wir  $\xi := x - x_*$  und finden wegen

$$|\xi_i \nabla^2 g^h| \leq |\nabla^2(\xi_i g^h)| + |\nabla g^h|$$

die Beziehung

$$\begin{aligned} \|\sigma \nabla^2 g^h\|^2 &= \sum_{i=1}^2 \|\xi_i \nabla^2 g^h\|^2 + \kappa^2 h^2 \|\nabla^2 g^h\|^2 \\ &\leq \sum_{i=1}^2 \{\|\nabla^2(\xi_i g^h)\|^2 + \|\nabla g^h\|^2\} + \kappa^2 h^2 \|\nabla^2 g^h\|^2. \end{aligned}$$

Mit Hilfe der üblichen a priori  $L^2/H^2$ -Abschätzung (3.4.100) folgt

$$\begin{aligned} \|\nabla^2(\xi_i g^h)\| &\leq \|\Delta(\xi_i g^h)\| \leq \|\xi_i \Delta g^h\| + \|\nabla g^h\| \\ &\leq \|\xi_i \delta^h\| + \|\nabla g^h\| \leq c + cL(h)^{1/2}. \end{aligned}$$

Die vorausgehenden Abschätzungen implizieren dann

$$\|\sigma \nabla^2 g^h\| \leq cL(h)^{1/2},$$

woraus (3.4.113) folgt. Dies vervollständigt den Beweis von (3.4.113). Q.E.D.

iv) Wir kehren nun zum Beweis des Satzes zurück und schätzen wie folgt ab:

$$\int_{\Omega} |\nabla(g^h - g_h^h)| dx \leq \|\sigma^{-1}\| \|\sigma \nabla(g^h - g_h^h)\| \leq cL(h)^{1/2} \|\sigma \nabla(g^h - g_h^h)\|.$$

Durch Ausdifferenzieren folgt weiter

$$\begin{aligned} \|\sigma \nabla(g^h - g_h^h)\|^2 &= (\nabla(g^h - g_h^h), \nabla(\sigma^2(g^h - g_h^h))) - (\nabla(g^h - g_h^h), (g^h - g_h^h) \nabla \sigma^2) \\ &=: E_1 - E_2. \end{aligned}$$

Die Terme  $E_1$  und  $E_2$  werden im folgenden separat abgeschätzt. Im Hinblick auf die Galerkin-Orthogonalität gilt

$$E_1 = (\nabla(g^h - g_h^h), \nabla(\sigma^2(g^h - g_h^h) - \psi_h))$$

mit der Knoteninterpolierenden  $\psi_h := I_h(\sigma^2(g^h - g_h^h)) \in V_h^{(1)}$ . Dies wird abgeschätzt durch

$$\begin{aligned} E_1 &\leq \sum_{T \in \mathbb{T}_h} \|\sigma \nabla(g^h - g_h^h)\|_T \|\sigma^{-1} \nabla(\sigma^2(g^h - g_h^h) - \psi_h)\|_T \\ &\leq \frac{1}{4} \|\sigma \nabla(g^h - g_h^h)\|^2 + c \sum_{T \in \mathbb{T}_h} \|\sigma^{-1} \nabla(\sigma^2(g^h - g_h^h) - \psi_h)\|_T^2. \end{aligned}$$

Mit Hilfe der Interpolationsfehlerabschätzung (3.4.103) werden die einzelnen Summanden wie folgt abgeschätzt:

$$\begin{aligned} \|\sigma^{-1} \nabla(\sigma^2(g^h - g_h^h) - \psi_h)\|_T^2 &\leq \underline{\sigma}_T^{-2} \|\nabla(\sigma^2(g^h - g_h^h) - \psi_h)\|_T^2 \\ &\leq c \underline{\sigma}_T^{-2} h_T^2 \|\nabla^2(\sigma^2(g^h - g_h^h))\|_T^2 \\ &\leq c \underline{\sigma}_T^{-2} h_T^2 \{ \|g^h - g_h^h\|_T^2 + \|\sigma \nabla(g^h - g_h^h)\|_T^2 + \|\sigma^2 \nabla^2 g^h\|_T^2 \} \\ &\leq c \|g^h - g_h^h\|_T^2 + c \underline{\sigma}_T^{-2} \bar{\sigma}_T^2 h_T^2 \{ \|\nabla(g^h - g_h^h)\|_T^2 + \|\sigma \nabla^2 g^h\|_T^2 \}. \end{aligned}$$

Dies ergibt wegen (3.4.111):

$$E_1 \leq \frac{1}{4} \|\sigma \nabla(g^h - g_h^h)\|^2 + c \|g^h - g_h^h\|^2 + c h^2 \{ \|\nabla(g^h - g_h^h)\|^2 + \|\sigma \nabla^2 g^h\|^2 \}.$$

Die schon bekannten  $L^2$ -Fehlerabschätzungen liefern weiter

$$E_1 \leq \frac{1}{4} \|\sigma \nabla(g^h - g_h^h)\|^2 + ch^2 \{ h^2 \|\nabla^2 g^h\|^2 + \|\sigma \nabla^2 g^h\|^2 \},$$

sowie unter Beachtung der Abschätzungen von Hilfssatz 3.7

$$E_1 \leq \frac{1}{4} \|\sigma \nabla(g^h - g_h^h)\|^2 + ch^2 |\ln(h)|.$$

Für den zweiten Term gilt wegen  $|\nabla \sigma^2| \leq c\sigma$ :

$$E_2 \leq c \|\sigma \nabla(g^h - g_h^h)\| \|g^h - g_h^h\| \leq \frac{1}{4} \|\sigma \nabla(g^h - g_h^h)\|^2 + c \|g^h - g_h^h\|^2$$

sowie mit den Argumenten von eben:

$$E_2 \leq \frac{1}{4} \|\sigma \nabla(g^h - g_h^h)\|^2 + ch^2.$$

Kombination der Abschätzungen für  $E_1$  und  $E_2$  ergibt schließlich

$$\|\sigma \nabla(g^h - g_h^h)\|^2 \leq \frac{1}{2} \|\sigma \nabla(g^h - g_h^h)\|^2 + ch^2 L(h),$$

und damit die Behauptung. Q.E.D.

Der logarithmische Term  $L(h) := |\ln(h)| + 1$  in der Abschätzung (3.4.104) lässt sich auf allgemeinen Gittern nicht vermeiden. Dies wird durch numerische Tests und auch durch theoretische Analyse bestätigt. Auf gleichförmigen Gittern (Je zwei benachbarte Dreiecke bilden ein Parallelogramm.) erhält man unter der stärkeren Glättungsbedingung  $u \in C^{2+\alpha}(\bar{\Omega})$  allerdings die optimale Konvergenzordnung  $O(h^2)$ . Vergleicht man dieses

Resultat mit dem entsprechenden für die Differenzenapproximation mit dem 5-Punkte-Operator, so stellen wir eine deutliche Abschwächung der Regularitätsanforderungen um fast zwei Stufen fest.

Auf Polygonegebieten ist die Bedingung  $u \in H^{2,\infty}(\Omega)$  für die schwache Lösung von (3.4.97) im allgemeinen unrealistisch. Bei den Ecken können die zweiten Ableitungen Singularitäten haben. Auf konvexen Polygonegebieten ist gerade die Voraussetzung  $u \in H^2(\Omega)$  natürlich, d. h. stets erfüllt, wogegen  $u \in H^{2,\infty}(\Omega_1)$  nur auf Teilgebieten  $\Omega'_1 \subset \Omega$  mit positivem Abstand zu den Eckpunkten gilt. In diesem Fall haben wir das folgende „lokale“ Resultat:

**Satz 3.11 (Lokales Fehlerverhalten):** Sei  $\Omega_1 \subset \Omega$  ein Teilgebiet mit positivem Abstand  $\delta_1$  zu den Eckpunkten von  $\Omega$  und  $u \in H^2(\Omega) \cap C^2(\bar{\Omega}_1)$ . Dann gilt auf jedem zweiten Teilgebiet  $\Omega_2 \subset \Omega_1$  mit Abstand  $\delta_2 > \delta_1$  zu den Eckpunkten die Fehlerabschätzung

$$\sup_{\Omega_2} |e| \leq ch^2 \left\{ L(h) \sup_{\Omega_1} |\nabla^2 u| + \|\nabla^2 u\|_\Omega \right\}. \quad (3.4.116)$$

**Beweis:** Der technisch aufwendige Beweis kann im Rahmen dieses Textes nicht geführt werden; wir verweisen dafür auf die entsprechende Literatur. Q.E.D.

**Bemerkung 3.9:** Zum Abschluss bemerken wir noch, dass sich für finite Elemente höherer Ordnung (Polynomgrad  $m-1 \geq 2$ ) zu (3.4.104) analoge punktweise Fehlerabschätzungen unter der Voraussetzung  $u \in C^m(\bar{\Omega})$  herleiten lassen:

$$\sup_{\Omega} |e| \leq ch^m \sup_{\Omega} |\nabla^m u|. \quad (3.4.117)$$

Bemerkenswerterweise tritt dabei ab Polynomordnung  $m-1 = 2$  der störende logarithmische Term  $L(h)$  nicht auf. Dasselbe gilt auch für den niedrigsten Ansatzgrad  $m-1 = 1$ , wenn der maximale Fehlergradient betrachtet wird:

$$\sup_{\Omega} |\nabla e| \leq ch \sup_{\Omega} |\nabla^2 u|. \quad (3.4.118)$$

## 3.5 Implementierungsaspekte

Im folgenden wollen wir einige Fragen im Zusammenhang mit der praktischen Realisierung der Finite-Elemente-Methode diskutieren. Dazu betrachten wir als Modellfall die 1. RWA eines (elliptischen) Differentialoperators,

$$Lu := -\nabla \cdot \{a \nabla u\} = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega, \quad (3.5.119)$$

mit möglicherweise variablem Koeffizienten  $a = a(x) \geq \alpha > 0$  auf einem (konvexen) Polygonegebiet  $\Omega \subset \mathbb{R}^2$ . Die Diskretisierung erfolgt wieder auf Ansatzräumen  $V_h \subset H_0^1(\Omega)$

zu einer Folge von (gleichmäßig) regulären Zerlegungen  $\mathbb{T}_h = \{T\}$  des Grundgebiets  $\Omega \subset \mathbb{R}^d$  ( $d = 2, 3$ ),

$$a(u, \varphi) = l(\varphi) \quad \forall \varphi_h \in V_h. \quad (3.5.120)$$

mit den bilinearen bzw. linearen Formen

$$a(u, \varphi) := (a\nabla u_h, \nabla \varphi_h), \quad l(\varphi) := (f, \varphi_h).$$

Die Aufstellung des zugehörigen linearen Gleichungssystems erfordert in der Regel die Anwendung numerischer Integration, was zu einem zusätzlichen Fehler führt.

### 3.5.1 Aufbau der Systemmatrizen und Vektoren

Im Gegensatz zu den Differenzenverfahren auf strukturierten Gittern lassen sich die algebraischen Gleichungssysteme der Finite-Elemente-Methode auf allgemeinen Zerlegungen  $\mathbb{T}_h$  in der Regel nicht explizit „per Hand“ aufstellen.

Mit der Knotenbasis  $\{\varphi_h^{(n)}, n = 1, \dots, N\}$  des Finite-Elemente-Raumes  $V_h \subset H_0^1(\Omega)$  sind die Systemmatrizen (unter Weglassung des Index  $h$ )  $A = (a_{nm})_{n,m=1}^N$  („Steifigkeitsmatrix“) und  $M = (m_{nm})_{n,m=1}^N$  („Massenmatrix“) sowie der „Lastvektor“  $b = (b_n)_{n=1}^N$  gebildet gemäß

$$a_{nm} = a(\varphi_h^{(m)}, \varphi_h^{(n)}), \quad m_{nm} = (\varphi_h^{(m)}, \varphi_h^{(n)}), \quad b_n = l(\varphi_h^{(n)}).$$

Beide Matrizen sind konstruktionsgemäß symmetrisch und positiv-definit. Ihre größten und kleinsten Eigenwerte seien  $\lambda_{\max}(A)$ ,  $\lambda_{\min}(A)$  bzw.  $\lambda_{\max}(M)$ ,  $\lambda_{\min}(M)$ , und die zugehörigen Speztralkonditionen:

$$\kappa_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}, \quad \kappa_2(M) = \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}.$$

Wir rekapitulieren die folgenden Beziehungen zwischen einer Finite-Elemente-Funktion  $v_h \in V_h$  und ihrem zugehörigen Knotenvektor  $\xi = (\xi_n)_{n=1}^N \in \mathbb{R}^N$ :

$$v_h = \sum_{n=1}^N \xi_n \varphi_h^{(n)}, \quad \|v_h\|^2 = \langle M\xi, \xi \rangle, \quad a(v_h, v_h) = \langle A\xi, \xi \rangle,$$

wobei  $\langle \cdot, \cdot \rangle$  das euklidische Skalarprodukt bezeichnet; die euklidische Vektornorm ist  $|\cdot|$ . Der Knotenvektor  $\xi$  ist bestimmt durch das lineare Gleichungssystem

$$A\xi = b \quad (3.5.121)$$

Zur Aufstellung des Systems (3.5.121) bedient man sich in der Praxis eines zell-orientierten Prozesses, des sog. „Assemblierens“ (englischen „assembling“). Dabei wird das Konzept der „Element-(Last)-Vektoren“ und „Element-(Steifigkeits)-Matrizen“ verwendet. Für einen Knotenvektor  $\xi \in \mathbb{R}^N$  und eine Zelle  $T \in \mathbb{T}_h$  ist dabei  $\xi_T = (\xi_1^T, \dots, \xi_n^T, \dots, \xi_N^T)^T$

mit

$$\xi_n^T := \xi_n \quad \text{falls } \varphi_h^{(n)} \not\equiv 0 \text{ auf } T, \quad \xi_n^T := 0 \quad \text{falls } \varphi_h^{(n)} \equiv 0 \text{ auf } T.$$

Die zugehörigen Element-Matrizen und -Vektoren haben die Form

$$\begin{aligned} A_T &= (a_{nm}^T)_{n,m=1}^N := ((a \nabla \varphi_h^{(m)}, \nabla \varphi_h^{(n)})_T)_{n,m=1}^N, \\ M_T &= (m_{nm}^T)_{n,m=1}^N := ((\varphi_h^{(m)}, \varphi_h^{(n)})_T)_{n,m=1}^N, \\ b_T &= (b_n^T)_{n=1}^N := ((f, \varphi_h^{(n)})_T)_{n=1}^N. \end{aligned}$$

Dabei werden natürlich nur die wesentlichen, von Null verschiedenen Elemente von  $A_T$ ,  $M_T$  und  $b_T$  gespeichert. Die einzelnen Gesamtmatrizen und Vektoren werden dann gebildet durch Assemblierung der entsprechenden Element-Matrizen und -Vektoren gemäß :

$$A = \sum_{T \in \mathbf{T}_h} A_T, \quad M = \sum_{T \in \mathbf{T}_h} M_T, \quad b = \sum_{T \in \mathbf{T}_h} b_T.$$

Entsprechend gilt

$$\langle A\xi, \xi \rangle = \sum_{T \in \mathbf{T}_h} \langle A_T \xi_T, \xi_T \rangle, \quad \langle M\xi, \xi \rangle = \sum_{T \in \mathbf{T}_h} \langle M_T \xi_T, \xi_T \rangle.$$

Die Element-Beiträge werden in der Regel durch Transformation auf ein Referenzelement berechnet. Wir diskutieren hier nur den Fall von Dreieckszerlegungen. Sei also wieder  $\sigma_T : \hat{T} \rightarrow T$  die affin-lineare Abbildung des Einheitsdreiecks  $\hat{T}$  auf das Dreieck  $T$ :

$$x = \sigma_T(\hat{x}) = B_T \hat{x} + b_T, \quad \hat{x} = \sigma_T^{-1}(x) = B_T^{-1}x - B_T^{-1}b_T.$$

Die charakteristischen Parameter von  $T$  sowie  $\hat{T}$  sind mit  $h_T, \rho_T$  bzw.  $\hat{h}, \hat{\rho}$  bezeichnet. Für transformierte Funktionen  $\hat{v}(\hat{x}) = v(x)$  gilt dann

$$\int_T v(x) dx = |\det B_T| \int_{\hat{T}} \hat{v}(\hat{x}) d\hat{x},$$

woraus insbesondere  $|\det B_T| = |T||\hat{T}|^{-1} \approx h_T^d$  folgt. Ferner ist mit der Inversen  $B_T^{-1} = (b_{ij}^{(-1)})_{ij=1}^d$ :

$$\widehat{\partial_i v}(\hat{x}) = \partial_i v(x) = \partial_i \hat{v}(\hat{x}) = \sum_{j=1}^d \hat{\partial}_j \hat{v}(\hat{x}) \partial_i \hat{x}_j = \sum_{j=1}^d \hat{\partial}_j \hat{v}(\hat{x}) b_{ji}^{(-1)}.$$

bzw.  $\widehat{\nabla v}(\hat{x}) = B_T^{-T} \hat{\nabla} \hat{v}(\hat{x})$ . Die Elemente der Element-Matrizen  $A_T$  und  $M_T$  sowie des Element-Vektors  $b_T$  transformieren sich wie folgt:

$$\begin{aligned}
a_{nm}^T &= \int_T a \nabla \varphi_h^{(m)} \nabla \varphi_h^{(n)} dx = |\det B_T| \int_{\hat{T}} \hat{a} \widehat{\nabla \varphi_h^{(m)}} \widehat{\nabla \varphi_h^{(n)}} d\hat{x} \\
&= |\det B_T| \int_{\hat{T}} \hat{a} B_T^{-T} \hat{\nabla} \hat{\varphi}_h^{(m)} B_T^{-T} \hat{\nabla} \hat{\varphi}_h^{(n)} d\hat{x} =: |\det B_T| \hat{a}_{nm}^T, \\
m_{nm}^T &= \int_T \varphi_h^{(m)} \varphi_h^{(n)} dx = |\det B_T| \int_{\hat{T}} \hat{\varphi}_h^{(m)} \hat{\varphi}_h^{(n)} d\hat{x} =: |\det B_T| \hat{m}_{nm}. \\
b_n^T &= \int_T f \varphi_h^{(n)} dx = |\det B_T| \int_{\hat{T}} \hat{f} \hat{\varphi}_h^{(n)} d\hat{x} =: |\det B_T| \hat{b}_n.
\end{aligned}$$

Die Werte  $\hat{a}_{nm}^T$ ,  $\hat{m}_{nm}^T$  und  $\hat{b}_n^T$  auf der Referenzzelle werden nun mit Hilfe von Quadraturformeln auf  $\hat{T}$  berechnet. Dazu werden wir weiter unten noch mehr Details angeben. Wichtig ist, dass diese Quadratur nur auf der Referenzzelle stattfindet und die tatsächlich verwendeten Größen  $a_{nm}$ ,  $m_{nm}$  und  $b_n$  im wesentlichen durch Skalierung mit  $|\det B_T|$  gewonnen werden.

### 3.5.2 Konditionierung der Systemmatrix

Wir wollen zunächst die Stabilität (für  $h \rightarrow 0$ ) der diskreten Finite-Elemente-Gleichung (3.5.121) gegenüber Störungen der Daten untersuchen. Diese treten z. B. auf durch die Fehler bei der Berechnung der Elemente  $a_{nm}$  und  $m_{nm}$  bei Verwendung von numerischer Intergration. Durch rein algebraische Argumente erhalten wir zunächst eine Stabilitätsabschätzung für allgemeine lineare Gleichungssysteme (s. den Band „Numerik 0 – Einführung in die Numerische Mathematik“).

**Satz 3.12 (Allgemeiner Störungssatz):** Seien Störungen  $\delta A$  der Matrix  $A$  und  $\delta b$  der rechten Seite  $b$  gegeben, so dass  $\mu := \kappa_2(A) \|\delta A\| / \|A\| < 1$ . Dann gilt die Fehlerabschätzung

$$\frac{|\delta \xi|}{|\xi|} \leq \frac{\kappa_2(A)}{1 - \mu} \left\{ \frac{\|\delta A\|}{\|A\|} + \frac{|\delta b|}{|b|} \right\}. \quad (3.5.122)$$

Zur quantitativen Auswertung dieser Abschätzung müssen wir die Spektralkondition der Steifigkeitsmatrix  $A$  in Abhängigkeit von der Gitterweite  $h$  abschätzen. Dazu nehmen wir an, dass die betrachtete Familie von Zerlegungen  $(T_h)_{h>0}$  gleichmäßig „form- und größen-regulär“ ist, d. h.:

$$\sup_{h>0} \left( \max_{T \in T_h} \frac{h_T}{\rho_T} \right) \leq c, \quad \sup_{h>0} \left( \frac{\max_{T \in T_h} h_T}{\min_{T \in T_h} h_T} \right) \leq c.$$

Dann ergibt sich analog zum Differenzenverfahren das folgende allgemeine Resultat.

**Satz 3.13 (Konditionierung):** Auf einer Folge von (gleichmäßig) regulären Zerlegungen  $\mathbb{T}_h$  gilt für die Spektralkonditionen der (symmetrischen und positiv definiten) Steifigkeitsmatrizen  $A$  und der Massenmatrizen  $M$ :

$$\kappa_2(A) = \mathcal{O}(h^{-2}), \quad \kappa_2(M) = \mathcal{O}(1) \quad (h \rightarrow 0). \quad (3.5.123)$$

**Beweis:** i) Für die größten und kleinsten Eigenwerte von  $M$  gilt

$$\lambda_{\min}(M) = \min_{\xi \in \mathbb{R}^N} \frac{\langle M\xi, \xi \rangle}{|\xi|^2} \leq \max_{\xi \in \mathbb{R}^N} \frac{\langle M\xi, \xi \rangle}{|\xi|^2} = \lambda_{\max}(M).$$

In der folgenden Argumentation werden wieder die Element-Matrizen  $M_T$ . Ferner bezeichne  $d_{\min}$  und  $d_{\max}$  die kleinste bzw. die größte Anzahl von Zellen, die in einem Knoten der Zerlegung  $\mathbb{T}_h$  zusammentreffen. Mit dieser Notation ergibt sich:

$$\begin{aligned} \langle M\xi, \xi \rangle &= \sum_{T \in \mathbb{T}_h} \langle M_T \xi_T, \xi_T \rangle \geq \min_{\xi \in \mathbb{R}^N, T \in \mathbb{T}_h} \frac{\langle M_T \xi_T, \xi_T \rangle}{|\xi_T|^2} \sum_{T \in \mathbb{T}_h} |\xi_T|^2 \geq \min_{T \in \mathbb{T}_h} \{\lambda_{\min}(M_T)\} d_{\min} |\xi|^2, \\ \langle M\xi, \xi \rangle &= \sum_{T \in \mathbb{T}_h} \langle M_T \xi_T, \xi_T \rangle \leq \max_{\xi \in \mathbb{R}^N, T \in \mathbb{T}_h} \frac{\langle M_T \xi_T, \xi_T \rangle}{|\xi_T|^2} \sum_{T \in \mathbb{T}_h} |\xi_T|^2 \leq \max_{T \in \mathbb{T}_h} \{\lambda_{\max}(M_T)\} d_{\max} |\xi|^2. \end{aligned}$$

Mit Hilfe der Beziehung  $|\det B_T| \approx h_T^d$  ergibt sich mit der (festen) Matrix  $\hat{M} := (\hat{m}_{nm})_{n,m=1}^N$ :

$$\lambda_{\max}(M_T) = |\det B_T| \lambda_{\max}(\hat{M}) \leq ch_T^d, \quad \lambda_{\min}(M_T) = |\det B_T| \lambda_{\min}(\hat{M}) \geq ch_T^d,$$

und folglich  $\kappa_2(M) = \mathcal{O}(1)$ .

ii) Für die kleinsten und größten Eigenwerte von  $A$  gilt:

$$\begin{aligned} \lambda_{\min}(A) &\geq \min_{\xi \in \mathbb{R}^N} \frac{\langle A\xi, \xi \rangle}{\langle M\xi, \xi \rangle} \min_{\xi \in \mathbb{R}^N} \frac{\langle M\xi, \xi \rangle}{|\xi|^2} = \min_{v_h \in V_h} \frac{a(v_h, v_h)}{\|v_h\|^2} \lambda_{\min}(M), \\ \lambda_{\max}(A) &\leq \max_{\xi \in \mathbb{R}^N} \frac{\langle A\xi, \xi \rangle}{\langle M\xi, \xi \rangle} \max_{\xi \in \mathbb{R}^N} \frac{\langle M\xi, \xi \rangle}{|\xi|^2} = \max_{v_h \in V_h} \frac{a(v_h, v_h)}{\|v_h\|^2} \lambda_{\max}(M). \end{aligned}$$

Wir schätzen weiter ab durch

$$\min_{v_h \in V_h} \frac{a(v_h, v_h)}{\|v_h\|^2} \geq \min_{v \in H_0^1(\Omega)} \frac{a(v, v)}{\|v\|^2} =: \lambda_{\min}(L)$$

mit dem kleinste Eigenwert des Differentialoperators  $L$  auf dem Gebiet  $\Omega$ . Ferner ergibt sich mit Hilfe der „inversen Beziehung“ für Finite-Elemente-Funktionen:

$$a(v_h, v_h) \leq \|a\|_\infty \sum_{T \in \mathbb{T}_h} \|\nabla v_h\|_T^2 \leq c \|a\|_\infty \sum_{T \in \mathbb{T}_h} \rho_T^{-2} \|v_h\|_T^2 \leq c \|a\|_\infty \max_{T \in \mathbb{T}_h} \rho_T^{-2} \|v_h\|^2,$$

bzw.  $\lambda_{\max}(A) \leq c \max_{T \in \mathbb{T}_h} \rho_T^2 \lambda_{\max}(M)$ . Wir gewinnen so die Abschätzung

$$\lambda_{\min}(L) \lambda_{\min}(M) \leq \lambda_{\min}(A) \leq \lambda_{\max}(A) \leq c \max_{T \in \mathbb{T}_h} \rho_T^{-2} \lambda_{\max}(M).$$

Also ist  $\kappa_2(A) \leq c \max_{T \in \mathbb{T}_h} \rho_T^{-2}$ , was im Hinblick auf die Gleichförmigkeitsannahmen an die Zerlegungsfolge den Beweis vervollständigt. Q.E.D.

**Bemerkung 3.10:** Wir betonen, dass die Asymptotik  $O(h^{-2})$  der Kondition der Steifigkeitsmatrix durch die Ordnung des zugrunde liegenden Differentialoperators  $L$  bestimmt ist; sie hat *nichts* mit dem Polynomgrad  $m - 1$  des Finite-Elemente-Ansatzes oder der Raumdimension  $d$  zu tun. In der Tat wurde der Beweis von Satz 3.13 allgemein für  $d \geq 1$  und für beliebigen Polynomgrad  $m - 1 \geq 1$  geführt. Die Abhängigkeit von  $\rho := \min_{T \in \mathbb{T}_h} \rho_T$  kommt über die Verwendung der inversen Beziehung zur Abschätzung von  $\lambda_{\max}(A)$  ins Spiel. Dabei ergibt offenbar jede Ableitungsstufe in der Energieform  $a(\cdot, \cdot)$  genau eine negative  $\rho$ -Potenz. Der Exponent  $-2$  ist also gerade durch die Ordnung des betrachteten Differentialoperators bestimmt. Bei der Finite-Elemente-Diskretisierung von Differentialoperatoren höherer Ordnung  $2r \geq 2$  verhält sich die Kondition der Steifigkeitsmatrix dementsprechend wie  $\kappa_2(A) = O(h^{-2r})$ . Zum Beispiel treten in der Plattenstatik Randwertaufgaben vierter Ordnung ( $r = 2$ ) mit dem biharmonischen Operator  $\Delta^2$  auf. In diesem Fall verhält sich die Kondition der zugehörigen Steifigkeitsmatrix wie  $O(h^{-4})$ . Für eine Gitterweite der Größenordnung  $h \sim 10^{-2}$  in zwei Dimensionen ergibt sich damit  $\kappa_2(A) \sim 10^8$ , was Rechnung in mindestens doppelt-genauer Arithmetik nahe legt.

Die Fehlerabschätzung (3.5.122) ist am Extremfall einer Störung des Eigenvektors  $w_{\max}$  in Richtung des Eigenvektors  $w_{\min}$  orientiert. Sie erfassen also den ungünstigsten Fehlereinfluss, wie er in der Praxis kaum auftreten wird. Tatsächlich erweist sich (3.5.122) als viel zu pessimistisch zur realistischen Erfassung des Einflusses von Datenfehlern bei der Lösung der Finite-Elemente-Gleichungen  $Ax = b$ . Zur Verdeutlichung betrachten wir im folgenden ausschließlich den Fall von Störungen in der rechten Seite  $b$ , welche durch fehlerhafte Auswertung (z. B. durch numerische Integration) der gegebenen rechten Seite  $f$  der Differentialgleichung entstehen.

**Satz 3.14 (Spezieller Störungssatz):** Auf einer Folge von (gleichmäßig) regulären Zerlegungen  $\mathbb{T}_h$  gilt die Fehlerabschätzung

$$\frac{|\delta\xi|}{|\xi|} \leq \frac{\kappa_2(M)}{\lambda} \frac{\|f\|}{\|u_h\|} \frac{|\delta b|}{|b|}, \quad (3.5.124)$$

mit dem kleinsten Eigenwert  $\lambda$  der 1. RWA des Differentialoperators  $L$  auf  $\Omega$ .

**Beweis:** Aus der Identität  $A \delta\xi = \delta b$  folgt  $|\delta\xi| \leq \|A^{-1}\| |\delta b| = \lambda_{\min}(A)^{-1} |\delta b|$ . Weiter ist

$$\begin{aligned}\lambda_{\min}(A) &= \min_{z \in \mathbb{R}^N} \frac{\langle Az, z \rangle}{|z|^2} \geq \min_{z \in \mathbb{R}^N} \frac{\langle Az, z \rangle}{\langle Mz, z \rangle} \min_{z \in \mathbb{R}^N} \frac{\langle Mz, z \rangle}{|z|^2} \\ &= \min_{v_h \in V_h} \frac{a(v_h, v_h)}{\|v_h\|^2} \lambda_{\min}(M) \geq \lambda \lambda_{\min}(M),\end{aligned}$$

und somit  $|\delta\xi| \leq \lambda^{-1} \lambda_{\min}(M)^{-1} |\delta b|$ . Mit Hilfe der Schwarzschen Ungleichung ergibt sich

$$|b|^2 = \sum_{n=1}^N (f, \varphi_h^{(n)})^2 = (f, \sum_{n=1}^N (f, \varphi_h^{(n)}) \varphi_h^{(n)}) \leq \|f\| \left\| \sum_{n=1}^N (f, \varphi_h^{(n)}) \varphi_h^{(n)} \right\|.$$

Wegen

$$\left\| \sum_{n=1}^N (f, \varphi_h^{(n)}) \varphi_h^{(n)} \right\|^2 = \sum_{n,m=1}^N (f, \varphi_h^{(m)}) (f, \varphi_h^{(n)}) (\varphi_h^{(m)}, \varphi_h^{(n)}) = \langle Mb, b \rangle \leq \lambda_{\max}(M) |b|^2$$

folgt dann  $|b| \leq \lambda_{\max}(M)^{1/2} \|f\|$ . Ferner gilt wegen  $\langle M\xi, \xi \rangle \leq \lambda_{\max}(M) \|\xi\|^2$ :

$$|\xi| \geq \lambda_{\max}(M)^{-1/2} \langle M\xi, \xi \rangle^{1/2} = \lambda_{\max}(M)^{-1/2} \|u_h\|.$$

Wir kombinieren die obigen Beziehungen und erhalten

$$\frac{|\delta\xi|}{|\xi|} \leq \lambda^{-1} \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)} \frac{|\delta b|}{|b|} \frac{\|f\|}{\|u_h\|},$$

was zu beweisen war.

Q.E.D.

Wir haben in Satz 3.13 gesehen, dass die Massenmatrix  $M$  eine gleichmäßig bzgl.  $h$  beschränkte Spektralkondition hat  $\kappa_2(M) = \mathcal{O}(1)$  ( $h \rightarrow 0$ ). Ferner folgt aus der Konvergenz des Verfahrens die Beziehung

$$\|u_h\| = \|u\| + \mathcal{O}(h) \quad (h \rightarrow 0).$$

Damit erhalten wir aus Satz 3.14 die folgende asymptotische Abschätzung für die Fehlerfortpflanzung im Finite-Elemente-Galerkin-Verfahren

$$\frac{|\delta\xi|}{|\xi|} \leq c(f, u, \Omega) \frac{|\delta b|}{|b|}, \tag{3.5.125}$$

mit einer nur von  $f$ ,  $u$  und  $\Omega$  abhängigen Konstante  $c(f, u, \Omega)$ . Dies besagt, dass die Finite-Elemente-Methode stabil ist (für  $h \rightarrow 0$ ) bzgl. Störungen der rechten Seite  $f$ . Die Ausdehnung dieser Aussage auf Störungen in der Matrix selbst ist noch offen.

### 3.5.3 Aufstellung der Systemmatrizen mit numerischer Integration

Im folgenden analysieren wir den zusätzlichen Fehler, welcher bei näherungsweiser Berechnung der Matrix- und Vektorelemente  $a_{nm}$  und  $b_n$  mittels numerischer Quadratur entsteht. Im Zuge der Matrix-Assemblierung müssen Integrale über Gitterzellen  $T \in \mathbb{T}_h$ ,

$$\int_T a(x) \nabla \varphi_h^{(i)}(x) \nabla \varphi_h^{(j)}(x) dx, \quad \int_T f(x) \varphi_h^{(j)}(x) dx, \quad (3.5.126)$$

berechnet werden, wobei  $\varphi_h^{(i)}(x)$  die Knotenbasisfunktionen sind. Wenn die Datenfunktionen  $a(x)$ ,  $f(x)$  konstant oder auf der Zelle durch Polynome approximiert sind, so sind Integrale der Form

$$\int_T x_1^p x_2^q dx, \quad p, q \in \mathbb{N}_0,$$

zu berechnen. Hierfür gibt es z. B. auf Dreiecken explizite Formeln. Sei  $T$  ein Dreieck in der  $(x, y)$ -Ebene mit den Eckpunkten  $(x_i, y_i)$ ,  $i = 1, 2, 3$ . Ist der Ursprung des Koordinatensystems im Schwerpunkt von  $T$ , d.h.  $x_1 + x_2 + x_3 = y_1 + y_2 + y_3 = 0$ , so gilt z. B.:

$$\begin{aligned} \int_T 1 d(x, y) &= |T|, \\ \int_T x d(x, y) &= \int_T y d(x, y) = 0, \\ \int_T x^2 d(x, y) &= \frac{|T|}{12} (x_1^2 + x_2^2 + x_3^2), \\ \int_T xy d(x, y) &= \frac{|T|}{12} (x_1 y_1 + x_2 y_2 + x_3 y_3), \\ \int_T y^2 d(x, y) &= \frac{|T|}{12} (y_1^2 + y_2^2 + y_3^2). \end{aligned}$$

Nicht exakt berechenbare Integrale werden durch numerische Integration angenähert. Dazu dienen sog. „Quadraturformeln“, welche analog wie auf 1-dimensionalen Intervallen auch auf 2- oder 3-dimensionalen Zellen über einen Interpolationsansatz erzeugt werden. Dies geschieht zunächst auf der Referenzzelle  $\hat{T}$  und ergibt dann durch Transformation  $\hat{T} \rightarrow T = \sigma_T(\hat{T}) = B_T \hat{x} + b_T$  auch Quadraturformeln auf den einzelnen Zellen  $T \in \mathbb{T}_h$ . Wir beschreiben diesen Prozess im folgenden nur für Quadratur basierend auf Lagrange-Interpolation.

Auf der Referenzzelle  $\hat{T}$  seien ein Polynomraum  $P(\hat{T})$  mit  $S := \dim P(\hat{T})$  sowie ein Satz von Stützpunkten  $\{\hat{x}_s \in \hat{T}, s = 1, \dots, S\}$  gewählt, welche unisolvent sind. Die Stützpunkte  $\hat{x}_s$  brauchen nicht mit den Knotenpunkten des Finite-Elemente-Ansatzes übereinzustimmen. Seien weiter  $\hat{L}_s \in P(\hat{T})$  die zugehörigen Lagrangeschen Basispolynome, welche durch die Eigenschaft  $\hat{L}_s(\hat{x}_r) = \delta_{sr}$  charakterisiert sind. Dies erlaubt wieder die explizite Darstellung des zu einer stetigen Funktion  $\hat{v}(\hat{x})$  gehörenden Interpolations-

polynoms durch

$$p(\hat{x}) = \sum_{s=1}^S \hat{v}(\hat{x}_s) \hat{L}_s(\hat{x}).$$

Dies führt zu folgendem Ansatz für eine Quadraturformel auf  $\hat{T}$ :

$$Q_T(\hat{v}) := \sum_{s=1}^S \hat{\omega}_s \hat{v}(\hat{x}_s), \quad \hat{\omega}_s := \int_{\hat{T}} \hat{L}_s(\hat{x}) d\hat{x}. \quad (3.5.127)$$

Die Stützpunkte  $\hat{x}_s$  werden so gewählt, dass Polynome von möglichst hohem Grad durch die Formel exakt integriert werden. Dabei ist aus Stabilitätsgründen wieder darauf zu achten, dass die Gewichte  $\hat{\omega}_s$  positiv sind.

Mit den Bezeichnungen

$$x_s := \sigma_T(\hat{x}_s), \quad \omega_s := |\det B_T| \hat{\omega}_s \quad (s = 1, \dots, S),$$

erhalten wir durch

$$Q_T(v) := \sum_{s=1}^S \omega_s v(x_s) := \sum_{s=1}^S |\det B_T| \hat{\omega}_s \hat{v}(\hat{x}_s) = |\det B_T| Q_{\hat{T}}(\hat{v}) \quad (3.5.128)$$

Quadraturformeln  $Q_T(\cdot)$  auf den einzelnen Zellen  $T \in \mathbb{T}_h$ . Die gebräuchlichsten solcher Formeln für Dreiecke sind in Abb. 3.22 zusammengestellt. Dabei bedienen wir uns der gebräuchlichen Schreibweise mit sog. „baryzentrischen Koordinaten“. Für ein  $d$ -Simplex  $T$  mit Eckpunkten  $\{a_0, \dots, a_d\}$  besitzt jeder Punkt  $x \in T$  eine eindeutige Darstellung als konvexe Linearkombination der Eckpunkte:

$$x = \sum_{i=0}^d \lambda_i a_i, \quad 0 \leq \lambda \leq 1, \quad \sum_{i=0}^d \lambda_i = 1.$$

Die Koeffizienten  $\{\lambda_0, \dots, \lambda_d\}$  sind dann die baryzentrischen Koordinaten von  $x$  im Simplex  $T$ . Zum Beispiel sind  $\{1, 0, \dots, 0\}$  die baryzentrischen Koordinaten des Eckpunkts  $a_0$  und  $\{\frac{1}{d+1}, \dots, \frac{1}{d+1}\}$  die des Mittelpunktes  $z_T$ . Die einfachsten Quadraturformeln auf Dreiecken/Simplizes sind die „Mittelpunktregel“ und die „Trapezregel“ (s. auch Abb. 3.22):

$$Q_T(v) := |T| v(z_T), \quad Q_T(v) := \frac{1}{d+1} |T| \sum_{i=0}^d v(a_i).$$

**Definition 3.6 (Quadraturformel):** Eine interpolatorische Quadraturformel der Art (3.5.127) auf einer Referenzzelle  $T$  heißt „von der Ordnung  $r$ “, wenn durch sie Polynome bis zum Grad  $r-1$  (und nicht höher) exakt integriert werden. Sie wird „zulässig“ für den Polynomansatz  $P(T)$  genannt, wenn ihre Stützstellenmenge reichhaltig genug ist, so dass

$$q \in P(T) : \quad \nabla q(x_s) = 0 \quad (s = 1, \dots, S) \quad \Rightarrow \quad q \equiv \text{konst.} \quad (3.5.129)$$

## NUMERICAL INTEGRATION FORMULAS FOR TRIANGLES

Order	Fig.	Error	Points	Triangular Co-ordinates	Weights $2W_k$
Linear		$R = O(h^2)$	$a$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	1
Quadratic		$R = O(h^4)$	$a$ $b$ $c$	$\frac{1}{3}, \frac{1}{3}, 0$ $0, \frac{1}{3}, \frac{1}{3}$ $\frac{1}{3}, 0, \frac{1}{3}$	$\frac{1}{3}$ $\frac{1}{3}$ $\frac{1}{3}$
Cubic		$R = O(h^6)$	$a$ $b$ $c$ $d$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ $0, \frac{1}{2}, \frac{1}{2}$ $\frac{1}{2}, 0, \frac{1}{2}$ $\frac{1}{2}, \frac{1}{2}, 0$ $0, \frac{1}{2}, 0$	$-\frac{27}{448}$ $-\frac{27}{448}$ $\frac{8}{448}$ $\frac{8}{448}$ $\frac{3}{448}$ $\frac{3}{448}$
This formula not recommended due to negative weight and round-off error					
Cubic		$R = O(h^8)$	$a$ $b$ $c$ $d$ $e$ $f$ $g$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ $\frac{1}{3}, \frac{1}{3}, 0$ $0, \frac{1}{2}, \frac{1}{2}$ $\frac{1}{2}, 0, \frac{1}{2}$ $1, 0, 0$ $0, 1, 0$ $0, 0, 1$	$-\frac{27}{64}$ $\frac{8}{64}$ $\frac{3}{64}$
Quintic		$R = O(h^{10})$	$a$ $b$ $c$ $d$ $e$ $f$ $g$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ $x_1, \beta_1, \beta_1$ $\beta_1, x_1, \beta_1$ $\beta_1, \beta_1, x_1$ $x_2, \beta_2, \beta_2$ $\beta_2, x_2, \beta_2$ $\beta_2, \beta_2, x_2$	0.225 with $x_1 = 0.05971587$ $\beta_1 = 0.47014206$ $x_2 = 0.79742699$ $\beta_2 = 0.10128651$

Abbildung 3.22: Beispiele von Quadraturformeln auf Dreiecken.

Auf Vierecken bzw. Hexaedern werden sog. „Tensorproduktformeln“ verwendet. Diese erhält man ebenfalls über Transformation von der Referenzzelle. Demgemäß lauten auf dem Quadrat/Quader mit Mittelpunkt  $z_T$  und Eckpunkten  $\{a_1, \dots, a_{2d}\}$  die Mittelpunktsregel sowie die (Tensorprodukt)-Trapezregel:

$$Q_T(v) := |T|v(z_T), \quad Q_T(v) := \frac{1}{2^d}|T| \sum_{i=0}^{2^d} v(a_i).$$

Zur Erzielung höherer Genauigkeit werden extrapolierte Trapez-Formeln (z. B. Tensorprodukt-Simpson-Regel) oder Tensorprodukt-Gauß-Formeln verwendet, welche ähnlich wie im eindimensionalen Fall konstruiert sind. Ein der gebräuchlichsten Formeln für Quadrate ist die 4-Punkt-Gauß-Formel

$$Q_T(v) := \frac{1}{4} \sum_{i=1}^4 v(\xi_i),$$

mit den Gauß-Punkten  $\xi_1 = (\eta_-, \eta_-)$ ,  $\xi_2 = (\eta_-, \eta_+)$ ,  $\xi_3 = (\eta_+, \eta_-)$  und  $\xi_4(\eta_+, \eta_+)$ , wobei  $\eta_\pm := \frac{1}{2}(1 \pm \sqrt{1/3})$ .

**Satz 3.15 (Quadraturfehler):** Für eine interpolatorische Quadraturformel  $Q_T(\cdot)$  der Ordnung  $r \geq d$  auf einer Zelle  $T \in \mathbb{T}_h$  angewendet auf eine Funktion  $v \in H^{r,1}(T)$  gilt

$$\left| \int_T v \, dx - Q_T(v) \right| \leq c_q h_T^r \int_T |\nabla^r v| \, dx, \quad (3.5.130)$$

mit einer von  $T$  und  $v \in H^{r,1}(T)$  unabhängigen Quadraturkonstanten  $c_q > 0$ .

**Beweis:** Der Beweis verwendet das Bramble-Hilbert-Lemma 3.5 in Verbindung mit dem Transformationsargument von Satz 3.6. Auf der Referenzzelle  $\hat{T}$  definieren wir das Fehlerfunktional

$$F(\hat{v}) := \left| \int_{\hat{T}} \hat{v}(\hat{x}) \, d\hat{x} - Q_{\hat{T}}(\hat{v}) \right|.$$

Zur Definition von  $F(\cdot)$  benötigen wir Punktwerte von  $\hat{v}$ . Diese sind aufgrund des Sobolewschen Einbettungssatzes in zwei Dimensionen für  $v \in H^{2,1}(\Omega)$  und in drei Dimensionen für  $v \in H^{3,1}(\Omega)$  wohl definiert. Es gilt dann

$$|F(\hat{v})| \leq c \|\hat{v}\|_{H^{r,1}}.$$

Ferner ist  $F(\cdot)$  offensichtlich sublinear und verschwindet nach Voraussetzung auf  $P_{r-1}$ . Nach der  $L^1$ -Variante des Bramble-Hilbert-Lemmas gilt dann

$$|F(\hat{v})| \leq c \|\hat{\nabla}^r \hat{v}\|_{L^1(\hat{T})}.$$

Sei nun  $\sigma_T$  wieder die affin-lineare Transformation von  $\hat{T}$  auf die Zelle  $T$ . Dann gilt für eine Funktion  $v \in H^{r,1}(\Omega)$  und ihre Transformierte  $\hat{v}(\hat{x}) := v(x)$ ,  $x = \sigma_T(\hat{x})$ :

$$\left| \int_T v(x) \, dx - Q_T(v) \right| = |\det B_T| |F(\hat{v})|,$$

sowie

$$\|\hat{\nabla}^r \hat{v}\|_{L^1(\hat{T})} \leq c |\det B_T|^{-1} h_T^r \|\nabla^r v\|_{L^1(T)}.$$

Kombination der letzten beiden Beziehungen ergibt die Behauptung. Q.E.D.

**Bemerkung 3.11:** Die Voraussetzung  $r \geq d$  in Satz 3.15 dient zur Vereinfachung der Formulierung des Resultats. Wegen der Einbettung  $H^{d,1}(T) \subset C(\bar{T})$  sind für Funktionen

$v \in H^{r,1}(T)$  die für die Anwendung der Quadraturformel erforderlichen Punktwerte wohl definiert. Dies ist aber auch gerade die „richtige“ Regularitätsstufe für die maximale Ausnutzung der Approximationsgüte der Quadraturformel. Dies ist wichtig für die folgende Untersuchung des Einflusses des Quadraturfehlers auf den Gesamtfehler. Im Übrigen ist jede der gebräuchlichen Quadraturformeln (z. B. Mittelpunkts- oder Trapezregel) mindestens von der Ordnung  $r = 2$ , so dass in zwei Dimensionen gar keine Einschränkung besteht. In drei Dimensionen bedarf der Fall  $r = 2$  eine gesonderte Betrachtung, die obwohl nicht schwer, hier nicht durchgeführt wird.

Im Folgenden werden wir uns auf die Quadratur auf Dreiecken bzw. Simplizes beschränken und nehmen außerdem an, dass der Polynomansatz ein voller Polynomraum ist:  $P(\hat{T}) = P_{m-1}(\hat{T})$ . Analoge Resultate gelten auch für die Quadratur auf Vierecken bzw. Hexaedern, doch erfordern deren Formulierung und Beweis eine aufwendigere Notation.

Die Anwendung von Quadraturformeln zur Berechnung der Zellintegrale (3.5.126) ergibt eine gestörte Bilinearform und rechte Seite

$$a_h(u, \varphi) := \sum_{T \in \mathbf{T}_h} Q_T(a \nabla u \nabla \varphi), \quad l_h(\varphi) := \sum_{T \in \mathbf{T}_h} Q_T(f \varphi)$$

sowie zugehörige gestörte Steifigkeitsmatrix- und Lastvektorelemente

$$\tilde{a}_{ij} := \sum_{T \in \mathbf{T}_h} Q_T(a \nabla \varphi_h^{(j)} \nabla \varphi_h^{(i)}), \quad \tilde{b}_j := \sum_{T \in \mathbf{T}_h} Q_T(f \varphi_h^{(i)}).$$

Statt der exakten Finite-Elemente-Lösung  $u_h \in V_h$  ist dann eine gestörte Approximation  $\tilde{u}_h \in V_h$  zu bestimmen durch

$$a_h(\tilde{u}_h, \varphi_h) = l_h(\varphi_h) \quad \forall \varphi_h \in V_h. \quad (3.5.131)$$

**Satz 3.16 (Numerische Integration):** *Die Quadraturformel auf der Referenzzelle  $Q_{\hat{T}}(\cdot)$  sei „zulässig“ für den Finite-Elemente-Ansatz  $P(\hat{T})$  und von der Ordnung  $r \geq d$ . Dann besitzen die gestörten Finite-Elemente-Gleichungen (3.5.131) eindeutige Lösungen  $\tilde{u}_h \in V_h$ . Ist  $a \in C^r(\bar{\Omega})$ , so gilt für das gestörte Finite-Elemente-Verfahren der Ordnung  $m \geq 2$  die Fehlerabschätzung*

$$\|u - \tilde{u}_h\| + h \|\nabla(u - \tilde{u}_h)\| \leq c h^{\min\{m, r+3-m\}} \|u\|_{H^m}. \quad (3.5.132)$$

Zur Erzielung einer maximalen Konvergenzordnung ist also  $r \geq 2m - 3$  zu wählen.

**Beweis:** i) Koerzitivität: Auf einer Zelle  $T \in \mathbf{T}_h$  gilt unter Verwendung der Transformation  $T = \sigma_T(\hat{T})$  wegen  $a \geq \alpha > 0$ :

$$\begin{aligned}
Q_T(a|\nabla v_h|^2) &= \sum_{s=1}^S \omega_s a(x_s) |\nabla v_h(x_s)|^2 \geq \alpha \sum_{s=1}^S \omega_s |\nabla v_h(x_s)|^2 \\
&= \alpha \sum_{s=1}^S |\det B_T| \hat{\omega}_s |B_T^{-1} \hat{\nabla} \hat{v}_h(\hat{x}_s)|^2 \geq \alpha \sum_{s=1}^S |\det B_T| \hat{\omega}_s \|B_T\|^{-2} |\hat{\nabla} \hat{v}_h(\hat{x}_s)|^2 \\
&\geq \alpha c |\det B_T| h_T^{-2} \sum_{s=1}^S \hat{\omega}_s |\hat{\nabla} \hat{v}_h(\hat{x}_s)|^2,
\end{aligned}$$

wobei  $\hat{v}_h(\hat{x}) = v_h(x)$ ,  $\hat{x} = \sigma_T^{-1}(x)$ . Analog gilt

$$\|\hat{\nabla} \hat{v}_h\|_T^2 \geq c |\det B_T|^{-1} h_T^2 \|\nabla v_h\|_T^2.$$

Durch

$$\|v_h\|_T := \left( \sum_{s=1}^S \hat{\omega}_s |\hat{\nabla} \hat{v}_h(\hat{x}_s)|^2 \right)^{1/2}$$

ist auf dem Quotientenraum  $P(\hat{T})/P_0$  eine Norm definiert. Dazu muss nur noch die Definitheit gezeigt werden. Aus  $\|v_h\|_T = 0$  folgt offenbar  $\hat{\nabla} \hat{v}_h(\hat{x}_s) = 0$  ( $s = 1, \dots, S$ ), so dass wegen der vorausgesetzten „Zulässigkeit“ der Quadraturformel notwendig  $v_h \equiv \text{konst.}$  ist. Da auch  $\|\hat{\nabla} \hat{v}_h\|_T$  auf  $P(\hat{T})/P_0$  eine Norm ist, gilt wegen der Äquivalenz aller Normen auf einem endlich dimensionalen Vektorraum mit einer Konstante  $\hat{c} > 0$ :

$$\|v_h\|_T \geq \hat{c} \|\hat{\nabla} \hat{v}_h\|_T.$$

Dies impliziert dann

$$Q_T(a|\nabla v_h|^2) \geq c |\det B_T| h_T^{-2} \|v_h\|_T^2 \geq c |\det B_T| h_T^{-2} \|\hat{\nabla} \hat{v}_h\|_T^2 \geq c \|\nabla v_h\|_T^2,$$

bzw. die gleichmäßige Koerzitivität der Bilinearformen  $a_h(\cdot, \cdot)$ :

$$a_h(v_h, v_h) \geq c \|\nabla v_h\|_T^2, \quad v_h \in V_h. \quad (3.5.133)$$

ii) Fehlerabschätzung: Mit Hilfe der Koerzitivitätsabschätzung (3.5.133) folgt für den Fehler  $\eta_h := u_h - \tilde{u}_h \in V_h$ :

$$\begin{aligned}
c \|\nabla \eta_h\|^2 &\leq a_h(u_h - \tilde{u}_h, \eta_h) = a_h(u_h, \eta_h) - a_h(\tilde{u}_h, \eta_h) \\
&\leq (a_h - a)(u_h, \eta_h) + (l - l_h)(\eta_h).
\end{aligned}$$

Dies impliziert

$$\|\nabla \eta_h\| \leq c \max_{v_h \in V_h} \left\{ \frac{|(a_h - a)(u_h, v_h)|}{\|\nabla v_h\|} + \frac{|(l - l_h)(v_h)|}{\|\nabla v_h\|} \right\}.$$

Zusammen mit der bekannten  $H^1$ -Fehlerabschätzung

$$\|\nabla(u - u_h)\| \leq ch^m \|u\|_{H^m}$$

folgt

$$\|\nabla(u - \tilde{u}_h)\| \leq ch^m \|u\|_{H^m} + c \max_{v_h \in V_h} \left\{ \frac{|(a_h - a)(u_h, v_h)|}{\|\nabla v_h\|} + \frac{|(l - l_h)(v_h)|}{\|\nabla v_h\|} \right\}.$$

Im nächsten Schritt werden wir zeigen, dass

$$\max_{v_h \in V_h} \left\{ \frac{|(a_h - a)(u_h, v_h)|}{\|\nabla v_h\|} + \frac{|(l - l_h)(v_h)|}{\|\nabla v_h\|} \right\} \leq c h^{\min\{m, r-m+2\}} \|u\|_{H^m}. \quad (3.5.134)$$

Dies impliziert dann den ersten Teil der Behauptung

$$\|\nabla(u - \tilde{u}_h)\| \leq ch^{\min\{m, r-m+2\}} \|u\|_{H^m}. \quad (3.5.135)$$

iii) Konsistenz: Als nächstes schätzen wir den „Abstand“ zwischen  $a(\cdot, \cdot)$  und  $a_h(\cdot, \cdot)$  sowie  $l(\cdot)$  und  $l_h(\cdot)$  ab. Für  $u_h, v_h \in V_h$  folgt mit Hilfe der Abschätzung des Integrationsfehlers in Satz 3.15:

$$\begin{aligned} |(a - a_h)(u_h, v_h)| &\leq \sum_{T \in \mathbb{T}_h} \left| \int_T a \nabla u_h \nabla v_h \, dx - Q_T(a \nabla u_h \nabla v_h) \right| \\ &\leq c_q \sum_{T \in \mathbb{T}_h} h_T^r \int_T |\nabla^r(a \nabla u_h \nabla v_h)| \, dx. \end{aligned}$$

Durch Ausdifferenzieren erhalten wir

$$|(a - a_h)(u_h, v_h)| \leq c \sum_{T \in \mathbb{T}_h} h_T^r \|\nabla u_h\|_{H^r(T)} \|\nabla v_h\|_{H^r(T)},$$

wobei die Konstante  $c$  wesentlich durch Schranken für die Ableitungen der Koeffizientenfunktion  $a(x)$  bestimmt ist. Bei Beachtung von  $v_{h|T} \in P_{m-1}$  ergibt sich

$$|(a - a_h)(u_h, v_h)| \leq c \sum_{T \in \mathbb{T}_h} h_T^r \|\nabla u_h\|_{H^{m-2}(T)} \|\nabla v_h\|_{H^{m-2}(T)}.$$

Mit Hilfe der „inversen Beziehung“ für finite Elemente gilt

$$\|\nabla v_h\|_{H^{m-2}(T)} \leq ch_T^{2-m} \|\nabla v_h\|_T,$$

womit folgt:

$$|(a - a_h)(u_h, v_h)| \leq ch^{r-m+2} \left( \sum_{T \in \mathbb{T}_h} \|\nabla u_h\|_{H^{m-2}(T)}^2 \right)^{1/2} \|\nabla v_h\|.$$

Die Terme in  $u_h$  werden unter Verwendung der Knoteninterpolierenden  $I_h u \in V_h$  und mit Hilfe der inversen Beziehung sowie der lokalen Interpolationsabschätzungen wie folgt abgeschätzt:

$$\begin{aligned}
\|\nabla u_h\|_{H^{m-2}(T)} &\leq \|u_h - I_h u\|_{H^{m-1}(T)} + \|I_h u - u\|_{H^{m-1}(T)} + \|u\|_{H^{m-1}(T)} \\
&\leq ch_T^{2-m} \|u_h - I_h u\|_{H^1(T)} + ch_T \|\nabla^m u\|_T + \|u\|_{H^{m-1}(T)} \\
&\leq ch_T^{2-m} \|u_h - u\|_{H^1(T)} + ch_T^{2-m} \|u - I_h u\|_{H^1(T)} + \|u\|_{H^m(T)} \\
&\leq ch_T^{2-m} \|u_h - u\|_{H^1(T)} + \|u\|_{H^m(T)}.
\end{aligned}$$

Zusammenfassen der bisherigen Abschätzungen ergibt

$$|(a - a_h)(u_h, v_h)| \leq ch^{r+2-m} \{ h^{2-m} \|u_h - u\|_{H^1} + \|u\|_{H^m} \} \|\nabla v_h\|,$$

und unter Verwendung der bekannten  $H^1$ -Fehlerabschätzung  $\|u_h - u\|_{H^1} \leq ch^{m-1} \|u\|_{H^m}$ :

$$|(a - a_h)(u_h, v_h)| \leq ch^{r+2-m} \|u\|_{H^m} \|\nabla v_h\|. \quad (3.5.136)$$

Auf analoge Weise erschließen wir

$$|(l - l_h)(v_h)| \leq ch^{r-m+2} \|\nabla v_h\| \|u\|_{H^m}. \quad (3.5.137)$$

Die Abschätzungen (3.5.136) und (3.5.137) ergeben (3.5.134) und damit das Resultat (3.5.135).

iv)  $L^2$ -Fehlerabschätzung: Zur Abschätzung des Fehlers in der  $L^2$ -Norm bedienen wir uns wieder eines Dualitätsarguments. Sei  $z \in V$  die (eindeutige) Lösung des Hilfsproblems

$$-\nabla \cdot \{a \nabla z\} = u_h - \tilde{u}_h \quad \text{in } \Omega, \quad z = 0 \quad \text{auf } \partial\Omega.$$

Wegen der angenommenen Glattheit von  $a$  und der Konvexität von  $\Omega$  ist  $z \in H^2(\Omega)$  und genügt der a priori Abschätzung

$$\|z\|_{H^2} \leq c \|u_h - \tilde{u}_h\|.$$

Sei  $z_h \in V_h$  die Ritz-Projektion von  $z$ . Mit dieser Konstruktion erhalten wir mit Hilfe der Galerkin-Orthogonalität:

$$\begin{aligned}
\|u_h - \tilde{u}_h\|^2 &= a(u_h - \tilde{u}_h, z) = a(u_h - \tilde{u}_h, z_h) \\
&= a(u_h, z_h) - (a - a_h)(\tilde{u}_h, z_h) - a_h(\tilde{u}_h, z_h) \\
&= (l - l_h)(z_h) - (a - a_h)(\tilde{u}_h, z_h).
\end{aligned}$$

Die beiden Störungsterme werden nun analog wie unter (iii) abgeschätzt. Zunächst gilt wieder mit Hilfe der inversen Beziehung:

$$\begin{aligned}
|(a - a_h)(\tilde{u}_h, z_h)| &\leq c \sum_{T \in \mathbf{T}_h} h_T^r \|\nabla \tilde{u}_h\|_{H^{m-2}(T)} \|\nabla z_h\|_{H^{m-2}(T)} \\
&\leq ch^{r+3-m} \left( \sum_{T \in \mathbf{T}_h} \|\nabla \tilde{u}_h\|_{H^{m-2}(T)}^2 \right)^{1/2} \left( \sum_{T \in \mathbf{T}_h} \|\nabla z_h\|_{H^1(T)}^2 \right)^{1/2}.
\end{aligned}$$

Für die beiden Summen erhalten wir analog wie oben unter (iii):

$$\left( \sum_{T \in \mathbb{T}_h} \|\nabla \tilde{u}_h\|_{H^{m-2}(T)}^2 \right)^{1/2} \leq ch_T^{2-m} \|\tilde{u}_h - u\|_{H^1} + \|u\|_{H^m},$$

sowie

$$\left( \sum_{T \in \mathbb{T}_h} \|\nabla z_h\|_{H^1(T)}^2 \right)^{1/2} \leq c\|z\|_{H^2} \leq c\|u_h - \tilde{u}_h\|.$$

Kombination dieser Abschätzungen und Berücksichtigung der schon bewiesenen  $H^1$ -Fehlerabschätzung ergibt dann:

$$|(a - a_h)(\tilde{u}_h, z_h)| \leq ch^{r-m+3} \|u_h - \tilde{u}_h\| \|u\|_{H^m}. \quad (3.5.138)$$

Analog erschließen wir

$$|(l - l_h)(z_h)| \leq ch^{r-m+3} \|u_h - \tilde{u}_h\|. \quad (3.5.139)$$

Kombinieren aller vorausgehenden Beziehungen ergibt

$$\|u_h - \tilde{u}_h\| \leq ch^{r-m+3} \|u\|_{H^m}.$$

Zusammen mit der bekannten  $L^2$ -Fehlerabschätzung für  $e_h = u - u_h$  erhalten wir schließlich

$$\|u - \tilde{u}_h\| \leq ch^{\min\{m, r-m+3\}} \|u\|_{H^m},$$

was den Beweis vervollständigt. Q.E.D.

**Bemerkung 3.12:** Die Aussage von Satz 3.16 kann so interpretiert werden, dass zur Vermeidung einer Reduzierung der Konvergenzordnung  $\mathcal{O}(h^{m-1})$  in der Energie-Norm eine Quadraturformel der Ordnung  $r \geq 2m - 3$  verwendet werden sollte. Damit würden dann im Falle eines konstanten Koeffizienten  $a$  die Matrixelemente  $a_{nm}$  exakt berechnet werden. Für bloße Konvergenz  $\tilde{u}_h \rightarrow u$  ( $h \rightarrow 0$ ) wäre die Wahl  $r \geq \max\{d, m - 2\}$  ausreichend, vorausgesetzt die Zulässigkeitsbedingung ist erfüllt.

### 3.6 A posteriori Fehleranalyse und Gittersteuerung

Eine wichtige Rolle bei der Lösung von partiellen Differentialgleichungen spielen die beiden Aspekte Fehlerkontrolle und Gittersteuerung. Hat man eine approximative Lösung  $u_h$  berechnet, ist es von Interesse, den Fehler zwischen dieser Näherung und der exakten Lösung  $u$  bzgl. eines geeigneten Maßes abzuschätzen. Zu diesem Zweck dienen sog. „a posteriori Fehlerschätzer“, welche im besten Fall ausschließlich von berechneten Größen und den Daten  $f$  abhängen. In diesem Abschnitt wollen wir einfache residuen-basierte Fehlerschätzer für das Modellproblem der 1. RWA des Laplace-Operators herleiten.

Eine globale Verfeinerung des ganzen Rechengebietes ist in drei Dimensionen in der Regel nicht realisierbar, da der Speicherplatz auf den verfügbaren Rechnern dazu nicht

ausreicht. Deshalb versucht man, nur dort lokal zu verfeinern, wo es die Lösungsstruktur bzw. die Genauigkeitsanforderungen verlangen. Unter *optimaler Gittersteuerung* wird dabei verstanden, möglichst wenige markierte Elemente des Gitters zu verfeinern, so dass der Fehler in möglichst wenigen Schritten unter eine vorgegebene Toleranz gedrückt wird. Die Wahl der zu verfeinernden Elemente trifft man aufgrund sog. „lokaler Fehlerindikatoren“, aus denen sich der globale Fehlerschätzer zusammensetzt. Solche Gittersteuerungsmethoden werden wir im zweiten Teil dieses Abschnitts diskutieren.

### 3.6.1 Allgemeine a posteriori Fehlerabschätzung

Aus der a priori Fehlerschätzung aus Abschnitt 3.4 konnten wir die Konvergenzordnung des Diskretisierungsfehlers  $e_h = u - u_h$  schon für verschiedene Normen herleiten:

- Energienorm-Fehlerabschätzung:  $\|\nabla e_h\| \leq c_I c_S h \|u\|_{H^2}$ ,
- $L^2$ -Norm-Fehlerabschätzung:  $\|e_h\| \leq c_I c_S h^2 \|u\|_{H^2}$ ,
- $L^\infty$ -Abschätzung:  $\max_{\bar{\Omega}} |e_h| \leq c_I c_S h^2 L(h) M_2(u)$ ,

mit *lokalen* Interpolationskonstanten  $c_I$  und *globalen* Stabilitätskonstanten  $c_S$ . Leider sind diese Abschätzungen für eine quantitative Fehlerkontrolle nicht zu gebrauchen, da die nötigen Informationen über die höheren Ableitungen der exakten Lösung  $u$  fehlen und insbesondere präzise Abschätzungen für die Stabilitätskonstanten  $c_S$  i. Allg. nicht zur Verfügung stehen. Ist aber der Charakter einer lokalen Singularität der Lösung bekannt, wie z. B. im Fall von „Ecken- oder Kantensingularitäten“, so kann diese Information zur vorab Anpassung des Gitters verwendet werden. Dabei wird die Gitterweite  $h$  in Richtung auf die singuläre Stelle hin systematisch verkleinert, etwa gemäß  $h(r) \approx h_0 r^\alpha$  wobei der Exponent  $\alpha > 1$  aus dem bekannten singulären Verhalten der Ableitungen der Lösung abgeleitet wird.

In allgemeinen Situationen muss man sich aber heuristischer Methoden zur Bestimmung der Regularität der Lösung bedienen. Dies läuft (in Anlehnung an die traditionelle, abschneidefehler-basierte Vorgehensweise bei Differenzenverfahren) auf die Schätzung der lokalen Glattheit der unbekannten Lösung aus der berechneten numerischen Approximation hinaus. Zum Beispiel kann man versuchen, auf einem Zellblock (etwa aus  $2^d$  Zellen) aus einer linearen Näherungslösung  $u_h$  durch Anwendung eines Differenzenquotienten  $\nabla_h^2$  zweiter Ordnung im Schwerpunkt  $z_T$  einer Zelle  $T \in \mathbb{T}_h$  eine Schätzung der zweiten Ableitungen von  $u$  auf  $T$  zu gewinnen:

$$\max_T |\nabla^2 u|_T \approx \eta_T := |\nabla_h^2 u_h(z_T)|_T. \quad (3.6.140)$$

Auf der Basis dieses Indikators ließen sich dann Strategien zu lokaler Gitterverfeinerung oder -vergrößerung aufstellen: Ist zum Beispiel  $\eta_T$  auf einer Zelle  $T \in \mathbb{T}_h$  überdurchschnittlich groß, so wird diese in Teilzellen zerlegt. Diese Strategie der *ad hoc* Gitteranpassung erfordert keinen großen Aufwand und funktioniert in der Praxis in vielen Fällen erstaunlich gut. Daher sind Varianten dieser Strategie derzeit auch in vielen kommerziellen

Programmen realisiert, wenn diese überhaupt Gitteradaption beinhalten. Dabei bestehen aber die folgenden grundsätzlichen Schwächen:

- Die Auswertung von (3.6.140) liefert keine Aussage über die tatsächliche Größe des Fehlers  $e_h = u - u_h$ .
- Die auf den lokalen Indikatoren  $\eta_T$  basierende Gitterverfeinerungsstrategie geht davon aus, dass der „gemessene“ Fehler in  $T$  auch dort entstanden ist und durch lokale Verfeinerung von  $T$  reduziert werden kann. Dies ist aber i. Allg. nicht richtig, da dabei das Phänomen der globalen „Fehlerakkumulation“ (auch „pollution effect“ genannt) vernachlässigt wird.

Ziel der folgenden Diskussion ist es, über eine a posteriori Fehleranalyse via Dualitätsargumente systematisch auswertbare und (asymptotisch) zuverlässige Fehlerschätzer zu entwickeln, aus denen auch effiziente Kriterien zur lokalen Gitteranpassung abgeleitet werden können.

Wir betrachten dazu wieder das Modellproblem

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega, \quad (3.6.141)$$

auf einem (nicht notwendig konvexen) Polygon- oder Polyedergebiet  $\Omega \subset \mathbb{R}^d$ . Die durch die Variationsgleichung

$$(\nabla u, \nabla \varphi) = (f, \varphi) \quad \forall \varphi \in V = H_0^1(\Omega), \quad (3.6.142)$$

definierte schwache Lösung  $u \in H_0^1(\Omega)$  wird durch ein Galerkin-Finite-Elemente-Verfahren approximiert. Wir konzentrieren uns im folgenden auf die Approximation mit „linearen“ finiten Elementen. Die Näherungslösung  $u_h = R_h u \in V_h^{(1)} \subset V$  ist bestimmt durch die Gleichung

$$(\nabla u_h, \nabla \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h. \quad (3.6.143)$$

Bei der Schätzung des Fehlers  $e_h$  muss man sich für ein geeignetes „Fehlermaß“ entscheiden, welches sich am Bedarf der betrachteten Anwendung orientieren sollte. Beispiele sind etwa wie schon oben erwähnt die traditionellen Maße „Energienorm“,  $L^2$ -Norm sowie  $L^\infty$ -Norm. Weitere Beispiele sind:

- Mittelwerte:  $|(\epsilon_h, \psi)_\Omega|$ ,  $\psi \in C(\bar{\Omega})$ ,
- Linienintegrale:  $|(\epsilon_h, \psi)_\Gamma|$ ,  $\psi \in C(\partial\Omega)$ ,
- Ableitungswerte:  $|\partial_i \epsilon_h(a)|$ ,  $a \in \Omega$ .

Um alle diese Sonderfälle im Rahmen einer einheitlichen Theorie behandeln zu können, führen wir zunächst den Begriff des „Fehlerfunktional“ ein. Dies ist ein (der Einfachheit halber als linear angenommenes) Funktional

$$J(\cdot) : V \rightarrow \mathbb{R},$$

bzgl. dem der Fehler  $e_h$  geschätzt werden soll; d. h.: Gesucht ist eine berechenbare Schranke für  $J(e_h) = J(u) - J(u_h)$ . Zu einem solchen Fehlerfunktional gehört eine (eindeutige) „duale Lösung“  $z \in V$ , welche als Lösung des zugehörigen „dualen Problems“ bestimmt ist:

$$(\nabla \varphi, \nabla z) = J(\varphi) \quad \forall \varphi \in V. \quad (3.6.144)$$

Testen wir in (3.6.144) mit  $\varphi = e_h \in V$ , so ergibt sich unter Berücksichtigung der Galerkin-Orthogonalität die Fehleridentität

$$J(e_h) = (\nabla e_h, \nabla z) = (\nabla e_h, \nabla(z - \psi_h)), \quad \psi_h \in V_h. \quad (3.6.145)$$

Dies wird nun weiter mit Hilfe von elementweise partieller Integration umgeformt zu

$$\begin{aligned} J(e_h) &= \sum_{T \in T_h} \left\{ -(\Delta e_h, z - \psi_h)_T + (\partial_n e_h, z - \psi_h)_{\partial T} \right\} \\ &= \sum_{T \in T_h} \left\{ (f + \Delta u_h, z - \psi_h)_T - (\partial_n u_h, z - \psi_h)_{\partial T} \right\}. \end{aligned}$$

Da  $z - \psi_h$  stetige Spuren entlang  $\partial T$  hat und in der obigen Summe jede Kante zweimal auftritt mit wechselnder Richtung der Normalableitung  $\partial_n u_h$ , erhalten wir

$$J(e_h) = \sum_{T \in T_h} \left\{ (f + \Delta u_h, z - \psi_h)_T - \frac{1}{2}([\partial_n u_h], z - \psi_h)_{\partial T} \right\}, \quad (3.6.146)$$

wobei auf „inneren“ Kanten  $[\partial_n u_h]_{|\Gamma} := \partial_n u_h|_T + \partial_{n'} u_h|_T$ , jeweils den Sprung der Normalableitung über  $\Gamma \subset \partial T$  zur Nachbarzelle bedeutet. Entlang von Kanten am Rand,  $\Gamma \subset \partial\Omega$ , setzen wir  $[\partial_n u_h]_{|\Gamma} := 2\partial_n u_h$ . Diese Fehleridentität beinhaltet Information über die Abhängigkeit des Fehlerterms  $J(e_h)$  von den lokalen „Residuen“  $R_T := (f + \Delta u_h)|_T$  und  $r_{\partial T} := -\frac{1}{2}[\partial_n u_h]|_{\partial T}$ :

$$\frac{\partial J(e_h)}{\partial R_T} \approx (z - \psi_h)|_T, \quad \frac{\partial J(e_h)}{\partial r_{\partial T}} \approx (z - \psi_h)|_{\partial T}. \quad (3.6.147)$$

Damit können wir hoffen, auf das spezielle Zielfunktional  $J(e_h)$  bezogene Kriterien für lokale Gitterverfeinerung und damit Reduzierung dieser Residuen zu gewinnen. Dabei „misst“ das „Zellresiduum“  $R(u_h)$  das Erfülltsein der Differentialgleichung durch die Näherungslösung  $u_h$  und das „Kantenresiduum“  $r(u_h)$  deren „Glattheit“.

Von der exakten *Fehlerdarstellung* (3.6.146) können wir nun in mehreren Schritten weitere, gegebenenfalls leichter auswertbare *Fehlerabschätzungen* ableiten. Zunächst ist

$$|J(e_h)| \leq \left| \sum_{T \in T_h} \left\{ (f + \Delta u_h, z - \psi_h)_T - \frac{1}{2}([\partial_n u_h], z - \psi_h)_{\partial T} \right\} \right|, \quad (3.6.148)$$

und weiter

$$|J(e_h)| \leq \sum_{T \in \mathbf{T}_h} |(f + \Delta u_h, z - \psi_h)_T - \frac{1}{2}([\partial_n u_h], z - \psi_h)_{\partial T}|. \quad (3.6.149)$$

Bei Beachtung von  $z - \psi_h|_{\partial\Omega} = 0$  folgt dann mit Hilfe der Hölderschen Ungleichung:

$$|J(e_h)| \leq \sum_{T \in \mathbf{T}_h} \{ \|f + \Delta u_h\|_T \|z - \psi_h\|_T + \frac{1}{2} \|[\partial_n u_h]\|_{\partial T \setminus \partial\Omega} \|z - \psi_h\|_{\partial T} \}. \quad (3.6.150)$$

Die letzte Abschätzung schreiben wir in der kompakten Form

$$|J(e_h)| \leq \eta(u_h) := \sum_{T \in \mathbf{T}_h} \{ \rho_T(u_h) \omega_T(z) + \rho_{\partial T}(u_h) \omega_{\partial T}(z) \} \quad (3.6.151)$$

mit den „Zellresiduentermen“

$$\rho_T(u_h) := \|f + \Delta u_h\|_T, \quad \rho_{\partial T}(u_h) := \frac{1}{2} h_T^{-1/2} \|[\partial_n u_h]\|_{\partial T \setminus \partial\Omega},$$

und den „Zellgewichten“

$$\omega_T(z) := \|z - I_h z\|_T, \quad \omega_{\partial T}(z) := h_T^{1/2} \|z - I_h z\|_{\partial T}.$$

Die Fehlerdarstellung (3.6.146) bzw. die Fehlerabschätzung (3.6.151) sind nicht unmittelbar auswertbar. Zwar sind die „Residuenterme“  $R(u_h) = f + \Delta u_h$  und  $r(u_h) = -\frac{1}{2}[\partial_n u_h]$  aus der Näherung  $u_h$  berechenbar, doch die Wichtungsfaktoren  $z - \psi_h$  sind nur implizit über das duale Problem (3.6.144) gegeben und müssen gesondert bestimmt werden. Mit dieser kritischen Frage werden wir uns im Folgenden noch eingehender beschäftigen.

Die Zellgewichte lassen sich bei Wahl  $\psi_h := I_h z$  mit Hilfe der lokalen Interpolationsfehlerabschätzungen aus Abschnitt 3.3 weiter abschätzen durch

$$\omega_T(z) \leq c_I h_T^3 \max_T \{ |\nabla^2 z| \}. \quad (3.6.152)$$

Hierbei muss natürlich vorausgesetzt werden, dass die duale Lösung  $z \in H^{2,\infty}(\Omega)$  ist. Wir werden später sehen, wie man sich von dieser sehr einschränkenden Annahme in gewisser Weise befreien kann. Die Interpolationskonstante  $c_I$  ist von verschiedenen Faktoren abhängig: Von dem Polynomgrad der verwendeten Interpolierenden, vom Referenzelement  $\hat{T}$  sowie von den Transformationen  $\sigma_T : \hat{T} \rightarrow T$ . Sie stellt einen Unsicherheitsfaktor dar. Genaueres Nachrechnen ergibt eine Größenordnung von  $c_I \approx 0,1 - 10$ .

Die a posteriori Fehlerabschätzung (3.6.151) ist „zuverlässig“ im Sinne, dass sie eine sichere obere Schranke für den Fehler  $|J(e_h)|$  liefert, vorausgesetzt, es liegen zuverlässige Werte für die Gewichte  $\omega_T(z)$  und  $\omega_{\partial T}(z)$  vor. Sie wäre auch „effizient“, wenn für den zugehörigen „Effektivitätsindex“ („Überschätzungs faktor“) gilt:

$$I_{\text{eff}} := \frac{\eta(u_h)}{|J(e_h)|} \sim 1 \quad (h \rightarrow 0). \quad (3.6.153)$$

**Bemerkung 3.13:** Es ist zu bemerken, dass bereits der Übergang von der „Fehleridentität“ (3.6.146) zur „Fehlerabschätzung“ (3.6.148) in gewissen Fällen zu einer groben Überschätzung des tatsächlichen Fehlers führen kann. Betrachten wir hierzu folgendes Beispiel: Auf dem Quadrat  $\Omega = (-1, 1) \times (-1, 1)$  gilt es, für die Poisson-Gleichung den Punktfehler im Ursprung zu finden:  $J(e_h) = e_h(0)$ . Es ist möglich, die rechte Seite  $f$  und das Gitter  $\mathbb{T}_h$  so anti-symmetrisch bzgl.  $x = 0$  zu konstruieren, dass  $u(0) = 0 = u_h(0)$ . In diesem Fall wird aber in der Regel  $\eta(u_h) \approx h^2 > 0$  sein, d. h.:  $I_{\text{eff}} = \infty$ .

### 3.6.2 Spezielle a posteriori Fehlerschätzer

Der oben abgeleitete allgemeine Fehlerschätzer wird nun für verschiedene spezielle Fehlerfunktionale ausgewertet. In den betrachteten Fällen kann man die Gewichte  $\omega_T(z)$ ,  $\omega_{\partial T}(z)$  analytisch abschätzen. Wir beschränken uns hier auf die Betrachtung von  $P_1$ -Elementen.

#### a) Energienorm-Fehlerschätzer:

Zur Abschätzung des „Energienormfehlers“  $\|\nabla e_h\|$  wählen wir das Fehlerfunktional

$$J(\varphi) := \|\nabla e_h\|^{-1}(\nabla \varphi, \nabla e_h),$$

so dass wir für  $\varphi = e_h$  automatisch  $J(e_h) = \|\nabla e_h\|$  haben. Für die zugehörige Lösung  $z \in V$  des dualen Problems

$$(\nabla \varphi, \nabla z) = J(\varphi) \quad \forall \varphi \in V \tag{3.6.154}$$

gilt dann

$$\|\nabla z\|^2 = \|\nabla e_h\|^{-1}(\nabla z, \nabla e_h) \leq \|\nabla z\|,$$

woraus sich die einfache a priori Abschätzung  $\|\nabla z\| \leq 1$  ergibt. Ausgehend von der allgemeinen Fehlerabschätzung (3.6.150) erhalten wir

$$\|\nabla e_h\| \leq \sum_{T \in \mathbb{T}_h} \left\{ \|f + \Delta u_h\|_T \|z - \psi_h\|_T + \frac{1}{2} \|[\partial_n u_h]\|_{\partial T \setminus \partial \Omega} \|z - \psi_h\|_{\partial T} \right\}.$$

Wir wählen nun  $\psi_h := I_h z$  als eine „verallgemeinerte“ Knoteninterpolierende von  $z$ , für welche die folgende lokale Fehlerabschätzung gilt (Für deren technisch aufwendige Konstruktion verweisen wir auf die einschlägige Literatur, z. B. Brenner/Scott [13]):

$$\|z - I_h z\|_T + h_T^{1/2} \|z - I_h z\|_{\partial T} \leq \tilde{c}_i h_T \|\nabla z\|_{\tilde{T}}, \tag{3.6.155}$$

wobei  $\tilde{T} := \cup\{T' \in \mathbb{T}_h : T' \cap T \neq \emptyset\}$ . Damit folgt dann mit den Residuen  $\rho_T(u_h)$  und  $\rho_{\partial T}(u_h)$ :

$$\begin{aligned}\|\nabla e_h\| &\leq \tilde{c}_i \sum_{T \in \mathbb{T}_h} h_T \{ \|f + \Delta u_h\|_T + \frac{1}{2} h_T^{-1/2} \|[\partial_n u_h]\|_{\partial T \setminus \partial \Omega} \} \|\nabla z\|_{\tilde{T}} \\ &\leq c \left( \sum_{T \in \mathbb{T}_h} h_T^2 \{ \rho_T(u_h)^2 + \rho_{\partial T}(u_h)^2 \} \right)^{1/2} \left( \sum_{T \in \mathbb{T}_h} \|\nabla z\|_T^2 \right)^{1/2}.\end{aligned}$$

Wegen

$$\left( \sum_{T \in \mathbb{T}_h} \|\nabla z\|_T^2 \right)^{1/2} \leq c \|\nabla z\| \leq c,$$

ergibt sich schließlich das folgende Resultat:

**Satz 3.17 (a posteriori Energienormfehler):** Für den Fehler  $e_h := u - u_h$  gilt die a posteriori Abschätzung bzgl. der Energienorm:

$$\|\nabla e_h\| \leq \eta_E(u_h) := c \left( \sum_{T \in \mathbb{T}_h} h_T^2 \{ \rho_T(u_h)^2 + \rho_{\partial T}(u_h)^2 \} \right)^{1/2}. \quad (3.6.156)$$

Diese Abschätzung ist in folgendem Sinne ‘‘scharf’’:

$$\eta_E(u_h) \leq c \|\nabla e_h\| + c \left( \sum_{T \in \mathbb{T}_h} h_T^2 \|f\|_T^2 \right)^{1/2}. \quad (3.6.157)$$

**Beweis:** Es bleibt, die Abschätzung (3.6.157) zu beweisen. Auf jeder Zelle  $T \in \mathbb{T}_h$  ist  $u_{h|T} \in P_1(T)$  und folglich

$$\rho_T(u_h)^2 = \|f + \Delta u_h\|_T^2 = \|f\|_T^2.$$

Es bezeichne wieder  $\partial \mathbb{T}_h$  die Menge aller (inneren) Kanten  $\Gamma$  der Triangulierung  $\mathbb{T}_h$ . Wegen der Formregularitätsbedingung gilt  $ch_T \leq h_T := |\Gamma| \leq c'h_T$ ,  $\Gamma \subset \partial T$ , gleichmäßig für alle Zellen  $T \in \mathbb{T}_h$  mit Kanten  $\Gamma \in \partial \mathbb{T}_h$  und  $h \in \mathbb{R}_+$ . Folglich ist

$$\sum_{T \in \mathbb{T}_h} h_T^2 \rho_{\partial T}(u_h)^2 \leq c \sum_{\Gamma \in \partial \mathbb{T}_h} h_\Gamma \|[\partial_n u_h]\|_{\Gamma \setminus \partial \Omega}^2.$$

Für eine Kante  $\Gamma \in \partial \mathbb{T}_h$  seien  $T$  und  $T'$  die beiden benachbarten Zellen mit  $\Gamma$  als gemeinsamer Kante. Bezogen auf den Mittelpunkt  $a_\Gamma$  von  $\Gamma$  definieren wir die  $C^\infty$ -Funktion

$$\hat{\omega}_\Gamma(x) := \begin{cases} \exp \left( -\frac{|x-a_\Gamma|^2}{r^2 - |x-a_\Gamma|^2} \right), & x \in B_r(a_\Gamma), \\ 0, & x \in B_r(a_\Gamma)^c, \end{cases} \quad \omega_\Gamma(x) := h_\Gamma \left( \int_\Gamma \hat{\omega}_\Gamma(s) ds \right)^{-1} \hat{\omega}_\Gamma(x),$$

wobei  $r = \gamma h_\Gamma$  so gewählt ist, dass  $B_r(a_\Gamma) := \{x \in \mathbb{R}^2 \mid |x - a_\Gamma| \leq r_\Gamma\} \subset T \cup T'$ . Diese Funktion hat konstruktionsgemäß die Eigenschaften:

$$\begin{aligned}\omega_\Gamma|_{\partial(T \cup T')} &= 0, \quad \partial_n \omega_\Gamma|_{\partial T} = \partial_n \omega_\Gamma|_{\partial T'} = \partial_n \omega_\Gamma|_\Gamma = 0, \\ \int_\Gamma \omega_\Gamma(x) ds &= h_\Gamma = \int_\Gamma ds, \\ \|\omega_\Gamma\|_\infty &\leq c, \quad \|\nabla \omega_\Gamma\|_\infty \leq ch_\Gamma^{-1}, \quad \|\nabla^2 \omega_\Gamma\|_\infty \leq ch_\Gamma^{-2},\end{aligned}$$

wobei die Konstanten unabhängig von  $T$  und  $h$  gewählt werden können. Entlang der Kante  $\Gamma$  ist der Sprungterm  $[\partial_n u_h]$  konstant und es gilt daher wegen der Stetigkeit von  $\partial_n u$  über die Kante  $\Gamma$  mit der gerade definierten Funktion  $\omega_\Gamma$

$$\begin{aligned}\|[\partial_n u_h]\|_\Gamma^2 &= [\partial_n u_h]_\Gamma^2 \int_\Gamma ds = [\partial_n u_h]_\Gamma^2 \int_\Gamma \omega_\Gamma ds = [\partial_n u_h]_\Gamma \int_\Gamma [\partial_n(u_h - u)] \omega_\Gamma ds \\ &= [\partial_n u_h]_\Gamma \left\{ \int_{\partial T} \partial_n(u_h - u) \omega_\Gamma ds + \int_{\partial T'} \partial_n(u_h - u) \omega_\Gamma ds \right\}.\end{aligned}$$

Partielle Integration ergibt dann

$$\begin{aligned}\|[\partial_n u_h]\|_\Gamma^2 &= [\partial_n u_h]_\Gamma \int_{T \cup T'} \{ \Delta(u_h - u) \omega_\Gamma(x) - \nabla(u_h - u) \nabla \omega_\Gamma \} dx \\ &= [\partial_n u_h]_\Gamma \int_{T \cup T'} \{ -f \omega_\Gamma(x) - \nabla(u_h - u) \nabla \omega_\Gamma \} dx.\end{aligned}$$

Unter Beachtung der obigen Eigenschaften der Funktion  $\omega_\Gamma$  folgt

$$\begin{aligned}\|[\partial_n u_h]\|_\Gamma^2 &\leq |[\partial_n u_h]_\Gamma| \{ \|f\|_{T \cup T'} \|\omega_\Gamma\|_{T \cup T'} + \|\nabla e_h\|_{T \cup T'} \|\nabla \omega_\Gamma\|_{T \cup T'} \} \\ &\leq c |[\partial_n u_h]_\Gamma| \{ h_T \|f\|_{T \cup T'} + \|\nabla e_h\|_{T \cup T'} \}.\end{aligned}$$

Dies impliziert dann die folgende Abschätzung:

$$\begin{aligned}\sum_{\Gamma \in \partial T_h} h_\Gamma \|[\partial_n u_h]\|_{\Gamma \setminus \partial \Omega}^2 &\leq c \sum_{\Gamma \in \partial T_h} h_\Gamma |[\partial_n u_h]_\Gamma| \{ h_T \|f\|_{T \cup T'} + \|\nabla e_h\|_{T \cup T'} \} \\ &\leq c \left( \sum_{\Gamma \in \partial T_h} h_\Gamma^2 |[\partial_n u_h]_\Gamma|^2 \right)^{1/2} \left( \sum_{\Gamma \in \partial T_h} \{ h_T^2 \|f\|_{T \cup T'}^2 + \|\nabla e_h\|_{T \cup T'}^2 \} \right)^{1/2} \\ &\leq c \left( \sum_{\Gamma \in \partial T_h} h_\Gamma |[\partial_n u_h]_\Gamma|^2 \right)^{1/2} \left( \sum_{T \in T_h} \{ h_T^2 \|f\|_T^2 + \|\nabla e_h\|_T^2 \} \right)^{1/2}\end{aligned}$$

bzw.

$$\sum_{\Gamma \in \partial T_h} h_\Gamma \|[\partial_n u_h]\|_{\Gamma \setminus \partial \Omega}^2 \leq c \sum_{T \in T_h} h_T^2 \|f\|_T^2 + c \|\nabla e_h\|^2.$$

Dies ergibt die behauptete Abschätzung (3.6.157). Q.E.D.

**Bemerkung 3.14:** Mit einer etwas aufwendigeren Argumentation kann man die folgende verschärzte a posteriori Abschätzung herleiten:

$$\eta_E(u_h) \leq c \left( \sum_{T \in \mathbf{T}_h} h_T^2 \left\{ h_T^2 \|\nabla f\|_T^2 + \rho_{\partial T}(u_h)^2 \right\} \right)^{1/2}.$$

Diese besagt, dass im Fehlerschätzer  $\eta_E(u_h)$  in der Regel der Einfluss der Gleichungsresiduen  $\|f + \Delta u_h\|_T$  gegenüber den Regularitätsresiduen  $\|[\partial_n u_h]\|_{\partial T \setminus \partial \Omega}$  vernachlässigt werden können. Dies ist in dieser Form allerdings nur richtig für *lineare* bzw. *bilineare* finite Elemente.

### b) $L^2$ -Norm-Fehlerschätzer:

Zur Abschätzung des  $L^2$ -Fehlers  $\|e_h\|$  wählen wir das Fehlerfunktional

$$J(\varphi) := \|e_h\|^{-1}(\varphi, e_h),$$

womit automatisch  $J(e) = \|e_h\|$  gilt. Die zugehörige Lösung  $z \in V$  des dualen Problems

$$(\nabla \varphi, \nabla z) = J(\varphi) \quad \forall \varphi \in V \tag{3.6.158}$$

ist dann auf dem konvexen Gebiet  $\Omega$  auch in  $H^2(\Omega)$ , und es gilt die a priori Abschätzung  $\|\nabla^2 z\| \leq \|\Delta z\| = 1$ . Ausgehend von der Fehlerabschätzung (3.6.150) erhalten wir wieder

$$\|e_h\| \leq \sum_{T \in \mathbf{T}_h} \left\{ \|f + \Delta u_h\|_T \|z - \psi_h\|_T + \frac{1}{2} \|[\partial_n u_h]\|_{\partial T \setminus \partial \Omega} \|z - \psi_h\|_{\partial T} \right\}.$$

Wir können nun  $\psi_h := I_h z$  als die normale Knoteninterpolierende von  $z$  wählen:

$$\|z - I_h z\|_T + h_T^{1/2} \|z - I_h z\|_{\partial T} \leq c_I h_T^2 \|\nabla^2 z\|_T. \tag{3.6.159}$$

Damit folgt dann wieder mit den oben definierten Residuen  $\rho_T(u_h)$  und  $\rho_{\partial T}(u_h)$ :

$$\begin{aligned} \|e_h\| &\leq c_I \sum_{T \in \mathbf{T}_h} h_T^2 \left\{ \|f + \Delta u_h\|_T + \frac{1}{2} h_T^{-1/2} \|[\partial_n u_h]\|_{\partial T \setminus \partial \Omega} \right\} \|\nabla^2 z\|_T \\ &\leq c \left( \sum_{T \in \mathbf{T}_h} h_T^4 \{\rho_T(u_h)^2 + \rho_{\partial T}(u_h)^2\} \right)^{1/2} \left( \sum_{T \in \mathbf{T}_h} \|\nabla^2 z\|_T^2 \right)^{1/2}. \end{aligned}$$

Wegen

$$\left( \sum_{T \in \mathbf{T}_h} \|\nabla^2 z\|_T^2 \right)^{1/2} = \|\nabla^2 z\| \leq 1,$$

ergibt sich schließlich das folgende Resultat:

**Satz 3.18 (a posteriori  $L^2$ -Normfehler):** Für den Fehler  $e_h := u - u_h$  gilt die a posteriori Abschätzung bzgl. der  $L^2$ -Norm:

$$\|e_h\| \leq \eta_{L^2}(u_h) := c \left( \sum_{T \in \mathbb{T}_h} h_T^4 \{ \rho_T(u_h)^2 + \rho_{\partial T}(u_h)^2 \} \right)^{1/2}. \quad (3.6.160)$$

Diese Abschätzung ist ebenfalls „scharf“ in folgendem Sinne:

$$\eta_{L^2}(u_h) \leq c \|e_h\| + c \left( \sum_{T \in \mathbb{T}_h} h_T^4 \|f\|_T^2 \right)^{1/2}. \quad (3.6.161)$$

**Beweis:** Der Beweis verläuft analog wie bei der Energienorm-Fehlerabschätzung. Ein kleiner Zusatzschritt ist

$$\int_{T \cup T'} \{ \Delta(u_h - u) \omega_\Gamma(x) - \nabla(u_h - u) \nabla \omega_\Gamma \} dx = \int_{T \cup T'} \{ \Delta(u_h - u) \omega_\Gamma(x) + (u_h - u) \Delta \omega_\Gamma \} dx,$$

wobei der Randterm wegen  $\partial_n \omega_\Gamma|_{\partial T} = \partial_n \omega_\Gamma|_{\partial T'} = 0$  verschwindet. Die weiteren Details seien als Übungsaufgabe gestellt. Q.E.D.

### c) Punktfehler-Schätzer:

Zur Abschätzung des Fehlers  $e_h(P)$  in einem festen Punkt  $P \in \Omega$  würden wir gern das Fehlerfunktional  $J(\varphi) := \varphi(P)$  wählen. Dieses ist zwar auf stetigen Funktionen, aber nicht auf dem ganzen Lösungsraum  $V = H_0^1(\Omega)$  definiert, so dass die oben entwickelte allgemeine Theorie nicht unmittelbar anwendbar wäre. Deshalb wählen wir eine Kugelumgebung  $B_\varepsilon := \{x \in \mathbb{R}^d : |x - P| < \varepsilon\}$  des Punktes  $P$  und arbeiten statt dessen mit dem regularisierten Funktional

$$J_\varepsilon(\varphi) := |B_\varepsilon|^{-1} \int_{B_\varepsilon} \varphi dx,$$

welches sicher auf ganz  $V$  definiert und beschränkt ist. Der Regularisierungsparameter  $\varepsilon$  wird üblicherweise gleich einer vorgegebenen Fehlertoleranz  $\text{TOL}$  gesetzt. Für hinreichend glatte Funktionen  $\varphi$  ist dabei (Fehler der Mittelpunktregel)

$$|\varphi(P) - J_\varepsilon(\varphi)| \leq c \varepsilon^2. \quad (3.6.162)$$

Wir betrachten den Fall  $d = 2$ . Die Lösung  $z_\varepsilon \in V$  des zugehörigen dualen Problems

$$(\nabla \varphi, \nabla z_\varepsilon) = J_\varepsilon(\varphi) \quad \forall \varphi \in V, \quad (3.6.163)$$

ist dann eine „regularisierte“ Greensche Funktion und verhält sich wie

$$z_\varepsilon(x) \approx |\log(|x - P| + \varepsilon)|, \quad |\nabla^2 z_\varepsilon| \approx ||x - P| + \varepsilon|^{-2}.$$

Die Gewichte in der a posteriori Fehlerabschätzung (3.6.151) gestatten daher die Abschätzung

$$\omega_T(z_\varepsilon) + \omega_{\partial T}(z_\varepsilon) \leq ch_T^2 \|\nabla^2 z_\varepsilon\|_T \leq c h_T^3 d_T^{-2},$$

mit  $d_T := \max_{x \in T} \{|x - a| + \varepsilon\}$ . Wir finden also als obere Schranke für den Punktfehler:

$$|e_h(P)| \leq c \sum_{T \in \mathcal{T}_h} \rho_T(u_h) h_T^2 \|\nabla^2 z_\varepsilon\|_T + c\varepsilon^2 \leq c \sum_{T \in \mathcal{T}_h} \rho_T(u_h) \frac{h_T^3}{d_T^2} + c\varepsilon^2.$$

In diesem Fall ist aber die a-priori Bestimmung der Konstante  $c$  praktisch unmöglich.

#### d) Ein hyper-singulärer Fall

Einen kuriosen Sonderfall stellt die folgende Auswertungsgröße dar:

$$J(u) := \int_{\partial\Omega} \partial_n u \, ds \quad \left( = \int_{\Omega} \Delta u \, dx = - \int_{\Omega} f \, dx \right).$$

Dieser „mittlere Normalfluss“ ließe sich offenbar auch direkt aus den Daten  $f$  berechnen. Zur Illustration wollen wir aber annehmen, dass diese Information nicht ausgenutzt wird, sondern statt dessen  $J(u)$  durch  $J(u_h)$  approximiert wird. Das Funktional  $J(\cdot)$  ist wieder nicht auf dem ganzen Lösungsraum  $V$  definiert (wohl aber auf der reguläreren Lösung  $u \in V \cap H^2(\Omega)$ ). Zur Anwendung unserer allgemeinen Theorie muss das Funktional daher zunächst regularisiert werden. Für das Folgende nehmen wir der Einfachheit halber an, dass  $\Omega$  der Einheitskreis im  $\mathbb{R}^2$  ist. Für  $\varepsilon = \text{TOL}$  setzen wir  $S_\varepsilon := \{x \in \Omega, \text{dist}\{x, \partial\Omega\} < \varepsilon\}$  und erhalten für glattes  $\varphi$ :

$$J_\varepsilon(\varphi) := \varepsilon^{-1} \int_{S_\varepsilon} \partial_r \varphi \, dx = \int_{\partial\Omega} \partial_r \varphi \, ds + \mathcal{O}(\varepsilon).$$

Dabei ist  $\partial_n$  auf  $S_\varepsilon$  auf natürliche Weise durch Fortsetzung von  $\partial\Omega$  definiert. Die Lösung  $z_\varepsilon \in V$  des zugehörigen dualen Problems

$$(\nabla \varphi, \nabla z_\varepsilon) = J_\varepsilon(\varphi) \quad \forall \varphi \in V,$$

ist dann gegeben durch

$$z_\varepsilon = -1 \quad \text{in } \Omega \setminus S_\varepsilon, \quad z_\varepsilon(x) = -\varepsilon^{-1}(1 - |x|) \quad \text{auf } S_\varepsilon.$$

Hieraus ergibt sich die Fehlerabschätzung

$$|J_\varepsilon(e_h)| \leq c_I \sum_{T \in \mathcal{T}_h} h_T^2 \rho_T(u_h) \|\nabla^2 z_\varepsilon\|_T \approx \sum_{T \cap S_\varepsilon \neq \emptyset} \dots,$$

d. h.: Die Zellen im Innern von  $\Omega$  tragen nicht zum Gesamtfehler bei. Daher wäre die beste Strategie zur Gitterverfeinerung, in jedem Verfeinerungszyklus jeweils nur die Zellen entlang des Randes  $\partial\Omega$  zu verfeinern. Dies setzt aber voraus, dass die Elemente des Lastvektors,  $b_n = (f, \varphi_h^{(n)})_\Omega$ , exakt berechnet werden. Abbildung 3.23 zeigt ein automa-

tisch verfeinertes Gitter nach 7 Verfeinerungszyklen sowie die numerisch bestimmte duale Lösung auf diesem Gitter. Die zugehörigen Ergebnisse sind in Tabelle 3.1 gelistet.

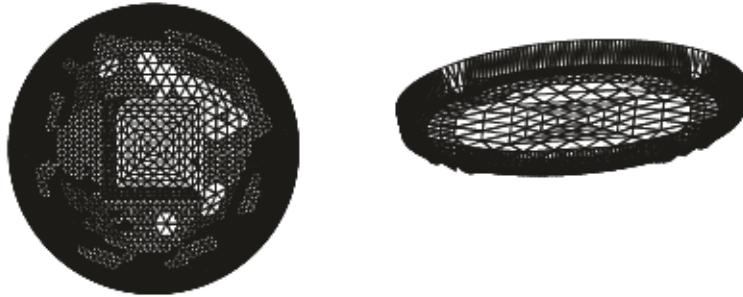


Abbildung 3.23: Verfeinertes Gitter nach 7 Verfeinerungszyklen (links) und die auf diesem Gitter numerisch bestimmte Lösung (rechts).

Tabelle 3.1: Resultate mit dem Energienorm-Fehlerschätzer  $\eta_E(u_h)$  und der „optimalen“ Strategie.

$\eta_E$			„optimale“ Strategie		
$L$	N	$J(e_h)$	$L$	N	$J(e_h)$
1	1024	$2.91 + 0$	1	559	$2.90 + 0$
2	4096	$1.50 + 0$	2	1294	$1.50 + 0$
3	16384	$7.49 - 1$	3	3079	$7.48 - 1$
4	64768	$3.77 - 1$	4	7174	$3.62 - 1$
5	253858	$1.73 - 1$	5	16738	$1.67 - 1$
6	memory	exhausted	6	38146	$6.93 - 2$

### Numerische Auswertung des a-posteriori Fehlerschätzers

Wir wollen nun kurz die numerische Auswertung der Fehleridentität (3.6.146) diskutieren:

$$J(e_h) = \eta(u_h) := \sum_{T \in \mathcal{T}_h} \{(f + \Delta u_h, z - \psi_h)_T - \frac{1}{2}([\partial_n u_h], z - \psi_h)_{\partial T}\}. \quad (3.6.164)$$

Wir wählen dazu  $\psi_h := I_h z$ , die Knoteninterpolierende von  $z$ . Die Qualität der Auswertung wird durch den sog. „Effektivitätsindex“ gemessen:

$$I_{\text{eff}} := \left| \frac{\eta(u_h)}{J(e_h)} \right|.$$

Alle Methoden der Auswertung von (3.6.164) bedienen sich einer numerischen Lösung  $z_h \in V_h$  des dualen Problems, welche im einfachsten Fall durch direkte Diskretisierung von (3.6.164) mit Hilfe des vorliegenden Verfahrens gewonnen wird:

$$(\nabla \varphi_h, \nabla z_h) = J(\varphi_h) \quad \forall \varphi_h \in V_h. \quad (3.6.165)$$

Wir unterscheiden zwei Anwendungen der Fehleridentität (3.6.164):

- Überprüfung der Genauigkeit einer auf dem Gitter  $T_h$  berechneten Näherungslösung  $u_h$  und Abbruchkriterium für einen Gitterverfeinerungsprozess:  $\eta(u_h) \leq TOL$ ?
- Grundlage zur Gewinnung von Kriterien („Verfeinerungsindikatoren“  $\eta_T$ ) zur lokalen Gitteranpassung.

Die folgenden Strategien zur Auswertung von  $\eta(u_h)$  kommen in Betracht (hier beschrieben für den Fall  $d$ -linearer Ansätze auf Rechtekgittern im  $\mathbb{R}^d$ ):

1. Die duale Lösung  $z$  wird durch eine Näherung höherer Ordnung approximiert. Zum Beispiel liefert die Approximation  $z \approx z_h^{(2)}$  mit der  $d$ -quadratischen Ritz-Projektion das gewünschte asymptotische Verhalten  $\lim_{TOL \rightarrow 0} I_{\text{eff}} = 1$ .
2. Die duale Lösung  $z$  wird durch eine zellblockweise (je  $2^d$  Zellen)  $d$ -quadratische Interpolierende der  $d$ -linearen Ritzapproximation approximiert (s. Abb. 3.24):  $z \approx I_h^{(2)} z_h$ . In diesem Fall beobachtet man  $\lim_{TOL \rightarrow 0} I_{\text{eff}} < 1$  ( $\approx 0.5 - 0.9$ ).

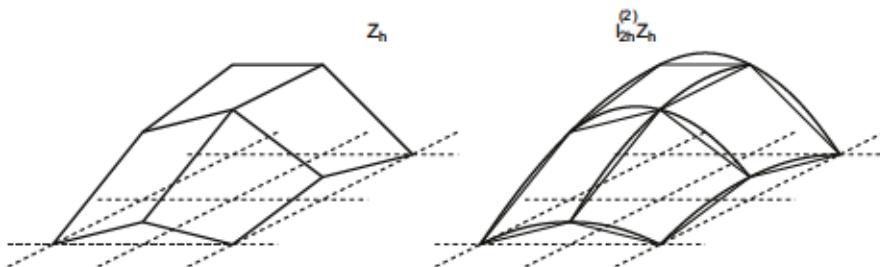


Abbildung 3.24: Blockweise bi-quadratische Interpolation (rechts) bilinearer Knotenwerte (links).

3. Die Differenz  $z - I_h z$  wird in Anlehnung an die oben zitierte Interpolationsfehlerabschätzung mit Hilfe eines skalierten Differenzenquotienten der  $d$ -linearen Ritz-Projektion approximiert:  $|z - I_h z| \approx c_{i,T} h_T^2 |\nabla_h^2 z_h F|$ . Dies führt in der Regel auf eine grobe Überschätzung des Fehlers.

In der Praxis wird meist die „billige“ und ausreichend genaue Prozedur (2) verwendet. Dabei brauchen insbesondere keine Interpolationskonstanten  $c_I$  spezifiziert zu werden.

### 3.6.3 Strategien zur Gittersteuerung

Ziel ist es, den Fehler in dem durch das Fehlerfunktional  $J(\cdot)$  beschriebenen Maß unter eine gewisse vorgegebene Toleranz  $|J(e_h)| \leq \text{TOL}$  zu bringen und dabei mit dem vorhandenen Speicher auszukommen. Es können also nur eine bestimmte Anzahl von Zellen  $N \leq N_{\max}$  verwaltet werden. Die Gitterverfeinerung erfolgt dabei standardmäßig durch „Kantenbisektion“, d. h. durch Unterteilung einer Zelle  $T$  in  $2^d$  Teilzellen. Bei Gittervergrößerung werden mehrere Zellen zu einer Makrozelle zusammengefaßt. Wir werden im folgenden hauptsächlich Vierecks- oder Hexaedergitter betrachten. Bei Unterteilung einer Zelle entstehen sog. „hängende“ Knoten auf den Zellrändern, so dass das neue Gitter zunächst nicht zulässig wäre. Diese Unzulässigkeit kann durch verschiedenen Methoden aufgehoben werden. Die meist gebräuchliche ist die Elimination der zu den hängenden Knoten gehörenden Freiheitsgraden durch lineare Interpolation, wodurch der entstehende diskrete Ansatzraum wieder konform wird:  $V_h \subset V$  (s. Abb. 3.25). Dabei wird der fiktive Knotenwert  $v_h(P)$  in einem hängenden Knoten  $P$  auf einer Kante  $\Gamma \subset \partial T$  mit Endpunkten  $P', P''$  durch die Interpolationsvorschrift  $v_h(P) := \frac{1}{2}\{v_h(P') + v_h(P'')\}$  festgelegt. Alle im folgenden gezeigten Beispielrechnungen bedienen sich dieser Technik.

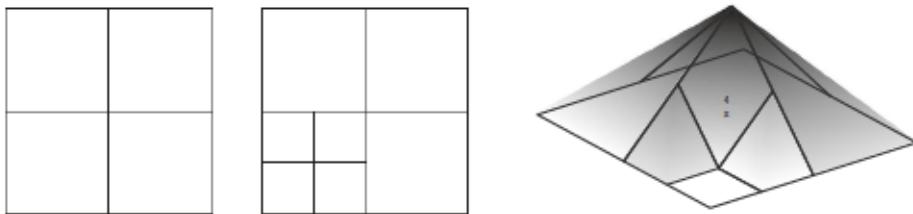


Abbildung 3.25: Verfeinertes Rechteckkitter mit „hängendem“ Knoten sowie die Basisfunktion zum Mittelknoten.

Ausgangspunkt zur Gewinnung von Kriterien für die lokale Gitteranpassung ist eine a-posteriori Fehlerschätzung der Form

$$|J(e_h)| \approx \sum_{T \in \mathbb{T}_h} \eta_T , \quad (3.6.166)$$

mit lokalen „Fehlerindikatoren“  $\eta_T$ , welche durch Auswertung der Fehleridentität (3.6.146) mit Hilfe einer der oben beschriebenen Methoden gewonnen werden. Wir erhalten bei Verwendung von Methode (2):

$$\eta_T := |(f + \Delta u_h, I_h^{(2)} z_h - z_h)_T - \frac{1}{2}([\partial_n u_h], I_h^{(2)} z_h - z_h)_{\partial T \setminus \partial \Omega}| \quad (3.6.167)$$

und mit Methode (3):

$$\eta_T := c_I h_T^2 \{\rho_T(u_h) + \rho_{\partial T}(u_h)\} \|\nabla_h^2 z_h\|_T . \quad (3.6.168)$$

Wir beschreiben im folgenden das allgemeine Schema eines auf lokalen Fehlerindikatoren  $\eta_T$  basierenden Gitteranpassungsprozesses:

Nachdem über das Gebiet  $\Omega$  ein Grobgitter  $T_0 := T_{h_0}$  mit einer Gitterweitenverteilung  $h_0$  gelegt worden ist, sei der Prozess der adaptiven Gittersteuerung schon  $L$ -mal durchlaufen worden. Wir bestimmen nun auf dem Gitter  $T_L := T_{h_L}$  die approximative Lösung  $u_L \in V_L := V_{h_L}$ . Ebenso wird die diskrete duale Lösung  $z_L \in V_L$  auf  $T_L$  berechnet. Nun können die lokalen Fehlerindikatoren  $\eta_T$  gemäß (3.6.167) oder (3.6.168) für jede Zelle  $T \in T_L$  ausgewertet werden. Wir bezeichnen mit  $N_L \approx \dim(V_L)$  die Anzahl der Zellen des Gitters  $T_L$ .

**Abbruchabfrage:** Ist das Abbruchkriterium  $\eta(u_h) \leq \frac{1}{2}\text{TOL}$  auf dem Gitter  $T_L$  erfüllt, wird der Adoptionsprozess abgebrochen und  $u_L$  als Näherung zu  $u$  akzeptiert, welche die Zielgröße  $J(u)$  durch  $J(u_h)$  mit der gewünschten Genauigkeit  $\text{TOL}$  approximiert. Andernfalls wird ein neuer Adoptionszyklus begonnen.

**Adoptionszyklus:** Der Übergang vom Gitter  $T_L$  zum nächsten Gitter  $T_{L+1}$  erfolgt nach einer der im folgenden beschriebenen „Adoptionsstrategien“. Zunächst werden die Zellen  $T \in T_L$  nach der Größe ihrer Indikatorwerte  $\eta_T$  angeordnet,

$$\{T_i, i = 1, \dots, N_L\} : \quad \eta_{T,1} \geq \dots \geq \eta_{T,i} \geq \eta_{T,i+1} \geq \dots \geq \eta_{T,N_L}.$$

I) *Fehlerbalancierungs-Strategie:* Das Ziel ist, die Indikatorwerte so zu balancieren, dass

$$\eta_T \approx \frac{\text{TOL}}{N_L}, \quad T \in T_L. \quad (3.6.169)$$

Dann würde wie gewünscht gelten:

$$|J(e_L)| \approx \eta(u_L) \approx \sum_{T \in T_L} \frac{\text{TOL}}{N_L} = \text{TOL}.$$

Das Problem bei dieser Strategie ist zum einen, dass die Zahl  $N_L$  sich während des Verfeinerungsprozesses ständig ändert, und zum anderen, dass die Balancierungsvorschrift (3.6.169) die delikate Wahl von Parametern  $0 < \alpha < \beta < 1$  erfordert:

$$\alpha \frac{\text{TOL}}{N_L} \leq \eta_T \leq \beta \frac{\text{TOL}}{N_L}.$$

Für die folgende Diskussion setzen wir der Einfachheit halber  $\alpha = 1/2$ ,  $\beta = 1$ . Im ersten Schritt testet man, ob  $\eta_{T,1} \leq \text{TOL}/N_L$  ist. Wenn „ja“, ist das Abbruchkriterium erfüllt, wenn „nein“, wird die Zelle  $T_1$  verfeinert. Dies führt zu einer Erhöhung von  $N_L$  auf  $N_L + 3$ . Danach wäre es möglich, dass für alle übrigen Zellen  $T_{L,i}$  gilt:

$$\frac{\text{TOL}}{N_L + 3} \leq \eta_{T,i} \leq \frac{\text{TOL}}{N_L}.$$

Der Verfeinerungszyklus wäre also bereits nach dem ersten Schritt zu beenden und auf

dem erhaltenen Gitter  $T_{L+1}$  eine neue Näherungslösung  $u_{L+1}$  zu berechnen. Dies wäre aber sehr ineffizient. Der Verfeinerungsprozess sollte daher beschleunigt werden. Dazu überprüft man von dem Indikator  $\eta_{T,i}$  ausgehend und beginnend mit  $j = 0$ , ob

$$\eta_{T,i} \leq \frac{\text{TOL}}{N_L + 3j}.$$

Ist dies nicht erfüllt, wird das Element  $T_i$  geteilt, die Zähler  $j$  und  $i$  um Eins erhöht, und man geht zum nächstkleineren  $\eta_{T,i}$  über. Ist die Bedingung allerdings erfüllt, hat man das neue Gitter  $T_{L+1}$  gefunden. Diese Balancierungsstrategie ist zwar potentiell optimal, jedoch mit aufwendigen Abfragen verbunden. Mögliche alternative Strategien, die einfacher sind, aber weniger gute Ergebnisse liefern, sind die folgenden.

II) „Fest-Raten-Strategie“: Ziel ist es, in jedem Adoptionszyklus die Anzahl der Gitterzellen  $N_L$  mit einer festen Rate zu erhöhen oder den Fehlerschätzwert  $\eta(u_L)$  mit einer festen Rate zu verkleinern. Ausgangspunkt ist wieder eine Anordnung der Zellen von  $T_L$  nach der Größe der zugehörigen Indikatorwerte. Zu vorgewählten Prozentsätzen  $X\%$  und  $Y\%$  werden die Zellen so gruppiert, dass (and der Zellanzahl orientierte Strategie)

$$\#\{T_1, \dots, T_{N^*}\} \approx \frac{X}{100} N_L, \quad \#\{T_{N_L-N^*+1}, \dots, T_{N_L}\} \approx \frac{Y}{100} N_L,$$

oder alternativ (am Schätzwert orientierte Strategie)

$$\sum_{i=1}^{N^*} \eta_{T,i} \approx Y \eta(u_L), \quad \sum_{i=N_L-N^*+1}^{N_L} \eta_{T,i} \approx X \eta(u_L).$$

Dann werden die Zellen  $T_1, \dots, T_{N^*}$  verfeinert und die Zellen  $T_{N_L-N^*+1}, \dots, T_{N_L}$  vergröbert.

### „Exakte“ Gitteroptimierung

Zum Abschluss wollen wir noch diskutieren, dass die Fehlerschätzung (3.6.151)

$$|J(e_h)| \approx \eta := \sum_{T \in \mathbf{T}_h} \rho_T(u_h) \omega_T(z_h) \tag{3.6.170}$$

im Prinzip auch zur direkten Bestimmung eines „optimalen“ Gitters mit Gitterweitenfunktion  $h = h(x)$  verwendet werden könnte. Dies wird als Nebenprodukt auch eine Rechtfertigung für die „Fehlerbalancierungsstrategie“ liefern. Zu lösen sind die folgenden Optimierungsaufgaben:

(OP I) Bei vorgegebener Fehlertoleranz TOL soll die zu deren Erreichung benötigte Anzahl von Zellen  $N$  (d. h. der numerische Aufwand) minimiert werden:

$$N \rightarrow \text{MIN}, \quad \eta \leq \text{TOL}. \tag{3.6.171}$$

(OP II) Bei vorgegebener maximalen Zellzahl  $N_{\max}$  (d. h. begrenzter Speicherkapazität) soll der Fehler (genauer der Fehlerschätzer) minimiert werden:

$$\eta \rightarrow \text{MIN}, \quad N \leq N_{\max}. \quad (3.6.172)$$

Wir diskutieren im Folgenden nur die für die Praxis relevantere Fragestellung (OP II). Das Optimierungsproblem (OP I) lässt sich mit analogen Argumenten behandeln (Übungsaufgabe). Die grundlegende Annahme ist, dass die Größen  $\rho_T(u_h)$  sowie  $\omega_T(z_h)$  (nach geeigneter Skalierung) für  $\text{TOL} \rightarrow 0$  zunehmend bessere, lokale Approximationen gewisser kontinuierlicher Funktionen  $\Phi(x)$  bzw.  $\Psi(x)$  sind:

$$h_T^{-1} \rho_T(u_h) = h_T^{-1} \|f + \Delta u_h\|_T + \frac{1}{2} h_T^{-1/2} \|h_T^{-1} [\partial_n u_h]\|_{\partial T \setminus \partial \Omega} \rightarrow \Phi_u(x_T), \quad (3.6.173)$$

$$h^{-3} \omega_T(z_h) = \max \left\{ h_T^{-3} \|z - I_h z\|_T, h_T^{-5/2} \|z - I_h z\|_{\partial T} \right\} \rightarrow \Psi_z(x_T), \quad (3.6.174)$$

wobei  $x_T$  ein gemeinsamer Punkt einer Folge von Zellen  $T$  ist. Diese Annahme ist natürlich im strengen Sinne nicht realistisch, da auf allgemeinen Gitterfolgen  $(\mathbb{T}_h)_{h \rightarrow 0}$  ein sehr unregelmäßiges Konvergenzverhalten auftreten kann. Unter der obigen Annahme gilt mit der Funktion  $A(x) := \Phi(x)\Psi(x)$ :

$$\sum_{T \in \mathbb{T}_h} \rho_T(u_h) \omega_T(z_h) = \sum_{T \in \mathbb{T}_h} h_T^4 \{h_T^{-4} \rho_T(u_h) \omega_T(z_h)\} \approx \int_{\Omega} h(x)^2 A(x) dx. \quad (3.6.175)$$

Für die Zellzahl  $N$  haben wir die Darstellung

$$N = \sum_{T \in \mathbb{T}_h} 1 = \sum_{T \in \mathbb{T}_h} h_T^2 h_T^{-2} \approx \int_{\Omega} h(x)^{-2} dx. \quad (3.6.176)$$

Damit können wir die Optimierungsaufgabe (OP II) näherungsweiser in kontinuierlicher Form schreiben als:

$$F(h) := \int_{\Omega} h(x)^2 A(x) dx \rightarrow \text{MIN}, \quad N(h) := \int_{\Omega} h(x)^{-2} dx = N_{\max}. \quad (3.6.177)$$

Hierbei haben wir o.B.d.A. die Ungleichungsbedingung  $N \leq N_{\max}$  durch eine Gleichungsbedingung  $N = N_{\max}$  ersetzt, da sich ein minimaler Fehler sicherlich unter maximaler Ausnutzung der möglichen Zellzahl ergibt. Dieses restriktierte Optimierungsproblem wird nun mit dem Lagrange-Formalismus der Variationsrechnung gelöst. Dazu definieren wir die „Lagrange-Funktion“

$$L(h, \lambda) := F(h) + \lambda \{N(h) - N_{\max}\}$$

mit einem skalaren Lagrange-Parameter  $\lambda \in \mathbb{R}$ . Jede Lösung des Optimierungsproblems ist dann notwendig stationärer Punkt (Sattelpunkt) von  $L$ . Zur Bestimmung eines solchen Sattelpunkts machen wir den folgenden Ansatz

$$\frac{d}{dt} L(h + t\varphi, \lambda + t\mu)|_{t=0} = 0 \quad \forall \varphi \in C(\bar{\Omega}), \forall \mu \in \mathbb{R}.$$

Auswertung dieser Beziehung ergibt

$$2 \int_{\Omega} h(x) A(x) \varphi(x) dx - 2\lambda \int_{\Omega} h^{-3}(x) \varphi(x) dx = 0 \quad \forall \varphi \in C(\bar{\Omega})$$

sowie

$$\mu \left\{ \int_{\Omega} h(x)^{-2} dx - N_{\max} \right\} = 0 \quad \forall \mu \in \mathbb{R},$$

bzw. notwendig

$$h(x) A(x) - \lambda h^{-3}(x) = 0, \quad \int_{\Omega} h(x)^{-2} dx = N_{\max}.$$

Hieraus ergibt sich, dass

$$h(x) = \left( \frac{A}{\lambda} \right)^{-1/4},$$

und weiter

$$W := \int_{\Omega} A(x)^{1/2} dx = \lambda^{1/2} N_{\max}.$$

Damit erhalten wir einerseits den Lagrange-Parameter,

$$\lambda = \left( \frac{W}{N_{\max}} \right)^2, \quad (3.6.178)$$

und andererseits eine Gleichung für die gesuchte „optimale“ Gitterweitenverteilung

$$h_{\text{opt}}(x) = \left( \frac{W}{N_{\max}} \right)^{1/2} A(x)^{-1/4}. \quad (3.6.179)$$

Als Nebenprodukt dieser Rechnung ergibt sich für die optimale Gitterweite die Beziehung

$$\eta_T = h_T^4 A_T = \lambda \equiv \text{konst.}, \quad (3.6.180)$$

d. h.: Die an der Äquilibrierung der lokalen Fehlerindikatoren  $\eta_T$  orientierten Gittersteuerungsstrategien führen tatsächlich auf optimale Gitterweitenverteilungen.

Wir erwähnen noch, dass das zu (OP II) „duale“ Optimierungsproblem (OP I) auf die folgende Lösung führt:

$$h_{\text{opt}}(x) = \left( \frac{TOL}{W} \right)^{1/2} A(x)^{-1/4}. \quad (3.6.181)$$

Die Größe  $W$  ist bei den üblichen Fehlerfunktionalen wohl definiert; dies beinhaltet sogar so singuläre Fälle wie die Punktauswertung von Ableitungen  $J(u) = \partial_i u(a)$ ,  $A(x) \sim |x - a|^{-3}$ . Erst die Auswertung von zweiten Ableitungen  $J(u) = \partial_i^2 u(a)$  macht hier Probleme.

### 3.6.4 Ein Testbeispiel

Wir wollen die bisher erzielten Ergebnisse anhand eines konkreten Beispiels illustrieren. Dazu wird wieder das übliche Modellproblem betrachtet:

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega. \quad (3.6.182)$$

Wir wollen den Punktwert  $\partial_1 u(P)$  für ein  $P \in \Omega$  berechnen. Das zugehörige Funktional ist wieder nicht auf dem ganzen Lösungsraum  $V = H_0^1(\Omega)$  definiert und muss regularisiert werden. Wir setzen dazu mit  $\varepsilon := \text{TOL}$ :

$$J_\varepsilon(\varphi) := |B_\varepsilon|^{-1} \int_{B_\varepsilon} \partial_1 u \, dx. \quad (3.6.183)$$

**Testfall 1:** Sei  $\Omega := (-1, 1)^2$  und  $P = 0$  (siehe Abb. 3.26). Die zu dem Funktional  $J_\varepsilon(\cdot)$  gehörige duale Lösung  $z_\varepsilon$  verhält sich dann wie  $|\nabla^2 z_\varepsilon(x)| \approx (|x| + \varepsilon)^{-3}$ . Dies impliziert

$$|\partial_1 e_h(0)| \approx c_I \sum_{T \in \mathbb{T}_h} \frac{h_T^4}{d_T^3} \rho_T(u_h), \quad (3.6.184)$$

mit  $d_T := |x_T| + \varepsilon$  und dem Mittelpunkt  $x_T$  von  $T$ . Wir wollen hierfür eine optimale Gitterweitenverteilung bestimmen. Ausgangspunkt ist die Äquilibrierungsbedingung

$$\eta_T \approx \frac{h_T^4}{d_T^3} \approx \frac{\text{TOL}}{N} \quad \Rightarrow \quad h_T^2 \approx d_T^{3/2} \left( \frac{\text{TOL}}{N} \right)^{1/2}.$$

Hieraus ergibt sich

$$N = \sum_{T \in \mathbb{T}_h} h_T^2 h_T^{-2} = \left( \frac{N}{\text{TOL}} \right)^{1/2} \sum_{T \in \mathbb{T}_h} h_T^2 d_T^{-3/2} \approx \left( \frac{N}{\text{TOL}} \right)^{1/2}$$

und folglich  $N_{\text{opt}} \approx \text{TOL}^{-1}$ . Bei Verwendung des Energienorm-Fehlerschätzers (3.6.156) zur Gitterverfeinerung ergibt sich dagegen die Gitterkomplexität  $N_{\text{opt}} \approx \text{TOL}^{-2}$ .

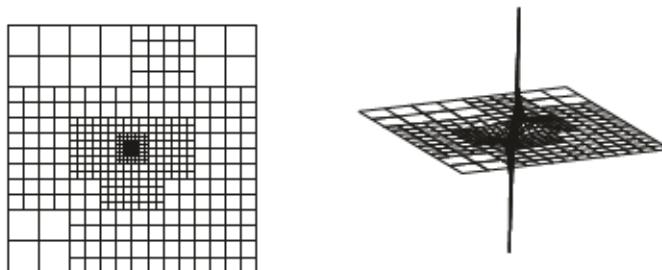


Abbildung 3.26: Verfeinerte Gitter und numerisch bestimmte duale Lösung zur Berechnung von  $\partial_1 u(0)$  auf dem Quadrat bei Verwendung des a posteriori Fehlerschätzers  $\eta_{\text{weight}}(u_h)$  mit  $\text{TOL} = 4^{-4}$ .

Tabelle 3.2: Resultate der Berechnung von  $\partial_1 u(0)$  unter Verwendung des gewichteten a posteriori Fehlerschätzers  $\eta_{\text{weight}}(u_h)$  für verschiedenen Verfeinerungslevel  $L$ ; der „Effektivitätsindex“ ist definiert durch  $I_{\text{eff}} := |\eta_{\text{weight}}(u_h)/\partial_1 e_h(0)|$ .

TOL	N	L	$ \partial_1 e_h(0) $	$\eta_{\text{weight}}(u_h)$	$I_{\text{eff}}$
$4^{-2}$	148	6	7.51e-1	5.92e-2	0.08
$4^{-3}$	940	9	4.10e-1	1.42e-2	0.03
$4^{-4}$	4912	12	4.14e-3	3.50e-3	0.65
$4^{-5}$	20980	15	2.27e-4	9.25e-4	4.16
$4^{-6}$	86740	17	5.82e-5	2.38e-4	4.16

Das vorhergesagte Verhalten der verschiedenen Gitterverfeinerungsprozesse wird durch die numerischen Ergebnisse in Tabelle 3.2 gut bestätigt. Wir betonen, dass die richtige Wahl des Regularisierungsparameters  $\varepsilon = \text{TOL}$  in (3.6.183) für den Lösungsprozess wichtig ist. Die triviale Alternative  $\varepsilon = h_{\min}$  ist automatisch realisiert, wenn zur Berechnung der Gewichte  $\omega_T(z)$  die numerisch bestimmte, diskrete, duale Lösung  $z_h \in V_h$  verwendet wird (siehe Abb. 3.26). Dies kann bei sehr „singulären“ Funktionalen zu starker lokaler Überverfeinerung, d. h. zu unnötig vielen Verfeinerungsschritten, führen.

**Testfall 2:** Sei nun  $\Omega$  das Rechteckgebiet  $\Omega = (-1, 1) \times (-1, 3)$  mit Schlitz bei  $(0, 0)$  (siehe Abb. 3.27). Die Präsenz der (einspringenden) scharfen Ecke mit Innenwinkel  $\omega = 2\pi$  bewirkt in der schwachen Lösung eine „Eckensingularität“ der Form  $s = \psi(\theta)r^{1/2}$ , d. h. eine Singularität im Gradienten. Wir wollen illustrieren, wie diese Eckensingularität mit der durch die Gewichte  $\omega_T(z)$  im Fehlerschätzer induzierten zusammenspielt.

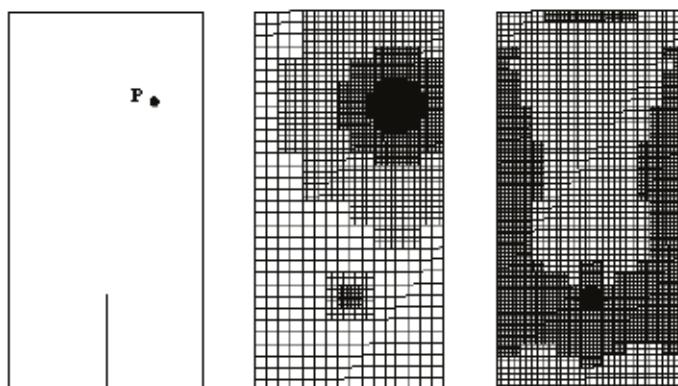


Abbildung 3.27: Verfeinerte Gitter mit ungefähr 5000 Zellen zur Berechnung von  $\partial_1 u(P)$ , erzeugt mit dem gewichteten Fehlerschätzer  $\eta_{\text{weight}}(u_h)$  (mittig) sowie dem Energienorm-Fehlerschätzer  $\eta_E(u_h)$  (rechts).

Wir sehen in Abb. 3.27, dass der gewichtete Fehlerschätzer Zellen sowohl bei der Schlitzspitze (zur Unterdrückung des „Pollutionseffekts“), aber auch beim Auswertungspunkt konzentriert, während der Energienorm-Fehlerschätzer wesentlich stärker an der Schlitzspitze und natürlich gar nicht im Auswertungspunkt verfeinert.

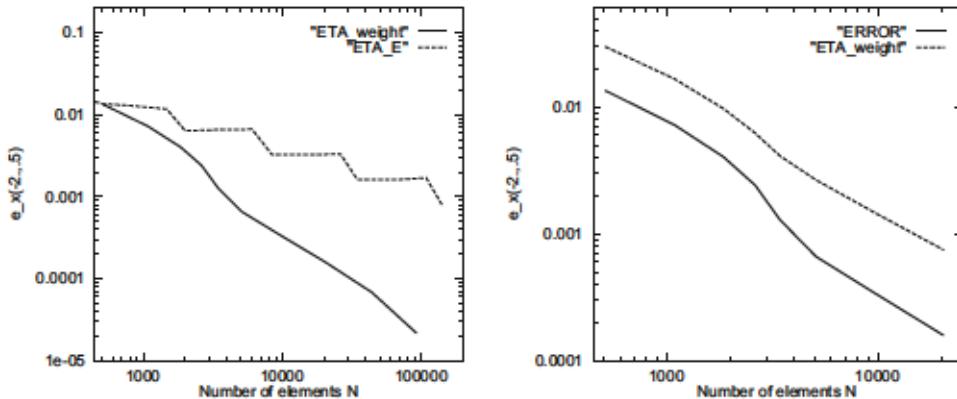


Abbildung 3.28: Vergleich der Effizienz von  $\eta_E(u_h)$  und  $\eta_{\text{weight}}(u_h)$  auf dem Schlitzgebiet.

### 3.7 Übungen

**Übung 3.1:** Man formuliere das Ritzsche Verfahren mit endlich dimensionalen Teilräumen geeigneter Sobolew-Räume  $H^m(\Omega)$  für die folgenden Aufgabenstellungen:

a) Neumannsche RWA des Laplace-Operators:

$$-\Delta u = f \quad \text{in } \Omega, \quad \partial_n u = g \quad \text{auf } \partial\Omega.$$

b) Eigenwertproblem des Laplace-Operators:

$$-\Delta u = \lambda u \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega.$$

c) Dirichletsche RWA des „biharmonischen“ Operators:

$$-\Delta^2 u = f \quad \text{in } \Omega, \quad u = \partial_n u = 0 \quad \text{auf } \partial\Omega.$$

Dabei seien jeweils das Gebiet  $\Omega$  sowie die Daten  $f, g$  als genügend regulär und kompatibel angenommen. Man versuche, hierfür analog zur Vorlesung „Bestapproximations“-Aussagen herzuleiten.

**Übung 3.2:** Seien  $V$  ein Hilbert-Raum mit Skalarprodukt  $(\cdot, \cdot)_V$  und zugehöriger Norm

$\|\cdot\|_V$  und  $a(\cdot, \cdot)$  sowie  $l(\cdot)$  bilineare bzw. lineare Formen auf  $V$  mit den Eigenschaften

$$\begin{aligned} |a(v, w)| &\leq \alpha \|v\|_V \|w\|_V, \quad v, w \in V \quad (\text{Beschränktheit}), \\ |a(v, v)| &\geq \kappa \|v\|_V^2, \quad v, w \in V \quad (V\text{-Elliptizität}), \\ |l(v)| &\leq \beta \|v\|_V, \quad v \in V \quad (\text{Beschränktheit}). \end{aligned}$$

Dann hat die Variationsgleichung

$$a(u, \varphi) = l(\varphi) \quad \forall \varphi \in V$$

nach dem Satz von Lax-Milgram eine eindeutige Lösung  $u \in V$ . Für endlich dimensionale Teilräume  $V_h \subset V$  werden approximative Lösungen  $u_h \in V_h$  bestimmt durch die „Galerkin-Gleichungen“

$$a(u_h, \varphi_h) = l(\varphi_h) \quad \varphi_h \in V_h.$$

- a) Man zeige hierfür die (eindeutige) Existenz der „diskreten“ Lösungen  $u_h \in V_h$  und die Fehlerabschätzung

$$\|u - u_h\|_V \leq \frac{\alpha}{\kappa} \inf_{\varphi_h \in V_h} \|u - \varphi_h\|_V.$$

- b) Man wende das Resultat von (a) zum Nachweis der Konvergenz der Galerkin-Approximation des Diffusions-Konvektions-Problems

$$-\Delta u + \partial_1 u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega,$$

auf einem regulären Gebiet  $\Omega \subset \mathbb{R}^2$  mit rechter Seite  $f \in L^2(\Omega)$  an.

- c) Man versuche das Resultat von (a) für den Fall zu verallgemeinern, dass die Bilinearform  $a(\cdot, \cdot)$  nicht  $V$ -elliptisch sondern nur „koerzitiv“ ist, d. h.:

$$\sup_{\varphi \in V \setminus \{0\}} \frac{a(v, \varphi)}{\|\varphi\|_V} \geq \gamma \|v\|, \quad \sup_{\varphi \in V \setminus \{0\}} \frac{a(\varphi, v)}{\|\varphi\|_V} \geq \gamma \|v\|, \quad v \in V.$$

Worin bestehen hierbei die Probleme?

### Übung 3.3: Das Modellproblem

$$-\Delta u = f \quad \text{in } \Omega, \quad u|_{\partial\Omega} = 0,$$

auf dem Einheitsquadrat  $\Omega = (0, 1)^2$  werde auf einem äquidistanten, kartesischen Gitter mit der Gitterweite  $h$  mit Hilfe der Finite-Elemente-Methode mit stückweise bilinearen Ansatzfunktionen diskretisiert. Man stelle die zugehörigen Systemmatrizen auf:

- a) mit exakter Integration,  
b) unter Verwendung der 2-dimensionalen „Tensorprodukt-Trapezregel“

$$Q_T(f) := \frac{|T|}{4} \sum_{i=1}^4 f(a_i), \quad a_i \text{ Eckpunkte der Zelle } T.$$

Welche Besonderheit ergibt sich?

**Übung 3.4:** Man überlege (etwa durch Konstruktion von Beispielen), in wie weit die folgenden Bedingungen an eine Folge von (im Sinne des Textes regulären) Zerlegungen  $\{\mathbb{T}_h\}_{h>0}$  eines Gebiets  $\Omega \subset \mathbb{R}^2$  (in Dreiecke oder Vierecke) äquivalent sind:

a) Die inneren Winkel aller Zellen  $T \in \mathbb{T}_h$  sind gleichmäßig für  $T \in \mathbb{T}_h$  und  $h > 0$  von Null wegbeschränkt ("minimum angle condition").

b) Für die Inkreisradien  $\rho_T$  und Umkreisradien  $h_T$  der Zellen  $T \in \mathbb{T}_h$  gilt ("uniform shape condition"):

$$\sup_{T \in \mathbb{T}_h, h > 0} \left\{ \frac{h_T}{\rho_T} \right\} < \infty.$$

c) Für die Seiten  $\Gamma \subset \partial T$  jeder Zelle  $T \in \mathbb{T}_h$  gilt

$$\max_{\Gamma \subset \partial T} |\Gamma| \leq c \min_{\Gamma \subset \partial T} |\Gamma|$$

mit einer von  $T$  unabhängigen Konstante  $c$ .

**Übung 3.5:** Sei  $V_h^{(1)} \subset H^1(\Omega)$  der Raum der stückweise linearen finiten Elemente bzgl. einer regulären Triangulierung von  $\bar{\Omega} \subset \mathbb{R}^2$ . Mit Hilfe des  $L^2$ -Skalarprodukts  $(\cdot, \cdot)$  ist die sog. "L<sup>2</sup>-Projektion"  $P_h : L^2(\Omega) \rightarrow V_h^{(1)}$  definiert durch die Vorschrift

$$(P_h u, \varphi_h) = (u, \varphi_h) \quad \forall \varphi_h \in V_h^{(1)}.$$

a) Man leite eine Fehlerabschätzung für die  $L^2$ -Norm  $\|u - P_h u\|$  her, zunächst unter der Annahme  $v \in H^2(\Omega)$ . Was würde man unter den schwächeren Regularitätsannahmen  $v \in H^1(\Omega)$  oder  $v \in L^2(\Omega)$  erhalten?

b) Welche verbesserte Abschätzung erhält man bzgl. der „negativen“ Sobolew-Norm

$$\|u - P_h u\|_{-1} := \sup_{\varphi \in H_0^1(\Omega)} \frac{(u - P_h u, \varphi)}{\|\nabla \varphi\|} \leq ch^\gamma \|v\|_{H^2} ?$$

**Übung 3.6:** Man begründe anhand der eindimensionalen Poisson-Gleichung

$$-u''(x) = f(x) \quad \text{in } \Omega = (0, 1), \quad u(0) = u(1) = 0,$$

dass für die Ritz-Projektion  $R_h : H_0^1(\Omega) \rightarrow V_h^{(1)}$  eine Abschätzung wie in Aufgabe 6.3b nicht gelten kann. (Hinweis: Man verifiziere, dass in diesem Fall die Ritz-Projektion  $R_h u$  identisch mit der Knoteninterpolierenden  $I_h u$  ist.)

**Übung 3.7:** Die Dirichletsche Randwertaufgabe

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega,$$

auf einem konvexen Polygonegebiet  $\Omega \subset \mathbb{R}^2$  werde mit einem Galerkin-Verfahren mit Ansatzräumen  $V_h^{(1)} \subset V := H_0^1(\Omega)$  von „linearen“ finiten Elementen approximiert. Mit welcher Ordnung konvergieren

- i) der gewichtete Mittelwert über  $\Omega$  ( $\omega \in H^1(\Omega)$  eine glatte Gewichtsfunktion):

$$\int_{\Omega} u_h \omega \, dx \rightarrow \int_{\Omega} u \omega \, dx \quad (h \rightarrow 0) ?$$

- ii) der quadratischen Mittelwert über einen glatten geschlossenen Weg  $\Gamma \subset \Omega$ :

$$\int_{\Gamma} u_h^2 \, ds \rightarrow \int_{\Gamma} u^2 \, ds \quad (h \rightarrow 0) ?$$

(Hinweis: Zur Erzielung eines optimalen Resultats verwende man die Variante des Spurlemmas für Funktionen in  $H^{1,1}(\Omega)$  und die Fehlerabschätzungen aus dem Text.)

**Übung 3.8:** Man untersuche, ob die folgenden Sätze von Funktionalen für die angegebenen Polynomräume „unisolvant“ sind. Dabei bezeichnen  $a_i$  die Ecken,  $m_i$  die Seitenmitten sowie  $b_{ij}$  ( $j = 1, 2$ ) jeweils zwei Gauß-Punkte auf der Kante  $\Gamma_i$ ,  $i = 1, \dots, d+1$ , und  $z$  den Schwerpunkt des Elements  $T$ .

- i)  $T$  (kartesisches) Einheitsdreieck:

$$\begin{aligned} P(T) &= P_3(T), \quad p(a_i), \nabla p(a_i), p(z); \\ P(T) &= P_3(T), \quad p(a_i), p(b_{ij}), p(z); \\ P(T) &= P_5(T), \quad p(a_i), \nabla p(a_i), \nabla^2 p(a_i), \partial_n p(m_i). \end{aligned}$$

- ii)  $T$  (kartesisches) Einheitsquadrat:

$$\begin{aligned} P(T) &= \tilde{Q}_1(T) := P_1(T) \oplus \text{span}\{x^2 - y^2\}, \quad p(m_i); \\ P(T) &= \tilde{Q}_3(T) := P_3(T) \oplus \text{span}\{x^3y, xy^3\}, \quad p(a_i), \nabla p(a_i). \end{aligned}$$

**Übung 3.9:** Man leite mit den Argumenten des Textes für die Knoteninterpolierende  $I_h : C(\bar{\Omega}) \rightarrow V_h^{(1)}$  zu stückweise linearen finiten Elementen auf einer quasi-gleichförmigen Folge regulärer Triangulierungen  $(T_h)_{h \in \mathbb{R}_+}$  eines Polygonegebiets (in 2D) bzw. Polyeders (in 3D) die folgenden Fehlerabschätzungen her:

$$\max_{\bar{\Omega}} |v - I_h v| \leq \begin{cases} ch \|\nabla^2 v\|_{\Omega}, & \text{in } 2D, \\ ch^2 \|\nabla^2 v\|_{\Omega}, & \text{in } 3D. \end{cases}$$

**Übung 3.10:** Welche der folgenden (zellweisen) „inversen Abschätzungen“ für Finite-Elementefunktionen  $v_h \in V_h$  sind in 2 Dimensionen gültig und welche nicht (mit Be-

gründung):

- (i)  $\|\nabla^2 v_h\|_{L^2(T)} \leq ch_T^{-2} \|v_h\|_{L^2(T)},$
- (ii)  $\|\partial_n v_h\|_{L^2(\partial T)} \leq ch_T^{-1/2} \|\nabla v_h\|_{L^2(T)},$
- (iii)  $\|\nabla v_h\|_{L^\infty(T)} \leq ch_T^{-2} \|v_h\|_{L^2(T)},$
- (iv)  $\|v_h\|_{L^2(T)} \leq ch_T^{-1} \|v_h\|_{L^1(T)}.$

**Übung 3.11:** Sei  $\Omega \subset \mathbb{R}^2$  ein Polygonegebiet und  $T_h = \{T\}$  eine Triangulierung von  $\bar{\Omega}$ . Man betrachte die beiden mit den oben definierten kubischen Ansätzen gebildeten Ansatzräumen  $S_h^{(3)}$  und  $\tilde{S}_h^{(3)}$  und bestimme deren Dimensionen asymptotisch in Abhängigkeit von  $h$  auf gleichmäßigen Triangulierungen. Wieviele von Null verschiedene Elemente haben die zugehörigen Steifigkeitsmatrizen pro Zeile bei der Approximation der Laplace-Gleichung?

**Übung 3.12:** Die Neumannsche Randwertaufgabe

$$-\Delta u + u = f \quad \text{in } \Omega, \quad \partial_n u = 0 \quad \text{auf } \partial\Omega,$$

auf einem konvexen Polygonegebiet  $\Omega \subset \mathbb{R}^2$  soll mit einem Galerkin-Verfahren mit Ansatzräumen  $V_h^{(1)} \subset V := H_0^1(\Omega)$  von „linearen“ finiten Elementen approximiert werden.

- a) Man formuliere diese Approximation, d. h. die Ansatzräume  $V_h \subset V := H^1(\Omega)$  und die zugehörigen Variationsgleichungen.
- b) Man leite Fehlerabschätzungen in der  $H^1$ - und der  $L^2$ -Norm her.
- c) Wie muss dieser Ansatz im Fall eines krumm berandeten Gebiets und einer inhomogenen Neumann-Randbedingung  $\partial_n u = g$  auf  $\partial\Omega$  zur Erzielung einer *konformen* Approximation modifiziert werden?

**Übung 3.13:** Die Randwertaufgabe

$$-\Delta u = f \quad \text{in } \Omega, \quad u|_{\partial\Omega} = 0,$$

auf einem Gebiet  $\Omega \subset \mathbb{R}^2$  mit (stückweise) kubisch (z. B. CAD-Daten mit kubischen Spline-Funktionen) parametrisiertem  $C^2$ -Rand  $\partial\Omega$  soll mit einem (isoparametrischen) kubischen Finite-Elemente-Ansatz diskretisiert werden. In diesem Fall hat die Lösung mindestens die Regularitätsstufe  $u \in H^3(\Omega)$  (i. Allg. aber  $u \notin H^4(\Omega)$ ), und es gilt die a priori Abschätzung

$$\|u\|_{2+k} \leq c \|\Delta u\|_k, \quad k = 0, 1.$$

- a) Man gebe einen geeigneten Ansatzraum an. Welche Konvergenzordnungen sind dafür zu erwarten bzgl. der Energie-Norm ( $H^1$ -Seminorm) und der  $L^2$ -Norm, und welche Regularität muss dabei jeweils für die Lösung  $u$  vorausgesetzt werden?

b) Welche Fehlerordnung lässt sich mit Hilfe eines Dualitätsarguments für die Mittelwerte zeigen, d. h.:

$$\left| \int_{\Omega} u \, dx - \int_{\Omega} u_h \, dx \right| \leq ch^2 \|u\|_2 ?$$

**Übung 3.14:** Sei  $T \subset \mathbb{R}^2$  ein Dreieck mit Durchmesser  $h_T$  und Inkreisradius  $\rho_T$  mit  $h_T \leq c\rho_T$ , und sei  $I_h v$  die lineare Interpolierende mit den Funktionswerten in den Eckpunkten von  $T$  als Knotenwerte. Man zeige mit den Mitteln des Textes die Abschätzung

$$\|v - I_h v\|_{\partial T} \leq ch_T^{3/2} \|\nabla^2 v\|_T.$$

Hierdurch wird auch die Abschätzung

$$\|v - I_h v\|_{\partial T} \leq ch_T^{1/2} \|\nabla v\|_T$$

nahegelegt. Kann diese gelten?

**Übung 3.15:** Man gebe die bestmöglichen  $h$ -Potenzen in den folgenden Interpolationsfehlerabschätzungen für die Lagrange-Interpolation in  $P(T) := P_2(T)$  an:

- (i)  $\|\nabla^2(v - I_T v)\|_T \leq c_I h_T^2 \|\nabla^3 v\|_T;$
- (ii)  $|(v - I_T v)(a)| \leq c_I h_T^2 \|\nabla^3 v\|_T;$
- (iii)  $\|\partial_n(v - I_T v)\|_{\partial T} \leq c_I h^2 \|\nabla^3 v\|_T;$
- (iv)  $\|v - I_T v\|_T \leq c_I h_T^2 \|\nabla^2 v\|_T.$

**Übung 3.16:** Zur Finite-Elemente-Approximation des sog. „Plattenbiegeproblems“ (Randwertproblem des „biharmonischen“ Operators)

$$\Delta^2 u = f \quad \text{in } \Omega, \quad u = \partial_n u = 0 \quad \text{auf } \partial\Omega,$$

wird wieder von einer zugehörigen variationellen Formulierung ausgegangen. Diese ist auf natürliche Weise im Sobolew-Raum  $V := H_0^2(\Omega) = \{v \in H^2(\Omega) | v|_{\partial\Omega} = \partial_n v|_{\partial\Omega} = 0\}$  definiert: Finde  $u \in V$  mit

$$a(u, \varphi) := (\Delta u, \Delta \varphi) = (f, \varphi) \quad \forall \varphi \in V.$$

Die Bilinearform ist  $a(\cdot, \cdot)$   $V$ -elliptisch, so dass die betrachtete RWA eine eindeutige „schwache“ Lösung  $u \in V$  besitzt. Auf einem Rechteck  $\Omega$  ist diese schwache Lösung sogar in  $H^4(\Omega)$  und genügt der a priori Abschätzung

$$\|u\|_{H^4} \leq c\|f\|.$$

Für einen konformen Finite-Elemente-Ansatzraum  $V_h$  muss nun gelten  $V_h \subset V$ , d. h.: Die stückweise polynomialen Ansatzfunktionen müssen global stetig differenzierbar sein. Das Grundgebiet  $\Omega$  sei als konvex polygonal angenommen.

- a) Einen konformen Finite-Elemente-Ansatzraum  $V_h \subset V$  erhält man mit Hilfe des quintischen Argyris-Elements. Man gebe hierfür eine Fehlerabschätzung in der „Energienorm“ sowie in der  $L^2$ -Norm an. (Hinweis: Dualitätsargument.)
- b) Welche Spektral-Kondition in Abhängigkeit von der Gitterweite  $h$  ist für die zugehörige Steifigkeitsmatrix  $A_h$  zu erwarten?

**Übung 3.17:** Die Steifigkeitsmatrix und der Lastvektor eines kubischen FE-Ansatzes zur Approximation der RWA

$$-\Delta u = 0 \quad \text{in } \Omega, \quad u|_{\partial\Omega} = 0,$$

auf einer Triangulierung von  $\overline{\Omega} \subset \mathbb{R}^2$  werde mit Hilfe numerischer Quadratur berechnet.

- a) Von welcher Ordnung sollte die verwendete Quadraturformel sein, damit (i) Konvergenz des resultierenden Verfahrens bzgl. der Energienorm garantiert ist, und (ii) seine Ordnung optimal ist? Man gebe jeweils ein Verfahrensbeispiel (Quadraturformel) mit möglichst geringer Komplexität (Anzahl der Funktionsauswertungen) an.

- b) Zur Illustration der Notwendigkeit bzw. Nichtnotwendigkeit der im Text abgeleiteten Bedingungen an die numerische Quadratur betrachte man die Approximation der 1. RWA des Laplace-Operators auf dem Einheitsquadrat mit bilinearen finiten Elementen auf einem äquidistanten, kartesischen Gitter. Welches Differenzenschema erhält man, wenn die Elemente der Systemmatrix mit der Mittelpunktsregel berechnet werden? Diese Situation ist durch die Theorie aus der Vorlesung nicht abgedeckt. Ist das resultierende Verfahren dennoch konvergent?

**Übung 3.18:** Für die Systemmatrix  $A_h$  einer Finite-Elemente-Diskretisierung für die 1. RWA des Laplace-Operators

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega,$$

wurde in der Vorlesung für „quasi-gleichförmige“ Triangulierungsfolgen  $(T_h)_{h>0}$  die Kondition  $\text{cond}_2(A_h) = \mathcal{O}(h^{-2})$  gezeigt.

- (i) Man rekapituliere, was für eine Triangulierungsfolge „quasi-gleichförmig“ bedeutet.
- (ii) Wie hängt die Kondition von  $A_h$  von den Inkreis- bzw. Inkugelradien der Zellen ab, wenn die Triangulierungsfolge nicht „formregulär“ ist?
- (iii) Man untersuche die Abhängigkeit der Kondition der Systemmatrix  $\kappa_2(A_h)$  der 5-Punkte-Differenzendiskretisierung auf äquidistanten Tensorproduktgittern mit unterschiedlichen Gitterweiten  $h_x \neq h_y$  vom Seitenverhältnis  $h_x/h_y$  („aspect ratio“). (Hinweis: Man verallgemeinere die in einer früheren Übungsaufgabe hergeleiteten expliziten Formeln für die Eigenwerte der 5-Punkte-Matrix für die vorliegende Situation.)

**Übung 3.19:** Man rekapituliere den Beweis der Konditionsabschätzung  $\kappa_2(A_h) = \mathcal{O}(h^{-2})$  aus dem Text für die Systemmatrix einer FE-Diskretisierung der 1. RWA (Dirichletsche

RWA) des Laplace-Operators auf einer „quasi-gleichförmigen“ Triangulierungsfolge für die FE-Diskretisierung der 2. RWA (Neumannsche RWA):

$$-\Delta u = f \quad \text{in } \Omega, \quad \partial_n u = 0 \quad \text{auf } \partial\Omega.$$

**Übung 3.20:** Es werde die inhomogene Neumannsche RWA

$$-\nabla \cdot (\alpha \nabla u) + \gamma u = f \quad \text{in } \Omega, \quad n \cdot (\alpha \nabla u) = g \quad \text{auf } \partial\Omega,$$

auf einem konvexen Polygongebiet  $\Omega \subset \mathbb{R}^2$  betrachtet. Die Daten  $\alpha$ ,  $\gamma$ ,  $f$  und  $g$  seien glatt und ferner  $\alpha, \gamma > 0$ . Mit Hilfe der FEM mit stückweise linearen Ansatzfunktionen wird eine Näherung  $u_h \in V_h^{(1)} \subset H^1(\Omega)$  berechnet.

a) Man gebe die zugehörige variationelle Formulierung an.

b) Man leite eine *a posteriori* Fehlerabschätzung für den Energie-Norm-Fehler her:

$$\|e_h\|_E := ((\alpha \nabla e_h, \nabla e_h)_\Omega + (\gamma e_h, e_h)_\Omega)^{1/2}.$$

c) Man leite eine *a posteriori* Fehlerabschätzung für den  $L^2$ -Norm-Fehler  $\|e_h\|_2$  her.

**Übung 3.21:** Die 1. RWA des Laplace-Operators auf einem konvexen Polygongebiet  $\Omega \subset \mathbb{R}^2$  werde durch „lineare“ finite Elemente auf einer Triangulierung  $T_h$  von  $\bar{\Omega}$  approximiert. Man zeige, dass die in der Vorlesung abgeleitete *a posteriori* Abschätzung für den  $L^2$ -Norm-Fehler,

$$\|u - u_h\|_2 \leq c \eta_{L^2}(u_h)$$

mit dem „Schätzer“

$$\eta_{L^2}(u_h) := \left( \sum_{T \in T_h} h_T^4 \{ \|f + \Delta u_h\|_{2;T}^2 + \frac{1}{2} h_T^{-1} \|[\partial_n u_h]\|_{2;\partial T \setminus \partial\Omega}^2 \} \right)^{1/2},$$

asymptotisch optimal ist, d.h.:

$$\eta_{L^2}(u_h) \leq c \|u - u_h\|_2 + \left( \sum_{T \in T_h} h_T^4 \|f + \Delta u_h\|_2^2 \right)^{1/2}.$$

Dazu verwende man die „Spur-Abschätzung“

$$\|\partial_n v\|_{2;\partial T} \leq c h_T^{1/2} \|\Delta v\|_{2;T} + h_T^{-3/2} \|v\|_{2;T},$$

deren genauen  $h$ -Potenzen man mit Hilfe des üblichen Transformationsarguments aus der entsprechenden Ungleichung auf einer „Einheitszelle“ (Beweisskizze zur Wiederholung) erhält.

**Übung 3.22:** In Anlehnung an die Argumentationsweise des Textes leite man eine For-

mel für eine Gitterweitenfunktion  $h(x)$  her, welche in folgendem Sinne „optimal“ ist:

$$N \rightarrow \min!, \quad \eta(u_h) \leq \text{TOL},$$

wobei

$$N := \int_{\Omega} h(x)^{-2} dx, \quad \eta(u_h) := \int_{\Omega} h(x)^2 A(x) dx.$$

**Übung 3.23:** Dem in der vorigen Aufgabe verwendeten Optimierungsargument liegt die Annahme zugrunde, dass sich die Zellresiduen im wesentlichen wie

$$\rho_T := \|f + \Delta u_h\|_T + \frac{1}{2} h_T^{1/2} \|[\partial_n u_h]\|_{\partial T} = \mathcal{O}(h_T)$$

verhalten. Man zeige, dass dies im Fall linearer Ansatzfunktionen auf einer quasi-gleichförmigen (regulär mit “uniform shape”- und “uniform-size”-Eigenschaft) Folge von Triangulierungen im  $\mathbb{R}^2$  mit  $h_T \sim h$  tatsächlich der Fall ist. Dazu verwende man die bekannte a priori Fehlerabschätzung

$$\|\nabla(u - u_h)\|_{\infty} \leq ch \|\nabla^2 u\|_{\infty}$$

unter der Annahme  $u \in W^{2,\infty}(\Omega)$ , sowie die gleichfalls bekannten lokalen Interpolationsfehlerabschätzungen und „inversen“ Beziehungen für Finite-Elemente-Funktionen. (Bem.: Die Annahme der Quasi-Gleichförmigkeit ist natürlich unrealistisch, da ja im Endeffekt gerade lokal verfeinerte Gitter erzeugt werden sollen.)

**Übung 3.24:** Man gebe die bestmöglichen  $h$ -Potenzen in den folgenden Interpolationsfehlerabschätzungen für die Lagrange-Interpolation in  $P(T) := P_3(T)$  auf formregulären Gittern an (Begründung nicht erforderlich):

- (i)  $\|\nabla^2(v - I_T v)\|_T \leq c_i h_T^2 \|\nabla^4 v\|_T;$
- (ii)  $|v - I_T v)(a)| \leq c_i h_T^2 \|\nabla^4 v\|_T;$
- (iii)  $\|\partial_n(v - I_T v)\|_{\partial T} \leq c_i h_T^2 \|\nabla^4 v\|_T;$
- (iv)  $\|v - I_T v\|_T \leq c_i h_T^2 \|\nabla^2 v\|_T.$

**Übung 3.25:** Es werde die Dirichletsche RWA

$$-\nabla \cdot (\alpha \nabla u) + \gamma u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega,$$

auf einem konvexen Polygongebiet  $\Omega \subset \mathbb{R}^2$  betrachtet. Die Daten  $\alpha$ ,  $\gamma$  und  $f$  seien hinreichend glatt und ferner  $\alpha > 0, \gamma \geq 0$ . Mit Hilfe der FEM mit stückweise linearen Ansatzfunktionen wird eine Näherung  $u_h \in V_h^{(1)} \subset H_0^1(\Omega)$  berechnet.

- a) Man gebe die zugehörige variationelle Formulierung an.
- b) Man gebe eine *a posteriori* Fehlerabschätzung für den Energie-Norm-Fehler an:

$$\|e_h\|_E := \sqrt{(\alpha \nabla e_h, \nabla e_h)_\Omega + (\gamma e_h, e_h)_\Omega} \leq ?$$

- c) Man skizziere die Herleitung einer *a posteriori* Fehlerabschätzung für den  $L^2$ -Norm-Fehler:

$$\|e_h\| \leq ?$$

**Übung 3.26:** Zur Lösung der 1. RWA der Laplace-Gleichung

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega,$$

auf dem Einheitsquadrat  $\Omega = (0, 1)^2$  werde auf einer Folge äquidistanter, kartesischer Gitter  $T_l$  mit Gitterweiten  $h_l = 2^{-l}$  mit Hilfe bilinearer finiter Elemente approximiert. Die diskrete Gleichung auf Gitterlevel  $l$  werde dabei mit einem MG-Verfahren gelöst, wobei das Richardson-Verfahren zur Glättung, die natürliche Einbettung zur Prolongation und die  $L^2$ -Projektion zur Restriktion verwendet werden. Die Anzahl der Vor- und Nachglättungsschritte sei  $\nu = 2$  und  $\mu = 0$ . Wieviele a. Op. kosten dann ungefähr ein V-Zyklus und ein W-Zyklus ausgedrückt in Vielfachen der Dimension  $N_l = \dim V_l$ ?

**Übung 3.27:** Zur Überprüfung des bisherigen Lernerfolgs versuche man, ohne Rückgriff auf den Text die folgenden Fragen zu beantworten:

- Welches Differenzenschema erhält man, wenn bei der Approximation der 1. RWA des Laplace-Operators auf dem Einheitsquadrat mit bilinearen finiten Elementen auf einem äquidistanten, kartesischen Gitter die Elemente der Systemmatrix mit der Tensorprodukt-Trapezregel berechnet werden?
- Was unterscheidet bei Familien von Triangulierungen  $\{T_h\}_{h>0}$  von Gebieten  $\Omega \subset \mathbb{R}^2$  die „Minimalwinkelbedingung“ von der „Maximalwinkelbedingung“, und wie hängt das mit den Bedingungen „formregulär“ und „größenregulär“ zusammen?
- Auf einem Tetraeder  $T \in \mathbb{R}^3$  seien ein Polynomraum  $P(T)$  und ein Satz von Funktionale  $\chi_r : C^1(\bar{\Omega}) \rightarrow P(T)$  ( $r = 1, \dots, R$ ) gegeben. Was bedeutet die Aussage, dass  $\{\chi_r\}_{r=1,\dots,R}$  „unisolvent“ bzgl.  $P(T)$  ist?
- Welche Dimensionen haben auf dem  $\mathbb{R}^2$  die Polynomräume  $P_2$ ,  $P_5$  und  $Q_2$ ?
- Die 1. RWA des Laplace-Operators auf dem Einheitswürfel  $\Omega = (0, 1)^3 \subset \mathbb{R}^3$  sei mit stückweise quadratischen finiten Elementen auf einer regulären Folge von Gittern  $T_h$  approximiert. Welche Spektralkondition haben die zugehörigen Systemmatrizen (der Knotenbasen) in Abhängigkeit von der Gitterweite  $h$ ?



## 4 Lösung der FE-Gleichungen

In diesem Kapitel werden iterative Lösungsverfahren für die durch Anwendung einer Finite-Differenzen- oder Finite-Elemente-Diskretisierung entstehenden (linearen) Gleichungssysteme diskutiert. Dies sind neben den traditionellen Fixpunktiterationen vor allem sog. „PCG-Verfahren“ (preconditioned conjugate gradient methods) und die modernen „Mehrgittermethoden“. Zugrunde gelegt wird dabei meist wieder das Modellproblem der 1. RWA des Laplace-Operators

$$Lu := -\Delta u = f \quad \text{in } \Omega, \quad u = g \quad \text{auf } \partial\Omega, \quad (4.0.1)$$

auf einem (konvexen) Polygonegebiet  $\Omega \subset \mathbf{R}^2$ . Erweiterungen für Probleme mit variablen Koeffizienten, anderen Randbedingungen, Unsymmetrien sowie auf drei Raumdimensionen werden wieder in Bemerkungen berücksichtigt. Die zugehörigen algebraischen Systeme haben die Form

$$Ax = b, \quad (4.0.2)$$

mit Matrizen  $A = (a_{nm})_{n,m=1}^N$  und Vektoren  $b = (b_n)_{n=1}^N$  der Dimension  $N$ . In der Praxis ist meist  $N \gg 1000$ , so dass neben dem Rechenaufwand auch der Speicherbedarf ein wichtiger Aspekt ist. Je nach der gewählten Numerierung der Gitterpunkte bzw. Knoten haben die Matrizen in der Regel Bandstruktur und sind extrem dünn besetzt. Ist das kontinuierliche Problem selbstadjungiert sowie definit, wie z. B. im Fall der 1. RWA des Laplace-Operators, so übertragen sich diese Eigenschaften bei FE-Diskretisierungen direkt auf die Systemmatrizen  $A$ .

### 4.1 Krylow-Raum-Methoden

Wir diskutieren jetzt sog. „Krylow<sup>1</sup>-Raum-Methoden“, zu denen auch das klassische Verfahren der konjugierten Gradienten („CG-Verfahren“) gehört. Im folgenden werden euklidisches Skalarprodukt und Norm auf  $\mathbf{R}^N$  mit  $\langle x, y \rangle$  bzw.  $|x|$  bezeichnet. Die Koeffizientenmatrix  $A \in \mathbf{R}^{N \times N}$  sei zunächst als symmetrisch und positiv definit angenommen. Dann lässt sich die Gleichung (4.0.2) äquivalent charakterisieren durch ein quadratisches Minimierungsproblem:

$$Ax = b \quad \Leftrightarrow \quad Q(x) = \min_{y \in \mathbf{R}^N} Q(y) \quad (4.1.3)$$

mit

$$Q(y) := \frac{1}{2} \langle Ay, y \rangle - \langle b, y \rangle.$$

---

<sup>1</sup>Aleksei Nikolaeovich Krylov (1863–1945): Russischer Mathematiker; Prof. an der Sov. Akademie der Wissensch. in St. Petersburg; Beiträge zu Fourier-Analysis und Differentialgleichungen, Anwendungen in der Schiffstechnik.

Wegen  $\nabla^2 Q \equiv A$  folgt aus der Positiv-Definitheit von  $A$  die Existenz eines eindeutig bestimmten Minimums von  $Q(\cdot)$ , welches notwendig Lösung von (4.0.2) ist. Diese Konstruktion ist analog zu der beim Nachweis von „schwachen“ Lösungen der 1. RWA des Laplace-Operators. Wir halten fest, dass der Gradient von  $Q$  in einem Punkt  $y \in \mathbb{R}^n$  gegeben ist durch

$$\nabla Q(y) = \frac{1}{2} (A + A^T)y - b = Ay - b. \quad (4.1.4)$$

Dies ist gerade der „Defekt“ im Punkt  $y$ . Für jede symmetrische, positiv definite Matrix  $B \in \mathbb{R}^{N \times N}$  ist durch  $\|y\|_B := \langle By, y \rangle^{1/2}$  eine sog. „Energie-Norm“ definiert. Mit dieser Notation gilt dann

$$\begin{aligned} 2Q(y) &= \langle Ay, y \rangle - 2\langle b, y \rangle \\ &= \langle Ay, y \rangle - \langle b, y \rangle - \langle Ay, A^{-1}b \rangle + \langle b, A^{-1}b \rangle - \langle b, A^{-1}b \rangle \\ &= \langle Ay - b, y - A^{-1}b \rangle - \langle b, A^{-1}b \rangle \\ &= \langle A^{-1}(Ay - b), Ay - b \rangle - \langle b, A^{-1}b \rangle = |Ay - b|_{A^{-1}}^2 - |b|_{A^{-1}}^2 \\ &= \langle y - A^{-1}b, A(y - A^{-1}b) \rangle - \langle AA^{-1}b, A^{-1}b \rangle = |y - x|_A^2 - |x|_A^2, \end{aligned}$$

d. h.: Die Minimierung des Funktionals  $Q(\cdot)$  ist äquivalent zur Minimierung der Defektnorm  $|Ay - b|_{A^{-1}}$  bzw. der Energie-Norm  $|y - x|_A$ .

Die sog. „Abstiegsverfahren“ bestimmen nun ausgehend von einem geeigneten Startvektor  $x^{(0)} \in \mathbb{R}^n$  eine Folge von Iterierten  $x^{(t)}$ ,  $t \in \mathbb{N}$ , durch

$$x^{(t+1)} = x^{(t)} + \alpha_t d^{(t)}. \quad (4.1.5)$$

Dabei sind die  $d^{(t)}$  vorgegebene oder auch erst im Verlauf der Iteration berechnete „Abstiegsrichtungen“, und die „Schrittweiten“  $\alpha_t \in \mathbb{R}$  sind durch die Vorschrift bestimmt (sog. „line search“):

$$Q(x^{(t+1)}) = \min_{\alpha \in \mathbb{R}} Q(x^{(t)} + \alpha d^{(t)}). \quad (4.1.6)$$

Die notwendige Optimalitätsbedingung

$$\frac{d}{d\alpha} Q(x^{(t)} + \alpha d^{(t)}) = \nabla Q(x^{(t)} + \alpha d^{(t)}) \cdot d^{(t)} = \langle Ax^{(t)} - b, d^{(t)} \rangle + \alpha \langle Ad^{(t)}, d^{(t)} \rangle = 0$$

ergibt mit dem Residuum  $r^{(t)} := b - Ax^{(t)} = -\nabla Q(x^{(t)})$ :

$$\alpha_t = \frac{\langle r^{(t)}, d^{(t)} \rangle}{\langle Ad^{(t)}, d^{(t)} \rangle}.$$

Das allgemeine Abstiegsverfahren lautet also wie folgt:

$$\begin{aligned} \text{Startwert: } \quad & x^{(0)} \in \mathbb{R}^N, \\ \text{für } t \geq 0: \quad & \text{Iterierte } x^{(t)}, \quad \text{Residuum } r^{(t)} = b - Ax^{(t)}, \quad \text{Abstiegsrichtung } d^{(t)}, \\ & \alpha_t = \frac{\langle r^{(t)}, d^{(t)} \rangle}{\langle Ad^{(t)}, d^{(t)} \rangle}, \quad x^{(t+1)} = x^{(t)} + \alpha_t d^{(t)}. \end{aligned}$$

Praktisch günstiger ist die folgende Schreibweise, bei der man eine Matrix-Vektor-Multiplikation spart, wenn man den Vektor  $Ad^{(t)}$  abspeichert:

$$\begin{aligned} \text{Startwert: } & x^{(0)} \in \mathbb{R}^n, \quad r^{(0)} := b - Ax^{(0)}, \\ \text{für } t \geq 0: & \text{ Iterierte } x^{(t)}, \quad \text{Residuum } r^{(t)}, \quad \text{Abstiegsrichtung } d^{(t)}, \\ & \alpha_t = \frac{\langle r^{(t)}, d^{(t)} \rangle}{\langle Ad^{(t)}, d^{(t)} \rangle}, \quad x^{(t+1)} = x^{(t)} + \alpha_t d^{(t)}, \quad r^{(t+1)} = r^{(t)} - \alpha_t Ad^{(t)}. \end{aligned}$$

Die verschiedenen Abstiegsverfahren unterscheiden sich im wesentlichen durch die jeweilige Wahl der Abstiegsrichtungen  $d^{(t)}$ . Die einfachste Möglichkeit wäre, die Richtungen  $d^{(t)}$  zyklisch die kartesischen Einheitsvektoren  $\{e^{(1)}, \dots, e^{(n)}\}$  durchlaufen zu lassen. Die so erhaltene iterative Methode wird „Koordinatenrelaxation“ genannt. Ein voller Relaxationszyklus ist äquivalent zum Gauß-Seidel-Verfahren (Übungsaufgabe).

Naheliegender ist die Wahl der Richtung des stärksten Abfalls von  $Q(\cdot)$  im Punkt  $x^{(t)}$ , d. h. des Gradienten bzw. Residuums, als Suchrichtung  $d^{(t)} = -g^{(t)} = -\nabla Q(x^{(t)}) = r^{(t)}$ . Diese „Gradientenverfahren“ ist aber relativ langsam (vergleichbar mit dem Jacobi-Verfahren). Je zwei aufeinander folgende Abstiegsrichtungen sind dabei orthogonal zu einander,  $(d^{(t+1)}, d^{(t)}) = 0$ , aber  $d^{(t+2)}$  braucht nicht einmal annähernd orthogonal zu  $d^{(t)}$  zu sein. Dies führt zu einem stark oszillatorischen Konvergenzverhalten des Gradientenverfahrens besonders bei Matrizen  $A$  mit weit auseinander liegenden Eigenwerten. Dies bedeutet etwa in Fall  $N = 2$ , dass das Funktional  $Q(\cdot)$  stark exzentrische Niveaulinien hat und sich die Iterierten in einem Zickzackkurs der Lösung annähern (s. Abb. 4.1).



Abbildung 4.1: Niveaulinien des quadratischen Funktionals in zwei Dimensionen und „Zickzacking“ des Gradientenverfahrens

#### 4.1.1 Verfahren der konjugierten Richtungen (CG-Verfahren)

Das Gradientenverfahren nutzt die Struktur des quadratischen Funktionals  $Q(\cdot)$ , d. h. die Verteilung der Eigenwerte der Matrix  $A$ , nur lokal von einem Schritt zum nächsten aus. Es wäre günstiger, wenn bei der Wahl der Abstiegsrichtungen auch die bereits gewonnenen Informationen über die globale Struktur von  $Q(\cdot)$  berücksichtigt würden, d. h. wenn etwa die Abstiegsrichtungen paarweise orthogonal wären. Dies ist die Grundidee des sog. „Ver-

fahrens der konjugierten Gradienten<sup>2</sup> nach Hestenes<sup>2</sup> und Stiefel<sup>3</sup> („conjugate gradient method“ oder kurz „CG-Verfahren“), welches sukzessive eine Folge von Abstiegsrichtungen  $d^{(t)}$  erzeugt, die bzgl. des Skalarprodukts  $(\cdot, \cdot)_A$  orthogonal sind („A-orthogonal“). Zur Konstruktion dieser Folge macht man den Ansatz

$$K_t = \text{span}\{d^{(0)}, \dots, d^{(t-1)}\}$$

und sucht, Iterierte in der Form

$$x^{(t)} = x^{(0)} + \sum_{i=0}^{t-1} \alpha_i d^{(i)} \in x^{(0)} + K_t \quad (4.1.7)$$

zu bestimmen, so dass

$$Q(x^{(t)}) = \min_{y \in K_t} Q(x^{(0)} + y).$$

Dies ist äquivalent zu den „Galerkin-Gleichungen“

$$\langle x^{(t)} - x, d^{(t)} \rangle_A = \langle Ax^{(t)} - b, d^{(j)} \rangle = 0, \quad j = 0, \dots, t-1, \quad (4.1.8)$$

bzw. zu  $r^{(t)} = b - Ax^{(t)} \perp K_t$ . Setzt man den Ansatz (4.1.7) in (4.1.8) ein, so erhält man das Gleichungssystem

$$\sum_{i=0}^{t-1} \alpha_i \langle Ad^{(i)}, d^{(j)} \rangle = \langle b - Ax^{(0)}, d^j \rangle, \quad j = 0, \dots, t-1, \quad (4.1.9)$$

mit der regulären Koeffizientenmatrix  $M_A = (\langle Ad^{(k)}, d^{(j)} \rangle)_{j,k=0}^{t-1}$ .

Eine natürliche Wahl der Ansatzräume  $K_t$  sind die sog. „Krylow-Räume“

$$K_t(r^{(0)}; A) := \text{span}\{r^{(0)}, Ar^{(0)}, \dots, A^{t-1}r^{(0)}\}$$

zum Residuum  $r^{(0)} = b - Ax^{(0)}$  des Startvektors  $x^{(0)}$ . Nach Konstruktion ist stets

$$\begin{aligned} r^{(t)} &= b - Ax^{(t)} = r^{(0)} - r^{(0)} + b - Ax^{(t)} \\ &= r^{(0)} + A(x^{(0)} - x^{(t)}) \in r^{(0)} + AK_t(r^{(0)}; A) \subset K_{t+1}(r^{(0)}; A). \end{aligned}$$

Da nach Konstruktion  $r^{(t)} \perp K_t$ , ist also stets

$$\langle r^{(t)}, r^{(i)} \rangle = 0, \quad i = 0, \dots, t-1.$$

Ferner folgt im Fall  $A^t r^{(0)} \in K_t(r^{(0)}; A)$  notwendig  $r^{(t)} = 0$  bzw.  $Ax^{(t)} = b$ .

Ausgehend von der Formulierung (4.1.8) konstruiert das CG-Verfahren eine A-orthogo-

---

<sup>2</sup>Magnus R. Hestenes (1906-1991): US-amerikanischer Mathematiker; Prof. an der UCLA, USA, fundamentale Beiträge u.a. zur Numerischen Linearen Algebra.

<sup>3</sup>Eduard Stiefel (1909–1978): Schweizer Mathematiker; Prof. an der ETH Zürich; fundamentale Beiträge u.a. zur Numerischen Linearen Algebra.

nale Folge von Abstiegsrichtungen  $d^{(i)}$ , die eine Basis des Krylow-Raumes  $K_t(r^{(0)}; A)$  bildet. Dies ließe sich etwa mit Hilfe des klassischen Gram<sup>4</sup>-Schmidt<sup>5</sup>-Algorithmus leisten, was aber numerisch sehr instabil ist. In Analogie zur vergleichbaren Situation bei der Konstruktion orthogonaler Polynome (z.B. die Legendre<sup>6</sup>-Polynome) ist aber zu erwarten, dass dasselbe durch eine zweistufige Rekursion erreichbar ist.

Ausgehend von einem Startpunkt  $x^{(0)}$  mit Residuum (negativer Gradient)  $r^{(0)} = b - Ax^{(0)}$  seien Iterierte  $x^{(i)}$  und zugehörige Abstiegsrichtungen  $d^{(i)} (i = 0, \dots, t-1)$  bestimmt, so dass  $\{d^{(0)}, \dots, d^{(t-1)}\}$  eine A-orthogonale Basis von  $K_t(d^{(0)}; A)$  ist. Zur Konstruktion des nächsten  $d^{(t)} \in K_{t+1}(d^{(0)}; A)$  mit der Eigenschaft  $d^{(t)} \perp_A K_t(d^{(0)}; A)$  machen wir den folgenden Ansatz:

$$d^{(t)} = r^{(t)} + \sum_{j=0}^{t-1} \beta_j^{t-1} d^{(j)} \in K_{t+1}(d^{(0)}; A). \quad (4.1.10)$$

Dabei wird o.B.d.A. angenommen, dass  $r^{(t)} = b - Ax^{(t)} \notin K_t(d^{(0)}; A)$  ist, da andernfalls  $r^{(t)} = 0$  bzw.  $x^{(t)} = x$  wäre. Zur Bestimmung der Koeffizienten  $\beta_j^{t-1}$  beachten wir für  $i = 0, \dots, t-1$ :

$$0 = \langle d^{(t)}, Ad^{(i)} \rangle = \langle r^{(t)}, Ad^{(i)} \rangle + \sum_{j=0}^{t-1} \beta_j^{t-1} \langle d^{(j)}, Ad^{(i)} \rangle = \langle r^{(t)} + \beta_i^{t-1} d^{(i)}, Ad^{(i)} \rangle.$$

Für  $i < t-1$  ist  $\langle r^{(t)}, Ad^{(i)} \rangle = 0$  wegen  $Ad^{(i)} \in K_t(d^{(0)}; A)$  und demnach  $\beta_i^{t-1} = 0$ . Für  $i = t-1$  führt die Bedingung

$$0 = \langle r^{(t)}, Ad^{(t-1)} \rangle + \beta_{t-1}^{t-1} \langle d^{(t-1)}, Ad^{(t-1)} \rangle \quad (4.1.11)$$

zu den Formeln

$$\beta_{t-1} := \beta_{t-1}^{t-1} = -\frac{\langle r^{(t)}, Ad^{(t-1)} \rangle}{\langle d^{(t-1)}, Ad^{(t-1)} \rangle}, \quad d^{(t)} = r^{(t)} + \beta_{t-1} d^{(t-1)}. \quad (4.1.12)$$

Die nächsten Iterierten  $x^{(t+1)}$  und  $r^{(t+1)} = b - Ax^{(t+1)}$  sind dann bestimmt durch

$$\alpha_t = \frac{\langle r^{(t)}, d^{(t)} \rangle}{\langle d^{(t)}, Ad^{(t)} \rangle}, \quad x^{(t+1)} = x^{(t)} + \alpha_t d^{(t)}, \quad r^{(t+1)} = r^{(t)} - \alpha_t Ad^{(t)}. \quad (4.1.13)$$

---

<sup>4</sup>Jørgen Pedersen Gram (1850–1916): Dänischer Mathematiker, Mitarbeiter und später Eigentümer einer Versicherungsgesellschaft, Beiträge zur Algebra (Invariantentheorie), Wahrscheinlichkeitstheorie, Numerik und Forstwissenschaft; das u.a. nach ihm benannte Orthogonalisierungsverfahren geht aber wohl auf Laplace zurück und wurde bereits von Cauchy 1836 verwendet.

<sup>5</sup>Erhard Schmidt (1876–1959): Deutscher Mathematiker, Prof. in Berlin, Gründer des dortigen Instituts für Angewandte Mathematik 1920, nach dem Krieg Direktor des Mathematischen Instituts der Akademie der Wissenschaften der DDR; Beiträge zur Theorie der Integralgleichungen und der Hilbert-Räume sowie später zur Topologie.

<sup>6</sup>Adrien-Marie Legendre (1752–1833): Französischer Mathematiker; Mitglied der Pariser Akademie der Wissensch.; Beiträge zur Himmelsmechanik, Zahlentheorie und Geometrie.

Dies sind die Rekursionsformeln des klassischen CG-Verfahrens. Wegen  $r^{(t)} \perp K_t(r^{(0)}; A)$  und  $K_t(r^{(0)}; A) = \text{span}\{d^{(0)}, \dots, d^{(t-1)}\}$  ist

$$\langle r^{(t)}, d^{(i)} \rangle = 0, \quad i = 0, \dots, t-1.$$

Damit lassen sich die Formeln für die Koeffizienten  $\alpha_t$  und  $\beta_t$  vereinfachen. Mit

$$|r^{(t)}|^2 = \langle d^{(t)} - \beta_{t-1} d^{(t-1)}, r^{(t+1)} + \alpha_t A d^{(t)} \rangle = \alpha_t \langle d^{(t)}, A d^{(t)} \rangle, \quad (4.1.14)$$

$$|r^{(t+1)}|^2 = \langle r^{(t)} - \alpha_t A d^{(t)}, r^{(t+1)} \rangle = -\alpha_t \langle A d^{(t)}, r^{(t+1)} \rangle. \quad (4.1.15)$$

ergibt sich

$$\alpha_t = \frac{|r^{(t)}|^2}{\langle d^{(t)}, A d^{(t)} \rangle}, \quad \beta_t = \frac{|r^{(t+1)}|^2}{|r^{(t)}|^2}, \quad (4.1.16)$$

solange die Iteration nicht mit  $r^{(t)} = 0$  abbricht. Diese Konstruktion führt auf den folgenden „CG-Algorithmus“:

$$\begin{aligned} \text{Startwert: } & x^{(0)} \in \mathbb{R}^n, \quad r^{(0)} := b - Ax^{(0)}, \\ \text{für } t \geq 0: & \alpha_t = \frac{|r^{(t)}|^2}{\langle Ad^{(t)}, d^{(t)} \rangle}, \\ & x^{(t+1)} = x^{(t)} + \alpha_t d^{(t)}, \quad r^{(t+1)} = r^{(t)} - \alpha_t A d^{(t)}, \\ & \beta_t = \frac{|r^{(t+1)}|^2}{|r^{(t)}|^2}, \quad d^{(t+1)} = r^{(t+1)} + \beta_t d^{(t)}. \end{aligned}$$

In jedem Iterationsschritt ist dabei eine Matrix-Vektor-Multiplikation und fünf Operationen der Mächtigkeit eines Skalarproduktbildung durchzuführen. Im Falle einer dünn besetzten Matrix vom Typ der Modellmatrix sind das etwa  $10N$  arithmetische Operationen. Das CG-Verfahren erzeugt eine Basis von Abstiegsrichtungen, so dass es zwangsläufig nach spätestens  $t = N - 1$  Schritten mit der Lösung  $x$  des Gleichungssystems (4.0.2) abbricht. Es handelt sich hierbei also formal um ein „direktes“ Lösungsverfahren. Bei großer Dimension  $N \geq 10^3$  geht die A-Orthogonalität der Folge  $\{d^{(0)}, \dots, d^{(t-1)}\}$  wegen des unvermeidlichen Rundungsfehlereinflusses schnell verloren, und das Verfahren wird zu einem nicht terminierenden, iterativen Prozess. Außerdem wäre die Durchführung von nahezu  $N - 1$  Iterationsschritten wegen der damit verbundenen Zahl von  $O(N^2)$  arithmetischen Operationen viel zu hoch. Man wird also mit einer wesentlich geringeren Anzahl von  $t \ll N$  Schritten auskommen müssen. Wir fassen die wichtigsten Eigenschaften des CG-Verfahrens in folgendem Satz zusammen.

**Satz 4.1 (CG-Konvergenz):** Das CG-Verfahren bricht für jeden Startvektor  $x^{(0)} \in \mathbb{R}^N$  nach spätestens  $N - 1$  Schritten mit  $x^{(t)} = x$  ab. Für  $0 \leq t < N - 1$  gilt die Fehlerabschätzung

$$|e^{(t)}|_A \leq 2 \left( \frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}} \right)^t |e^{(0)}|_A, \quad t \geq 1, \quad (4.1.17)$$

mit der Spektralkondition  $\kappa := \kappa_2(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$  von  $A$ . Zur Reduzierung des Anfangsfehlers um den Faktor  $\varepsilon$  sind höchstens

$$t(\varepsilon) \leq \frac{1}{2} \sqrt{\kappa} \ln \left( \frac{2}{\varepsilon} \right) + 1 \quad (4.1.18)$$

*Iterationsschritte erforderlich.*

Dieses Resultat zeigt die Wichtigkeit der Kondition  $\kappa(A)$  für die Konvergenzgeschwindigkeit des CG-Verfahrens. Für das Modellproblem ist  $\kappa(A) = \mathcal{O}(h^{-2})$ , was eine Gesamtlösungskomplexität von  $O(N^{3/2})$  bedeutet. Das CG-Verfahren ist daher i.a. ähnlich effizient wie das „optimale“ SOR-Verfahren, allerdings mit einer größeren Fehlerkonstanten. Im Gegensatz zu letzterem erfordert das CG-Verfahren aber nicht die Bestimmung eines optimalen Relaxationsparameters. Dafür ist das Resultat (4.1.17) auf den Fall einer symmetrischen, positiv definiten Matrix  $A$  beschränkt.

**Beweis:** i) Unter Beachtung der Beziehung

$$|x^{(t)} - x|_A = \min_{y \in x^{(0)} + K_t} |y - x|_A,$$

$$K_t := K_t(r^{(0)}; A) = \text{span}\{d^{(0)}, \dots, d^{(t-1)}\} = \text{span}\{A^0 r^{(0)}, \dots, A^{t-1} r^{(0)}\}$$

finden wir

$$|x^{(t)} - x|_A = \min_{p \in P_{t-1}} |x^{(0)} - x + p(A)r^{(0)}|_A.$$

Wegen  $r^{(0)} = b - Ax^{(0)} = A(x - x^{(0)})$  folgt weiter

$$\begin{aligned} |x^{(t)} - x|_A &= \min_{p \in P_{t-1}} |[I - p(A)A](x^{(0)} - x)|_A \\ &\leq \min_{p \in P_{t-1}} |I + Ap(A)|_A |x^{(0)} - x|_A \leq \min_{p \in P_t, p(0)=1} |p(A)|_A |x^{(0)} - x|_A, \end{aligned}$$

wobei die zu der Vektornorm  $|\cdot|_A$  assoziierte natürliche Matrizenorm der Einfachheit halber ebenfalls mit  $|\cdot|_A$  bezeichnet ist. Für beliebiges  $y \in \mathbb{R}^N$  haben wir mit einer Orthonormalbasis  $\{w^{(1)}, \dots, w^{(N)}\}$  aus Eigenvektoren von  $A$  die Entwicklung

$$y = \sum_{i=1}^N \gamma_i w^{(i)}, \quad \gamma_i = \langle y, w^{(i)} \rangle,$$

und folglich

$$|p(A)y|_A^2 = \sum_{i=1}^N \lambda_i p(\lambda_i)^2 \gamma_i^2 \leq M^2 \sum_{i=1}^N \lambda_i \gamma_i^2 = M^2 |y|_A^2,$$

wobei

$$M := \sup_{\lambda \leq \mu \leq \Lambda} |p(\mu)|, \quad \lambda := \lambda_{\min}(A), \quad \Lambda := \lambda_{\max}(A).$$

Dies impliziert dann

$$|p(A)|_A = \sup_{y \in \mathbb{R}^n, y \neq 0} \frac{|p(A)y|_A}{|y|_A} \leq M.$$

ii) Wir haben gefunden, dass

$$|x^{(t)} - x|_A \leq \min_{p \in P_t, p(0)=1} \left\{ \sup_{\lambda \leq \mu \leq \Lambda} |p(\mu)| \right\} |x^{(0)} - x|_A.$$

Dies ergibt die Behauptung, wenn wir zeigen können, dass

$$\min_{p \in P_t, p(0)=1} \left\{ \sup_{\lambda \leq \mu \leq \Lambda} |p(\mu)| \right\} \leq 2 \left( \frac{1 - \sqrt{\lambda/\Lambda}}{1 + \sqrt{\lambda/\Lambda}} \right)^t.$$

Dabei handelt es sich um ein Problem der Bestapproximation mit Polynomen bzgl. der Maximumsnorm (Tschebyscheff<sup>7</sup>-Approximation). Die Lösung  $\bar{p}$  ist gegeben durch

$$\bar{p}(\mu) = T_t \left( \frac{\Lambda + \lambda - 2\mu}{\Lambda - \lambda} \right) T_t \left( \frac{\Lambda + \lambda}{\Lambda - \lambda} \right)^{-1},$$

mit dem  $t$ -ten Tschebyscheff-Polynom  $T_t$  auf  $[-1, 1]$ . Dabei ist

$$\sup_{\lambda \leq \mu \leq \Lambda} \bar{p}(\mu) = T_t \left( \frac{\Lambda + \lambda}{\Lambda - \lambda} \right)^{-1}.$$

Aus der Darstellung

$$T_t(\mu) = \frac{1}{2} [(\mu + \sqrt{\mu^2 - 1})^t + (\mu - \sqrt{\mu^2 - 1})^t], \quad \mu \in (-\infty, \infty),$$

für die Tschebyscheff-Polynome folgt über die Identität

$$\frac{\kappa + 1}{\kappa - 1} \pm \sqrt{\left( \frac{\kappa + 1}{\kappa - 1} \right)^2 - 1} = \frac{\kappa + 1}{\kappa - 1} \pm \frac{2\sqrt{\kappa}}{\kappa - 1} = \frac{(\sqrt{\kappa} \pm 1)^2}{\kappa - 1} = \frac{\sqrt{\kappa} \pm 1}{\sqrt{\kappa} \mp 1}$$

die Abschätzung nach unten

$$T_t \left( \frac{\Lambda + \lambda}{\Lambda - \lambda} \right) = T_t \left( \frac{\kappa + 1}{\kappa - 1} \right) = \frac{1}{2} \left[ \left( \frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} \right)^t + \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^t \right] \geq \frac{1}{2} \left( \frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} \right)^t.$$

Also wird

$$\sup_{\lambda \leq \mu \leq \Lambda} \bar{p}(\mu) \leq 2 \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^t,$$

was (4.1.17) impliziert.

---

<sup>7</sup>Pafnuty Lvovich Tschebyscheff (russ.: Chebyshev) (1821–1894): Russischer Mathematiker; Prof. in St. Petersburg; Beiträge zur Zahlentheorie, Wahrscheinlichkeitstheorie und vor allem zur Approximationstheorie; entwickelte allgemeine Theorie orthogonaler Polynome.

iii) Zur Herleitung von (4.1.18) fordern wir

$$2 \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{t(\varepsilon)} < \varepsilon \quad \Rightarrow \quad t(\varepsilon) > \ln \left( \frac{2}{\varepsilon} \right) \ln \left( \frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} \right)^{-1}.$$

Wegen

$$\ln \left( \frac{x+1}{x-1} \right) = 2 \left\{ \frac{1}{x} + \frac{1}{3} \frac{1}{x^3} + \frac{1}{5} \frac{1}{x^5} + \dots \right\} \geq \frac{2}{x}$$

ist dies erfüllt für  $t(\varepsilon) \geq \frac{1}{2} \sqrt{\kappa} \ln(2/\varepsilon)$ .

Q.E.D.

#### 4.1.2 CG-Verfahren für unsymmetrische und indefinite Probleme

Zur Lösung allgemeiner Gleichungssysteme  $Ax = b$  mit einer regulären, aber nicht notwendig symmetrisch und positiv definiten Matrix  $A \in \mathbb{R}^n$  mit Hilfe des CG-Verfahrens kann man etwa zu dem äquivalenten System

$$A^T A x = A^T b \quad (4.1.19)$$

mit der positiv definiten Matrix  $A^T A$  übergehen. Hierauf angewendet, schreibt sich das CG-Verfahren wie folgt:

$$\begin{aligned} \text{Startwerte: } \quad & x^{(0)} \in \mathbb{R}^N, \quad d^{(0)} = r^{(0)} = A^T(b - Ax^{(0)}), \\ \text{für } t \geq 0: \quad & \alpha_t = \frac{|r^{(t)}|^2}{|Ad^{(t)}|^2}, \\ & x^{(t+1)} = x^{(t)} + \alpha_t d^{(t)}, \quad r^{(t+1)} = r^{(t)} - \alpha_t A^T Ad^{(t)}, \\ & \beta_t = \frac{|r^{(t+1)}|^2}{|r^{(t)}|^2}, \quad d^{(t+1)} = r^{(t+1)} + \beta_t d^{(t)}. \end{aligned}$$

Die Konvergenzgeschwindigkeit ist dabei charakterisiert durch  $\kappa(A^T A)$ . Das ganze Verfahren beruht offenbar auf der Minimierung des Funktional

$$Q(y) := \frac{1}{2} \langle A^T A y, y \rangle - \langle A^T b, y \rangle = \frac{1}{2} |Ay - b|^2 - \frac{1}{2} |b|^2. \quad (4.1.20)$$

Da  $\kappa(A^T A) \sim \kappa(A)^2$  ist, muss man mit einer recht langsamen Konvergenz dieses „quadruierten“ CG-Verfahrens für nicht symmetrische Systeme rechnen.

Auf der Basis der Charakterisierung (4.1.3) ist das beschriebene CG-Verfahren auf symmetrische, positiv definite Matrizen beschränkt. Geht man allerdings von der „notwendigen Optimalitätsbedingung“ (4.1.8) aus, so ist dieser Ansatz auch für allgemeine Matrizen sinnvoll. Tatsächlich lassen sich auf diesem Wege leistungsfähige Verallgemeinerungen des CG-Verfahrens auch für unsymmetrische und indefinite Matrizen ableiten. Dabei werden in der Galerkin-Formulierung (4.1.8) als Ansatzräume meist wieder die Krylow-Räume

$$K_t = \text{span}\{r^{(0)}, Ar^{(0)}, \dots, A^{t-1}r^{(0)}\}$$

verwendet. Als „Testräume“ treten gleichfalls  $K_t^* = K_t$  oder auch

$$K_t^* = \text{span}\{r^{(0)}, A^T r^{(0)}, \dots, (A^T)^{t-1} r^{(0)}\}$$

auf. Die resultierenden Verfahren GMRES („Generalized Minimal Residual“ von Y. Saad und M. H. Schultz, 1986), ORTHOMIN („Orthogonalization-Minimization“ nach P. K. W. Vinsome, 1976, und Eisenstat et al., 1983), CRS („Conjugate Residual Squared“ nach P. Sonneveld, 1989), BiCGSTAB („Biconjugate Gradient Stabilized“ nach H. A. Van der Vorst, 1992) u.s.w., haben dann jeweils die eine oder die andere Eigenschaft des normalen CG-Verfahrens, lassen aber keine so vollständige Konvergenzanalyse zu.

#### 4.1.3 Vorkonditionierung (PCG-Verfahren)

Die Fehlerabschätzung für das CG-Verfahren garantiert eine besonders gute Konvergenz, wenn die Kondition der Matrix  $A$  nahe bei Eins liegt. Daher wird eine „Vorkonditionierung“ vorgenommen, d.h.: Das System  $Ax = b$  wird in ein äquivalentes umgeformt,  $\tilde{A}\tilde{x} = \tilde{b}$ , dessen Matrix  $\tilde{A}$  besser konditioniert ist. Sei  $C$  eine symmetrische, positiv definite Matrix, welche explizit in Produktform

$$C = KK^T \quad (4.1.21)$$

gegeben ist mit einer regulären Matrix  $K$ . Das System  $Ax = b$  wird dann in der äquivalenten Form geschrieben

$$\underbrace{\tilde{A}}_{\underbrace{K^{-1}A(K^T)^{-1}}_{\tilde{A}}} \underbrace{\tilde{x}}_{\tilde{x}} = \underbrace{K^{-1}b}_{\tilde{b}}. \quad (4.1.22)$$

Das CG-Verfahren wird nun auf das System  $\tilde{A}\tilde{x} = \tilde{b}$  angewendet. Die Beziehung

$$(K^T)^{-1}\tilde{A}K^T = (K^T)^{-1}K^{-1}A(K^T)^{-1}K^T = C^{-1}A \quad (4.1.23)$$

zeigt, dass für  $C \equiv A$  die Matrix  $\tilde{A}$  ähnlich zu  $I$ , d. h.  $\kappa(\tilde{A}) = \kappa(I) = 1$  wäre. Folglich wird man  $C^{-1}$  als möglichst gute Approximation von  $A^{-1}$  wählen, wobei natürlich die Zerlegung  $C = KK^T$  bekannt sein muss. Das CG-Verfahren für das transformierte System  $\tilde{A}\tilde{x} = \tilde{b}$  kann in den ursprünglichen Größen  $A$ ,  $b$  und  $x$  als „PCG-Verfahren“ geschrieben werden:

$$\begin{aligned} \text{Startwert: } & x^{(0)} \in \mathbb{R}^N, \quad d^{(0)} = r^{(0)} = b - Ax^{(0)}, \quad \rho^{(0)} = C^{-1}r^{(0)}, \\ \text{für } t \geq 0: \quad & \alpha_t = \frac{\langle r^{(t)}, \rho^{(t)} \rangle}{\langle Ad^{(t)}, d^{(t)} \rangle}, \\ & x^{(t+1)} = x^{(t)} + \alpha_t d^{(t)}, \quad r^{(t+1)} = r^{(t)} - \alpha_t Ad^{(t)}, \\ & \rho^{(t+1)} = C^{-1}r^{(t+1)}, \\ & \beta_t = \frac{\langle r^{(t+1)}, \rho^{(t+1)} \rangle}{\langle r^{(t)}, \rho^{(t)} \rangle}, \quad d^{(t+1)} = r^{(t+1)} + \beta_t d^{(t)}. \end{aligned}$$

Verglichen mit der einfachen CG-Iteration erfordert das PCG-Verfahren in jedem Schritt zusätzlich die Lösung des Systems  $C\rho^{(t+1)} = r^{(t+1)}$ , was unter Ausnutzung der Zerlegung  $C = KK^T$  erfolgt. Zur Erzielung einer Komplexität von  $\mathcal{O}(N)$  a.Op. pro Schritt sollte die Dreiecksmatrix  $K$  eine ähnliche Besetzungsstruktur wie der untere Dreiecksanteil  $L$  von  $A$  haben. Ausgehend von den oben betrachteten einfachen Fixpunktiterationen werden in der Praxis die folgenden Vorkonditionierer verwendet:

1) *Diagonal-Vorkonditionierung (Skalierung):*  $C = D^{1/2}D^{1/2}$ .

Die Skalierung bewirkt, dass die Elemente von  $A$  auf etwa gleiche Größenordnung gebracht werden, insbesondere wird  $\tilde{a}_{ii} = 1$ . Dies reduziert die Kondition, denn es gilt:

$$\kappa(A) \geq \frac{\max_{1 \leq i \leq N} a_{ii}}{\min_{1 \leq i \leq N} a_{ii}}. \quad (4.1.24)$$

Beispiel: Die Matrix  $A = \text{diag}\{\lambda_1 = \dots = \lambda_{N-1} = 1, \lambda_N = 10^k\}$  hat die Kondition  $\text{cond}_2(A) = 10^k$ . Die skalierte Matrix  $\tilde{A} = D^{-1/2}AD^{-1/2}$  hat dagegen die optimale Kondition  $\text{cond}_2(\tilde{A}) = 1$ .

2) *SSOR-Vorkonditionierung:* Mit einem Parameter  $\omega$  wird gesetzt

$$C = (D + \omega L)D^{-1}(D + \omega R) = \underbrace{(D^{1/2} + \omega LD^{-1/2})}_{K} \underbrace{(D^{1/2} + \omega D^{-1/2}R)}_{K^T}.$$

Offenbar besitzt die Dreiecksmatrix  $K$  dieselbe schwache Besetzung wie  $A$ . Pro Iterationsschritt erfordert das so vorkonditionierte Verfahren etwa doppelt so viel Aufwand wie das einfache Verfahren. Dagegen gilt für die Modellmatrix bei optimaler Wahl des Parameters  $\omega$  (i. Allg. nicht leicht zu bestimmen!)

$$\kappa(\tilde{A}) = \sqrt{\kappa(A)}.$$

3) *ICCG-Verfahren (Incomplete Cholesky Conjugate Gradient):*

Die symmetrische, positiv definite Matrix  $A$  besitzt eine Cholesky-Zerlegung  $A = LL^T$  mit einer unteren Dreiecksmatrix  $L = (l_{ij})_{i,j=1}^N$ . Die Elemente von  $L$  sind bestimmt durch die folgenden Rekursionsformeln:

$$\begin{aligned} l_{ii} &= \left( a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2 \right)^{1/2}, \quad i = 1, \dots, N, \\ l_{ji} &= \frac{1}{l_{ii}} \left( a_{ji} - \sum_{k=1}^{i-1} l_{jk} l_{ik} \right), \quad j = i+1, \dots, N. \end{aligned}$$

Die Matrix  $L$  hat i. Allg. innerhalb der Hülle von  $A$  von Null verschiedene Elemente, erfordert also in der Regel weit mehr Speicherplatz als  $A$  selbst. Dies wird jedoch dadurch ausgeglichen, dass man nur eine "unvollständige Cholesky-Zerlegung" vornimmt, d. h.: Im Cholesky-Algorithmus werden einige der  $l_{ji}$  Null gesetzt, z. B.:  $\tilde{l}_{ji} = 0$ , wenn  $a_{ji} = 0$ .

Dies ergibt dann eine Zerlegung

$$A = \tilde{L}\tilde{L}^T + E \quad (4.1.25)$$

mit einer unteren Dreiecksmatrix  $\tilde{L} = (\tilde{l}_{ij})_{i,j=1,\dots,N}$ , welche eine ähnliche (dünne) Besetzungsstruktur wie  $A$  besitzt. Man spricht von der  $ICCG(0)$ -Variante, wenn (4.1.25) gefordert wird. Werden im Fall einer Bandmatrix  $A$  weitere  $p$  Nebendiagonale mit von Null verschiedenen Elementen in  $\tilde{L}$  hinzugefügt bzw. weggestrichen, so nennt man dies die  $ICCG(+p)$  bzw.  $ICCG(-p)$ -Variante. Zur Vorkonditionierung verwendet die  $ICCG$ -Methode die Matrix

$$C = KK^T = \tilde{L}\tilde{L}^T. \quad (4.1.26)$$

Obwohl keine strenge theoretische Begründung für den Erfolg dieses Ansatzes vorliegt, so zeigen doch numerische Tests an Modellproblemen, welchen Einfluss diese Konditionierung auf die Verteilung der Eigenwerte der Matrix  $\tilde{A}$  hat. Zwar wird die Konditionszahl  $\kappa(\tilde{A})$  nicht deutlich kleiner als  $\kappa(A)$ , doch die Eigenwerte von  $\tilde{A}$  häufen sich im Gegensatz zu denen von  $A$  stark bei  $\lambda = 1$ . Dies bewirkt, wie eine feinere Analyse zeigt, eine deutliche Beschleunigung der Konvergenz.

4) *ADI-Vorkonditionierung:*  $C = (A_x + \omega I)(A_y + \omega I)(A_y^T + \omega I)(A_x^T + \omega I)$ .

Auch hier muss zur Bewahrung der Symmetrie von  $\tilde{A}$  eine symmetrisierte Variante des ADI-Matrix verwendet werden.

Eine gute Vorkonditionierung der Modellmatrix bewirkt eine Verbesserung der Konvergenzrate von  $\rho(A) = 1 - \mathcal{O}(h)$  auf  $\rho(\tilde{A}) = 1 - \mathcal{O}(h^{1/2})$  und damit auf die Lösungskomplexität  $\mathcal{O}(N^{5/4})$ . Besonders die ILU-Vorkonditionierung hat sich in der Praxis bei vielen Problemen als effizient und robust erwiesen. Sie reduziert zwar nicht die Kondition, doch bewirkt eine Konzentration der Eigenwerte um den Wert  $\lambda = 1$ , was ebenfalls eine deutliche Beschleunigung der CG-Konvergenz mit sich bringt.

## 4.2 Mehrgitterverfahren

Mehrgitterverfahren gehören zum Typ der (verallgemeinerten) Defektkorrekturiterationen und verwenden eine Folge von Subproblemen ähnlicher Struktur, aber sukzessive kleiner werdender Dimension. Sie sind speziell zugeschnitten auf die Lösung der Gleichungssysteme, wie sie bei der Diskretisierung partieller Differentialgleichungen mit Differenzen- oder Finite-Elemente-Verfahren entstehen. Die Idee ist die eines allen diskreten Problemen zugrunde liegenden übergeordneten, kontinuierlichen Problems und der fortgesetzten Aufspaltung von Fehlern und Defekten auf den verschiedenen Gittern in „niedrig- und hochfrequente“ Anteile, die separat behandelt werden. Bei richtiger Zusammenstellung der einzelnen Verfahrenskomponenten erhält man Idealfall das gewünschte „optimale“ Lösungsverfahren mit arithmetischem Aufwand  $\mathcal{O}(N)$  für die Berechnung der  $N$  Unbekannten auf dem feinsten Gitter. Die Grundidee des Mehrgitterverfahrens geht auf die

russischen Mathematiker R. Fedorenko<sup>8</sup> und N. Bachwalow<sup>9</sup> in den 1960-er Jahren zurück. Danach wurde der prinzipielle Ansatz in den späten 1970-er Jahren unabhängig voneinander von A. Brandt<sup>10</sup> und W. Hackbusch<sup>11</sup> zu einem allgemein anwendbaren Verfahren entwickelt. Die im Folgenden dargestellte Konvergenz- und Komplexitätsanalyse ist in ihrer rigorosen Form für Differenzenverfahren in Hackbusch [19] beschrieben.

Zum Einstieg betrachten wir das Gleichungssystem auf dem Gitter  $\mathbb{T}_h$ :

$$A_h x_h = b_h \quad (4.2.27)$$

und approximieren die Lösung mit dem Richardson-Verfahren

$$x_h^{(t+1)} = x_h^{(t)} + \theta_h(b_h - A_h x_h^{(t)}) = (I_h - \theta_h A_h)x_h^{(t)} + \theta_h b_h \quad (4.2.28)$$

mit einem Dämpfungsfaktor  $0 < \theta_h \leq 1$ . Die symmetrische, positiv definite Matrix  $A_h$  besitzt ein Orthonormalsystem von Eigenvektoren  $\{w_h^{(i)}, i = 1, \dots, N_h\}$  zu den geordneten Eigenwerten  $\lambda_{\min}(A_h) = \lambda_1 \leq \dots \leq \lambda_N = \lambda_{\max}(A_h) =: \Lambda_h$ . Entwickelt man den Anfangsfehler in der Form

$$e_h^{(0)} := x_h^{(0)} - x_h = \sum_{i=1}^{N_h} \varepsilon_i w_h^{(i)},$$

so gilt für die iterierten Fehler entsprechend

$$e_h^{(t)} = (I_h - \theta_h A_h)^t e_h^{(0)} = \sum_{i=1}^{N_h} \varepsilon_i (I_h - \theta_h A_h)^t w_h^{(i)} = \sum_{i=1}^{N_h} \varepsilon_i (1 - \theta_h \lambda_i)^t w_h^{(i)}.$$

---

<sup>8</sup>Radi Petrowitsch Fedorenko (1930–2009): Russischer Mathematiker; arbeitete ab 1953 am Keldysh-Institut für Angewandte Mathematik der Russischen Akademie der Wissenschaften; numerische Berechnungen für das sowjetische Kernwaffen- und Kerntechnikprojekt sowie für Luft- und Raumfahrt; erste Arbeiten unterlagen der Geheimhaltung, erste Veröffentlichung 1958 über ein Problem der Magnetohydrodynamik; Pionier der Mehrgittermethode mit Arbeiten Anfang der 1960-er Jahre in Zusammenhang mit der numerischen Lösung der Poisson-Gleichung in der Wettervorhersage.

<sup>9</sup>Nikolai Sergejewitsch Bachwalow (1934–2005): Russischer Mathematiker; Arbeiten zur Numerik; Promotion 1958 in Moskau bei A. Kolmogorow; ab 1966 Prof. an der Lomonossow-Universität und 1981 Abteilung Numerische Mathematik; Pionier der Mehrgitterverfahren und Beiträge zur Numerik von Wellenphänomenen und Verbundmaterialien (Methode der Homogenisierung und der „Fictitious Domain“-Methode); Autor verbreiteter russischer Lehrbücher der Numerik.

<sup>10</sup>Achi Brandt (1938–): Israelischer Mathematiker; Arbeiten über partielle Differentialgleichungen und Numerik; Prof. am Weizmann-Institut in Rechovot (Israel) und an der Univ. of California (Los Angeles, USA); einer der Pioniere der Mehrgittermethode (sog. „Full Approximation Scheme“, FAS, 1977); behauptet, jede partielle Differentialgleichung sei durch Mehrgitterverfahren effizient und schnell lösbar; Mitgründer der Softwarefirma VideoSurf.

<sup>11</sup>Wolfgang Hackbusch (1948–): Deutscher Mathematiker; Studium in Marburg; Promotion 1973 und Habilitation 1979 in Köln R. Bulirsch; Professuren für Praktische Mathematik in Bochum und Kiel; 1990/2000–2014 Direktor am MPI für Mathematik in den Naturwissenschaften in Leipzig; wichtige Beiträge zur Numerik von partiellen Differentialgleichungen und Integralgleichungen; am besten bekannt durch seine Arbeiten zur „Mehrgittermethode“, dem sog. „Panel Clustering“ und der „H-Matrizen“.

Folglich ist

$$|e_h^{(t)}|^2 = \sum_{i=1}^{N_h} \varepsilon_i^2 (1 - \theta_h \lambda_i)^{2t}. \quad (4.2.29)$$

Die Bedingung  $0 < \theta_h \leq \Lambda_h^{-1}$  ist hinreichend für die Konvergenz der Richardson-Iteration. Wegen  $|1 - \theta_h \lambda_i| \ll 1$  für große  $\lambda_i$  und  $|1 - \theta_h \lambda_1| \approx 1$  werden offenbar „hoch-frequente“ Komponenten des Fehlers sehr schnell, aber „niedrig-frequente“ nur sehr langsam gedämpft. Dasselbe gilt auch für das Residuum  $r_h^{(t)} = b_h - A_h x_h^{(t)} = A_h e_h^{(t)}$ , d.h.: Bereits nach wenigen Iterationen gilt:

$$|r_h^{(t)}|^2 \approx \sum_{i=1}^{[N/2]} \varepsilon_i^2 \lambda_i^2 (1 - \theta_h \lambda_i)^{2t}, \quad (4.2.30)$$

wobei  $[N/2] := \max\{n \in \mathbb{N} \mid n \leq N/2\}$  ist. Dies kann so interpretiert werden, dass der iterierte Defekt  $r_h^{(t)}$  auf dem Gitter  $T_h$  glatt ist. Daher sollte er auf einem größeren Gitter  $T_{2h}$  mit Gitterweite  $2h$  gut approximierbar sein. Die resultierende Defektgleichung zur Berechnung der Korrektur zur Näherung  $x_h^{(t)}$  auf  $T_h$  würde dann wegen ihrer geringeren Dimension  $N_{2h} \approx N_h/4$  auch weniger Aufwand kosten. Dieser Defektkorrekturprozess in Verbindung mit sukzessiver Vergrößerung kann weitergeführt werden bis zu einem grössten Gitter, auf dem die Defektgleichung dann exakt gelöst wird. Die wichtigsten Bestandteile eines solchen Mehrgitterprozesses sind die „Glättungsiteration“  $x_h^{(\nu)} = S_h^\nu(x_h^{(0)})$  sowie geeignete Transferoperationen zwischen den Finite-Elemente-Räumen auf den verschiedenen Gittern. Die Glättungsoperation  $S_h(\cdot)$  ist gewöhnlich gegeben in Form einer Fixpunktiteration (z.B. der Richardson-Iteration)

$$x_h^{(\nu+1)} = S_h(x_h^{(\nu)}) := (I_h - C_h^{-1} A_h)x_h^{(\nu)} + C_h^{-1} b_h,$$

mit der Iterationsmatrix  $S_h := I_h - C_h^{-1} A_h$ .

#### 4.2.1 Mehrgitteralgorithmus im Finite-Elemente-Kontext

Zur Formalisierung des Mehrgitterprozesses betrachten wir nun eine Folge von Gittern  $T_l = T_{h_l}$ ,  $l = 0, \dots, L$ , zunehmender Feinheit  $h_0 > \dots > h_l > \dots > h_L$  sowie zugehörige FE-Räume  $V_l := V_{h_l} \subset V$ . Der Einfachheit halber sei angenommen, dass die FE-Räume hierarchisch geordnet sind, d.h.:  $V_0 \subset V_1 \subset \dots \subset V_l \subset \dots \subset V_L$ . Diese Voraussetzung erleichtert die Analyse des Mehrgitterprozesses, ist aber nicht entscheidend für sein Funktionieren. Zwischen den Funktionen  $v_l \in V_l$  und den zugehörigen Knotenwertvektoren  $y_l \in \mathbb{R}^{N_l}$  gilt der übliche Zusammenhang  $v_l(a_n) = y_{l,n}$ ,  $n = 1, \dots, N_l$ . Wie üblich schreiben wir das kontinuierliche Problem und sein FE-Analogon in variationeller Form als

$$a(u, \varphi) = (f, \varphi) \quad \forall \varphi \in V, \quad (4.2.31)$$

bzw. auf dem feinsten Gitter  $\mathbb{T}_L$  als

$$a(u_L, \varphi_L) = (f, \varphi_L) \quad \forall \varphi_L \in V_L. \quad (4.2.32)$$

Dabei sind  $a(u, \varphi) := (Lu, \varphi)$  die zum (elliptischen) Operator  $L$  gehörende „Energieform“ und  $(f, \varphi)$  das  $L^2$ -Skalarprodukt auf dem Lösungsgebiet  $\Omega$ . Die „exakte“ diskrete Lösung  $u_L \in V_L$  genügt der a priori Fehlerabschätzung

$$\|u - u_L\| \leq c h_L^2 \|f\|. \quad (4.2.33)$$

Ziel ist es, einen Lösungsprozess zu finden, der eine Approximation  $\bar{u}_L \approx u_L$  liefert mit

$$\|u_L - \bar{u}_L\| \leq c h_L^2 \|f\|. \quad (4.2.34)$$

Ist der dazu erforderliche Aufwand  $\mathcal{O}(N_L)$  und zwar gleichmäßig bzgl.  $L$ , so nennt man diesen Prozess „komplexitäts-optimal“. Wir werden sehen, dass der Mehrgitteralgorithmus bei richtiger Wahl der Verfahrenskomponenten in diesem Sinne „optimal“ ist.

Sei  $u_L^{(0)} \in V_L$  eine Schätzung für die exakte Lösung  $u_L \in V_L$  auf Gitterlevel  $L$ . Zunächst wird  $u_L^{(0)}$  „geglättet“. Dazu werden ausgehend von  $\bar{u}_L^{(0)} := u_L^{(0)}$  z. B.  $\nu$  Schritte des Richardson-Verfahrens durchgeführt. In variationeller Schreibweise lautet dies:

$$(\bar{u}_L^{(k)}, \varphi_L) = (\bar{u}_L^{(k-1)}, \varphi_L) + \theta_L \{(f, \varphi_L) - a(\bar{u}_L^{(k-1)}, \varphi_L)\} \quad \forall \varphi_L \in V_L, \quad (4.2.35)$$

wobei  $\theta_L = \lambda_{max}(A_h)^{-1}$ . Mit der geglätteten Approximation wird der Defekt  $d_L \in V_L$  gebildet (ohne ihn wirklich zu berechnen):

$$(d_L, \varphi_L) := (f, \varphi_L) - a(\bar{u}_L^{(\nu)}, \varphi_L), \quad \varphi_L \in V_L. \quad (4.2.36)$$

Wegen  $V_{L-1} \subset V_L$  erhält man auf dem nächst gröberen Gitter  $\mathbb{T}_{L-1}$  die Defektgleichung („Grobgittergleichung“)

$$a(q_{L-1}, \varphi_{L-1}) = (d_L, \varphi_{L-1}) = (f, \varphi_{L-1}) - a(\bar{u}_L^{(\nu)}, \varphi_{L-1}) \quad \forall \varphi_{L-1}. \quad (4.2.37)$$

Die Korrektur  $q_{L-1} \in V_{L-1}$  wird nun entweder exakt berechnet (etwa mit einem „direkten“ Löser) oder nur näherungsweise mit Hilfe einer Defektkorrekturiteration  $q_{L-1}^{(0)} \rightarrow q_{L-1}^{(R)}$  unter Verwendung der noch gröberen Gitter  $\mathbb{T}_{L-2}, \dots, \mathbb{T}_0$ . Das Ergebnis  $q_{L-1}^{(R)} \in V_{L-1}$  wird dann als Element von  $V_L$  interpretiert und zur Korrektur von  $\bar{u}_L^{(\nu)}$  verwendet:

$$\bar{u}_L^{(0)} := \bar{u}_L^{(\nu)} + \omega_L q_{L-1}^{(R)}. \quad (4.2.38)$$

Dabei wird der Dämpfungsparameter  $\omega_L \in (0, 1)$  verwendet, um das Residuum von  $\bar{u}_L$  zu minimieren. Auf diesen in der Praxis sehr nützlichen Trick wollen wir hier nicht weiter eingehen. Die erhaltene korrigierte Näherung  $\bar{u}_L$  wird nun nochmals  $\mu$ -mal „nachgeglättet“. Ausgehend von  $\bar{u}_L^{(0)} := \bar{u}_L$  wird etwa wieder mit dem Richardson-Verfahren iteriert:

$$(\bar{u}_L^{(k)}, \varphi_L) = (\bar{u}_L^{(k-1)}, \varphi_L) + \theta_L \{(f, \varphi_L) - a(\bar{u}_L^{(k-1)}, \varphi_L)\} \quad \forall \varphi_L \in V_L. \quad (4.2.39)$$

Das Ergebnis wird schließlich als die nächste Mehrgitteriterierte  $u_L^{(1)} := \bar{u}_L^{(\mu)}$  akzeptiert. Damit haben wir einen Schritt des Mehrgitterverfahrens (einen „Zyklus“) auf dem Gitterlevel  $L$  beschrieben. Jeder solche Zyklus beinhaltet also neben  $\nu + \mu$  Richardson-Schritten (auf Level  $L$ ), welche jeweils eine Inversion der Massematrizen erfordern, die Lösung des „Grobgitterproblems“ (4.2.37).

Wir wollen nun den beschriebenen Mehrgitteralgorithmus in etwas abstrakterer Form darstellen, um seine Struktur besser zu verstehen und ihn auch leichter analysieren zu können. Zu den Matrizen  $A_l = A_{h_l}$  auf den Gittern  $T_l$  sind Operatoren  $\mathcal{A}_l : V_l \rightarrow V_l$  assoziiert durch

$$(\mathcal{A}_l v_l, w_l) = a(v_l, w_l) = (A_l y_l, z_l) \quad \forall v_l, w_l \in V_l. \quad (4.2.40)$$

Weiter seien  $\mathcal{S}_l(\cdot)$  die zugehörigen Glättungsoperationen mit (linearen) Iterationsoperatoren  $\mathcal{S}_l$ . Beim Richardson-Verfahren ist der Iterationsoperator  $\mathcal{S}_l = \mathcal{I}_l - \theta_l \mathcal{A}_l$ . Schließlich führen wir noch Transferoperatoren zwischen aufeinander folgenden Räumen ein:

$$r_l^{l-1} : V_l \rightarrow V_{l-1} \text{ (Restriktion)}, \quad p_l^l : V_{l-1} \rightarrow V_l \text{ (Prolongation)}.$$

Im Finite-Elemente-Kontext ist natürlicherweise  $r_l^{l-1} = P_{l-1}$  die  $L^2$ -Projektion und  $p_l^l = id$ . die natürliche Einbettung. Wir beschreiben nun den Mehrgitterprozeß zur Berechnung der Lösung des Systems

$$\mathcal{A}_L u_L = f_L \quad (4.2.41)$$

auf dem „feinsten“ Gitter  $T_L$ .

*Mehrgitterprozess:* Ausgehend von einem Startwert  $u_L^{(0)} \in V_L$  werden Iterierte  $u_L^{(t)}$  durch den folgenden rekursiven Prozess

$$u_L^{(t+1)} = MG(L, u_L^{(t)}, f_L) \quad (4.2.42)$$

erzeugt. Sei also die  $t$ -te Mehrgitteriterierte  $u_L^{(t)}$  bestimmt.

*Grobgitterlösung:* Für  $l = 0$  bedeutet  $MG(0, \cdot, g_0)$  stets die exakte Lösung des Systems  $\mathcal{A}_0 v_0 = g_0$  (z.B. mit Hilfe eines direkten Lösungsverfahrens), d. h.:

$$v_0 = MG(0, \cdot, g_0) = \mathcal{A}_0^{-1} g_0. \quad (4.2.43)$$

*Rekursion:* Sei für ein  $1 \leq l \leq L$  das System  $\mathcal{A}_l v_l = g_l$  zu lösen. Mit Parameterwerten  $\nu, \mu \geq 1$  ist dann

$$MG(l, v_l^{(0)}, g_l) := v_l^{(1)} \approx v_l \quad (4.2.44)$$

rekursiv definiert durch die folgenden Schritte:

i) *Vorglättung:*

$$\bar{v}_l := \mathcal{S}_l^\nu(v_l^{(0)}); \quad (4.2.45)$$

ii) *Defektbildung:*

$$d_l := g_l - \mathcal{A}_l \bar{v}_l; \quad (4.2.46)$$

iii) *Restriktion:*

$$\tilde{d}_{l-1} := r_l^{l-1} d_l; \quad (4.2.47)$$

iv) *Defektgleichung:* Ausgehend von  $q_{l-1}^{(0)} := 0$  wird für  $1 \leq r \leq R$  iteriert:

$$q_{l-1}^{(r)} := MG(l-1, q_{l-1}^{(r-1)}, \tilde{d}_{l-1}); \quad (4.2.48)$$

v) *Prolongation:*

$$q_l := p_{l-1}^l q_{l-1}^{(R)}; \quad (4.2.49)$$

vi) *Korrektur:* Mit einem Dämpfungsparameter  $\omega_l \in (0, 1]$  wird gesetzt:

$$\bar{v}_l := \bar{v}_l + \omega_l q_l; \quad (4.2.50)$$

vii) *Nachglättung:*

$$v_l^{(1)} := S_l^\mu(\bar{v}_l); \quad (4.2.51)$$

Im Falle  $l = L$  wird schließlich gesetzt:

$$u_L^{(t+1)} := v_L^{(1)}. \quad (4.2.52)$$

Schematische Darstellung des Mehrgitterschritts  $u_L^{(t)} \rightarrow u_L^{(t+1)}$ :

$$\begin{aligned} u_L^{(t)} \rightarrow \bar{u}_L^{(t)} &= S_L^\nu(u_L^{(t)}) \rightarrow d_L = f_L - \mathcal{A}_L \bar{u}_L^{(t)} \\ &\downarrow \quad \tilde{d}_{L-1} = r_L^{L-1} d_L \quad (\text{Restriktion}) \\ q_{L-1} &= \tilde{\mathcal{A}}_{L-1}^{-1} \tilde{d}_{L-1} \quad (R\text{-malige Defektkorrektur}) \\ &\downarrow \quad \tilde{q}_L = p_{L-1}^L q_{L-1} \quad (\text{Prolongation}) \\ \bar{u}_L^{(t)} &= \bar{u}_L^{(t)} + \omega_L \tilde{q}_L \rightarrow u_L^{(t+1)} = S_L^\mu(\bar{u}_L^{(t)}) \end{aligned}$$

Wenn die Defektgleichung  $\mathcal{A}_{L-1} q_{L-1} = \tilde{d}_{L-1}$  auf dem größeren Gitter  $T_{L-1}$  „exakt“ gelöst wird (z. B. durch Gauß-Elimination), spricht man von einer „Zweigittermethode“. In der Regel wird der Prozess aber rekursiv zum Mehrgitterverfahren bis zum grössten Gitter fortgesetzt. Dabei kann der vollständige Mehrgitterzyklus auf verschiedene Art organisiert werden. Seine Struktur ist im wesentlichen durch den Parameter  $R$  bestimmt, der festlegt, wie oft der Defektkorrekturprozess auf jedem Gitterlevel durchgeführt wird. In der Praxis spielen nur die Fälle  $R = 1$  oder  $R = 2$  eine Rolle. Dem entsprechen der im schematischen Bild gezeigte sog. „V-Zyklus“ und der sog. „W-Zyklus“. Dabei stehen die Punkte „•“ für Glättung und Defektkorrektur auf den Gittern  $T_l$ , und die Linie „—“ für den Transfer zwischen aufeinander folgenden Gitterniveaus.

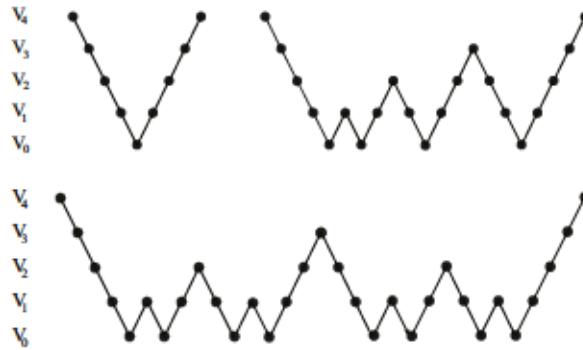


Abbildung 4.2: Schema eines Mehrgitterverfahrens mit V- (oben links) F- (oben rechts) und W-Zyklus (unten).

Der V-Zyklus ist sehr effizient (wenn er funktioniert), krankt aber oft an Instabilitäten, welche durch Irregularitäten im Problem (starke Unsymmetrien, Eckensingularitäten, Gitterunregelmäßigkeiten u.s.w.) hervorgerufen werden. Der W-Zyklus ist dagegen sehr robust, aber auch teurer. Methoden mit  $R \geq 3$  sind gewöhnlich zu ineffizient. Ein guter Kompromiss zwischen V-Zyklus und W-Zyklus stellt der sog. „F-Zyklus“ dar. Dieser wird gewöhnlich auf Gitter  $T_L$  gestartet mit einem beliebigen Startvektor  $u_L^{(0)}$  (meist  $u_L^{(0)} = 0$ ). Wird dieser Prozess allerdings zur Lösung eines nicht linearen Problems im Rahmen einer Newton-Iteration angewendet, so kann dieser Startwert zu ungenau sein, und die ganze Iteration divergiert. In einem solchen Fall startet man zur Generierung einer hinreichend genauen Anfangsapproximation den Mehrgitterprozess gewöhnlich auf dem grössten Gitter  $T_0$ . Wir beschreiben dieses sog. „geschachtelte“ Mehrgitter-Schema („nested multigrid“) für das *lineare* Problem.

*Geschachteltes MG:* Ausgehend von dem Startwert  $u_0 := \mathcal{A}_0^{-1} f_0$  auf dem grössten Gitter  $T_0$  werden für  $l = 1, \dots, L$  rekursiv Näherungen  $\tilde{u}_l \approx u_l$  berechnet nach der Vorschrift:

$$\begin{aligned} u_l^{(0)} &= p_{l-1}^l \tilde{u}_{l-1} \\ u_l^{(t)} &= MG(l, u_l^{(t-1)}, f_l), \quad t = 1, \dots, t_l, \quad \|u_l^{(t)} - u_l\| \leq \hat{c} h_l^2 \|f\|, \\ \tilde{u}_l &= u_l^{(t_l)}. \end{aligned}$$

Es gibt nicht „den Mehrgitteralgorithmus“. Die erfolgreiche Realisierung des Mehrgitterkonzepts erfordert eine sorgfältige Balance der verschiedenen Bestandteile wie Glätter  $S_l$  und Gitteroperatoren  $\mathcal{A}_l$  sowie der Gittertransfers  $r_l^{l-1}$  und  $p_{l-1}^l$  jeweils für das zu lösende Problem. Im folgenden werden wir diese Verfahrenskomponenten im Rahmen des Finite-Elemente-Kontexts diskutieren.

i) *Glätter:* „Glätter“ sind üblicherweise einfache Fixpunktiterationen, die auch als „Löser“ verwendet werden können, aber mit einer sehr schlechten Konvergenzrate. Sie werden auf jedem Gitterniveau nur ein paarmal angewendet ( $\nu, \mu \sim 1 - 4$ ), um die hochfrequenten Fehleranteile auszudämpfen. Wir betrachten im folgenden nur das klassische Richardson-

Verfahren,

$$\mathcal{S}_l := \mathcal{I}_l - \theta_l \mathcal{A}_l, \quad \theta_l = \lambda_{\max}(\mathcal{A}_l)^{-1}, \quad (4.2.53)$$

welches aber nur bei sehr „gutartigen“ Problemen funktioniert. Leistungsfähiger und robuster sind das Gauß-Seidel- und das ILU-Verfahren. Diese funktionieren auch noch gut, wenn das Problem gewisse Pathologien beinhaltet. Im Fall eines starken Advektionsterms besitzt die Systemmatrix bei Numerierung der Knotenpunkte in Transportrichtung einen dominanten unteren Dreicksanteil  $L$ , für den die Gauß-Seidel-Methode „exakt“ ist. Für Probleme mit degenerierten Koeffizienten in einer Raumrichtung sowie auf stark anisotropen Gittern besitzt die Systemmatrix einen dominaten Tridiagonalanteil, für den wiederum die ILU-Iteration „exakt“ ist. Für echt indefinite Probleme werden spezielle, der jeweiligen Struktur des Problems angepasste Glätter verwendet, deren Diskussion aber außerhalb des Rahmens dieses einführenden Textes liegt. Auf lokal verfeinerten Gittern darf die Glättung im Wesentlichen nur auf den jeweils neu hinzugekommenen Zellen operieren, da sonst der arithmetische Aufwand pro Gitterlevel zu groß wird.

ii) *Gittertransfers:* Im Kontext einer Finite-Elemente-Diskretisierung mit geschachtelten Ansatzräumen  $V_0 \subset V_1 \subset \dots \subset V_l \subset \dots \subset V_L$  ist die generische Wahl für die Prolongation  $p_{l-1}^l : V_{l-1} \rightarrow V_l$  die zellweise Einbettung und für die Restriktion  $r_l^{l-1} : V_l \rightarrow V_{l-1}$  die  $L^2$ -Projektion. Bei anderen Diskretisierungen (z. B. Differenzenschemata) verwendet man geeignete Interpolationsprozesse (z. B. bilineare Interpolation).

iii) *Grobgitteroperatoren:* Die Operatoren  $\mathcal{A}_l$  auf den verschiedenen Gitterniveaus müssen nicht notwendig zur selben Diskretisierung des Ausgangsproblems gehören. Dies wird z. B. wichtig bei der Berücksichtigung von gitterweitenabhängiger künstlicher Diffusion („upwinding“) zur Behandlung von Transporttermen. Wir beschränken uns hier aber auf den Idealfall, dass alle  $\mathcal{A}_l$  durch dieselben FE-Diskretisierungen auf der Gitterhierarchie  $\{\mathbb{T}_l\}_{l=0,\dots,L}$  erzeugt sind. In diesem Fall gilt die für die theoretische Analyse nützliche Beziehung

$$\begin{aligned} (\mathcal{A}_{l-1} v_{l-1}, w_{l-1}) &= a(v_{l-1}, w_{l-1}) \\ &= a(p_{l-1}^l v_{l-1}, p_{l-1}^l w_{l-1}) \\ &= (\mathcal{A}_l p_{l-1}^l v_{l-1}, p_{l-1}^l w_{l-1}) = (r_l^{l-1} \mathcal{A}_l p_{l-1}^l v_{l-1}, w_{l-1}), \end{aligned}$$

d. h.:  $\mathcal{A}_{l-1} = r_l^{l-1} \mathcal{A}_l p_{l-1}^l$ .

iv) *Korrekturschritt:* Im Korrekturschritt wird ein Dämpfungsparameter  $\omega_l \in (0, 1]$  verwendet, der im einfachsten Fall  $\omega_l = 1$  gesetzt ist. Es hat sich als sehr wirksam erwiesen, ihn so zu wählen, dass der Defekt  $\mathcal{A}_l \bar{v}_l - \bar{d}_{l-1}$  minimal wird. Dies führt auf die Formel

$$\omega_l = \frac{(\mathcal{A}_l \bar{v}_l, \bar{d}_{l-1} - \mathcal{A}_l \bar{v}_l)}{\|\mathcal{A}_l \bar{v}_l\|^2}. \quad (4.2.54)$$

In der folgenden Analyse werden wir der Einfachheit halber stets  $\omega_l = 1$  setzen.

### 4.2.2 Konvergenz- und Aufwandsanalyse

Die klassische Analyse des Mehrgitteralgorithmus basiert auf seiner Interpretation als eine Defektkorrekturiteration und dem Konzept einer rekursiven Anwendung des Zweigitterverfahrens. Zur Vereinfachung nehmen wir an, dass nur Vorglättung angewendet wird (d. h.:  $\nu > 0$ ,  $\mu = 0$ ) und dass im Korrekturschritt keine Dämpfung erfolgt (d. h.:  $\omega_l = 1$ ). Der Zweigitterprozess lässt sich dann in der folgenden Form schreiben:

$$\begin{aligned} u_L^{(t+1)} &= S_L^\nu(u_L^{(t)}) + p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1} (f_L - \mathcal{A}_L S_L^\nu(u_L^{(t)})) \\ &= S_L^\nu(u_L^{(t)}) + p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L (u_L - S_L^\nu(u_L^{(t)})). \end{aligned}$$

Für den Iterationsfehler  $e_L^{(t)} := u_L^{(t)} - u_L$  gilt daher

$$e_L^{(t+1)} = (\mathcal{I}_L - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L) (S_L^\nu(u_L^{(t)}) - u_L). \quad (4.2.55)$$

Die Glättungsoperation ist gegeben in der (affin-linearen) Form

$$S_L(v_L) := S_L v_L + g_L$$

und erfüllt als Fixpunktiteration die Bedingung  $S_L(u_L) = u_L$ . Daraus erschließt man rekursiv, dass

$$S_L^\nu(u_L^{(t)}) - u_L = S_L(S_L^{\nu-1}(u_L^{(t)}) - u_L) = \dots = S_L^\nu e_L^{(t)}.$$

Mit dem sog. „Zweigitteroperator“

$$ZG_L(\nu) := (\mathcal{I}_L - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L) S_L^\nu$$

gilt daher

$$e_L^{(t+1)} = ZG_L(\nu) e_L^{(t)}. \quad (4.2.56)$$

**Satz 4.2 (Zweigitterkonvergenz):** Für hinreichend häufige Glättung,  $\nu > 0$ , ist der Zweigitteralgorithmus konvergent mit einer bzgl.  $L$  gleichmäßigen  $L^2$ -Konvergenzrate:

$$\|ZG_L(\nu)\| \leq \rho_{ZG}(\nu) = c \nu^{-1} < 1. \quad (4.2.57)$$

**Beweis:** Wir schreiben

$$ZG_L(\nu) = (\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}) \mathcal{A}_L S_L^\nu \quad (4.2.58)$$

und schätzen ab:

$$\|ZG_L(\nu)\| \leq \|\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}\| \|\mathcal{A}_L S_L^\nu\|. \quad (4.2.59)$$

Der erste Term rechts beschreibt die Qualität der Approximation der Feingitterlösung auf dem größeren Gitter, während der zweite Term den Glättungseffekt enthält. Die Idee für die weitere Analyse ist nun, zu zeigen, dass der Glätter  $S_L(\cdot)$  die sog. „Glättungseigen-

schaft“,

$$\|\mathcal{A}_L \mathcal{S}_L^\nu v_L\| \leq c_s \nu^{-1} h_L^{-2} \|v_L\|, \quad v_L \in V_L, \quad (4.2.60)$$

und die Grobgitterkorrektur die sog. „Approximationseigenschaft“ besitzt,

$$\|(\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}) v_L\| \leq c_a h_L^2 \|v_L\|, \quad v_L \in V_L, \quad (4.2.61)$$

mit positiven Konstanten  $c_s, c_a$  gleichmäßig bzgl.  $L$ . Kombination dieser beiden Abschätzungen ergibt dann die behauptete Ungleichung (4.2.57). Für hinreichend häufige Glättung ist  $\rho_{ZG} := c\nu^{-1} < 1$ , und der Zweigitteralgorithmus konvergiert gleichmäßig bzgl.  $L$ . Alle im Folgenden auftretenden Konstanten sind unabhängig von  $L$ .

i) *Glättungseigenschaft:* Der selbstadjungierte Operator  $\mathcal{A}_L$  besitzt reelle, positive Eigenwerte  $0 < \lambda_1 \leq \dots \leq \lambda_i \leq \dots \leq \lambda_{N_L} =: \Lambda_L$  mit einem zugehörigen  $L^2$ -Orthonormalsystem von Eigenfunktionen  $\{w^{(1)}, \dots, w^{(N_L)}\}$ , so dass sich jedes  $v_L \in V_L$  in der Form

$$v_L = \sum_{i=1}^{N_L} \gamma_i w^{(i)}, \quad \gamma_i = (v_L, w^{(i)}) \quad (4.2.62)$$

darstellen lässt. Für den Richardson-Iterationsoperator

$$\mathcal{S}_L := \mathcal{I}_L - \theta_L \mathcal{A}_L : V_L \rightarrow V_L, \quad \theta_L = \Lambda_L^{-1}, \quad (4.2.63)$$

gilt dann

$$\mathcal{A}_L \mathcal{S}_L^\nu v_L = \sum_{i=1}^{N_L} \gamma_i \lambda_i \left(1 - \frac{\lambda_i}{\Lambda_L}\right)^\nu w^{(i)}, \quad (4.2.64)$$

und folglich:

$$\begin{aligned} \|\mathcal{A}_L \mathcal{S}_L^\nu v_L\|^2 &= \sum_{i=1}^{N_L} \gamma_i^2 \lambda_i^2 \left(1 - \frac{\lambda_i}{\Lambda_L}\right)^{2\nu} \\ &\leq \Lambda_L^2 \max_{1 \leq i \leq N_L} \left\{ \left(\frac{\lambda_i}{\Lambda_L}\right)^2 \left(1 - \frac{\lambda_i}{\Lambda_L}\right)^{2\nu} \right\} \sum_{i=1}^{N_L} \gamma_i^2 \\ &= \Lambda_L^2 \max_{1 \leq i \leq N_L} \left\{ \left(\frac{\lambda_i}{\Lambda_L}\right)^2 \left(1 - \frac{\lambda_i}{\Lambda_L}\right)^{2\nu} \right\} \|v_L\|^2. \end{aligned}$$

Mit Hilfe der Beziehung (Übungsaufgabe)

$$\max_{0 \leq x \leq 1} \{x^2(1-x)^{2\nu}\} \leq (1+\nu)^{-2} \quad (4.2.65)$$

ergibt sich

$$\|\mathcal{A}_L \mathcal{S}_L^\nu v_L\|^2 \leq \Lambda_L^2 (1+\nu)^{-2} \|v_L\|^2. \quad (4.2.66)$$

Die Beziehung  $\Lambda_L \leq ch_L^{-2}$  liefert dann schließlich die behauptete Ungleichung für den Richardson-Iterationsoperator

$$\|\mathcal{A}_L S_L^\nu\| \leq c_s \nu^{-1} h_L^{-2}, \quad \nu \geq 1. \quad (4.2.67)$$

ii) *Approximationseigenschaft:* Wir erinnern daran, dass im vorliegenden Kontext geschachtelter FE-Räume Prolongationen und Restriktionen gegeben sind durch

$$p_{L-1}^L = \text{id. (Identität)}, \quad r_L^{L-1} = P_{L-1} \text{ (} L^2\text{-Projektion).}$$

Ferner erfüllt der Operator  $\mathcal{A}_L : V_L \rightarrow V_L$  definitionsgemäß

$$(\mathcal{A}_L v_L, \varphi_L) = a(v_L, \varphi_L), \quad v_L, \varphi_L \in V_L.$$

Für ein beliebiges, aber fest gewähltes  $f_L \in V_L$  gilt demnach für die Funktionen  $v_L := \mathcal{A}_L^{-1} f_L$  und  $v_{L-1} := \mathcal{A}_{L-1}^{-1} r_L^{L-1} f_L$ :

$$\begin{aligned} a(v_L, \varphi_L) &= (f_L, \varphi_L) \quad \forall \varphi_L \in V_L, \\ a(v_{L-1}, \varphi_{L-1}) &= (f_L, \varphi_{L-1}) \quad \forall \varphi_{L-1} \in V_{L-1}. \end{aligned}$$

Der Funktion  $v_L \in V_L$  ordnen wir eine Funktion  $v \in V \cap H^2(\Omega)$  zu als Lösung der Randwertaufgabe

$$Lv = f_L \text{ in } \Omega, \quad v = 0 \text{ auf } \partial\Omega, \quad (4.2.68)$$

bzw. in „schwacher“ Formulierung

$$a(v, \varphi) = (f_L, \varphi) \quad \forall \varphi \in V. \quad (4.2.69)$$

Dafür gilt die a priori Abschätzung

$$\|\nabla^2 v\| \leq c \|f_L\|. \quad (4.2.70)$$

Dann ist

$$\begin{aligned} a(v_L, \varphi_L) &= (f_L, \varphi_L) = a(v, \varphi_L), \quad \varphi_L \in V_L, \\ a(v_{L-1}, \varphi_{L-1}) &= (f_L, \varphi_{L-1}) = a(v, \varphi_{L-1}), \quad \varphi_{L-1} \in V_{L-1}, \end{aligned}$$

d. h.:  $v_L$  und  $v_{L-1}$  sind gerade die Ritz-Projektionen von  $v$  auf  $V_L$  bzw.  $V_{L-1}$ . Für diese gelten die  $L^2$ -Fehlerabschätzungen

$$\|v_L - v\| \leq ch_L^2 \|\nabla^2 v\|, \quad \|v_{L-1} - v\| \leq ch_{L-1}^2 \|\nabla^2 v\|. \quad (4.2.71)$$

Damit erhalten wir wegen  $h_{L-1} \leq 4h_L$  und der a priori Abschätzung (4.2.70):

$$\|v_L - v_{L-1}\| \leq ch_L^2 \|\nabla^2 v\| \leq ch_L^2 \|f_L\|. \quad (4.2.72)$$

Dies bedeutet mit der obigen Setzung, dass

$$\|\mathcal{A}_L^{-1}f_L - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1} f_L\| \leq ch_L^2 \|f_L\|. \quad (4.2.73)$$

Damit folgt die gewünschte Abschätzung

$$\|\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}\| \leq ch_L^2. \quad (4.2.74)$$

Dies vervollständigt den Beweis. Q.E.D.

Das Resultat für den Zweigitteralgorithmus wird nun verwendet zum Nachweis der Konvergenz des vollen Mehrgitteralgorithmus.

**Satz 4.3 (Mehrgitterkonvergenz):** *Es sei angenommen, dass der Zweigitteralgorithmus konvergiert mit einer  $L^2$ -Konvergenzrate  $\rho_{ZG}(\nu) \rightarrow 0$  für  $\nu \rightarrow \infty$ , gleichmäßig bzgl.  $L$ . Dann konvergiert für hinreichend häufige Glättung der Mehrgitteralgorithmus mit  $R \geq 2$  (W-Zyklus) mit einer von  $L$  unabhängigen  $L^2$ -Konvergenzrate  $\rho_{MG} < 1$ ,*

$$\|u_L - MG(L, u_L^{(t)}, f_L)\| \leq \rho_{MG} \|u_L - u_L^{(t)}\|. \quad (4.2.75)$$

**Beweis:** Der Beweis wird durch Induktion nach dem Gitterlevel  $L$  geführt. Wir betrachten nur den relevanten Fall  $R = 2$  (W-Zyklus) und werden uns der Einfachheit halber nicht bemühen, die auftretenden Konstanten zu optimieren. Sei  $\nu$  so groß, dass die Konvergenzrate des Zweigitteralgorithmus  $\rho_{ZG} \leq \frac{1}{8}$  ist. Wir wollen zeigen, dass dann die Konvergenzrate des Mehrgitteralgorithmus  $\rho_{MG} \leq \frac{1}{4}$  ist, gleichmäßig bzgl.  $L$ . Für  $L = 1$  ist dies dann offenbar richtig. Sei nun auch  $\rho_{MG} \leq \frac{1}{4}$  für Gitterlevel  $L - 1$ . Auf Gitterlevel  $L$  gilt dann ausgehend von der Iterierten  $u_L^{(t)}$  mit der approximativen Lösung  $q_{L-1}^{(2)}$  (nach 2-maliger Anwendung der Grobgitterkorrektur) und der exakten Lösung  $\hat{q}_{L-1}$  der Defektgleichung auf Level  $L - 1$ :

$$\begin{aligned} u_L^{(t+1)} &= MG(L, u_L^{(t)}, f_L) = S_L^\nu(u_L^{(t)}) + p_{L-1}^L q_{L-1}^{(2)} \\ &= S_L^\nu(u_L^{(t)}) + p_{L-1}^L \hat{q}_{L-1} + p_{L-1}^L (q_{L-1}^{(2)} - \hat{q}_{L-1}) \\ &= ZG(L, u_L^{(t)}, f_L) + p_{L-1}^L (q_{L-1}^{(2)} - \hat{q}_{L-1}) \end{aligned} \quad (4.2.76)$$

Nach Induktionsvoraussetzung ist (Man beachte, dass der Startwert der Mehrgitteriteration auf Level  $L - 1$  gleich Null ist und  $\hat{\rho}_{L-1} = \mathcal{A}_{L-1}^{-1} r_L^{L-1} d_L$ ):

$$\|\hat{q}_{L-1} - q_{L-1}^{(2)}\| \leq \rho_{MG}^2 \|\hat{q}_{L-1}\| = \rho_{MG}^2 \|\mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L S_L^\nu(u_L - u_L^{(t)})\|. \quad (4.2.77)$$

Kombination der letzten Beziehungen ergibt für den Iterationsfehler  $e_L^{(t)} := u_L^{(t)} - u_L$ :

$$\|e_L^{(t+1)}\| \leq (\rho_{ZG} + \rho_{MG}^2 \|\mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L S_L^\nu\|) \|e_L^{(t)}\|. \quad (4.2.78)$$

Die Norm rechts ist bereits im Zusammenhang mit der Konvergenz des Zweigitteralgorithmus abgeschätzt worden. Mit dem Zweigitteroperator  $ZG_L = (\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}) \mathcal{A}_L S_L^\nu$

gilt

$$\mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L S_L^\nu = S_L^\nu - (\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}) \mathcal{A}_L S_L^\nu = S_L^\nu - ZG_L$$

und somit

$$\|\mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L S_L^\nu\| \leq \|S_L^\nu\| + \|ZG_L\| \leq 1 + \rho_{ZG} \leq 2. \quad (4.2.79)$$

Damit erhalten wir schließlich

$$\|e_L^{(t+1)}\| \leq (\rho_{ZG} + 2\rho_{MG}^2) \|e_L^{(t)}\|. \quad (4.2.80)$$

Mit Hilfe der Annahme über  $\rho_{ZG}$  und der Induktionsannahme folgt

$$\|e_L^{(t+1)}\| \leq \left(\frac{1}{8} + 2\frac{1}{16}\right) \|e_L^{(t)}\| \leq \frac{1}{4} \|e_L^{(t)}\|, \quad (4.2.81)$$

was den Induktionsbeweis vervollständigt.

Q.E.D.

Für „gutartige“ Probleme (symmetrischer, positiv definiter Operator, glatte Koeffizienten, quasi-gleichförmige Gitter u.s.w.) erreicht man in der Regel Mehrgitterkonvergenzraten im Bereich  $\rho_{MG} = 0,1 - 0,3$ . Die obige Analyse ist nur für den W-Zyklus gültig, da im Beweisteil (ii)  $R \geq 2$  benötigt wird. Der V-Zyklus kann nicht auf Basis nur einer Zweigitteranalyse behandelt werden. In der Literatur finden sich allgemeinere Ansätze, die Konvergenz von Mehrgitterverfahren auch in weniger regulären Situationen garantieren.

Als nächstes diskutieren wir die numerische Komplexität des Mehrgitteralgorithmus. Dabei werden die folgenden Bezeichnungen verwendet:

$OP(T)$  := Anzahl der a.Op. zur Durchführung einer Operation  $T$ ,

$R$  := Anzahl der Defektkorrekturschritte auf den einzelnen Gitterniveaus,

$N_l := \dim V_l \approx h_l^{-d}$  ( $d$  = Raumdimension),

$\kappa := \max_{1 \leq l \leq L} N_{l-1}/N_l < 1$ ,

$C_0 := OP(\mathcal{A}_0^{-1})/N_0$ ,

$C_s := \max_{1 \leq l \leq L} \{OP(S_l)/N_l\}, \quad C_d := \max_{1 \leq l \leq L} \{OP(d_l)/N_l\}$ ,

$C_r := \max_{1 \leq l \leq L} \{OP(r_l)/N_l\}, \quad C_p := \max_{1 \leq l \leq L} \{OP(p_l)/N_l\}$ .

In der Praxis ist meist  $\kappa \approx 2^{-d}$ ,  $C_s \approx C_d \approx C_r \approx C_p \approx \#\{a_{nm} \neq 0\}$  und  $C_0 N_0 \ll N_L$ .

**Satz 4.4 (Mehrgitterkomplexität):** Unter der Bedingung  $q := R\kappa < 1$  gilt für einen Mehrgitterzyklus  $MG_L$ :

$$OP(MG_L) \leq C_L N_L \quad (4.2.82)$$

mit

$$C_L = \frac{(\nu + \mu)C_s + C_d + C_r + C_p}{1 - q} + C_0 q^L,$$

Der Mehrgitteralgorithmus liefert die  $N_L$ -dimensionale diskrete Lösung  $u_L \in V_L$  auf dem Gitter  $\mathbb{T}_L$  im Rahmen der Diskretisierungsgenauigkeit  $\mathcal{O}(h_L^2)$  bzgl. der  $L^2$ -Norm mit  $\mathcal{O}(N_L \ln(N_L))$  a.Op. und hat damit (fast) optimale Komplexität.

**Beweis:** Wir setzen  $C_l := OP(MG_l)/N_l$ . Ein Mehrgitterschritt beinhaltet die  $R$ -fache Anwendung desselben Algorithmus auf dem nächst größeren Gitter. Bei Beachtung von  $N_{l-1} \leq \kappa N_l$  gilt mit  $\hat{C} := (\nu + \mu)C_s + C_d + C_r + C_p$ :

$$C_L N_L = OP(MG_L) \leq \hat{C} N_L + R \cdot OP(MG_{L-1}) = \hat{C} N_L + R \cdot C_{L-1} N_{L-1} \leq \hat{C} N_L + q C_{L-1} N_L,$$

und folglich  $C_L \leq \hat{C} + q C_{L-1}$ . Rekursive Anwendung dieser Beziehung liefert

$$C_L \leq \hat{C}(1 + q + q^2 + \dots + q^{L-1}) + q^L C_0 \leq \frac{\hat{C}}{1-q} + q^L C_0.$$

Dies impliziert die behauptete Abschätzung (4.2.82). Die Komplexität des Gesamtalgorithmus ergibt sich dann aus den Beziehungen

$$\rho_{MG}^t \approx h_L^2 \approx N_L^{-2/d}, \quad t \approx -\frac{\ln(N_L)}{\ln(\rho_{MG})}.$$

Dies vervollständigt den Beweis.

Q.E.D.

Wir bemerken, dass im Beweis der Aussage (4.2.82) die Bedingung

$$q := R\kappa = R \max_{1 \leq l \leq L} N_{l-1}/N_l < 1$$

wesentlich ist. Dies besagt für den W-Zyklus ( $R = 2$ ), dass sich beim Übergang vom Gitter  $\mathbb{T}_{l-1}$  zum nächst feineren  $\mathbb{T}_l$  die Anzahl der Gitterpunkte (bzw. Freiheitsgrade) hinreichend stark erhöhen muss, etwa wie bei einer gleichförmigen Verfeinerung

$$N_l \approx 4N_{l-1}.$$

Bei einem adaptiv gesteuerten Verfeinerungsprozess mit teilweise nur lokaler Gitterverfeinerung ist dies meist nicht erfüllt; selbst bei Verwendung der „Fest-Raten“-Strategie ist z. B. oft nur  $N_l \approx 2N_{l-1}$ . In solchen Fällen muss der Mehrgitterprozess zur Aufwandsersparnis modifiziert werden. Dies geschieht dadurch, dass die kostenintensive Glättung sowie die anderen Operationen nur jeweils auf den beim Übergang von  $\mathbb{T}_{l-1}$  zu  $\mathbb{T}_l$  neu hinzugekommenen Gitterpunkten durchgeführt werden. Bei der Implementierung eines Mehrgitteralgorithmus auf lokal verfeinerten Gittern ist viel Fingerspitzengefühl erforderlich, wenn der resultierende Gesamtalgorithmus komplexitäts-optimal sein soll.

Für das geschachtelte MG-Schema erhält man sogar im strengen Sinne „optimale“ Lösungskomplexität  $O(N_L)$ , da auf jedem Gitterniveau bestmögliche Startwerte verwendet werden.

**Satz 4.5 (Geschachteltes Mehrgitterverfahren):** Das geschachtelte MG-Schema ist

*komplexitäts-optimal*, d.h.: Es liefert die diskrete Lösung  $u_L \in V_L$  auf dem feinsten Gitter  $\mathbb{T}_L$  im Rahmen der Diskretisierungsgenauigkeit  $\mathcal{O}(h_L^2)$  bzgl. der  $L^2$ -Norm mit einem Aufwand von  $\mathcal{O}(N_L)$  a.Op.

**Beweis:** Die Genauigkeitsanforderung für die Mehrgitteriteration auf Gitterlevel  $\mathbb{T}_L$  ist

$$\|e_L^{(t)}\| \leq \hat{c}h_L^2\|f\|. \quad (4.2.83)$$

i) Wir wollen zunächst zeigen, dass (4.2.83) beim geschachtelten MG-Schema unter den Voraussetzungen des Mehrgitterkonvergenzsatzes 4.3 auf jedem Level  $L$  mit einer (hinreichend großen) festen Zahl  $t_*$  von Mehrgitterschritten erreichbar ist. Sei  $e_L^{(t)} := u_L^{(t)} - u_L$  wieder der Iterationsfehler auf Level  $L$ . Nach Annahme ist  $e_0^{(t)} = 0$ ,  $t \geq 1$ . Im Fall  $u_L^{(0)} := u_{L-1}^{(t)}$  gilt dann

$$\begin{aligned}\|e_L^{(t)}\| &\leq \rho_{MG}^t \|e_L^{(0)}\| = \rho_{MG}^t \|u_{L-1}^{(t)} - u_L\| \\ &\leq \rho_{MG}^t (\|u_{L-1}^{(t)} - u_{L-1}\| + \|u_{L-1} - u\| + \|u - u_L\|) \\ &\leq \rho_{MG}^t (\|e_{L-1}^{(t)}\| + ch_L^2\|f\|).\end{aligned}$$

Rekursive Anwendung dieser Beziehung für  $L \geq l \geq 1$  ergibt dann (wegen  $h_l \leq \kappa^{l-L}h_L$ )

$$\begin{aligned}\|e_L^{(t)}\| &\leq \rho_{MG}^t (\rho_{MG}^t (\|e_{L-1}^{(t)}\| + ch_{L-1}^2\|f\|) + ch_L^2\|f\|) \\ &\vdots \\ &\leq \rho_{MG}^{Lt} \|e_0^{(t)}\| + (c\rho_{MG}^t h_L^2 + c\rho_{MG}^{2t} h_{L-1}^2 + \dots + c\rho_{MG}^{Lt} h_1^2) \|f\| \\ &= ch_L^2 \kappa^2 (\rho_{MG}^t \kappa^{-2-1} + \rho_{MG}^{2t} \kappa^{-2-2} + \dots + \rho_{MG}^{Lt} \kappa^{-2L}) \|f\| \\ &\leq ch_L^2 \kappa^2 \|f\| \frac{\kappa^{-2} \rho_{MG}^t}{1 - \kappa^{-2} \rho_{MG}^t},\end{aligned}$$

vorausgesetzt  $\kappa^{-2} \rho_{MG}^t < 1$ . Offenbar gibt es also ein  $t_*$ , so dass (4.2.83) für  $t \geq t_*$  erfüllt ist, und zwar gleichmäßig bzgl.  $L$ .

ii) Wir kommen nun zur Aufwandsanalyse. Satz 4.4 besagt, dass ein Zyklus des „einfachen“ Mehrgitteralgorithmus  $MG(l, \cdot, \cdot)$  auf dem  $l$ -ten Level  $W_l \leq c_* N_l$  a.Op. benötigt (gleichmäßig bzgl.  $l$ ). Sei nun  $\hat{W}_l$  die Anzahl der a.Op. des geschachtelten Schemas auf Gitterlevel  $l$ . Dann gilt konstruktionsgemäß:

$$\hat{W}_L \leq \hat{W}_{L-1} + t_* W_L.$$

Durch Iteration dieser Beziehung erhalten wir mit  $\kappa := \max_{1 \leq l \leq L} N_{l-1}/N_l < 1$ :

$$\hat{W}_L \leq t_* c_* \{N_L + \dots + N_0\} \leq ct_* c_* N_L \{1 + \dots + \kappa^L\} \leq \frac{ct_* c_*}{1 - \kappa} N_L,$$

was zu beweisen war.

Q.E.D.

### 4.3 Übungen

**Übung 4.1:** Das allgemeine „Abstiegsverfahren“ zur iterativen Lösung des Gleichungssystems  $Ax = b$  mit symmetrischer, positiv-definiter Matrix  $A \in \mathbb{R}^{N \times N}$  lautet:

$$\begin{aligned} \text{Startwert: } & x^{(0)} \in \mathbb{R}^n, \quad r^{(0)} := b - Ax^{(0)}, \\ \text{für } t \geq 0: & \text{ Abstiegsrichtung } d^{(t)}, \\ & \alpha_t = \frac{\langle r^{(t)}, d^{(t)} \rangle}{\langle Ad^{(t)}, d^{(t)} \rangle}, \\ & x^{(t+1)} = x^{(t)} + \alpha_t d^{(t)}, \quad r^{(t+1)} = r^{(t)} - \alpha_t Ad^{(t)}. \end{aligned}$$

Die sog. „Koordinatenrelaxation“ erhält man durch zyklische Wahl der Abstiegsrichtungen  $d^{(t)}$  aus den kartesischen Einheitsvektoren  $\{e^{(1)}, \dots, e^{(N)}\}$ . Man zeige, dass jeder  $N$ -Zyklus der Koordinatenrelaxation äquivalent ist zum üblichen Gauß-Seidel-Verfahren.

**Bemerkung:** Für eine typische FE-Matrix hat die zyklische Koordinatenrelaxation also das Konvergenzverhalten:

$$|x^{(tN)} - x| \leq cq^t, \quad q \approx 1 - \text{cond}_2(A)^{-1} \approx 1 - h^2.$$

**Übung 4.2:** Die erste RWA des Laplace-Operators

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega,$$

auf einem „regulären“ Gebiet  $\Omega \subset \mathbb{R}^2$  werde mit einem FE-Verfahren mit stückweise linearen Ansatzfunktionen auf einer Folge von quasi-gleichförmigen Gittern (d. h. größen- und form-regulär) der Weite  $h$  diskretisiert. Dies führt auf lineare Gleichungssysteme  $Ax = b$  mit symmetrischen, positiv definiten  $(N \times N)$ -Matrizen  $A$ , wobei  $N$  die Anzahl der Knotenpunkte ist.

Welchen arithmetischen Aufwand (ausgedrückt in Potenzen von  $h$ ) erfordert dabei die Lösung dieser Gleichungssysteme mit dem CG-Verfahren mit der Genauigkeit des Diskretisierungsfehlers gemessen in der „Energie-Norm“  $\|\nabla(u - u_h)\|$ ? Dazu verwende man die folgende bekannte Fehlerabschätzung für das CG-Verfahren:

$$|x - x^t|_A \leq 2q^t |x - x^0|_A, \quad q := \frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}},$$

mit der diskreten „Energienorm“  $\|x\|_A := (Ax, x)^{1/2}$  und der Spektralkondition  $\kappa := \kappa_2(A)$  von  $A$ .

**Hinweis:** Man verwende die bekannte Beziehung für die Spektralkondition von  $A$  sowie die aus der obigen Fehlerabschätzung abgeleitete Abschätzung für die Anzahl der Iterationsschritte. Der Aufwand pro CG-Schritt entspricht etwas der zweimaligen Defektberechnung  $x \rightarrow d := Ax - b$ .

**Übung 4.3:** Man versuche, den Beweis der Konvergenz des Zweigitterverfahrens ZG aus dem Text für den Fall zu verallgemeinern, dass die Restriktion  $r_l^{l-1} : V_l \rightarrow V_{l-1}$  mit Hilfe

lokaler, bilinearer Interpolation (anstelle der  $L^2$ -Projektion) auf dem Gitter  $T_{l-1}$  definiert ist. Wo ist dabei das Problem, und wie kann man damit fertig werden?

**Übung 4.4:** Die FE-Diskretisierung des Konvektions-Diffusionsproblems

$$-\Delta u + \partial_1 u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega,$$

führt auf unsymmetrische Systemmatrizen  $A_h$ . In diesem Fall erfordert die Analyse des Mehrgitterverfahrens einige Modifikationen. Man übertrage den Beweis aus dem Text für die Konvergenz des Zweigitteralgorithmus, wenn als Glätter wieder das Richardson-Verfahren

$$x_h^{t+1} = x_h^t - \theta_t(A_h x_h^t - b_h), \quad t = 0, 1, 2, \dots,$$

mit den Dämpfungsparametern  $\theta_t := \frac{1}{2}\|A_h\|^{-1}$  verwendet wird.

**Übung 4.5:** Zur Lösung der 1. RWA der Laplace-Gleichung

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega,$$

auf dem Einheitsquadrat  $\Omega = (0, 1)^2$  werde auf einer Folge äquidistanter, kartesischer Gitter  $T_l$  mit Gitterweiten  $h_l = 2^{-l}$  mit Hilfe bilinearer finiter Elemente approximiert. Die diskrete Gleichung auf Gitterlevel  $l$  werde dabei mit einem MG-Verfahren gelöst, wobei das Richardson-Verfahren zur Glättung, die natürliche Einbettung zur Prolongation und die lokale bilineare Interpolation zur Restriktion verwendet werden. Die Anzahl der Vor- und Nachglättungsschritte sei  $\nu = 2$  und  $\mu = 0$ . Wieviele a. Op. kosten dann ungefähr ein V-Zyklus und ein W-Zyklus ausgedrückt in Vielfachen der Dimension  $N_l = \dim V_l$ ?

**Übung 4.6:** Die erste RWA des Laplace-Operators

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega,$$

auf einem „regulären“ Gebiet  $\Omega \subset \mathbb{R}^2$  werde mit einem FE-Verfahren mit stückweise linearen Ansatzfunktionen auf einer Folge von quasi-gleichförmigen Gittern (d. h. größen- und form-regulär) der Weite  $h$  diskretisiert. Dies führt auf lineare Gleichungssysteme  $Ax = b$  mit symmetrischen, positiv definiten  $(N \times N)$ -Matrizen  $A$ , wobei  $N$  die Anzahl der Knotenpunkte ist. Das „Richardson-Verfahren“ iteriert zur Lösung des Gleichungssystems  $Ax = b$  ausgehend von einem Startwert  $x^0 \in \mathbb{R}^N$  mit einem Dämpfungsparameter  $\theta \in \mathbb{R}$  gemäß

$$x^{t+1} = x^t - \theta(Ax^t - b), \quad t \in \mathbb{N}_0.$$

Im Falle, dass  $A$  nur reelle positive Eigenwerte  $0 < \lambda_{\min} \leq \dots \leq \lambda_{\max}$  besitzt, ist der Spektralradius der Iterationsmatrix  $B_\theta = I - \theta A$  gegeben durch

$$\rho(B_\theta) = \max\{|1 - \theta\lambda_{\min}|, |1 - \theta\lambda_{\max}|\}.$$

Für welches  $\theta$  wird  $\rho(B_\theta)$  minimal, d. h. konvergiert die Iteration am besten, und für welches  $\theta$  hat die Iteration die beste Glättungseigenschaft?

**Übung 4.7:** Man beschäftige sich mit den folgenden Fragen:

- a) Wie lautet die variationelle („schwache“) Formulierung der Randwertaufgabe

$$-\nabla \cdot (a \nabla u) + bu = f \text{ in } \Omega, \quad u|_{\partial\Omega} = 0,$$

auf einem konvexen Polyeder  $\Omega \subset \mathbb{R}^3$  und unter welchen Bedingungen an die Koeffizientenfunktionen  $a \in C^1(\bar{\Omega})$  und  $b \in C(\bar{\Omega})$  ist diese „wohl gestellt“.

- b) Die Randwertaufgabe in a) werde durch einen konformen „quadratischen“ Finite-Elemente-Ansatz mit Gitterweite  $h \in \mathbb{R}_+$  diskretisiert. Man beschreibe die einzelnen Schritte (Gitter, Knotenbasis, Systemmatrix) zur Aufstellung der zugehörigen algebraischen Gleichungssysteme

$$A_h x_h = b_h.$$

- c) Man gebe für die Diskretisierung in b) optimale a priori Fehlerabschätzungen in der Energie- und der  $L^2$ -Norm an. Wie hängt in diesem Fall die Kondition der Systemmatrix  $A_h$  von der Gitterweite  $h$  ab?
- d) Man formuliere i) das Gauß-Seidel-Verfahren und ii) das Gradienten-Verfahren zur iterativen Lösung des Gleichungssystems in b). Wieviele Iterationsschritte sind mit diesen Verfahren in Abhängigkeit von der Anzahl der Unbekannten  $N_h := \dim V_h$  notwendig, um den Anfangsfehler um  $10^{-3}$  zu reduzieren?



## 5 Verfahren für parabolische Probleme

Wir diskutieren zunächst wieder die klassischen Differenzenapproximationen zur Lösung parabolischer Anfangs-Randwert-Aufgaben (ARWA). Der Übersichtlichkeit halber beschränken wir uns dabei auf das Modellproblem der Wärmeleitungsgleichung in zwei Ortsdimensionen mit Dirichletschen Randbedingungen, d. h. auf die 1. ARWA:

$$\partial_t u + L u = f \quad \text{in } \Omega \times (0, T), \quad u|_{\partial\Omega} = 0, \quad u|_{t=0} = u^0, \quad (5.0.1)$$

mit einem elliptischen Operator  $L$ , der hier exemplarisch als  $L := -a\Delta$  mit einer Konstanten  $a > 0$  gewählt wird.

Das Definitionsgebiet  $\Omega \in \mathbb{R}^2$  wird wieder als glatt berandet oder als konkaves Polygonegebiet vorausgesetzt. Die Problemdaten  $f, g, u^0$  sind ebenfalls glatt und kompatibel, so dass die Lösung ebenfalls als glatt angenommen werden kann. Erweiterungen für Probleme mit weniger regulären Daten oder anderen Randbedingungen sowie auf den dreidimensionalen Fall werden gegebenenfalls in Bemerkungen berücksichtigt. Gelegentlich wird auch das eindimensionale Analogon von (5.0.1) betrachtet. Als Basis von Finite-Elemente-Diskretisierungen dient wieder die variationelle Formulierung von (5.0.1):

$$(\partial_t u, \varphi) + a(u, \varphi) = (f, \varphi) \quad \forall \varphi \in V, \quad t > 0, \quad u|_{t=0} = 0, \quad (5.0.2)$$

mit dem üblichen Sobolew-Raum  $V := H_0^1(\Omega)$  und der symmetrischen und positiv definiten „Energie-Form“  $a(u, \varphi) := (a\nabla u, \nabla \varphi)_\Omega$ .

Bei der Diskretisierung von instationären Problemen gibt es drei verschiedene Vorgehensweisen, die wir im Folgenden kurz beschreiben wollen.

**i) „Linienmethode“:** Zunächst wird eine Diskretisierung bzgl. der Ortsvariablen vorgenommen, d.h.: mit Hilfe eines Finite-Differenzen- oder Finite-Elemente-Ansatzes werden „diskrete“ Funktionen  $u_h(t) = u_h(\cdot, t)$  bestimmt aus der Gleichung

$$u'_h(t) + \mathcal{A}_h u_h(t) = f_h(t), \quad t \geq 0, \quad u_h(0) = u_h^0. \quad (5.0.3)$$

Im Falle eines Differenzenverfahrens auf einem Ortsgitter  $\{x_i\}_{i=1,\dots,N}$  ist die diskrete Funktion  $u_h(t) = (u_n(t))_{n=1}^N$  der Vektor der Knotenwerte  $u_n(t) \approx u(x_n, t)$ ,  $\mathcal{A}_h = A_h : \mathbb{R}^N \rightarrow \mathbb{R}^N$  die zum verwendeten Differenzenoperator korrespondierende Matrix und  $f_h = b_h = (f(x_n))_{n=1}^N$ . Beim Finite-Elemente-Ansatz ist  $u_h(t) \in V_h$  eine Finite-Elemente-Funktion,  $\mathcal{A}_h =: V_h \rightarrow V_h$  das durch

$$(\mathcal{A}_h v_h, \varphi_h) = a(v_h, \varphi_h), \quad v_h, \varphi_h \in V_h,$$

definierte diskrete Analogon zum Differentialoperators  $L$  und  $f_h = P_h f \in V_h$  die  $L^2$ -Projektion der rechten Seite  $f$  auf  $V_h$ . Die Aufgabe (5.0.3) lautet demgemäß in variationaler Form wie folgt:

$$(u'_h(t), \varphi_h) + a(u_h(t), \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h, \quad t \in I, \quad u_h(0) = P_h u^0. \quad (5.0.4)$$

Nach Einführung einer Knotenbasis  $\{\varphi_h^{(n)}, n = 1, \dots, N = \dim(V_h)\}$  geht dieses Problem über in ein System für den Vektor  $U_h(t) = (U_n(t))_{n=1}^N$  der zugehörigen Knotenwerte,

$$M_h U'_h(t) + A_h U_h(t) = b_h(t), \quad t \geq 0, \quad U_h(0) = U_h^0, \quad (5.0.5)$$

mit der „Steifigkeitsmatrix“ und „Massenmatrix“

$$A_h = (a(\varphi_h^{(n)}, \varphi_h^{(m)}))_{n,m=1}^N, \quad M_h = ((\varphi_h^{(n)}, \varphi_h^{(m)}))_{n,m=1}^N$$

der Finite-Elemente-Basis. In beiden Fällen, (5.0.3) oder (5.0.5), handelt es sich um ein System von (linearen) gewöhnlichen Differentialgleichungen. Dieses wird nun mit einem der üblichen Schemata bzgl. der Zeit diskretisiert. Nach Wahl einer (zunächst konstanten) Zeitschrittweite  $k$  werden zu den „diskreten“ Zeitleveln  $t_m = mk$  Approximationen  $U_h^m = (U_n^m)_{n=1}^N$  zu  $u(\cdot, t_m)$  bestimmt. Wir sprechen von einem „Einschritt-“ bzw. einem „Zweischrittverfahren“, wenn  $U_h^m$  aus den vorausgehenden Werten gemäß einer Formel der Form

$$U_h^m = F(U_h^m, U_h^{m-1}) \quad \text{bzw.} \quad U_h^m = F(U_h^m, U_h^{m-1}, U_h^{m-2})$$

berechnet wird. Im Falle

$$U_h^m = F(U_h^{m-1}) \quad \text{bzw.} \quad U_h^m = F(U_h^{m-1}, U_h^{m-2})$$

heißt die Methode „explizit“. Zur Durchführung einer nicht expliziten, d.h. „impliziten“, Methode müssen in jedem Zeitschritt Gleichungssysteme gelöst werden. Die hohe Dimension des Systems,  $N = \#\{\text{Gitterpunkte } a_n\}$  bzw.  $N = \dim(V_h)$ , mit  $N \sim 10^3 - 10^8$  impliziert im Hinblick auf die Lösungsökonomie Einschränkungen bei der Wahl der Verfahren. Es kommen in der Regel nur Schemata einfacher Struktur, d.h. mit wenigen Matrix-Vektor-Multiplikationen, und niedriger Ordnung  $r = 1 - 4$  in Frage. Eine weitere wesentliche Einschränkung besteht in der generischen Steifheit des Systems. Die Systemmatrix  $A_h$  hat in Abhängigkeit von der (gleichförmigen) Gitterfeinheit  $h$  die Kondition

$$\kappa_2(A_h) \approx h^{-2}.$$

Bei *expliziten* Zeitschrittschemata sind also einschneidende Schrittweitenrestriktionen einzuhalten, welche deren Verwendung in der Regel verbietet. Der formale Vorteil der expliziten Verfahren, dass in den einzelnen Zeitschritten keine impliziten Gleichungssysteme zu lösen sind, wird besonders in höheren Raumdimensionen ( $d = 2, 3$ ) durch die hohe Zahl von durchzuführenden Zeitschritten (besonders bei Verwendung lokal verfeinerter Ortsgitter) schnell aufgehoben.

**Beispiel numerischer Instabilität:** Wir wollen dies anhand einer einfachen Modellsituation illustrieren. Die eindimensionale, homogene Version der ARWA (5.0.1)

$$\partial_t u - \partial_x^2 u = 0 \quad \text{in } \Omega = (0, 1), \quad u|_{\partial\Omega} = 0, \quad u|_{t=0} = u^0 \quad (5.0.6)$$

wird auf einem äquidistanten Gitter  $0 = x_0 < \dots < x_n < \dots < x_{N+1} = 1$  mit Hilfe

zentraler Differenzenquotienten 2. Ordnung,

$$\partial_x^2 u(x_n, t) \approx h^{-2} \{ U_{n-1}(t) - 2U_n(t) + U_{n+1}(t) \}.$$

diskretisiert. Die Vektorfunktion  $U_h(t) = (U_n(t))_{n=1}^N$  genügt dann dem System gewöhnlicher Differentialgleichungen

$$U'_n(t) - h^{-2} \{ U_{n-1}(t) - 2U_n(t) + U_{n+1}(t) \} = 0,$$

wobei bei Berücksichtigung der Randbedingungen  $U_0 \equiv U_{N+1} \equiv 0$  gesetzt ist. Die Anfangswerte sind naturgemäß  $U_n(0) = u^0(x_n)$ . Dies kann kompakt geschrieben werden als

$$U'_h + A_h U_h(t) = 0, \quad t \geq 0, \quad U_h(0) = U^0, \quad (5.0.7)$$

mit der  $(N \times N)$ -Matrix

$$A_h = h^{-2} \begin{bmatrix} -2 & 1 & & & 0 \\ 1 & -2 & & & \\ & \ddots & \ddots & \ddots & \\ & & & -2 & 1 \\ 0 & & & 1 & -2 \end{bmatrix}.$$

Diese Matrix hat, wie wir bereits wissen, die Eigenwerte

$$0 < \lambda_1 \leq \dots \leq \lambda_N = \frac{4}{h^2} + \mathcal{O}(h^2), \quad (5.0.8)$$

d. h.: Das nach Ortsdiskretisierung entstandene System (5.0.7) wird für kleines  $h$  zunehmend steif mit Steifigkeitsrate  $\kappa = \mathcal{O}(h^{-2})$ . Beim expliziten Euler<sup>1</sup>-Schema („Polygonzugmethode“) ist z. B. aus Stabilitätsgründen die Schrittweitenbedingung

$$-\lambda_n k \in [-2, 0] \quad \Leftrightarrow \quad k \leq \frac{1}{2}h^2 \quad (5.0.9)$$

einzuhalten. Diese Schrittweitenbeschränkung für explizite Verfahren hat entscheidende praktische Bedeutung. Wir wollen das Phänomen der numerischen Instabilität illustrieren. Dazu betrachten wir als einfachstes explizites Zeitschrittschema das klassische Euler-Verfahren (Polygonzugmethode) mit äquidistanter Schrittweite  $k$ . Dies führt auf die folgenden Differenzengleichungen für die Approximationen  $U_n^m \approx u(x_n, t_m)$ :

$$U_n^{m+1} = U_n^m + \frac{k}{h^2} \left( U_{n-1}^m - 2U_n^m + U_{n+1}^m \right). \quad (5.0.10)$$

---

<sup>1</sup>Leonhard Euler (1707–1783), geb. in Basel: universeller Mathematiker und Physiker; bedeutendster und produktivster Mathematiker seiner Zeit; wirkte in Berlin und St. Petersburg; Arbeiten zu allen mathematischen Gebieten seiner Zeit.

Für  $k = h^2$  ist dann

$$U_n^{m+1} = U_{n-1}^m - U_n^m + U_{n+1}^m.$$

Im Fall oszillierender Anfangsdaten  $u_n^0 = (-1)^n$  ergibt sich

$$U_n^1 = (-1)^{n-1} - (-1)^n + (-1)^{n+1} = -3(-1)^n = -3U_n^0,$$

und bei Fortsetzung dieses Arguments:

$$U_n^m = (-3)^m U_n^0, \quad m \geq 1, \quad n = 1, \dots, N. \quad (5.0.11)$$

Dieses Verhalten bedeutet numerische „Instabilität“. Es mag unrealistisch erscheinen, eine oszillierende Anfangsbedingung der Art  $U_n^0 = (-1)^n$  anzunehmen, doch bedingt durch Rundungsfehler könnte gelten:

$$U_n^0 = V_n^0 + \varepsilon(-1)^n \quad (5.0.12)$$

mit „glatten“ exakten Anfangsdaten  $V^0$ . Wegen der Linearität der betrachteten Differenzengleichungen folgt

$$U_n^m = V_n^m + \varepsilon(-3)^m(-1)^n, \quad (5.0.13)$$

so dass die anfänglich kleinen Anfangsstörungen schnell anwachsen; z. B. ist diese für  $\varepsilon = 10^{-15} > 3^{-32}$  bereits nach nur 32 Zeitschritten auf Größe  $\approx 1$  angewachsen und zwar unabhängig von der Größe von  $h$ .

**ii) „Rothe-Methode“:** Bei der Rothe<sup>2</sup>-Methode wird die Differentialgleichung als gewöhnliche Differentialgleichung für eine Hilbertraum-wertige Funktion  $U(t) \in V$  aufgefasst und zunächst mit einem A-stabilen Verfahren in der Zeit diskretisiert. Bei Verwendung z. B. des impliziten Euler-Schemas ergibt sich eine Folge von speziellen Randwertaufgaben

$$U^m + kLU^m = U^{m-1} + kf^m, \quad m \geq 1, \quad U^0(x) = u^0(x).$$

Diese Probleme werden nun nacheinander auf möglicherweise wechselnden, dem Lösungsverlauf angepaßten Ortsgritten diskretisiert. Das Problem ist dabei der adäquate Transfer der jeweiligen Startlösung  $U^{m-1}$  vom alten auf das neue Ortsgritter. Hier zeigt sich wieder der systematische Vorteil einer Finite-Elemente-Galerkin-Methode, bei der sich ganz automatisch als *richtige* Wahl die  $L^2$ -Projektion von  $U^{m-1}$  auf das neue Gitter ergibt.

**iii) Globale Orts-Zeit-Diskretisierung:** Ähnlich wie bei den Transportproblemen in zwei Dimensionen könnte auch bei der Wärmeleitungsgleichung eine simultane Diskretisierung (etwa mit einem Finite-Elemente-Galerkin-Verfahren) auf einem unstrukturierten Gitter der ganzen  $(x,t)$ -Ebene erfolgen. Dieser theoretisch durchaus attraktive Ansatz wird aber bei höher dimensionalnen Problemen wegen der globalen Kopplung aller Unbekannten sehr rechenaufwendig und spielt daher bei parabolischen Problemen in der Praxis keine wesentliche Rolle.

---

<sup>2</sup>Erich Rothe (1895–1988): Deutscher Mathematiker; Promotion und Habilitation in Berlin (1928), danach Assistent in Breslau, nach dem Krieg Prof. an der University of Michigan, USA.

Im Folgenden Abschnitt werden wir Differenzenapproximationen in Verbindung mit der Linienmethode betrachten. Die Rothe-Methode wird in Verbindung mit Finite-Elemente-Verfahren im Ort diskutiert. Dies mündet dann auch ohne Probleme in Galerkin-Diskretisierungen simultan in Ort und Zeit, den sog. „unstetigen“ oder „stetigen“ Galerkin-Verfahren (sog. „dG(r)-“ oder „cG(r)-Verfahren“).

## 5.1 Differenzenverfahren für parabolische Probleme

### 5.1.1 Zeitschrittverfahren

Wir beginnen mit der Diskussion der „Linienmethode“ zur Diskretisierung von parabolischen ARWAn der Art

$$\partial_t u + Lu = f \quad \text{in } Q_T := \Omega \times I, \quad u|_{\partial\Omega} = 0, \quad u|_{t=0} = u^0, \quad (5.1.14)$$

mit  $L := -a\Delta$  auf einem beschränkten (regulär berandeten) Gebiet  $\Omega \subset \mathbb{R}^2$  und einem Zeitintervall  $I = [0, T]$ . Der Einfachheit halber wird die rechte Seite  $f$  gelegentlich als Null angenommen.

Ortsdiskretisierung von (5.1.14) mit einem der üblichen Differenzenverfahren (z. B. dem 5-Punkte-Operator mit geeigneter Randapproximation) führt auf ein System gewöhnlicher Differentialgleichungen

$$U'_h(t) + A_h U_h(t) = 0, \quad t > 0, \quad U_h(0) = U^0, \quad (5.1.15)$$

für den Vektor  $U_h(t) \in \mathbb{R}^N$  der Knotenwerte. Da die Eigenwerte der Systemmatrix  $A_h$  alle reell sind (oder wenigstens nahe an der reellen Achse liegen), käme zur stabilen Integration des Systems (5.1.15) jede  $A(0)$ -stabile Formel in Frage. Dabei muss aber der hohe numerische Aufwand bei der Durchführung komplizierter impliziter Verfahren hoher Ordnung berücksichtigt werden. Durch Übertragung der klassischen Zeitschrittformeln für gewöhnliche Differentialgleichungen auf das System (5.1.15) erhalten wir unter Benutzung der oben eingeführten Bezeichnungen die folgenden einfachsten Einschrittverfahren:

1) *Explizites Euler-Verfahren (Polygonzugmethode):*

$$k^{-1} \{U_h^m - U_h^{m-1}\} + A_h U_h^{m-1} = f^{m-1}, \quad m \geq 1,$$

2) *Implizites Euler-Verfahren:*

$$k^{-1} \{U_h^m - U_h^{m-1}\} + A_h U_h^m = f^m, \quad m \geq 1,$$

3) Crank<sup>3</sup>-Nicolson<sup>4</sup>-Verfahren (Trapezregel):

$$k^{-1}\{U_h^m - U_h^{m-1}\} + \frac{1}{2}A_h(U_h^m + U_h^{m-1}) = \frac{1}{2}(f^m + f^{m-1}), \quad m \geq 1,$$

und die Zweischrittverfahren: 4) BDF(2)-Verfahren (Rückwärtsdifferenzenformel):

$$\frac{1}{2}k^{-1}\{3U_h^m - 4U_h^{m-1} + U_h^{m-2}\} + A_hU_h^m = f^m, \quad m \geq 2,$$

5) Mittelpunkts-Verfahren:

$$\frac{1}{2}k^{-1}\{U_h^m - U_h^{m-2}\} + A_hU_h^{m-1} = f^{m-1}, \quad m \geq 2,$$

6) Simpson-Verfahren:

$$\frac{1}{2}k^{-1}\{U_h^m - U_h^{m-2}\} + \frac{2}{3}A_h\{U_h^m + 4U_h^{m-1} + U_h^{m-2}\} = f^{m-1}, \quad m \geq 2.$$

Als Startwerte werden gewöhnlich (im Fall glatter Anfangsdaten) einfach die Restriktionen  $U_n^0 = u^0(a_n)$  verwendet. Bei den Zweischrittverfahren wird der zweite erforderliche Startwert  $U_h^1$  durch Anwendung einer Einschrittformel entsprechender Ordnung gewonnen. Wegen ihrer inhärenten Instabilität (triviales Stabilitätsgebiet) kommen die Mittelpunktsformel und die Simpson-Formel für die praktische Anwendung nicht in Frage.

Wie bei der Analyse von Differenzenverfahren üblich verwenden wir den „Abschneidefehler“  $\tau_{h,k}^m = (\tau_n^m)_{n=1}^N$  der Differenzenformeln. Diesen erhält man wieder durch formales Auswerten der Differenzenformeln auf der exakten Lösung:

$$k \tau_{h,k}^m := u^m - F(u^m, u^{m-1}, u^{m-2}).$$

Bei einer Ortsdiskretisierung der Ordnung  $p$  verhält sich der Abschneidefehler dann gemäß

$$\tau_{h,k}^m = \mathcal{O}(h^p + k^q),$$

wobei  $q$  die „Ordnung“ des Zeitschrittverfahrens ist. Von der Fehleranalyse der Zeitschrittverfahren für gewöhnliche Differentialgleichungen wissen wir bereits, dass die einfachen Euler-Verfahren die Ordnung  $q = 1$  und das Crank-Nicolson- sowie das BDF(2)-Verfahren die Ordnung  $q = 2$  haben. Später werden wir noch Verfahren der Ordnung  $q = 3, 4$  kennenlernen. Bei der Analyse dieser Zeitschrittschemata für parabolische Probleme ist die genaue Abhängigkeit des Abschneidefehlers von der örtlichen und zeitlichen Regularität der Lösung interessant.

---

<sup>3</sup>John Crank (1916–2006): Englischer Mathematiker; Prof. an der Brunel University, Uxbridge, England; Arbeiten zur Numerik partieller Differentialgleichungen.

<sup>4</sup>Phyllis L. Nicolson (1917–1968): Englische Physikerin; Lecturer in Leeds und Manchester.

**Hilfssatz 5.1 (Konsistenz):** Für die ARWA (5.1.14) genügen die Abschneidefehler der betrachteten Differenzenverfahren den folgenden (scharfen) Abschätzungen:

i) Explizites und implizites Euler-Verfahren:

$$\max_{Q_T} |\tau_{h,k}^m| \leq \max_{Q_T} |\tau_h^m| + \frac{1}{2}k \max_{Q_T} |\partial_t^2 u|; \quad (5.1.16)$$

ii) Crank-Nicolson-Verfahren:

$$\max_{Q_T} |\tau_{h,k}^m| \leq \max_{Q_T} |\tau_h^m| + \frac{1}{12}k^2 \max_{Q_T} |\partial_t^3 u|; \quad (5.1.17)$$

iii) BDF(2)-Verfahren:

$$\max_{\tilde{Q}_T} |\tau_{h,k}^m| \leq \max_{\tilde{Q}_T} |\tau_h^m| + \frac{2}{3}k^2 \max_{\tilde{Q}_T} |\partial_t^3 u|. \quad (5.1.18)$$

Dabei ist  $\tau_h^m = \mathcal{O}(h^2)$  der Abschneidefehler der Ortsdiskretisierung.

**Beweis:** Der Abschneidefehler der Ortsdiskretisierung genügt i. Allg. der Abschätzung

$$|\tau_h^m| = |Lu^m - L_h u^m| \leq ch^2 M_4^m(u),$$

wobei  $L_h$  der Ortsdifferenzenoperator ist und  $M_4(u) := \max_{\Omega} |\nabla^4 u^m|$ . Speziell in einer Raumdimension mit  $\Omega = (0, 1)$  gilt

$$|\tau_h^m| = |\partial_x^2 u^m - L_h u^m| \leq \frac{1}{12}h^2 \max_{[0,1]} |\partial_x^4 u^m|.$$

i) Für die explizite Euler-Formel gilt

$$\begin{aligned} |k^{-1}(u^m - u^{m-1}) + L_h u^{m-1}| &= \left| k^{-1} \int_{t_{m-1}}^{t_m} \partial_t u \, dt + L_h u^{m-1} \right| \\ &= \left| k^{-1} \int_{t_{m-1}}^{t_m} \partial_t u \, dt - \partial_t u^{m-1} - Lu^{m-1} + L_h u^{m-1} \right| \\ &\leq k^{-1} \left| \int_{t_{m-1}}^{t_m} \{\partial_t u - \partial_t u^{m-1}\} \, ds \right| + |Lu^{m-1} - L_h u^{m-1}| \\ &\leq k^{-1} \int_{t_{m-1}}^{t_m} (t - t_{m-1}) \, dt \max_{[t_{m-1}, t_m]} |\partial_t^2 u| + |\tau_h^{m-1}| \end{aligned}$$

Es folgt

$$\max_{Q_T} |\tau_{h,k}^m| \leq \frac{1}{2}k \max_{Q_T} |\partial_t^2 u| + \max_{Q_T} |\tau_h^{m-1}|.$$

Dieselbe Abschätzung gilt auch für die implizite Euler-Formel.

ii) Für die Crank-Nicolson-Formel gilt

$$\begin{aligned} |k^{-1}(u^m - u^{m-1}) + \frac{1}{2}L_h(u^m + u^{m-1})| &= \left| k^{-1} \int_{t_{m-1}}^{t_m} \partial_t u \, dt - \frac{1}{2}(\partial_t u^m + \partial_t u^{m-1}) \right. \\ &\quad \left. + \frac{1}{2}(Lu^m - L_h u^m) + \frac{1}{2}(Lu^{m-1} - L_h u^{m-1}) \right| \\ &\leq k^{-1} \left| \int_{t_{m-1}}^{t_m} \frac{1}{2}(t - t_m)(t - t_{m-1}) \, dt \right| \max_{[t_{m-1}, t_m]} |\partial_t^3 u| \\ &\quad + \frac{1}{2}(|\tau_h^m| + |\tau_h^{m-1}|) \end{aligned}$$

Wir erhalten damit

$$\max_{Q_T} |\tau_{h,k}^m| \leq \frac{1}{12}k^2 \max_{Q_T} |\partial_t^3 u| + \max_{Q_T} |\tau_h^m|.$$

iii) Für die BDF(2)-Formel gilt

$$\begin{aligned} \frac{1}{2}k^{-1}\{3u^m - 4u^{m-1} + u^{m-2} + L_h u^m\} &= \frac{1}{2}k^{-1}\{3u^m - 4u^{m-1} + u^{m-2} - 2k\partial_t u^m\} \\ &\quad + L_h u^m - Lu^m. \end{aligned}$$

Taylor-Entwicklung um  $t_m$  liefert

$$3u^m - 4u^{m-1} + u^{m-2} - 2k\partial_t u^m = \frac{4}{3}k^3 \partial_t^3 u(\cdot, \eta^m)$$

mit gewissen Zwischenstellen  $\eta^m \in [t_{m-2}, t_m]$ . Damit erhalten wir

$$\max_{Q_T} |\tau_{h,k}^m| \leq \frac{2}{3}k^2 \max_{Q_T} |\partial_t^3 u| + \max_{Q_T} |\tau_h^m|.$$

Dies vervollständigt den Beweis. Q.E.D.

Die Lösung der ARWA (5.1.14) besitzt die explizite Darstellung

$$u(x, t) = \sum_{n=1}^{\infty} u_n^0 v^{(n)}(x) e^{-\lambda_n t}, \quad (x, t) \in Q_T, \tag{5.1.19}$$

mit den Eigenwerten und (orthonormierten) Eigenfunktionen des regulären „elliptischen“ Operators  $L = -a\Delta : V \subset L^2(\Omega) \rightarrow L^2(\Omega)$ ,

$$0 < \lambda_1 < \dots \leq \lambda_n \leq \dots \quad (n \in \mathbb{N}), \quad v^{(n)}(x) \in V : \quad Lv^{(n)} = \lambda_n v^{(n)},$$

und den Entwicklungskoeffizienten der Startwerte

$$u^0(x) = \sum_{n=0}^{\infty} u_n^0 v^{(n)}(x), \quad u_n^0 = (u^0, v^{(n)})_{\Omega}.$$

Diese Darstellung lässt sich wegen der gleichmäßigen Konvergenz der Reihen wie folgt

umformen:

$$\begin{aligned} u(x, t) &= \sum_{n=1}^{\infty} u_n^0 v^{(n)}(x) \left( \sum_{i=0}^{\infty} (-1)^i \frac{\lambda_n^i t^i}{i!} \right) = \sum_{i=0}^{\infty} (-1)^i \frac{t^i}{i!} \left( \sum_{n=1}^{\infty} u_n^0 \lambda_n^i v^{(n)}(x) \right) \\ &= \sum_{i=0}^{\infty} (-1)^i \frac{t^i}{i!} \left( \sum_{n=1}^{\infty} u_n^0 L^i v^{(n)}(x) \right) = \sum_{i=0}^{\infty} (-1)^i \frac{t^i}{i!} L^i \left( \sum_{n=1}^{\infty} u_n^0 v^{(n)}(x) \right) \\ &= \left( \sum_{i=0}^{\infty} (-1)^i \frac{1}{i!} L^i \right) u^0(x) =: e^{-tL} u^0(x). \end{aligned}$$

Die Definition der Operatorfunktion  $e^{-tL}$  über eine konvergente Taylor-Reihe lässt sich auf beliebige analytische Funktionen übertragen. Wir betonen, dass eine solche kompakte Lösungsdarstellung nur im Fall zeitlich konstanter Koeffizienten  $a$  möglich ist. Auf dem diskreten Zeitgitter gilt dann

$$u(\cdot, t_m) = e^{-kL} u(\cdot, t_{m-1}), \quad m \in \mathbb{N}. \quad (5.1.20)$$

Dies legt es nahe, den Zeitschritt  $t_{m-1} \rightarrow t_m$  mit Hilfe einer rationalen Approximation  $R(z) \approx e^z$  der Exponentialfunktion der „Ordnung“  $q+1$  anzusetzen,

$$R(z) = \frac{P(z)}{Q(z)} = e^z + \mathcal{O}(|z|^{q+1}), \quad z \leq 0, \quad (5.1.21)$$

mit geeigneten Polynomen  $P \in P_r$  und  $Q \in P_s$ , wobei natürlich  $Q$  auf  $z \in \mathbb{R}_-$  keine Nullstellen haben darf. Das Diskretisierungsschema lautet dann

$$U_h^m = R(-kA_h)U_h^{m-1} \quad \text{bzw.} \quad Q(-kA_h)U_h^m = P(-kA_h)U_h^{m-1}. \quad (5.1.22)$$

Die oben betrachteten Einschrittverfahren lassen sich in diesen Rahmen einordnen gemäß:

$$\begin{aligned} \text{„Expliziter Euler“ : } \quad R(z) &= 1 + z, \\ \text{„Impliziter Euler“ : } \quad R(z) &= (1 - z)^{-1}, \\ \text{„Crank-Nicolson“ : } \quad R(z) &= (1 + \frac{1}{2}z)(1 - \frac{1}{2}z)^{-1}. \end{aligned}$$

Durch die Ordnungsbedingung

$$e^z Q_{rs}(z) - P_{rs}(z) = \mathcal{O}(|z|^{r+s+1}), \quad z \leq 0, \quad (5.1.23)$$

für den Ansatz  $P_{rs} \in P_r$ ,  $Q_{rs} \in P_s$  wird man auf die sog. „Padé<sup>5</sup>-Schemata“ geführt. Diese sind eindeutig bestimmt und werden gewöhnlich in der sog. „Padé-Tafel“ dargestellt:

---

<sup>5</sup>Henri Eugène Padé (1785–1836): Französischer Mathematiker; Prof. in Poitiers und Bordeaux; entwickelte die sog. Padé-Approximation.

$\frac{1}{1}$	$\frac{1+z}{1}$	$\frac{1+z+\frac{1}{2}z^2}{1}$	$\frac{1+x+\frac{1}{2}z^2+\frac{1}{6}z^3}{1}$	...
$\frac{1}{1-z}$	$\frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}$	$\frac{1+\frac{2}{3}z+\frac{1}{6}z^2}{1-\frac{1}{3}z}$	$\frac{1+\frac{3}{4}z+\frac{1}{4}z^2+\frac{1}{24}z^3}{1-\frac{1}{4}z}$	...
$\frac{1}{1-z+\frac{1}{2}z^2}$	$\frac{1+\frac{1}{3}z}{1-\frac{2}{3}z+\frac{1}{6}z^2}$	$\frac{1+\frac{1}{2}z+\frac{1}{12}z^2}{1-\frac{1}{2}z+\frac{1}{12}z^2}$	...	...
...	...	...	$\frac{1+\frac{1}{2}z+\frac{1}{10}z^2+\frac{1}{120}z^3}{1-\frac{1}{2}z+\frac{1}{10}z^2-\frac{1}{120}z^3}$	...
...	...	...	...	...

Abbildung 5.1: Padé-Tafel.

Offensichtlich sind alle bisher betrachteten Einschrittschemata Padé-Formeln und damit in diesem Sinne ordnungsoptimal. Aus der Padé-Tafel erhalten wir nun weitere Zeitschrittverfahren höherer Ordnung. Dabei kommen aus Ökonomiegründen nur die „diagonalen“ oder „subdiagonalen“ Padé-Schemata in Frage; z. B. die folgenden impliziten Verfahren 3. bzw. 4. Ordnung:

$$(I + \frac{1}{3}kA_h)U_h^m = (I - \frac{2}{3}kA_h + \frac{1}{6}k^2A_h^2)U_h^{m-1} \quad (q=3), \\ (I + \frac{1}{2}kA_h + \frac{1}{12}k^2A_h^2)U_h^m = (I - \frac{1}{2}kA_h + \frac{1}{12}k^2A_h^2)U_h^{m-1} \quad (q=4).$$

Wir bemerken für die weitere Analyse, dass eine rationale Approximation  $R(z)$  der Exponentialfunktion (der Ordnung  $r \geq 1$ ) die folgende Eigenschaft hat:

$$|R(z)| \leq e^{\delta z}, \quad -1 \leq z \leq 0, \quad (5.1.24)$$

mit einem geeigneten  $\delta > 0$ . Die Wirkung der Zeitschrittschemata des Typs (5.1.22) lässt sich mit Hilfe der Spektralzerlegung der Matrix  $A_h$  wieder beschreiben durch:

$$U_h^m = \sum_{n=1}^N U_n^0 R(-k\lambda_n)^m v^{(n)}, \quad m \geq 1,$$

bzw. (mit der Euklidischen Vektornorm  $|\cdot|$ )

$$|U_h^m|^2 = \sum_{n=1}^N |U_n^0|^2 |R(-k\lambda_n)|^{2m}.$$

Ihr qualitatives Verhalten lässt sich also weitgehend durch die Eigenschaften der verwendeten rationalen Funktion  $R(z)$  charakterisieren. Wir stellen einige wichtige Bedingungen für die folgende Analyse zusammen.

i) Die *A-Stabilität*

$$|R(z)| \leq 1, \quad z \leq 0.$$

sichert die „Stabilität“ der Zeititeration  $\sup_{m \geq 0} |U_h^m| < \infty$ .

ii) Die *strenge A-Stabilität*

$$|R(z)| \leq 1 - ck, \quad z \leq -1,$$

sichert die Beschränktheit der diskreten Lösung auch im Fall inhomogener rechter Seiten,  $\sup_{m \geq 0} |U_h^m| < c \sup_{m \geq 0} |f^m|$ .

iii) Die *starke A-Stabilität*

$$|R(z)| \leq \kappa < 1, \quad z \leq -1,$$

sichert die (exponentielle) Dämpfung „hochfrequenter“ Lösungsanteile und macht das Verfahren robust gegenüber lokalen Störungen der Daten („Glättungseigenschaft“).

iv) Zur korrekten Wiedergabe von Schwingungsprozessen (im Ort oszillierenden Lösungen) sollte

$$R(\pm i) \sim 1$$

sein, um diese Schwingungen möglichst wenig zu dämpfen („numerischen Dissipativität“).

Offensichtlich können nur *implizite* Verfahren die gelisteten Eigenschaften haben. Das implizite Euler-Schema (und genauso alle sub-diagonalen Padé-Schemata) ist stark A-stabil (mit Limes  $\kappa = 0$ ), neigt aber zur Überdämpfung:  $|((1+i)^{-1})| = 1/\sqrt{2}$ . Dagegen ist das Crank-Nicolson-Schema (und genauso alle diagonalen Padé-Schemata) nur einfach A-stabil,

$$\lim_{z \rightarrow -\infty} \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z} = -1,$$

besitzt aber praktisch auch keine numerische Dissipation:  $|((1-i/2)(1+i/2)^{-1})| = 1$ . Die fehlende starke A-Stabilität hat nachteilige Konsequenzen im Fall von irregulären Anfangswerten  $u^0$  (z.B.: lokalen Temperaturspitzen). Die durch diese Anfangsdaten induzierten hochfrequenten Fehleranteile werden durch das Crank-Nicolson-Schema nur unzureichend ausgedämpft, so dass sich ein unphysikalisches Lösungsverhalten zeigen kann. Es sei daran erinnert, dass der kontinuierliche Differentialoperator stark dämpfend ist:

$$\|u(t)\| \leq e^{-\lambda_{\min} t} \|u^0\|, \quad t \geq 0,$$

mit dem kleinsten Eigenwert des Ortsoperators,  $\lambda_{\min} > 0$ .

Bei Verwendung des Crank-Nicolson-Schemas für Rechnungen über lange Zeiträume sollte es stabilisiert werden, um wenigstens *strenge A-Stabilität* zu sichern. Dies kann ohne Reduktion der Konsistenzordnung durch einen leichten  $k$ -abhängigen Schift erfolgen:

$$(I + \frac{1}{2}(1+ck)kA_h) U_h^m = (I - \frac{1}{2}(1-ck)kA_h) U_h^{m-1}. \quad (5.1.25)$$

Verfahren höherer Ordnung erfordern die Invertierung der Operatorfunktion  $Q(-kA_h)$ . Dies ist in der Regel zu teuer. Einerseits ist die Besetzungstruktur von  $Q(-kA_h)$  selbst bei Polynomgrad  $j = 2$  bereits deutlich dichter als die von  $A_h$ , andererseits würde das Arbeiten mit der Linearfaktorzerlegung  $Q(z) = (z - \mu)(z - \bar{\mu})$  die Verwendung (kostspielige)

liger) komplexer Arithmetik erfordern. Geeignet wären dagegen Schemata, bei denen das Nennerpolynom in reelle Linearfaktoren zerfällt:  $Q(z) = \prod_{j=1}^s (z - \mu_j)$ ,  $\mu_j \in \mathbb{R}$ . Durch diesen Ansatz sollten sich systematisch Verfahren mit günstigeren Eigenschaften als die der einfachen Basisschemata gewinnen lassen.

Ein Beispiel für einen solchen Ansatz ist die parameter-abhängige rationale Funktion

$$R_\theta(z) = \frac{(1 + \alpha\theta' z)(1 + \beta\theta z)^2}{(1 - \alpha\theta z)^2(1 - \beta\theta' z)} = e^z + O(|z|^3), \quad z \leq 0,$$

mit  $\theta = 1 - \frac{1}{2}\sqrt{2} = 0,292893\dots$ ,  $\theta' = 1 - 2\theta$  und beliebigen Werten  $\alpha \in (\frac{1}{2}, 1]$ ,  $\beta = 1 - \alpha$ . Das auf dieser rationalen Funktion basierende Schema ist wegen

$$|R_\theta(z)| < 1, \quad z < 0, \quad \lim_{z \rightarrow -\infty} |R_\theta(z)| = \frac{\beta}{\alpha} < 1.$$

stark A-stabil. Die Entwicklung

$$R_\theta(z) = 1 + z + \frac{1}{2}z^2\{1 - (\alpha - \beta)(2\theta^2 - 4\theta + 1)\} + \frac{1}{6}r(\theta, \alpha)z^3 + \mathcal{O}(|z|^4)$$

zeigt, dass für die obige Parameterwahl von der Ordnung  $\mathcal{O}(k^2)$  ist. Für die Güte dieser Approximation im Vergleich zu der des Crank-Nicolson-Schemas ist die Größe der führenden Fehlerkonstante  $r(\alpha)$  bestimmd. Eine Taylor-Entwicklung ergibt

$$\begin{aligned} r(\theta, \alpha) &= (18\theta' + 24\theta^3)\alpha^3 + (42\theta^2\theta' + 12\theta\theta'^2 + 30\theta^3)\alpha^2\beta \\ &\quad + (12\theta^3 + 30\theta^2\theta' + 24\theta\theta'^2 + 6\theta'^3)\alpha\beta^2 + (6\theta^2\theta' + 12\theta\theta'^2 + 6\theta'^3). \end{aligned}$$

Im betrachteten Bereich  $\{0,5 < \alpha \leq 1\}$  ist  $|r(\theta, \alpha)| \leq 0,5$ . Damit ist die Fehlerkonstante dieser Approximation nur in akzeptablem Maß größer als die entsprechende Fehlerkonstante  $\frac{1}{12}$  der Trapezregel. Das zugehörige Verfahren lässt sich in Form eines Teilschrittschemas schreiben (hier für den inhomogenen Fall),

*Teilschritt-θ-Verfahren (Fractional-Step-θ-Method):*

$$(I + \alpha\theta kA_h)U^{m-1+\theta} = (I - \beta\theta kA_h)U^{m-1} + \theta k f_h^{m-1}, \quad (5.1.26)$$

$$(I + \beta\theta' kA_h)U^{m-\theta} = (I - \alpha\theta' kA_h)U^{m-1+\theta} + \theta' k f_h^{m-\theta}, \quad (5.1.27)$$

$$(I + \alpha\theta kA_h)U^m = (I - \beta\theta kA_h)U^{m-\theta} + \theta k f_h^{m-\theta}. \quad (5.1.28)$$

Jeder der Teilschritte hat die Form eines geschifteten Crank-Nicolson-Schritts, so dass der Gesamtaufwand pro Zeitschritt dem von drei Crank-Nicolson-Schritten entspricht. Für den speziellen Wert

$$\alpha = (1 - 2\theta)(1 - \theta)^{-1} = 0,585786\dots$$

ist  $\alpha\theta = \beta\theta'$ , so dass die zu invertierenden Matrizen in den drei Teilschritten übereinstimmen, was z.B. bei der direkten Lösung der Gleichungssysteme ausgenutzt werden kann. Eine genaue Analyse des Abschneidefehlers des FS-Schemas zeigt, dass seine führende

Fehlerkonstante nur wenig größer als die von drei kombinierten Crank-Nicolson-Schritten ist:

$$\tau_k^m = \hat{c}k^2 + \mathcal{O}(k^3), \quad \hat{c}_{FS} \sim \hat{c}_{3 \times CN}.$$

Dies bedeutet, dass das FS-Schema gegenüber dem CN-Schema bzgl. Genauigkeit und Aufwand gleichwertig ist, aber über eine höhere Robustheit verfügt. Das FS-Schema hat sich in der Praxis als besonders geeignet zur Behandlung von parabolischen Problemen mit nicht notwendig regulären Daten und geringer natürlicher Eigendissipation erwiesen.

### 5.1.2 Stabilität und Konvergenz

Wir wollen nun die Stabilität und Konvergenz von Diskretisierungen der Wärmeleitungsgleichung untersuchen. Dabei bedienen wir uns exemplarisch verschiedener Techniken, die alle diskrete Analoga von Analysemethoden beim kontinuierlichen Problem sind.

#### i) „Maximumprinzipmethode“

Eine einfache, direkte Variante der Maximumprinzipmethode kann bei gewissen expliziten Differenzenschemata angewendet werden. Die Ortsdiskretisierung führt auf eine M-Matrix  $A_h$ . Für die explizite Euler-Formel

$$U_h^m = U_h^{m-1} - kA_h U_h^{m-1}$$

gilt dann wegen der Diagonaldominanz von  $A_h$ :

$$\begin{aligned} |U_n^m| &= |1 - ka_{nn}| |U_n^{m-1}| + k \sum_{\nu \neq n} |a_{n\nu}| |U_\nu^{m-1}| \\ &\leq |1 - ka_{nn}| |U_n^{m-1}| + ka_{nn} \max_\nu |U_\nu^{m-1}|. \end{aligned}$$

Unter der Schrittweitenbedingung

$$k \leq \max_n \{a_{nn}^{-1}\} \sim ch^2 \quad (5.1.29)$$

folgt daher die  $L^\infty$ -Stabilität des Verfahrens

$$\max_n |U_n^m| \leq \max_n |U_n^{m-1}| \leq \dots \leq \max_n |U_n^0|, \quad m \geq 1. \quad (5.1.30)$$

Im Fall des 5-Punkte-Schemas ist  $a_{nn} = 4h^{-2}$ , so dass die Stabilitätsbedingung (5.1.29) die Form

$$k \leq \frac{1}{4}h^2 \quad (5.1.31)$$

erhält.

**Satz 5.1 (Explizites Euler-Verfahren):** Unter der Schrittweitenbedingung (5.1.29) gilt für das explizite Euler-Verfahren die Konvergenzabschätzung

$$\max_{\bar{Q}_T} |U_h^m - u(\cdot, t_m)| \leq T \left\{ \frac{1}{2} k \max_{\bar{Q}_T} |\partial_t^2 u| + \max_{\bar{Q}_T} |\tau_h^m| \right\}. \quad (5.1.32)$$

mit dem örtlichen Abschneidefehler  $\tau_h^m = \mathcal{O}(h^q)$ .

**Beweis:** Der Fehler  $e^m := u(\cdot, t_m) - U^m$  genügt der Gleichung

$$k^{-1}(e^m - e^{m-1}) + A_h e^m = \tau_{h,k}^m$$

mit dem Abschneidefehler

$$\max_{\bar{Q}_T} |\tau_{h,k}^m| \leq \frac{1}{2} k \max_{\bar{Q}_T} |\partial_t^2 u| + \max_{\bar{Q}_T} |\tau_h^m|.$$

Das bei der Herleitung der Stabilitätsbedingung (5.1.29) verwendete Argument liefert

$$\max_{\Omega} |e^m| \leq \max_{\Omega} |e^{m-1}| + k \max_{\bar{Q}_T} |\tau_{h,k}^m|$$

Durch Iteration dieser Abschätzung folgt weiter wegen  $e^0 = 0$ :

$$\begin{aligned} \max_{\Omega} |e^m| &\leq k \sum_{\mu=1}^m |\tau_{h,k}^\mu| \\ &\leq \frac{1}{2} t_m k \max_{\bar{Q}_T} |\partial_t^2 u| + t_m \max_{\bar{Q}_T} |\tau_h^m|. \end{aligned}$$

Dies impliziert die behauptete Fehlerabschätzung. Q.E.D.

Eine wichtige Eigenschaft des kontinuierlichen Wärmeleitungsoperators ist seine „inverse Monotonie“, d. h.: Lösungen zu nicht-negativen Anfangsdaten und rechter Seite bleiben nicht-negativ. Diese Eigenschaft überträgt sich auf die diskretisierten Probleme, wenn die Systemmatrix  $A_h$  M-Matrix ist.

i) Für das explizite Euler-Verfahren folgt unter der Schrittweitenbedingung (5.1.29) aus  $U_n^{m-1} \geq 0$  und  $f_n^m \geq 0$  notwendig auch

$$U_n^m = (1 - ka_{nn})U_n^{m-1} + k \sum_{\nu \neq n} |a_{n\nu}| U_\nu^{m-1} + kf_n^m \geq 0.$$

ii) Für das implizite Euler-Verfahren ist im Falle  $U_n^{m-1} \geq 0$  und  $f_n^m \geq 0$

$$(I_h + kA_h)U_h^m = U_h^{m-1} + kf_n^m \geq 0.$$

Da mit  $A_h$  natürlich auch  $I_h + kA_h$  M-Matrix ist, gilt  $(I_h + kA_h)^{-1} \geq 0$ . Es folgt  $U_h^m \geq 0$ .

In beiden Fällen ist also auch das diskrete Schema „*verso-monoton*“. Dies ist i.a. für das Crank-Nicolson-Schema nicht der Fall.

ii) „Von Neumannsche Methode“ (Fourier-Methode)

Wir beschränken uns bei der Beschreibung der auf von Neumann<sup>6</sup> zurückgehenden Analysemethode auf den örtlich eindimensionalen Fall mit  $\Omega = (-\pi, \pi)$ ,

$$\partial_t u - \partial_x^2 u = 0 \quad \text{in } Q_T,$$

mit „periodischen“ Dirichlet-Randbedingungen

$$u(-\pi, t) = u(\pi, t), \quad t \geq 0.$$

In diesem Fall kann die Lösung der ARWA nach trigonometrischen Funktionen entwickelt werden (Fourier-Entwicklung). In komplexer Schreibweise lautet dies

$$u(x, t) = \sum_{\nu=0}^{\infty} a_{\nu}^0 e^{i\nu x} e^{\nu^2 t}, \quad (5.1.33)$$

mit den Entwicklungskoeffizienten  $a_{\nu}^0$  der Anfangsbedingung. Auf einem äquidistanten Punktgitter  $\{x_n = -\pi + nh, n = 0, \dots, N = 2\pi/h\}$  machen wir für die diskrete Lösung  $U_h^m = \{U_n^m, n = 0, \dots, N\}$ ,  $m \geq 0$ , den analogen Entwicklungsansatz

$$U_n^m = \sum_{\nu=0}^N a_{\nu}^m e^{i\nu nh} =: \sum_{\nu=0}^N a_{\nu}^0 \omega_{\nu}^m e^{i\beta_{\nu} n} \quad (5.1.34)$$

mit  $\beta_{\nu} := \nu h$  und zu bestimmenden Parametern  $\omega_{\nu} \in \mathbb{C}$ . Wir fragen nach der Stabilität für  $m \rightarrow \infty$  der Differenzendiskretisierung bzgl. der diskreten Spektralnorm

$$\|U_h^m\|_h := \left( \sum_{n=1}^N |a_n^m|^2 \right)^{1/2}.$$

Die Wirkung des (linearen) Differenzenschemas kann für jede einzelne Fourier-Komponente separat untersucht werden. Gesucht sind Bedingungen an  $k$  und  $h$ , unter denen  $|\omega_{\nu}| \leq 1$  ist für alle möglichen  $\beta_{\nu}$ . Dann liegt Stabilität vor in dem Sinne, dass

$$\|U_h^m\|_h^2 = \sum_{n=1}^N |a_n^0|^2 |\omega_{\nu}|^{2m} \leq \sum_{n=1}^N |a_n^0|^2 = \|U_h^0\|_h^2. \quad (5.1.35)$$

Wir führen diese Analyse wieder exemplarisch für das explizite Euler-Schema durch. Mit  $r := kh^{-2}$  gilt

$$U_n^{m+1} = r U_{n-1}^m + (1 - 2r) U_n^m + r U_{n+1}^m.$$

---

<sup>6</sup>John von Neumann (1903–1957): US-Amerikanischer Mathematiker österreichisch-ungarischer Herkunft; Studium in Budapest, Berlin und Zürich; 1927 Privatdozent in Berlin; arbeitete dann mit Hilbert in Göttingen; ab 1933 Prof. in Princeton (USA); bedeutende Beiträge zur mathematischen Logik, Funktionalanalysis, Quantenmechanik und Spieltheorie; gilt als einer der Väter der Informatik.

Einsetzen von  $U_n^m := \omega^m e^{i\beta n}$  ergibt

$$\omega^{m+1} e^{i\beta n} = r\omega^m e^{i\beta(n-1)} + (1-2r)\omega^m e^{i\beta n} + r\omega^m e^{i\beta(n+1)},$$

und nach Vereinfachung

$$\omega = re^{-i\beta} + (1-2r) + re^{i\beta}.$$

Es liegt Stabilität vor, wenn  $|\omega| \leq 1$  für beliebiges  $\beta$ . Unter Ausnutzung der Beziehungen

$$e^{i\beta} = \cos(\beta) + i \sin(\beta), \quad \cos(\beta) = 1 - 2 \sin^2(\frac{1}{2}\beta),$$

folgt

$$\begin{aligned} \omega &= r(e^{i\beta} + e^{-i\beta}) + (1-2r) = r(\cos(\beta) + i \sin(\beta) + \cos(\beta) - i \sin(\beta)) + (1-2r) \\ &= r(2 - 4 \sin^2(\frac{1}{2}\beta)) + (1-2r) = 1 - 4r \sin^2(\frac{1}{2}\beta). \end{aligned}$$

Stabilität liegt vor für

$$-1 \leq 1 - 4r \sin^2(\frac{1}{2}\beta) \leq 1 \quad \forall \beta,$$

was äquivalent ist zu

$$r \sin^2(\frac{1}{2}\beta) \leq \frac{1}{2}.$$

Dies führt auf die schon bekannte Stabilitätsbedingung

$$k \leq \frac{1}{2}h^2. \tag{5.1.36}$$

Die Fourier-Methode kann auch für „exotischere“ Differenzenformeln angewendet werden. Wir demonstrieren dies anhand des klassischen „Du Fort<sup>7</sup>-Frankel<sup>8</sup>-Verfahren“ (1953):

$$\frac{1}{2k} \left( U_n^{m+1} - U_n^{m-1} \right) - \frac{1}{h^2} \left( U_{n-1}^m - (U_n^{m+1} + U_n^{m-1}) + U_{n+1}^m \right) = 0. \tag{5.1.37}$$

Sein Abschneidefehler verhält sich wie

$$\max_{n,m} |\tau_n^m| = O(k^2/h + k^2 + h^2). \tag{5.1.38}$$

Die von Neumannsche Stabilitätsanalyse liefert für die Verstärkungsfaktoren die Darstellung ( $r := k/h^2$ )

<sup>7</sup>E.C. Du Fort (????–????): US-Amerikanischer Physiker; Publ. mit S. P. Frankel: Stability conditions in the numerical treatment of parabolic differential equations, Math. Tables and other Aids to Comput. (jetzt Math. Comput.) 7, 135–152 (1953).

<sup>8</sup>Stanley Phillips Frankel (1919–1978): US-Amerikanischer Informatiker; Mitglied der theoretischen Abteilung des „Manhattan Project“ in Los Alamos 1943 (Bau der ersten Atombombe); arbeitete mit dem ENIAC-Computer und in verschiedenen Instituten an der Nutzung mehrerer früher Computer-Systeme; Gruppenleiter am California Institute of Technology (CalTech) in Pasadena, USA; Entwicklung der sog. „Monte-Carlo-Methode“ in der statistischen Physik.

$$\omega = \frac{2r \cos(\beta) \pm \sqrt{1 - 4r^2 \sin^2(\beta)}}{1 + 2r}. \quad (5.1.39)$$

Dies impliziert, dass  $|\omega| \leq 1$  für alle  $\beta$ , d.h.: Das DuFord-Frankel-Schema ist unbedingt stabil. Analog zeigt man, dass das sog. „Richardson-Verfahren“

$$\frac{1}{2k}(U_n^{m+1} - U_n^{m-1}) - \frac{1}{h^2}(U_{n-1}^m - 2U_n^m + U_{n+1}^m) = 0 \quad (5.1.40)$$

unbedingt *instabil* ist. Obwohl es die „optimale“ Konsistenzordnung  $O(h^2 + k^2)$  besitzt, ist es also praktisch unbrauchbar. Dies ist nicht verwunderlich, da dieses Schema ein Derivat der Mittelpunktsregel mit dem Stabilitätspolynom  $\pi(z, h\lambda)$  und den Wurzeln  $z_{1,2} = h\lambda \pm (h^2\lambda^2 + 1)^{1/2}$  ist.

Die von Neumann'sche Fourier-Methode zur Stabilitätsanalyse von Differenzenschemata ist auf den Fall periodischer Dirichlet-Randbedingungen bzw. den Grenzfall von „Ganzraum-Problemen“ ( $\Omega = \mathbb{R}^1$ ) beschränkt und erfordert äquidistante Ortsgritter. Für allgemeinere Ortsdiskretisierungen anwendbar ist die im folgenden präsentierte „Spektral-methode“.

### iii) Spektral-Methode:

Die symmetrische, positiv definite Matrix  $A_h$  habe die Eigenwerte und zugehörigen ( $l_2$ -orthonormierten) Eigenvektoren

$$0 < \lambda_1 \leq \dots \leq \lambda_N, \quad \{w^{(n)}, n = 1, \dots, N\}.$$

Jede Gitterfunktion besitzt dann eine Entwicklung der Form

$$U_h^m = \sum_{n=1}^N a_n w^{(n)}, \quad a_n = \langle U_h^m, w^{(n)} \rangle.$$

Dabei ist das Skalarprodukt  $\langle \cdot, \cdot \rangle$  für eine FD-Diskretisierung im Ort wieder ein diskretes Analogon der kontinuierlichen  $L^2$ -Norm,

$$\langle v, w \rangle := \sum_{n=1}^N h_n^2 v_n w_n,$$

und für eine FE-Diskretisierung gerade diese:  $\langle v, w \rangle := (v, w)_\Omega$ . Entsprechend sind die zugehörigen Normen  $\|u\| := \langle v, v \rangle^{1/2}$  definiert. Wir analysieren im folgenden isoliert den Zeitschrittfehler im Rahmen der Linienmethode.

**Satz 5.2 (Glättungseigenschaft):** *Jedes stark A-stabile Einschrittschema vom Typ (5.1.22) der Ordnung r besitzt die Glättungseigenschaft:*

$$\|U_h^m - u_h^m\| \leq c \frac{k^r}{t_m^r} \|u_h^0\|, \quad m > 0. \quad (5.1.41)$$

**Beweis:** Nach Voraussetzung ist  $\sup_{z \geq 0} |R(-z)| \leq 1$ ,  $\lim_{z \rightarrow \infty} |R(-z)| \leq \omega < 1$  und

$$|R(-z) - e^{-z}| \leq c|z|^{r+1}, \quad 0 \leq z \leq 1.$$

O.B.d.A. nehmen wir an, dass  $|R(-z)| \leq \omega < 1$  für  $z \geq 1$ . Wir verwenden wieder das Spektralargument von oben. Mit den Eigenwerten  $0 < \lambda_1 \leq \dots \leq \lambda_N$  von  $A_h$  und einem zugehörigen Orthonormalsystem  $\{w^{(n)}, n = 1, \dots, N\}$  von Eigenvektoren gilt wieder für den Anfangswert

$$u_h^0 = \sum_{n=1}^N \alpha_n w^{(n)}$$

die Abschätzung ( $\tau_n := k\lambda_n$ )

$$\begin{aligned} |U_h^m - u_h^m|^2 &= \sum_{n=1}^N \alpha_n^2 \left| R(-k\lambda_n)^m - e^{-mk\lambda_n} \right|^2 \\ &= \sum_{\tau_n \leq 1} \dots + \sum_{\tau_n > 1} \dots \end{aligned}$$

Für die erste Summe rechts gilt mit einem geeigneten  $\delta > 0$ :

$$\begin{aligned} \sum_{\tau_n \leq 1} \dots &= \sum_{\tau_n \leq 1} \left| R(-\tau_n) - e^{-\tau_n} \right|^2 \left| \sum_{\mu=0}^{m-1} R(-\tau_n)^{m-1-\mu} e^{-\mu\tau_n} \right|^2 \alpha_n^2 \\ &\leq c \sum_{\tau_n \leq 1} \tau_n^{2r+2} m^2 e^{-2\delta(m-1)\tau_n} \alpha_n^2 \leq cm^{-2r} |u_h^0|^2. \end{aligned}$$

Für die zweite Summe rechts gilt entsprechend mit einem  $\delta > 0$ :

$$\begin{aligned} \sum_{\tau_n > 1} \dots &\leq 2 \sum_{\tau_n > 1} \alpha_n^2 \left\{ |R(-\tau_n)|^{2m} + e^{-2m\tau_n} \right\} \\ &\leq ce^{-\delta m} \sum_{\tau_n > 1} \alpha_n^2 \leq cm^{-2r} |u_h^0|^2. \end{aligned}$$

Kombination dieser beiden Abschätzungen liefert wegen  $m = t_m/k$ :

$$|U_h^m - u_h^m|^2 \leq c \frac{k^{2r}}{t_m^{2r}} |u_h^0|^2. \quad (5.1.42)$$

Dies vervollständigt den Beweis.

Q.E.D.

Das populäre Crank-Nicolson-Schema

$$U_h^m = (I_h + \frac{1}{2}kA_h)^{-1} (I_h - \frac{1}{2}kA_h) U_h^{m-1}$$

besitzt als nicht *stark A-stabiles* Schema nicht die volle Glättungseigenschaft. Wir wollen diesen Defekt anhand einer Modellbetrachtung erläutern.

Sei

$$U_h^0 = \sum_{n=1}^N \alpha_n^0 w^{(n)} \quad \Rightarrow \quad U_h^m = \sum_{n=1}^N \alpha_n^0 \left( \frac{1 - k\lambda_n/2}{1 + k\lambda_n/2} \right)^m w^{(n)}.$$

Die Lösungskomponente zur höchsten Frequenz  $\Lambda = \lambda_N$  verhält sich wie

$$\omega^m = \left( \frac{1 - k\Lambda/2}{1 + k\Lambda/2} \right)^m \sim e^{-t_m \Lambda},$$

was dem Abfall der „exakten“ Lösung entspricht.

i) Für  $k\Lambda < 2$  ( $\Leftrightarrow k \sim h^2$ ) ist

$$|\omega| \leq e^{-\delta}, \quad \delta > 0,$$

was den korrekten exponentiellen Abfall  $e^{-\delta m}$  impliziert.

ii) Im Fall  $k\Lambda \sim k/h^2 \sim 4/h$  ( $\Leftrightarrow k \sim h$ ) ist

$$\omega \sim -\frac{1 - h/2}{1 + h/2},$$

was oszillierendes Verhalten  $(-1)^m e^{-hm}$  impliziert.

Zur Dämpfung dieser Oszillationen in den „hochfrequenten“ Komponenten können folgende Strategien verwendet werden:

a) Mittelbildung:

$$\tilde{U}_h^1 = \frac{1}{4} \{ U_h^0 + 2U_h^1 + U_h^2 \} = \sum_{n=1}^N \alpha_n^0 \left\{ \frac{1}{4} + \frac{1}{2} \frac{1 - \frac{1}{2}k\lambda_n}{1 + \frac{1}{2}k\lambda_n} + \frac{1}{4} \left( \frac{1 - \frac{1}{2}k\lambda_n}{1 + \frac{1}{2}k\lambda_n} \right)^2 \right\} w^{(n)}.$$

Auswertung des Ausdrucks in der Klammer ergibt

$$\frac{1 + k\lambda_n + \frac{1}{2}k^2\lambda_n^2 + 2 - \frac{1}{2}k^2\lambda_n^2 + 1 - k\lambda_n + \frac{1}{4}k^2\lambda_n^2}{4(1 + \frac{1}{2}k\lambda_n)^2} = \frac{1}{(1 + \frac{1}{2}k\lambda_n)^2}$$

und somit

$$\tilde{U}_h^1 = \sum_{n=1}^N \frac{\alpha_n^0}{(1 + \frac{1}{2}k\lambda_n)^2} w^{(n)}.$$

b) Euler-Dämpfung: Der Zeitschrittprozess wird mit zwei impliziten Euler-Schritten mit halber Schrittlänge gestartet. Dies ergibt

$$\tilde{U}_h^1 = \sum_{n=1}^N \alpha_n^0 \left( \frac{1}{1 + \frac{1}{2}k\lambda_n} \right)^2 w^{(n)}.$$

**Satz 5.3 (Gedämpftes Crank-Nicolson-Verfahren):** Das durch zwei Euler-Schritte gedämpfte Crank-Nicolson-Verfahren besitzt die Glättungseigenschaft:

$$\|U_h^m - u_h^m\| \leq c \frac{k^2}{t_m^2} \|u_h^0\|. \quad (5.1.43)$$

**Beweis:** Für  $z \geq 0$  gilt

$$\left| e^{-z} - \frac{1 - \frac{1}{2}z}{1 + \frac{1}{2}z} \right| \leq c \frac{z^3}{1 + \frac{1}{2}z}, \quad \left| e^{-z} - \frac{1}{1+z} \right| \leq c \frac{z^2}{1+z}.$$

Wir verwenden dies in der folgenden Abschätzung:

$$\|U_h^m - u_h^m\|^2 = \sum_{n=1}^N \alpha_n^2 \left( \left( \frac{1 - \frac{1}{2}k\lambda_n}{1 + \frac{1}{2}k\lambda_n} \right)^{m-2} \left( \frac{1}{1 + \frac{1}{2}k\lambda_n} \right)^2 - e^{-mk\lambda_n} \right)^2$$

Wir bezeichnen den Inhalt der äußeren Klammer mit  $\sigma_n^m$  und setzen  $\tau_n := k\lambda_n$ . Es gilt

$$\begin{aligned} \sigma_n^m &= e^{-(m-2)\tau_n} \left\{ e^{-2\tau_n} - \left( \frac{1}{1 + \frac{1}{2}\tau_n} \right)^2 \right\} + \left\{ e^{-(m-2)\tau_n} - \left( \frac{1 - \frac{1}{2}\tau_n}{1 + \frac{1}{2}\tau_n} \right)^{m-2} \right\} \left( \frac{1}{1 + \frac{1}{2}\tau_n} \right)^2 \\ &= e^{-(m-2)\tau_n} \left( e^{-\tau_n} - \frac{1}{1 + \frac{1}{2}\tau_n} \right) \left( e^{-\tau_n} + \frac{1}{1 + \frac{1}{2}\tau_n} \right) \\ &\quad + \left( e^{-\tau_n} - \frac{1 - \frac{1}{2}\tau_n}{1 + \frac{1}{2}\tau_n} \right) \left( \frac{1}{1 + \frac{1}{2}\tau_n} \right)^2 \sum_{\mu=0}^{m-3} e^{-\mu\tau_n} \left( \frac{1 - \frac{1}{2}\tau_n}{1 + \frac{1}{2}\tau_n} \right)^{n-3-\mu}. \end{aligned}$$

i) Fall  $\tau_n \leq 2$ :

$$\frac{1 - \frac{1}{2}\tau_n}{1 + \frac{1}{2}\tau_n} \leq e^{-\frac{1}{2}\tau_n}.$$

Dies sieht man wie folgt: Wegen  $e^z \geq z$  gilt  $-1 + ze^{-z} \leq 0$ . Die Funktion  $f(z) := 1 - z - (1+z)e^{-z}$  hat die Eigenschaften  $f(0) = 0$  und  $f'(z) = -1 + ze^{-z} \leq 0$  und folglich  $f(z) \leq 0$ . Damit erschließen wir

$$\begin{aligned} |\sigma_n^m| &\leq c \left\{ e^{-m\tau_n} \tau_n^2 + \tau_n^3 e^{-m\tau_n/2} \sum_{\mu=0}^{m-3} e^{-\mu\tau_n/2} \right\} \\ &\leq c \left\{ e^{-m\tau_n} \tau_n^2 + \tau_n^3 e^{-m\tau_n/2} \frac{1 - e^{-(m-2)\tau_n/2}}{1 - e^{-\tau_n/2}} \right\} \leq \frac{c}{m^2} = c \frac{k^2}{t_m^2}. \end{aligned}$$

ii) Fall  $\tau_n > 2$ :

$$\left| \frac{1 - \tau_n/2}{1 + \tau_n/2} \right| \leq e^{-2/\tau_n}.$$

Damit erschließen wir:

$$\begin{aligned} |\sigma_n^m| &\leq c \left\{ e^{-m\tau_n} + e^{-2(m-2)/\tau_n} \frac{1}{\tau_n^2} \right\} \\ &\leq c \left\{ (m\tau_n)^2 e^{-m\tau_n} \frac{1}{m^2} + e^{-2m/\tau_n} \left(\frac{m}{\tau_n}\right)^2 \frac{1}{m^2} \right\} \leq c \frac{1}{m^2} = c \frac{k^2}{t_m^2}. \end{aligned}$$

Zusammenfassung der Resultate (i) und (ii) liefert nun:

$$\|U_h^m - u_h^m\|^2 \leq c \frac{k^4}{t_m^4} \sum_{n=1}^N \alpha_n^2. \quad (5.1.44)$$

Dies vervollständigt den Beweis.

Q.E.D.

Auch die Spektralmethode ist auf den Fall parabolischer Probleme mit zeitunabhängigen, selbstadjungierten Operatoren wie dem Laplace-Operator  $\Delta$  beschränkt. Die weitreichendste Analysetechnik ist die sog. „Energie-Methode“ (Hilbertraum-Methode), welche auch für Probleme mit unsymmetrischen Operatoren mit zeitabhängigen Koeffizienten anwendbar ist. Wir demonstrieren diese Technik hier aber nur für die vorliegende Modellsituation.

#### iv) Energie-Methode:

Wir betrachten das populäre Crank-Nicolson-Schema. Für Funktionen  $(v_n)_{n=1}^N$  auf einem äquidistanten Quadratgitter sind

$$(v, w)_h := h^d \sum_{n=1}^N v_n w_n, \quad \|v\|_h := (v, v)_h^{1/2},$$

diskrete Analoga des  $L^2$ -Skalarprodukts und der zugehörigen  $L^2$ -Norm.

**Satz 5.4 (Crank-Nicolson-Verfahren):** *Das Crank-Nicolson-Verfahren hat für hinreichend glatte Lösung  $u$  den globalen Diskretisierungsfehler*

$$\max_{Q_T} \|u - U_h\|_h \leq c(u) T \{h^2 + k^2\}, \quad (5.1.45)$$

mit einer Konstante  $c(u) \approx \max_{\bar{Q}_T} \{|\partial_t^3 u| + a|\nabla^4 u|\}$ .

**Beweis:** Für den Fehler  $e^m := u^m - U^m$  gilt

$$k^{-1}(e^m - e^{m-1}) + \frac{1}{2} A_h(e^m + e^{m-1}) = \tau_{h,k}^m.$$

Multiplikation dieser Identität mit  $e^m + e^{m-1}$  und Summation über  $m$  ergibt

$$k^{-1} \{\|e^m\|_h^2 - \|e^{m-1}\|_h^2\} + \frac{1}{2} (A_h(e^m + e^{m-1}), e^m + e^{m-1})_h = (\tau_{h,k}^m, e^m + e^{m-1})_h.$$

Der kleinste Eigenwert von  $A_h$  ist  $\lambda > 0$ . Damit erschließen wir

$$k^{-1} \{ \|e^m\|_h^2 - \|e^{m-1}\|_h^2 \} + \frac{1}{2} \lambda \|e^m + e^{m-1}\|_h^2 \leq \frac{1}{2} \lambda \|e^m + e^{m-1}\|_h^2 + \frac{1}{2} \lambda^{-1} \|\tau_{h,k}^m\|_h^2,$$

bzw.

$$\|e^m\|_h^2 \leq \|e^{m-1}\|_h^2 + \frac{1}{2} \lambda^{-1} k \|\tau_{h,k}^m\|_h^2.$$

Wir summieren nun über  $\mu = m, \dots, 1$  und erhalten

$$\|e^m\|_h^2 \leq \|e^0\|_h^2 + \frac{1}{2} \lambda^{-1} k \sum_{\mu=1}^m \|\tau_{h,k}^\mu\|_h^2.$$

Mit  $e^0 = 0$  und der obigen Abschätzung für den Abschneidefehler folgt schließlich die Behauptung. Q.E.D.

## 5.2 FE-Galerkin-Verfahren für parabolische Probleme

Wir diskutieren nun die Rothe-Methode zur Lösung des Problems

$$\partial_t u - \Delta u = f \quad \text{in } Q_T = \Omega \times [0, T], \quad (5.2.46)$$

mit den Nebenbedingungen  $u|_{t=0} = u^0$  und  $u|_{\partial\Omega} = 0$ . Da die folgende Analyse exemplarischen Charakter hat, betrachten wir nur das implizite Euler-Schema. Dieses lautet angewendet auf das kontinuierliche Problem (5.2.46)

$$k_m^{-1}(U^m - U^{m-1}) - \Delta U^m = \bar{f}^m, \quad U^0 := u^0, \quad (5.2.47)$$

wobei die rechte Seite im zeitlichen Mittel ausgewertet wird gemäß

$$\bar{f}^m := k_m^{-1} \int_{t_{m-1}}^{t_m} f(t) dt = f^m + \mathcal{O}(k_m).$$

Die Zeitschrittweite  $k_m := t_m - t_{m-1}$  darf hier variieren, um eine möglichst gute Anpassung an die Lösungseigenschaften zu erreichen. Mechanismen zur adaptiven Wahl der Zeitschrittweiten auf der Basis von a posteriori Fehlerabschätzungen werden weiter unten diskutiert.

Die einzelnen Zeitschritte seien mit Hilfe eines FE-Verfahrens mit Ansatzräumen  $V_h^m \subset V$  auf möglicherweise von Zeit zu Zeit wechselnden Gittern  $T_h^m$  diskretisiert:

$$(U_h^m, \varphi) + k_m(\nabla U_h^m, \nabla \varphi) = (U_h^{m-1}, \varphi) + k_m(\bar{f}^m, \varphi) \quad \forall \varphi \in V_h^m. \quad (5.2.48)$$

Die Varianz der Ortsdiskretisierungen im Verlaufe der Zeititeration ermöglicht die dynamische adaptive Anpassung der Ortsgitter an die momentane Lösungsstruktur. In Ope-

ratorschreibweise lautet das Schema (5.2.47)

$$(I_h^m + k_m \mathcal{A}_h^m) U_h^m = P_h^m U_h^{m-1} + k_m P_h^m \bar{f}^m, \quad U_h^0 = P_h^0 u^0, \quad (5.2.49)$$

mit der  $L^2$ -Projektion  $P_h^m$  auf  $V_h^m$ . Bezüglich der üblichen Knotenbasen  $\{\varphi_h^{m,n}, n = 1, \dots, N_m = \dim V_h^m\}$  der Räume  $V_h^m$  lässt sich dies als lineares Gleichungssystem zur Bestimmung der zugehörigen Knotenwertvektoren  $x_h^m \in \mathbb{R}^{N_m}$  schreiben. Dazu führen wir zusätzlich zu Massematrizen, Steifigkeitsmatrizen und Lastvektoren

$$M_h^m := ((\varphi_h^{m,i}, \varphi_h^{m,j}))_{i,j=1}^{N_m}, \quad A_h^m := ((\nabla \varphi_h^{m,i}, \nabla \varphi_h^{m,j}))_{i,j=1}^{N_m}, \quad b_h^m := ((\bar{f}^m, \varphi_h^{m,j}))_{j=1}^{N_m}$$

auf dem Gitter  $\mathbf{T}_h^m$  noch Transfermatrizen zwischen den Räumen  $V_h^{m-1}$  und  $V_h^m$  ein:

$$M_h^{m-1,m} := ((\varphi_h^{m-1,j}, \varphi_h^{m,n}))_{j,n=1}^{N_{m-1}, N_m}.$$

Damit schreibt sich

$$(U_h^{m-1}, \varphi_h^{m,n}) = \sum_{j=1}^{N_{m-1}} x_j^{m-1} (\varphi_h^{m-1,j}, \varphi_h^{m,n}) = M_h^{m-1,m} x_h^{m-1}$$

und folglich

$$(M_h^m + k_m A_h^m) x_h^m = M_h^{m-1,m} x_h^{m-1} + k_m M_h^m b_h^m. \quad (5.2.50)$$

Wir wollen dieses Verfahren im folgenden im Hinblick auf Stabilität, Konvergenz sowie a priori und a posteriori Fehlerabschätzung untersuchen.

### 5.2.1 A priori Konvergenzabschätzungen

Der natürliche Ansatz zur Analyse von FE-Diskretisierungen ist die „Energie-Methode“. Wir geben zunächst einen einfachen Beweis für das implizite Euler-Verfahren unter realistischen Annahmen an die Regularität der Lösung. Wir setzen (Die Zellen der Zerlegung  $\mathbf{T}_h^m$  werden ab jetzt mit  $K$  bezeichnet.)

$$h_m := \max_{K \in \mathbf{T}_h^m} \text{diam}(K), \quad k = \max_{1 \leq m \leq M} k_m$$

und  $e_h^m := U_h^m - u^m$  mit  $u^m := u(\cdot, t_m)$ .

**Satz 5.5 (Implizites Euler-Verfahren):** Für das implizite Euler-Schema in Verbindung mit einer FE-Diskretisierung 2. Ordnung gelten die folgenden Fehlerabschätzungen:

i) Für beliebig variierende Ortsdiskretisierung:

$$\max_{1 \leq m \leq M} \|e_h^m\| \leq c T^{1/2} \max_{0 \leq m \leq M} \left\{ \frac{h_m^2}{k_m^{1/2}} \|\nabla^2 u^m\| \right\} + c \left( \sum_{m=1}^M k_m^2 \int_{t_{m-1}}^{t_m} \|\nabla \partial_t u\|^2 dt \right)^{1/2}; \quad (5.2.51)$$

ii) Im Spezialfall  $V_h^m = V_h$  gleich für alle  $m$ :

$$\max_{1 \leq m \leq M} \|e_h^m\| \leq c T^{1/2} \max_{0 \leq m \leq M} \{h_m^2 \|\nabla^2 u^m\| + c \left( \sum_{m=1}^M k_m^2 \int_{t_{m-1}}^{t_m} \|\nabla \partial_t u\|^2 dt \right)^{1/2}\}. \quad (5.2.52)$$

**Beweis:** Wir bezeichnen mit  $R_h^m u \in V_h^m$  die „elliptische“ Ritz-Projektion der Lösung  $u^m$  zum Zeitlevel  $t_m$  auf den Finite-Elemente-Raum  $V_h^m$ , definiert durch

$$(\nabla R_h^m v, \nabla \varphi_h) = (\nabla v, \nabla \varphi_h) \quad \forall \varphi_h \in V_h^m. \quad (5.2.53)$$

Für deren Fehler gilt

$$\|v - R_h^m v\| + h_m \|\nabla(v - R_h^m v)\| \leq ch_m^2 \|\nabla^2 v\|. \quad (5.2.54)$$

Wir betrachten nun zunächst die Differenz  $\eta_h^m := U_h^m - R_h^m u^m$ . Für beliebiges  $\varphi_h \in V_h^m$  ist dann unter Ausnutzung der Identität

$$(U_h^m - U_h^{m-1}, \varphi_h) + k_m(\nabla U_h^m, \nabla \varphi_h) = k_m(\bar{f}^m, \varphi_h)$$

und der Projektionseigenschaft von  $R_h^m$ :

$$\begin{aligned} & (\eta_h^m - \eta_h^{m-1}, \varphi_h) + k_m(\nabla \eta_h^m, \nabla \varphi_h) \\ &= k_m(\bar{f}^m, \varphi_h) - (R_h^m u^m - R_h^{m-1} u^{m-1}, \varphi_h) - k_m(\nabla R_h^m u^m, \nabla \varphi_h) \\ &= k_m(\bar{f}^m, \varphi_h) - (R_h^m u^m - R_h^{m-1} u^{m-1}, \varphi_h) - k_m(\nabla u^m, \nabla \varphi_h). \end{aligned}$$

Wir setzen nun  $\varphi_h := \eta_h^m$  und erhalten mit Hilfe der Identität  $(a - b)a = \frac{1}{2}a^2 - \frac{1}{2}b^2 + \frac{1}{2}(a - b)^2$  die Beziehung

$$\begin{aligned} & \frac{1}{2} \|\eta_h^m\|^2 - \frac{1}{2} \|\eta_h^{m-1}\|^2 + \frac{1}{2} \|\eta_h^m - \eta_h^{m-1}\|^2 + k_m \|\nabla \eta_h^m\|^2 \\ &= k_m(\bar{f}^m, \eta_h^m) - (R_h^m u^m - R_h^{m-1} u^{m-1}, \eta_h^m) - k_m(\nabla u^m, \nabla \eta_h^m) \\ &= k_m(\bar{f}^m, \eta_h^m) - (u^m - u^{m-1}, \eta_h^m) - k_m(\nabla u^m, \nabla \eta_h^m) \\ &\quad + (u^m - R_h^m u^m, \eta_h^m) - (u^{m-1} - R_h^{m-1} u^{m-1}, \eta_h^{m-1}) \\ &\quad + (u^{m-1} - R_h^{m-1} u^{m-1}, \eta_h^{m-1} - \eta_h^m). \end{aligned} \quad (5.2.55)$$

Weiter haben wir

$$\begin{aligned} & k_m(\bar{f}^m, \eta_h^m) - (u^m - u^{m-1}, \eta_h^m) - k_m(\nabla u^m, \nabla \eta_h^m) \\ &= \int_{t_{m-1}}^{t_m} (f - \partial_t u, \eta_h^m) dt - k_m(\nabla u^m, \nabla \eta_h^m) \\ &= \int_{t_{m-1}}^{t_m} (\nabla u, \nabla \eta_h^m) dt - k_m(\nabla u^m, \nabla \eta_h^m) \\ &= \int_{t_{m-1}}^{t_m} (t - t_{m-1})(\nabla \partial_t u, \nabla \eta_h^m) dt \\ &\leq k_m \|\nabla \eta_h^m\|^2 + \frac{1}{4} k_m^2 \int_{t_{m-1}}^{t_m} \|\partial_t \nabla u\|^2 dt \end{aligned} \quad (5.2.56)$$

sowie

$$(u^{m-1} - R_h^{m-1} u^{m-1}, \eta_h^{m-1} - \eta_h^m) \leq \frac{1}{2} \|\eta_h^{m-1} - \eta_h^m\|^2 + ch_m^4 \|\nabla^2 u^{m-1}\|^2 \quad (5.2.57)$$

oder

$$(u^{m-1} - R_h^{m-1} u^{m-1}, \eta_h^{m-1} - \eta_h^m) \leq k_m^{-1} \|\eta_h^{m-1} - \eta_h^m\|^2 + ck_m h_m^4 \|\nabla^2 u^{m-1}\|^2 \quad (5.2.58)$$

i) Wir betrachten zunächst den Fall allgemein variiender Ortsdiskretisierung. Kombination der Beziehungen (5.2.55), (5.2.56) und (5.2.57) und Absorption von Termen in die linke Seite ergibt

$$\begin{aligned} \frac{1}{2} \|\eta_h^m\|^2 - \frac{1}{2} \|\eta_h^{m-1}\|^2 &\leq (u^m - R_h^m u^m, \eta_h^m) - (u^{m-1} - R_h^{m-1} u^{m-1}, \eta_h^{m-1}) \\ &\quad + \frac{1}{4} k_m^2 \int_{t_{m-1}}^{t_m} \|\partial_t \nabla u\|^2 dt + ch_m^4 \|\nabla^2 u^{m-1}\|^2. \end{aligned}$$

Wir wenden diese Abschätzung rekursiv für  $m, m-1, \dots, 1$  an und finden

$$\begin{aligned} \|\eta_h^m\|^2 &\leq \|\eta_h^0\|^2 + 2(u^m - R_h^m u^m, \eta_h^m) - 2(u^0 - R_h^0 u^0, \eta_h^0) \\ &\quad + c \sum_{\mu=1}^m \left\{ k_\mu^2 \int_{t_{\mu-1}}^{t_\mu} \|\partial_t \nabla u\|^2 dt + ch_\mu^4 \|\nabla^2 u^{\mu-1}\|^2 \right\} \end{aligned}$$

bzw.

$$\begin{aligned} \|\eta_h^m\|^2 &\leq \|\eta_h^0\|^2 + \frac{1}{2} \|\eta_h^m\|^2 + \frac{1}{2} \|u^m - R_h^m u^m\|^2 - (u^0 - R_h^0 u^0, \eta_h^0) \\ &\quad + c \sum_{\mu=1}^m k_\mu^2 \int_{t_{\mu-1}}^{t_\mu} \|\partial_t \nabla u\|^2 dt + ct_m \max_{0 \leq \mu \leq m} \{k_m^{-1} h_\mu^4 \|\nabla^2 u^\mu\|^2\}. \end{aligned}$$

Mit Hilfe der Abschätzung

$$\|u^\mu - R_h^\mu u^\mu\| + \|u^\mu - P_h^\mu u^\mu\| \leq ch_\mu^2 \|\nabla^2 u^\mu\|, \quad \mu = 1, \dots, m, \quad (5.2.59)$$

folgt

$$\|e_h^m\| \leq \|\eta_h^m\| + \|u^m - R_h^m u^m\| \leq \|\eta_h^m\| + ch_m^2 \|\nabla^2 u^m\|, \quad (5.2.60)$$

$$\|\eta_h^0\| \leq \|u^0 - R_h^0 u^0\| + \|u^0 - P_h^0 u^0\| \leq ch_0^2 \|\nabla^2 u^0\| \quad (5.2.61)$$

und damit schließlich die Fehlerabschätzung (5.2.51):

$$\|e_h^m\|^2 \leq ct_m \max_{0 \leq \mu \leq m} \{k_\mu^{-1} h_\mu^4 \|\nabla^2 u^\mu\|^2\} + c \sum_{\mu=1}^m k_\mu^2 \int_{t_{\mu-1}}^{t_\mu} \|\nabla \partial_t u\|^2 dt. \quad (5.2.62)$$

ii) Wir nehmen nun an, dass  $V_h^m = V_h$  bzw.  $R_h^m = R_h$  für  $m = 1, \dots, M$ . Kombination der Beziehungen (5.2.55), (5.2.56) und (5.2.58) und Absorption von Termen in die linke Seite ergibt:

$$\begin{aligned} \frac{1}{2}\|\eta_h^m\|^2 - \frac{1}{2}\|\eta_h^{m-1}\|^2 &\leq (u^m - R_h^m u^m, \eta_h^m) - (u^{m-1} - R_h^{m-1} u^{m-1}, \eta_h^{m-1}) \\ &\quad + \frac{1}{4}k_m^2 \int_{t_{m-1}}^{t_m} \|\partial_t \nabla u\|^2 dt + k_m^{-1}\|\eta_h^{m-1} - \eta_h^m\|^2 + ck_m h_m^4 \|\nabla^2 u^{m-1}\|^2. \end{aligned}$$

Wir wenden diese Abschätzung rekursiv für  $m, m-1, \dots, 1$  an und finden:

$$\begin{aligned} \|\eta_h^m\|^2 &\leq \|\eta_h^0\|^2 + 2(u^m - R_h^m u^m, \eta_h^m) - 2(u^0 - R_h^0 u^0, \eta_h^0) \\ &\quad + c \sum_{\mu=1}^m \left\{ k_\mu^2 \int_{t_{\mu-1}}^{t_\mu} \|\partial_t \nabla u\|^2 dt + k_m h_\mu^4 \|\nabla^2 u^{\mu-1}\|^2 + k_\mu^{-1} \|\eta_h^{\mu-1} - \eta_h^\mu\|^2 \right\} \end{aligned}$$

bzw. mit den Abschätzungen (5.2.59), (5.2.60) und (5.2.61),

$$\begin{aligned} \|e_h^m\|^2 &\leq ct_m \max_{0 \leq \mu \leq m} \{h_\mu^2 \|\nabla^2 u^\mu\|^2\} + \sum_{\mu=1}^m k_\mu^{-1} \|\eta_h^{\mu-1} - \eta_h^\mu\|^2 \\ &\quad + c \sum_{\mu=1}^m k_\mu^2 \int_{t_{\mu-1}}^{t_\mu} \|\partial_t \nabla u\|^2 dt. \end{aligned} \tag{5.2.63}$$

Im letzten Schritt schätzen wir die mittlere Summe rechts ab. Jetzt kann  $\varphi_h = \varphi_h^m := \eta_h^m - \eta_h^{m-1} \in V_h$  als Testfunktion verwendet werden, und wir erhalten wie oben

$$\begin{aligned} \|\varphi_h^m\|^2 &+ \frac{1}{2}k_m \|\nabla \eta_h^m\|^2 - \frac{1}{2}k_m \|\nabla \eta_h^{m-1}\|^2 + \frac{1}{2}k_m \|\nabla \varphi_h^m\|^2 \\ &= k_m(\bar{f}^m, \varphi_h^m) - (R_h^m u^m - R_h^{m-1} u^{m-1}, \varphi_h^m) - k_m(\nabla R_h^m u^m, \nabla \varphi_h^m) \\ &= k_m(\bar{f}^m, \varphi_h^m) - (u^m - u^{m-1}, \varphi_h^m) - k_m(\nabla u^m, \nabla \varphi_h^m) \\ &\quad + (u^m - u^{m-1} - R_h(u^m + u^{m-1}), \varphi_h^m). \end{aligned}$$

Weiter haben wir

$$\begin{aligned} k_m(\bar{f}^m, \varphi_h^m) - (u^m - u^{m-1}, \varphi_h^m) - k_m(\nabla u^m, \nabla \varphi_h^m) &= \int_{t_{m-1}}^{t_m} (f - \partial_t u, \varphi_h^m) dt - k_m(\nabla u^m, \nabla \varphi_h^m) \\ &= \int_{t_{m-1}}^{t_m} (\nabla u, \nabla \varphi_h^m) dt - k_m(\nabla u^m, \nabla \varphi_h^m) \\ &= \int_{t_{m-1}}^{t_m} (t - t_{m-1})(\nabla \partial_t u, \nabla \varphi_h^m) dt \\ &\leq \frac{1}{2}k_m \|\nabla \varphi_h^m\|^2 + \frac{1}{2}k_m^2 \int_{t_{m-1}}^{t_m} \|\partial_t \nabla u\|^2 dt \end{aligned}$$

sowie

$$\begin{aligned} (u^m - u^{m-1} - R_h(u^m - u^{m-1}), \varphi_h^m) &\leq \frac{1}{4}\|\varphi_h^m\|^2 + ch_m^2 \|\nabla(u^m - u^{m-1})\|^2 \\ &\leq \frac{1}{4}\|\varphi_h^m\|^2 + ch_m^2 k_m \int_{t_{m-1}}^{t_m} \|\nabla \partial_t u\|^2 dt. \end{aligned}$$

Kombination dieser Abschätzungen und Absorption von Termen in die linke Seite ergibt

$$\|\varphi_h^m\|^2 + k_m \|\nabla \eta_h^m\|^2 - k_m \|\nabla \eta_h^{m-1}\|^2 \leq \{k_m^2 + ch_m^2 k_m\} \int_{t_{m-1}}^{t_m} \|\nabla \partial_t u\|^2 dt.$$

Wir wenden diese Abschätzung wieder rekursiv für  $m, m-1, \dots, 1$  an und finden

$$\sum_{\mu=1}^m k_\mu^{-1} \|\varphi_h^m\|^2 + \|\nabla \eta_h^m\|^2 \leq \|\nabla \eta_h^0\|^2 + c \sum_{\mu=1}^m \{k_\mu + h_\mu^2\} \int_{t_{\mu-1}}^{t_\mu} \|\partial_t \nabla u\|^2 dt.$$

Mit  $\|\nabla \eta_h^0\|^2 \leq ch_0^4 \|\nabla^2 u^0\|^2$  folgt schließlich

$$\sum_{\mu=1}^m k_\mu^{-1} \|\eta_h^\mu - \eta_h^{\mu-1}\|^2 \leq ch_0^4 \|\nabla^2 u^0\|^2 + c \sum_{\mu=1}^m \{k_\mu + h_\mu^2\} \int_{t_{\mu-1}}^{t_\mu} \|\partial_t \nabla u\|^2 dt. \quad (5.2.64)$$

Wir setzen dies in (5.2.63) ein und erhalten die behauptete Abschätzung (5.2.52):

$$\|e_h^m\|^2 \leq ct_m \max_{0 \leq \mu \leq m} \{h_\mu^2 \|\nabla^2 u^\mu\|^2\} + \sum_{\mu=1}^m k_\mu^{-1} \|\eta_h^{\mu-1} - \eta_h^\mu\|^2 + c \sum_{\mu=1}^m k_\mu^2 \int_{t_{\mu-1}}^{t_\mu} \|\partial_t \nabla u\|^2 dt,$$

was den Beweis vervollständigt. Q.E.D.

Die Konvergenzordnung in (5.2.51) ist nicht optimal. Unter der Bedingung  $h_m^{4/3} \leq ck_m$  ergibt sich aber die zeit-optimale Konvergenzordnung  $\mathcal{O}(k_m)$ . Das optimale Resultat (5.2.52) lässt sich auch unter den weniger einschränkenden Bedingungen  $V_h^{m-1} \subset V_h^m$  oder  $h_m^2 k_m^{-1} \leq \kappa$  hinreichend klein beweisen.

**Hilfssatz 5.2 (A-priori Schranke):** Für die Lösung der ARWA (5.2.46) gilt die a priori Abschätzung

$$\max_{[0,T]} \|\nabla^2 u\| + \left( T^{-1} \int_0^T \|\nabla \partial_t u\|^2 dt \right)^{1/2} \leq c \|\nabla^2 u^0\| + c \max_{[0,T]} \{\|f\| + \|\partial_t f\|\}. \quad (5.2.65)$$

**Beweis:** Der Beweis verwendet die „Energie-Technik“, wird hier aber nicht ausgeführt. Q.E.D.

Wir wollen noch die Frage nach der „inversen Monotonie“ der Orts-Zeit-Diskretisierung diskutieren. Unter bestimmten Bedingungen an das Ortsgitter (z.B. alle Innenwinkel einer Triangulierung  $\omega \leq \pi/2$ ) ist die Steifigkeitsmatrix  $A_h$  eine M-Matrix. Die Systemmatrix  $M_h + kA_h$  muss aber nicht automatisch M-Matrix sein. Um dies dennoch sicherzustellen, werden die Elemente von  $M_h$ ,

$$m_{ij} = (\varphi_h^{m,i}, \varphi_h^{m,j}),$$

mit Hilfe der Trapezregel ausgewertet. Bei stückweise linearen Ansätzen liefert dies eine Diagonalmatrix  $\tilde{M}_h = M_h + \mathcal{O}(h^2)$  mit positiven Diagonalelementen, so dass  $\tilde{M}_h + kA_h$  M-Matrix wird. Dieser „Mass-Lumping“ genannte Prozess erhält die Konvergenzordnung des Gesamtschemas und stellt seine „inverse Monotonie“ sicher (Übungsaufgabe).

### 5.2.2 Fehlerkontrolle und Schrittweitensteuerung

Zur Herleitung von *a posteriori* Fehlerabschätzungen erweist sich eine globale Betrachtung simultan in Ort und Zeit (ohne die bisherige Aufspaltung in Orts- und Zeitdiskretisierung) als angemessen. Wir führen dazu das Konzept der „unstetigen Galerkin-Verfahren“ für parabolische ARWA ein, als deren einfacher Spezialfall das implizite Euler-Verfahren (5.2.48) erscheinen wird. Die Vorgehensweise entspricht der bereits von den gewöhnlichen AWAn her bekannten, ergänzt um die Aspekte der Ortsdiskretisierung.

Ausgehend von oben formulierten Diskretisierungen  $T_h^m = \{K_n^m\}$  des Ortsgebiets  $\bar{\Omega}$  und  $0 = t_0 < \dots < t_\mu < \dots < t_M = T$  des Zeitintervalls  $I = [0, T]$  führen wir die folgenden Bezeichnungen ein (Man beachte, dass die Zeitschrittweite nicht bzgl. des Orts variiert.):

$$\begin{aligned} I_m &:= (t_{m-1}, t_m], \quad k_m = t_m - t_{m-1}, \\ k &:= \max_{m=1, \dots, M} k_m, \quad h_m := \max_{K \in T_h^m} h_K, \quad h := \max_{m=1, \dots, M} h_m, \\ v^{m\pm} &:= \lim_{s \downarrow 0} v(\cdot, t_m \pm s), \quad [v]^m := v^{m+} - v^{m-}, \\ Q_n^m &:= K_n^m \times I_m, \quad \partial Q_n^m := \partial K_n^m \times I_m, \quad Q^m := \Omega \times I_m. \end{aligned}$$

Die ARWA (5.2.46) lässt sich äquivalent schreiben in der Form

$$\sum_{m=1}^M \left\{ \int_{I_m} \{(\partial_t u, \varphi) + (\nabla u \nabla \varphi)\} dt + ([u]^{m-1}, \varphi^{(m-1)+}) \right\} = \int_I (f, \varphi) dt \quad (5.2.66)$$

für beliebige in der Zeit stetige Testfunktion  $\varphi(\cdot, t) \in V$ , wobei die Anfangsbedingung durch die Setzung  $u^{0-} := u^0$  berücksichtigt ist. Jede (glatte) Lösung von (5.2.46) erfüllt offenbar die Beziehung (5.2.66), und umgekehrt muss jede Lösung von (5.2.66) in den Teilintervallen  $I_m$  der Wärmeleitungsgleichung genügen und bei  $t_m$  auch stetig sein. Damit folgt dann wieder, dass es sich um eine Lösung der ARWA handeln muss. Dieses Problem wird nun mit einem Galerkin-Ansatz auf dem ganzen Orts-Zeit-Zylinder  $Q_T$  diskretisiert. Dazu führen wir die folgenden Finite-Elemente-Räume ein:

$$V_h = \{v : \bar{Q}_T \rightarrow \mathbb{R} \mid v_{t \in I_m}(\cdot, t) \in V_h^m, v_{t \in I_m}(x, \cdot) \in P_r(I_m), (x, t) \in Q_T\}.$$

Die Funktionen in  $V_h$  sind also bzgl. des Ortes stückweise in  $V_h^m$  (d.h. linear oder bilinear und stetig) und bzgl. der Zeit stückweise polynomial vom Grad  $r$  (und unstetig). Der Galerkin-Ansatz sucht dann Approximationen  $U_h \in V_h$  zu bestimmen durch die Vorschrift

$$\sum_{m=1}^M \left\{ \int_{I_m} \{(\partial_t U_h, \varphi_h) + (\nabla U_h, \nabla \varphi_h)\} dt + ([U_h]^{m-1}, \varphi_h^{(m-1)+}) \right\} = \int_I (f, \varphi_h) dt \quad (5.2.67)$$

für beliebige Testfunktion  $\varphi_h \in V_h$  mit dem Anfangswert  $U_h^{0-} = P_h^0 u^0$  ( $P_h^0$  die  $L^2$ -Projektion auf  $V_h^0$ ). Da die Testfunktionen unstetig in der Zeit sein dürfen, zerfällt dieses formal zeitlich global gekoppelte System in lokale Teilprobleme auf jedem Zeitstreifen

$Q^m = \Omega \times I_m$ . Dieses Schema wird „unstetiges Galerkin-Verfahren“ (abgekürzt „ $dG(r)$ -Verfahren“) genannt. Wir wollen im folgenden nur den einfachsten Fall  $r = 0$ , d.h. das  $dG(0)$ -Verfahren, betrachten. In diesem Fall reduziert sich das globale Schema (5.2.67) auf die folgende Sequenz von lokalen Gleichungen auf den Zeitintervallen  $I_m$ ,  $m = 1, \dots, M$ :

$$\int_{I_m} \{(\partial_t U_h, \varphi_h) + (\nabla U_h, \nabla \varphi_h)\} dt + ([U_h]^{m-1}, \varphi_h) = \int_{I_m} (f, \varphi_h) dt \quad (5.2.68)$$

für beliebiges  $\varphi_h \in V_h^m$ . Mit der Setzung  $U_h^m := U_h^{m-1}$  ergibt sich wegen  $\partial_t U_h \equiv 0$  auf  $I_m$ :

$$k_m(\nabla U_h^m, \nabla \varphi_h) + (U_h^m - U_h^{m-1}, \varphi_h) = k_m(\bar{f}^m, \varphi_h) \quad (5.2.69)$$

für beliebiges  $\varphi_h \in V_h^m$ . Dies ist gerade das implizite Euler-Verfahren (5.2.48), welches sich also in diesem Rahmen als  $dG(0)$ -Verfahren interpretieren lässt. Diese Sicht ändert zwar nichts am Verfahren selbst, bietet jedoch einen systematischen Zugang zu seiner Fehleranalyse.

**Bemerkung 5.1:** Die  $dG(r)$ -Verfahren höheren Grades  $r \geq 1$  entsprechen keinem der oben diskutierten Zeitschritt-Schemata, sie sind vielmehr Varianten gewisser impliziter Runge<sup>9</sup>-Kutta<sup>10</sup>-Verfahren. Wir bemerken aber, dass sich auch das populäre Crank-Nicolson-Verfahren in den Rahmen der Galerkin-Verfahren einordnen lässt. Dazu macht man einen modifizierten Ansatz mit bzgl. der Zeit stückweise *linearen*, aber diesmal *global stetigen* Funktionen. Der Raum der Testfunktionen ist dagegen derselbe wie beim  $dG(0)$ -Verfahren. Dies wird dann ein „Petrow-Galerkin-Verfahren“ (sog.  $cG(1)$ -Verfahren), welches sich ähnlich analysieren lässt wie das  $dG(0)$ -Verfahren:

$$(U_h^m, \varphi_h) + \frac{1}{2} k_m(\nabla(U_h^m + U_h^{m-1}), \nabla \varphi_h) = (U_h^{m-1}, \varphi_h) + k_m(\bar{f}^m, \varphi_h), \quad (5.2.70)$$

für beliebige Testfunktion  $\varphi_h \in V_h^m$ .

Wir wollen jetzt eine a posteriori Fehlerabschätzung für das  $dG(0)$ -Verfahren ableiten, welche als Basis für eine simultane Anpassung des Ortsgitters und der Zeitschrittweite an den Lösungsverlauf dienen kann. Dazu setzen wir  $e_h := U_h - u$ . Ein typisches Beispiel ist die Kontrolle des  $L^2$ -Fehlers zum Endzeitpunkt  $J(e_h) := \|e_h^{N+}\|$ .

**Satz 5.6 (A posteriori Fehlerschranke):** Für das  $dG(0)$ -Verfahren gilt bei Kontrolle des örtlichen  $L^2$ -Fehlers zum Endzeitpunkt,  $J(e_h) := \|e_h^{N+}\|$ , die a posteriori Fehlerabschätzung

$$J(e_h) \leq \eta(U_h) := c_i \sum_{\mu=1}^m \sum_{K_n^\mu \in T_h^\mu} \rho_n^\mu(U_h) \omega_n^\mu(z), \quad (5.2.71)$$

<sup>9</sup>Carle David Tolmé Runge (1856–1927): Deutscher Mathematiker; Prof. in Hannover und Göttingen; Beiträge zur Spektraltheorie mit Anwendungen in der Atom-Physik.

<sup>10</sup>Martin Wilhelm Kutta (1867–1944): Deutscher Mathematiker; Prof. in Stuttgart; Beiträge zur Aerodynamik und zur Numerik von Differentialgleichungen.

mit den Residuentermen

$$\rho_n^\mu(U_h) := \|R(U_h)\|_{Q_n^\mu} + h_n^{-1/2}\|[\partial_n U_h]\|_{\partial K \times I^\mu} + k_\mu^{-1/2}\|[U_h]^{(\mu-1)+}\|_{K_n^\mu},$$

$R(U_h) := f - \partial_t U_h + \Delta U_h$ , und einer „Interpolationskonstante“  $c_i$ . Mit der Lösung  $z$  des „dualen Problems“

$$-\partial_t z - \Delta z = 0 \quad \text{in } \Omega \times [0, t_m], \quad z|_{\partial\Omega} = 0, \quad z|_{t=t_m} = \|e^{m+}\|^{-1} e^{m+}. \quad (5.2.72)$$

haben die Gewichte  $\omega_n^\mu(z)$  die Gestalt:

$$\omega_n^\mu(z) := k_\mu \|\partial_t z\|_{Q_n^\mu} + k_\mu^2 \|\partial_t^2 z\|_{Q_n^\mu} + h_n^2 \|\nabla^2 z\|_{Q_n^\mu}.$$

**Beweis:** i) Das erste „Standbein“ des Beweises ist ein (parabolisches) Dualitätsargument. Wir führen zunächst für ein festes  $t_m \in I$  das (kontinuierliche) „duale Problem“ ein:

$$-\partial_t z - \Delta z = \psi \quad \text{in } Q_m := \Omega \times [0, t_m], \quad z|_{\partial\Omega}, \quad z|_{t=t_m} = \chi, \quad (5.2.73)$$

bzw. in semi-variationeller Formulierung

$$-(\partial_t z, \varphi) + (\nabla z, \nabla \varphi) = (\psi, \varphi) \quad \forall \varphi \in V. \quad (5.2.74)$$

Umschreiben in die Form (5.2.66) ergibt

$$\sum_{\mu=1}^m \left\{ \int_{I_\mu} \{-(\varphi, \partial_t z) + (\nabla \varphi, \nabla z)\} dt + (\varphi^{(\mu-1)+}, [z]^\mu) \right\} = \int_{[0, t_m]} (\varphi, \psi) dt$$

für beliebige in der Zeit stetige Testfunktion  $\varphi(\cdot, t) \in V$ , wobei die Anfangsbedingung wieder durch die Vorschrift  $z^{(m+)} := \chi$  berücksichtigt ist. Durch partielle Integration in der Zeit wird dies umgeformt zu

$$\begin{aligned} \sum_{\mu=1}^m & \left\{ \int_{I_\mu} \{(\partial_t \varphi, z) + (\nabla \varphi, \nabla z)\} dt + ([\varphi]^{(\mu-1)-}, z^{(\mu-1)+}) \right\} \\ & = (\varphi^{(m+)}, \chi) + \int_{[0, t_m]} (\varphi, \psi) dt \end{aligned} \quad (5.2.75)$$

für beliebige in der Zeit (stückweise) differenzierbare Testfunktion  $\varphi(\cdot, t) \in V$ .

ii) Das zweite „Standbein“ des Beweises ist die Galerkin-Orthogonalität des Fehlers des  $dG(0)$ -Verfahrens. Durch Vergleich der beiden Gleichungen (5.2.66) für  $u$  und (5.2.67) für  $U_h$  erhalten wir für den Fehler  $e_h := U_h - u$  die „Galerkin-Orthogonalität“:

$$\sum_{m=1}^M \left\{ \int_{I_m} \{(\partial_t e_h, \varphi_h) + (\nabla e_h \nabla \varphi_h)\} dt + ([e_h]^{(m-1)-}, \varphi_h^{(m-1)+}) \right\} = 0 \quad (5.2.76)$$

für beliebige stetige Testfunktion  $\varphi_h(\cdot, t) \in V_h$ . Wir wählen nun in (5.2.75) die spezielle

Testfunktion  $\varphi := e_h$  und erhalten:

$$\sum_{\mu=1}^m \left\{ \int_{I_\mu} \{(\partial_t e_h, z) + (\nabla e_h, \nabla z)\} dt + ([e_h]^{\mu-1}, z^{(\mu-1)+}) \right\} = (e_h^{m+}, \chi) + \int_{[0, t_m]} (e_h, \psi) dt.$$

Unter Ausnutzung von (5.2.76) erhalten wir damit die allgemeine Fehleridentität:

$$(e_h^{m+}, \chi) + \int_{[0, t_m]} (e_h, \psi) dt = \sum_{\mu=1}^m \left\{ \int_{I_\mu} \{(\partial_t e_h, z - z_h) + (\nabla e_h, \nabla(z - z_h))\} dt + ([e_h]^{\mu-1}, (z - z_h)^{(\mu-1)+}) \right\} \quad (5.2.77)$$

mit einer beliebigen Approximation  $z_h \in V_h$  zur dualen Lösung  $z$ . Wir setzen nun im dualen Problem  $t_m = T$ ,  $\psi := 0$  und  $\chi := \|e^{N+}\|^{-1} e^{N+}$ . Aus der allgemeinen Fehleridentität (5.2.77) folgt

$$\|e_h^{m+}\| = \sum_{\mu=1}^m \left\{ \int_{I_\mu} \{(\partial_t e_h, z - z_h) + (\nabla e_h, \nabla(z - z_h))\} dt + ([e_h]^{\mu-1}, (z - z_h)^{(\mu-1)+}) \right\} \quad (5.2.78)$$

mit einer beliebigen Approximation  $z_h \in V_h$  zur dualen Lösung  $z$ . Wir nutzen nun die Lösungseigenschaften von  $u$  und integrieren auf jeder Zelle  $K_n^\mu$  partiell bzgl. des Orts:

$$\begin{aligned} \|e_h^{m+}\| &= \sum_{\mu=1}^m \sum_{K_n^\mu \in \mathbf{T}_h^\mu} \left\{ \int_{I_\mu} \{(\partial_t u - \partial_t U_h, z - z_h)_{K_n^\mu} - (\Delta u - \Delta U_h, z - z_h)_{K_n^\mu} \right. \\ &\quad \left. - (\partial_n(U_h - u), z - z_h)_{\partial K_n^\mu}\} dt + ([U_h - u]^{\mu-1}, (z - z_h)^{(\mu-1)+})_{K_n^\mu} \right\}. \end{aligned}$$

Dies ergibt

$$\begin{aligned} \|e_h^{m+}\| &= \sum_{\mu=1}^m \sum_{K_n^\mu \in \mathbf{T}_h^\mu} \left\{ \int_{I_\mu} \{(R(U_h), z - z_h)_{K_n^\mu} - \frac{1}{2}([\partial_n U_h], z - z_h)_{\partial K_n^\mu}\} dt \right. \\ &\quad \left. + ([U_h]^{\mu-1}, (z - z_h)^{(\mu-1)+})_{K_n^\mu} \right\}, \end{aligned}$$

mit dem Residuum  $R(U_h) := f - \partial_t U_h + \Delta U_h$  und dem Sprung  $[\partial_n U_h]$  von  $\partial_n U_h$  über die Zellkanten. Durch Anwendung der Hölder'schen Ungleichung erhalten wir

$$\begin{aligned} \|e_h^{m+}\| &\leq \sum_{\mu=1}^m \sum_{K_n^\mu \in \mathbf{T}_h^\mu} \left\{ \|R(U_h)\|_{Q_n^\mu} \|z - z_h\|_{Q_n^\mu} \right. \\ &\quad \left. + \frac{1}{2} \|[\partial_n U_h]\|_{\partial Q_n^\mu} \|z - z_h\|_{\partial Q_n^\mu} \right\} dt + \|([U_h]^{\mu-1})_{K_n^\mu}\| ((z - z_h)^{(\mu-1)+})_{K_n^\mu}, \quad (5.2.79) \end{aligned}$$

Wir wählen nun für  $z_h$  die natürliche Interpolierende  $I_{h,k} z \in V_h$ , welche zellweise definiert

ist durch die Vorschrift

$$I_{h,k} z|_{I_\mu} = k_\mu^{-1} \int_{I_\mu} z \, dt, \quad I_{h,k} z|_{K_n^\mu} = \text{konstante Interpolation von } z. \quad (5.2.80)$$

Dann gelten die Interpolationsabschätzungen (Beweis mit Hilfe einer Variante des Bramble-Hilbert-Lemmas)

$$\|z - I_{h,k} z\|_{Q_n^\mu} \leq c_i \omega_n^\mu(z), \quad (5.2.81)$$

$$h_n^{1/2} \|z - I_{h,k} z\|_{\partial Q_n^\mu} \leq c_i \omega_n^\mu(z), \quad (5.2.82)$$

$$k_\mu^{1/2} \|(z - I_{h,k} z)^{(\mu-1)+}\|_{K_n^\mu} \leq c_i \omega_n^\mu(z), \quad (5.2.83)$$

wobei

$$\omega_n^\mu(z) := k_\mu \|\partial_t z\|_{Q_n^\mu} + k_\mu^2 \|\partial_t^2 z\|_{Q_n^\mu} + h_n^2 \|\nabla^2 z\|_{Q_n^\mu}.$$

Die „Interpolationskonstante“  $c_i$ , welche in die a posteriori Fehlerabschätzung (5.2.71) eingeht, hat dabei in der Regel die Größe  $c_i \sim 0,1 - 1$ . Einsetzen dieser Abschätzungen in (5.2.79) und Anwendung der Schwarz'schen Ungleichung ergeben

$$\|e_h^{m+}\| \leq c_i \sum_{\mu=1}^m \sum_{K_n^\mu \in \mathbf{T}_h^\mu} \left\{ \|R(U_h)\|_{Q_n^\mu} + h_n^{-1/2} \|[\partial_n U_h]\|_{\partial Q_n^\mu} + k_\mu^{-1/2} \|[U_h]^{\mu-1}\|_{K_n^\mu} \right\} \omega_n^\mu(z),$$

Dies impliziert die Behauptung. Q.E.D.

Zur konkreten Auswertung der *a posteriori* Fehlerabschätzung (5.2.71) müssen die Gewichte  $\omega_n^m(z)$  berechnet werden. Dazu wird analog zum stationären, elliptischen Fall das duale Problem (5.2.72) numerisch auf dem aktuellen Gitter gelöst. Mit der resultierenden diskreten dualen Lösung  $z_h$  wird dann approximiert gemäß:

$$\omega_n^m(z) \approx \tilde{\omega}_n^m := k_\mu \|\partial_t^{-1} [z_h]^{\mu-1}\|_{K_n^\mu} + h_n^2 \|\nabla_h^2 z_h\|_{Q_n^\mu}, \quad (5.2.84)$$

wobei  $\nabla_h^2 z_h \sim \nabla^2 z$  ein geeigneter Differenzenquotient ist. Auf der Basis der approximativen *a posteriori* Fehlerabschätzung

$$\|e_h^{m+}\| \approx \tilde{\eta}(U_h) := c_i \sum_{\mu=1}^m \sum_{K_n^\mu \in \mathbf{T}_h^\mu} \rho_n^\mu(U_h) \tilde{\omega}_n^\mu(z) \quad (5.2.85)$$

können nun simultan das Ortsgitter und die Zeitschrittweite adaptiert werden. Dabei werden in einem iterativen Prozess die Ortsresiduen

$$\rho_n^{m,1}(U_h) := \|R(U_h)\|_{Q_n^\mu} + h_n^{-1/2} \|[\partial_n U_h]\|_{\partial Q_n^\mu},$$

und Zeitresiduen

$$\rho_n^{m,2}(U_h) := k_\mu^{-1/2} \|[U_h]^{\mu-1}\|_{K_n^\mu},$$

durch Anpassen von  $h_n$  und  $k_m$  balanciert, bis  $\tilde{\eta}(U_h) \leq \text{TOL}$  für eine vorgegebene Fehlertoleranz TOL.

Analoge Resultate gelten für andere Fehlermaße, z. B. den globalen  $L^2$ -Fehler auf  $Q_T$ :

$$J(e_h) := \|e_h\|_{Q_T} = \left( \int_I \|e_h\|^2 dt \right)^{1/2}. \quad (5.2.86)$$

In diesem Fall gilt die *a posteriori* Fehlerabschätzung (5.2.85) mit dem dualen Problem

$$-\partial_t z - \Delta z = \|e_h\|_{Q_T}^{-1/2} e_h \quad \text{in } \Omega \times [0, t_m], \quad z|_{\partial\Omega} = 0, \quad z|_{t=t_m} = 0. \quad (5.2.87)$$

Die Wirksamkeit der adaptiven Steuerung der Ortsgitterweite auf der Basis dieser *a posteriori* Fehlerabschätzungen wird anhand eines einfachen Beispiels demonstriert.

**Beispiel 5.1:** Wir lösen die inhomogene Wärmeleitungsgleichung

$$\partial_t u - \Delta u = f \quad \text{in } Q_T, \quad u|_{\partial\Omega} = 0, \quad u|_{t=0} = u^0, \quad (5.2.88)$$

auf dem Einheitsquadrat  $\Omega = (0, 1)^2$ . Als „exakte“ Lösung wird angesetzt:

$$u(x, t) := \frac{1}{1 + \alpha|x - x^0|^2}, \quad x^0 := \left(\frac{1}{2} + \frac{1}{4} \cos(2\pi t), \frac{1}{2} + \frac{1}{4} \sin(2\pi t)\right)^T,$$

woraus sich Anfangswert und rechte Seite ergeben zu

$$u^0(x) := u(x, 0), \quad f(x, t) := \partial_t u(x, t) - \Delta u(x, t).$$

Diese Lösungsfunktion beschreibt einen „Hügel“ im Gebiet  $(0, 1)^2$ , der während des Zeintervalls  $I = [0, 1]$  einmal im Kreis um den Punkt  $(\frac{1}{2}, \frac{1}{2})$  herumläuft. Die Größe bzw. Steigung des Hügels lässt sich durch Wahl des Parameters  $\alpha$  steuern. Für den Test wird  $\alpha = 50$  gesetzt. Die sich damit ergebenden Gitter sind in Abb. 5.2 und 5.3 gezeigt.

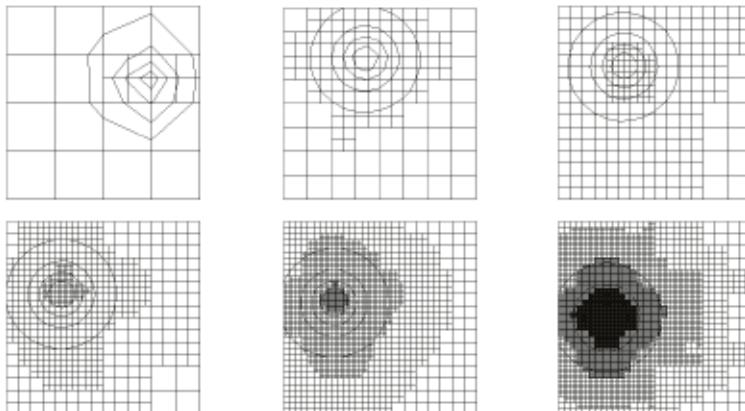


Abbildung 5.2: Gittersequenzen bei Kontrolle des Endzeit- $L^2$ -Fehlers  $\|e_M\|_\Omega$ ; Quelle: R. Hartmann, „A posteriori Fehlerschätzung ... für die Wärmeleitungsgleichung“, Diplomarbeit, Univ. Heidelberg, 1998.

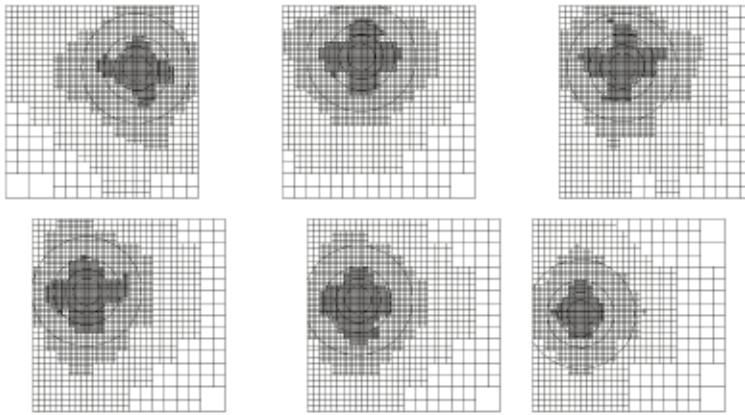


Abbildung 5.3: Gittersequenzen bei Kontrolle des globalen  $L^2$ -Fehlers  $\|e\|_{\Omega \times I}$  (unten); Quelle: R. Hartmann, „A posteriori Fehlerschätzung ... bei Galerkin-Verfahren für die Wärmeleitungsgleichung“, Diplomarbeit, Univ. Heidelberg, 1998.

### 5.3 Verallgemeinerungen und Lösungsaspekte

Wir haben als Modellfall die Wärmeleitungsgleichung

$$\partial_t u - a \Delta u = f \quad \text{in } Q_T = \Omega \times [0, T], \quad (5.3.89)$$

mit homogenen Dirichlet-Randbedingungen  $u|_{\partial\Omega} = 0$  und einem konstanten „Diffusionskoeffizienten“  $a > 0$  betrachtet.

#### i) Verallgemeinerungen:

Wir lassen nun zu, dass der Koeffizient  $a = a(x, t)$  vom Ort und von der Zeit abhängt. Das Problem schreibt sich dann in der Form

$$\partial_t u - \nabla \cdot \{a \nabla u\} = f \quad \text{in } Q_T = \Omega \times [0, T]. \quad (5.3.90)$$

Die oben betrachteten Zeitschrittverfahren lassen sich in der Regel leicht auf diesen allgemeineren Fall übertragen. Wir wollen dies anhand des Crank-Nicolson-Verfahrens diskutieren:

$$k^{-1} (U_h^m - U_h^{m-1}) + \frac{1}{2} (A_h^m U_h^m + A_h^{m-1} U_h^{m-1}) = \frac{1}{2} (f_h^m + f_h^{m-1}), \quad (5.3.91)$$

mit der Ortsdiskretisierung  $A_h(t)$  des Operators  $-\nabla \cdot (a(\cdot, t) \nabla)$ . Dabei werden für Funktionen  $w(t)$  die abkürzenden Bezeichnungen  $w^m := w(t_m)$ ,  $w^{m-1/2} := w(t_{m-1/2})$  und  $t_{m-1/2} := \frac{1}{2}(t_m + t_{m-1})$  verwendet. Dies entspricht der sog. „Sehnentrapezregel“; Anwen-

dung der „Tangententrapezregel“ führt auf das Schema:

$$k^{-1} (U_h^m - U_h^{m-1}) + \frac{1}{2} A_h^{m-1/2} (U_h^m + U_h^{m-1}) = f_h^{m-1/2}. \quad (5.3.92)$$

Beide Verfahrensvarianten sind *unbedingt stabil* und von der Konvergenzordnung  $\mathcal{O}(k^2 + h^2)$ . Zu Ihrer Analyse ist das elegante Spektralargument aus dem vorigen Abschnitt leider nicht mehr geeignet, da der Operator  $A_h(t)$  nun mit der Zeit variiert. Statt dessen verwendet man die flexiblere „Energietechnik“ und behält im wesentlichen dieselben Aussagen wie im autonomen Fall, allerdings mit wesentlich mehr Aufwand. Im (praktisch wichtigen) nichtlinearen Fall  $a = a(u(t))$  wird zweckmäßigerweise die Tangententrapezform des Crank-Nicolson-Schemas verwendet:

$$k^{-1} (U_h^m - U_h^{m-1}) + A_h \left( \frac{1}{2} (U_h^m + U_h^{m-1}) \right) = f_h^{m-1/2}. \quad (5.3.93)$$

Hier ist auch die BDF(2)-Formel gut anwendbar:

$$2k^{-1} (3U_h^m - 4U_h^{m-1} + U_h^{m-2}) + A_h (U_h^m) = f_h^m. \quad (5.3.94)$$

Eine Stabilitäts- und Konvergenzanalyse steht aber außerhalb des Rahmens dieses Textes.

## ii) Berechnung der Startwerte:

Bei der Durchführung jedes Zeitschritts ist die rechte Seite aufzubauen, welche die Information vom vorausgehenden Zeitlevel beinhaltet. Dabei ist unter Umständen eine  $L^2$ -Projektion auf das aktuelle Gitter vorzunehmen.

*a) Anfangswert:* Die Auswertung des Anfangswerts  $U_h^0$  kann meist durch einfache, lokale Interpolation (oder Restriktion) des kontinuierlichen Anfangswerts  $u^0$  auf das Gitter erfolgen. Im Fall eines irregulären Anfangswerts, etwa  $u^0 \notin C(\bar{\Omega})$ , ist jedoch Vorsicht geboten. Zur Gewährleistung der vollen „Glättungseigenschaft“ der Diskretisierung (im Ort sowie in der Zeit) sollte  $U_h^0$  als  $L^2$ -Projektion ausgewertet werden gemäß:

$$(U_h^0, \varphi_h) = (u^0, \varphi_h) \quad \varphi_h \in V_h. \quad (5.3.95)$$

*b) Ortsdiskretisierung:* Verändert sich die Ortsdiskretisierung vom Zeitlevel  $t_{m-1}$  zum Zeitlevel  $t_m$ , so muss die vorausgehende Näherung  $U_h^{m-1} \in V_h^{m-1}$  auf das neue Gitter transferiert werden. Im FE-Kontext geschieht dies zwangsläufig gemäß

$$(U_h^{m-1}, \varphi_h) \quad \forall \varphi_h \in V_h^m, \quad (5.3.96)$$

was gleichbedeutend mit der Auswertung der  $L^2$ -Projektion  $P_h^m U_h^{m-1} \in V_h^m$  ist. Dieser unscheinbare Schritt kann unter Umständen die „teure“ Komponente des ganzen Lösungsprozesses sein. Dies ist dann der Fall, wenn die beiden Gitter  $T_h^{m-1}$  und  $T_h^m$  völlig unabhängig voneinander erzeugt werden. Zur Auswertung von (5.3.96) müssen Zellintegrale über Produkte von Knotenbasisfunktionen berechnet werden:

$$\int_{K_n^m} \varphi_h^{m-1,j} \varphi_h^{m,n} dx.$$

Normalerweise geschieht dies mit Hilfe von Quadraturformeln. Da die Funktion  $\varphi_h^{m-1,j}$  auf der Zelle  $K_n^m$  aber in der Regel nur stückweise glatt ist, wäre das zu ungenau. Der dadurch in jedem Zeitschritt eingeschleppte Fehler würde im Verlaufe der Rechnung akkumulieren und das Ergebnis stark verfälschen. Die Verwendung von Quadraturformeln besonders hoher Ordnung (z.B.  $3 \times 3$ -Gauß-Formeln) behebt diese Schwierigkeit nicht, da letztere zur Erreichung ihrer hohen Genauigkeit natürlich auch eine entsprechend hohe Regularität des Integranden benötigen. Gerade diese ist aber im betrachteten Fall nicht gegeben. Es gibt im wesentlichen drei Wege zur Lösung dieses technischen Problems:

- Es werden „summierte“ Quadraturformeln auf den einzelnen Zellen  $K_n^m$  verwendet; etwa durch Unterteilung in  $4 - 16$  Unterzellen. Dies erhöht zwar nicht die Ordnung der Integration, vermindert aber die relevante Fehlerkonstante.
- Die Integration wird für den stückweise polynomialem Integranden „exakt“ durchgeführt. Dazu ist die Bestimmung aller Teilstücke von  $K_n^m$  erforderlich, auf denen  $\varphi_h^{m-1,j}$  glatt ist. Dies wird bei, unstrukturierten Gittern in 3-D allerdings sehr aufwendig.
- Gehören die Gitter  $T_h^{m-1}$  und  $T_h^m$  zu einer Familie von hierarchisch verfeinerten Gittern, kann diese Strukturinformation zur effizienten Berechnung der Integrale verwendet werden, da die Lage der kritischen Knicklinien von  $\varphi_h^{m-1,j}$  durch die reguläre Verfeinerung bestimmt ist.

### iii) Lösungskomplexität:

Wir betrachten wieder den Modelfall der homogenen Wärmeleitungsgleichung auf dem Einheitsquadrat  $\Omega = (0, 1)^2 \subset \mathbb{R}^2$ ,

$$\partial_t u - \Delta u = 0 \quad \text{in } Q_T, \tag{5.3.97}$$

welche auf einem äquidistanten Gitter mit dem 5-Punkte-Differenzenoperator diskretisiert sei.

– *Explizite* Verfahren (etwa das explizite Euler-Schema) erfordern in jedem Zeitschritt eine Matrix-Vektor-Multiplikation mit einem arithmetischen Aufwand  $\mathcal{O}(N)$ . Die Stabilitätsbedingung  $k \leq ch^2$  erzwingt etwa  $M \sim h^{-2} \sim N$  Zeitschritte pro Zeiteinheit. Dies bedeutet einen Gesamtaufwand von  $\mathcal{O}(N^2)$  OP.

– *Implizite* Verfahren erfordern in jedem Zeitschritt die Lösung eines linearen Gleichungssystems, erlauben aber größere Zeitschritte. Zur Überbrückung einer Zeiteinheit sind aus Genauigkeitsgründen in der Regel  $M \sim h^{-1}$  Zeitschritte erforderlich. Wir diskutieren exemplarisch das Crank-Nicolson-Verfahren. Bei zeilenweiser Numerierung der Gitterpunkte haben die resultierenden Gleichungssysteme

$$(I_h + \frac{1}{2}akA_h)U^m = (I_h - \frac{1}{2}akA_h)U^{m-1} \tag{5.3.98}$$

die Koeffizientenmatrix  $L_h := I_h + \frac{1}{2}akA_h$ , wobei wieder

$$A_h = \left[ \begin{array}{cccc} B_m & -I_m & & \\ -I_m & B_m & -I_m & \\ & -I_m & B_m & \ddots \\ & & \ddots & \ddots \end{array} \right]_N \quad B_m = \left[ \begin{array}{ccc} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & \ddots \\ & & \ddots & \ddots \end{array} \right]_m$$

mit der  $m \times m$ -Einheitsmatrix  $I_m$ . Die Eigenwerte dieser Matrix sind gegeben durch:

$$\lambda_{kl}(L_h) = 1 + \frac{1}{2}akh^{-2}\{4 - 2(\cos(kh\pi) + \cos(lh\pi))\}, \quad k, l = 1, \dots, m.$$

Damit ergibt sich ihre Spektralkondition zu

$$\text{cond}_2(L_h) = \frac{\lambda_{\max}(L_h)}{\lambda_{\min}(L_h)} \sim \frac{1 + 4akh^{-2}}{1 + ak\pi^2}. \quad (5.3.99)$$

Wir haben also unterschiedlich kritische Konditionierung abhängig von der Relation zwischen  $k$  und  $h$ . Im Hinblick auf eine Balancierung von Orts- und Zeitfehler ist die Wahl  $k \sim h$  sinnvoll. In diesem Fall ist dann

$$\kappa_2(L_h) = \mathcal{O}(h^{-1}). \quad (5.3.100)$$

Im parabolischen Fall ist in der Regel die Konditionierung der zu lösenden impliziten Gleichungssysteme also weniger kritisch als bei elliptischen Problemen. Bei Einhaltung der (natürlich unrealistischen) Schrittweitenrelation  $k \sim h^2$  wird sogar  $\kappa_2(L_h) = \mathcal{O}(1)$ , und der Lösungsaufwand der impliziten Verfahren nähert sich dem der expliziten.

Zur Lösung des Systems (5.3.98) können alle oben diskutierten Methoden verwendet werden. Da sich im autonomen Fall die Matrix  $L_h$  von Zeitschritt zu Zeitschritt nicht ändert, bietet sich die direkte Lösung mit Hilfe einer einmaligen Cholesky-Zerlegung zu Beginn der Rechnung an (wenn dies speichertechnisch möglich ist). Dieser Ansatz erfordert es aber, den Zeitschritt  $k$  konstant zu halten, was in den meisten praktischen Fällen nicht ökonomisch ist. I. Allg. muss in jedem Zeitschritt eine „neue“ Matrix invertiert werden, was die Verwendung iterativer Lösungsverfahren impliziert. Dabei steht mit  $U_h^{m-1}$  ein meist recht guter Startwert zur Verfügung. Bei Verwendung eines Mehrgitterverfahrens bietet sich daher die Organisation im F-Zyklus an. Häufig ist wegen der vergleichsweise moderaten Konditionierung der Matrizen  $L_h$  (bei kleiner Zeitschrittweite) zu ihrer Invertierung ein normales CG-Verfahren ausreichend schnell, so dass sich der Einsatz der komplizierten Mehrgitteriteration erübriggt. Dies hängt aber sehr von der jeweiligen konkreten Situation ab.

#### iv) Splitting-Methoden (ADI-Verfahren):

In höheren Raumdimensionen ist die Lösung der Gleichungssysteme in jedem Zeitschritt eines *impliziten* Verfahrens kostspielig und kann Rechnungen über sehr lange Zeitintervalle  $T \gg 1$  unmöglich machen. Der Übergang zu *expliziten* Schemata ist in der Regel wegen der damit verbundenen Zeitschrittrestriktion auch nicht sinnvoll. In dieser Situa-

tion stellen die sog. „Splitting-Methoden“ eine attraktive Alternative dar. Diese zerlegen die Lösung der vollen  $d$ -dimensionalen Gleichungssysteme in eine Folge von tridiagonalen Systemen (wie im 1-dimensionalen Fall), welche mit optimaler  $\mathcal{O}(N)$ -Komplexität gelöst werden können. Ein Vertreter dieses Verfahrenstyps ist das „ADI-Verfahren“ (Alternating Direction Implicit Iteration) nach Peaceman-Rachford, welches wir bereits im Zusammenhang mit iterativen Lösungsverfahren für spezielle „separable“ Gleichungssysteme kennengelernt haben. Bei mehrdimensionalen, parabolischen Problemen wird es nun als *Diskretisierungsverfahren* eingesetzt.

Wir betrachten wieder die Wärmeleitungsgleichung auf dem Einheitsquadrat  $\Omega = (0, 1)^2$ ,

$$\partial_t u - a \Delta u = 0 \quad \text{in } Q_T, \quad (5.3.101)$$

welche auf einem äquidistanten Gitter mit dem 5-Punkte-Differenzenoperator diskretisiert sei. Bei zeilenweiser Numerierung der Gitterpunkte haben die aus dem Crank-Nicolson-Verfahren resultierenden Gleichungssysteme die Gestalt

$$(I_h + \frac{1}{2}akA_h)U^m = (I_h - \frac{1}{2}akA_h)U^{m-1}. \quad (5.3.102)$$

Der Differenzenoperator  $A_h$  wird auf dem kartesischen Tensorprodukt-Gitter in seine Bestandteile bzgl. der einzelnen Koordinatenrichtungen zerlegt gemäß

$$A_h = A_{h,1} + A_{h,2}.$$

Entsprechend erhält das Gleichungssystem (5.3.102) des Crank-Nicolson-Schemas die Form

$$(I_h + \frac{1}{2}ak(A_{h,1} + A_{h,2}))U^m = (I_h - \frac{1}{2}ak(A_{h,1} + A_{h,2}))U^{m-1}$$

Die  $A_{h,i}$  sind Tridiagonalmatrizen. Es wird dann unter Einführung von Zwischenwerten  $U_h^{m-1/2}$  wie folgt iteriert:

$$\begin{aligned} (I_h + \frac{1}{2}akA_{h,1})U_h^{m-1/2} &= (I_h - \frac{1}{2}akA_{h,2})U_h^{m-1} \\ (I_h + \frac{1}{2}akA_{h,2})U_h^m &= (I_h - \frac{1}{2}akA_{h,1})U_h^{m-1/2}. \end{aligned}$$

Die ADI-Methode kann als ein Mehrschritt-Differenzenschema interpretiert werden, wobei allerdings die Zwischenwerte keine physikalische Relevanz haben. In jedem Teilschritt müssen Gleichungssysteme mit Tridiagonalgestalt gelöst werden. Wir wissen bereits von der Diskussion der iterativen Lösungsverfahren, dass der ADI-Algorithmus für jeden Wert des Parameters  $ak > 0$  gegen die Lösung  $U_h^\infty$  des Gleichungssystems  $A_h U_h^\infty = 0$  konvergiert. Dies ist auch „physikalisch“ sinnvoll, da ja im homogenen Fall ( $f \equiv 0$ ) auch für die exakte Lösung gilt  $u(t) \rightarrow 0$  ( $t \rightarrow \infty$ ). Damit erweist sich das ADI-Verfahren automatisch als *unbedingt numerisch stabil*. Durch Elimination des Zwischenwertes  $U_h^{m-1/2}$  erhalten wir

$$(I + \frac{1}{2}kA_{h,1})(I + \frac{1}{2}kA_{h,2})U_h^m = (I - \frac{1}{2}kA_{h,1})(I - \frac{1}{2}kA_{h,2})U_h^{m-1}. \quad (5.3.103)$$

Der Abschneidefehler dieser Differenzenformel erlaubt die Abschätzung

$$|\tau_{h,k}^m| \leq c \left\{ k^2 \max_{Q_T} |\partial_t^3 u| + h^2 \max_{Q_T} |\nabla^4 u| \right\}. \quad (5.3.104)$$

Eine Konvergenzanalyse ist leicht möglich, wenn wir wieder annehmen, dass die Zerlegungsmatrizen  $A_{h,1}$  und  $A_{h,2}$  kommutieren, d.h.:  $A_{h,1}A_{h,2} = A_{h,2}A_{h,1}$ . In diesem Fall besitzen sie ein gemeinsames ONS von Eigenvektoren  $\{v^{(n)}, n = 1, \dots, N\}$  zu Eigenwerten  $\lambda_n = \lambda_n(A_{h,1})$  und  $\mu_n = \mu_n(A_{h,2})$ . Die Koeffizienten in der Entwicklung

$$U_h^m = \sum_{\nu=1}^N \alpha_\nu^m v^{(\nu)}$$

werden dann durch das ADI-Schema wie folgt fortgepflanzt

$$\alpha_n^m = \frac{(1 - \frac{1}{2}k\lambda_n)(1 - \frac{1}{2}k\mu_n)}{(1 + \frac{1}{2}k\lambda_n)(1 + \frac{1}{2}k\mu_n)} \alpha_n^{m-1}. \quad (5.3.105)$$

Hieraus folgt wieder, analog zum Crank-Nicolson-Schema, direkt die unbedingte Stabilität des ADI-Schemas. Für die Fourier-Koeffizienten  $\alpha_n(t)$  der örtlich semi-diskreten Approximation  $u_h(t)$  gilt

$$\alpha_n(t_m) = e^{-k(\lambda_n + \mu_n)} \alpha_n(t_{m-1}). \quad (5.3.106)$$

Aus der Beziehung für  $z = z_1 + z_2$ ,  $z_i \leq 0$ ,

$$\frac{1 + \frac{1}{2}z_1}{1 - \frac{1}{2}z_1} \cdot \frac{1 + \frac{1}{2}z_2}{1 - \frac{1}{2}z_2} = \{e^{z_1} + \mathcal{O}(|z_1|^3)\} \{e^{z_2} + \mathcal{O}(|z_2|^3)\} = e^z + \mathcal{O}(|z|^3) \quad (5.3.107)$$

erhält man durch Adaption des Arguments, welches bei der Konvergenzanalyse der Padé-Verfahren verwendet wurde, den folgenden Satz.

**Satz 5.7 (ADI-Verfahren):** *Angewendet auf die 5-Punkte-Diskretisierung der Wärmeleitungsgleichung auf dem Einheitsquadrat ist das ADI-Verfahren für jede Zeitschrittweite  $k$  stabil und mit 2. Ordnung konvergent:*

$$\|U_h^m - u^m\|_h \leq c \left\{ k^2 \max_{Q_T} |\partial_t^3 u| + h^2 \max_{Q_T} |\nabla^4 u| \right\}. \quad (5.3.108)$$

**Beweis:** Der Beweis bedient sich wieder der Spektralmethode und wird ausgelassen.  
Q.E.D.

Der ADI-Ansatz ist generell auf (kartesischen) Tensorproduktgittern in beliebigen Raumdimensionen möglich, wenn der zugrunde liegende Differentialoperator „separabel“ ist, d.h. additiv in eindimensionale Operatoren zerfällt, wie z.B. der Operator

$$Lu = -\partial_1^2 u - \partial_2^2 u + \partial_1 u + \partial_2 u + u.$$

In allgemeineren Situationen (z. B. bei Auftreten gemischter Ableitungen  $\partial_1 \partial_2 u$  oder auf unstrukturierten Gittern) sind additive Zerlegungen von  $A_h$  in tridiagonale Teilmatrizen nicht mehr möglich, und der Wert des ADI-Ansatzes wird zweifelhaft.

## 5.4 Übungen

**Übung 5.1:** Das implizite Euler-Verfahren angewendet auf eine FE-Diskretisierung des (homogenen) Wärmeleitungsproblems

$$\partial_t u - \Delta u = 0, \quad t \geq 0, \quad u|_{t=0} = u^0, \quad u|_{\partial\Omega} = 0,$$

mit linearen oder bilinearen Ansatzfunktionen führt auf eine Folge von linearen Systemen

$$M_h U^m + k A_h U^m = M_h U^{m-1}, \quad m \geq 1, \quad U^0 = P_h u^0,$$

mit der zugehörigen Massematrix  $M_h$  und Steifigkeitsmatrix  $A_h$ . Man zeige:

- a) Die Auswertung der Elemente der Massematrix mit der Trapezregel ergibt eine Diagonalmatrix (sog. „Masche-Lumping“).
- b) Auf Triangulierungen ohne stumpfe Innenwinkel ist die durch Masche-Lumping entstehende Systemmatrix  $\tilde{M}_h + k A_h$  diagonal-dominant und vom nicht-negativen Typ, d. h. eine M-Matrix.
- c) Anspruchsvolle Zusatzaufgabe: Der Lumping-Prozess bewirkt einen Zusatzfehler der Größe  $|\tilde{U}^m - U^m| = \mathcal{O}(h^2)$ .

**Übung 5.2:** Man leite eine Bedingung für die  $L^2$ -Stabilität sowie die  $L^\infty$ -Stabilität des Wärmeleitungsproblems aus Aufgabe 13.1 auf dem Einheitswürfel im  $R^3$  her. Die Ortsdiskretisierung erfolge mit dem 7-Punkte-Differenzenoperator auf einem (äquidistanten) kartesischen Gitter mit Gitterweite  $h$ . (Hinweis: Man übertrage die Argumentation aus der Vorlesung von zwei auf drei Raumdimensionen.)

**Übung 5.3:** Die Wärmeleitungsgleichung

$$\partial_t u - a \Delta u = f, \quad t \geq 0, \quad u|_{t=0} = u^0, \quad u|_{\partial\Omega} = 0,$$

auf einem Polygonegebiet  $\Omega \subset \mathbb{R}^2$  mit Wärmeleitkoeffizient  $a > 0$  werde im Ort mit einem linearen FE-Ansatz auf einer quasi-gleichförmigen Folge von Triangulierungen der Gitterweite  $h$  und in der Zeit mit dem impliziten Euler-Schema mit Schrittweite  $k$  diskretisiert:

$$M_h U^m + k A_h U^m = M_h U^{m-1} + k b^m, \quad m \geq 1, \quad U^0 = P_h u^0.$$

Man untersuche die Abhängigkeit der Konditionierung der zugehörigen Systemmatrix  $M_h + k A_h$  von den Diskretisierungsparametern  $h$  und  $k$ . (Hinweis: Man betrachte zunächst den Spezialfall, dass  $\Omega$  das Einheitsquadrat mit einem gleichförmigen Rechteckgitter ist und die Massematrix  $M_h$  unter Anwendung der Trapezregel („Masche-Lumping“) nur näherungsweise berechnet wird.)

## 6 Verfahren für hyperbolische Probleme

Wir diskutieren zunächst wieder die klassischen Differenzenapproximationen zur Lösung hyperbolischer Anfangs-Randwert-Aufgaben (ARWA). Der Übersichtlichkeit halber beschränken wir uns dabei auf das Modellproblem der Wellengleichung in einer Ortsdimension mit Dirichletschen Randbedingungen, d.h. auf die 1. ARWA:

$$\begin{aligned}\partial_t^2 u - c^2 \partial_x^2 u &= f \quad \text{in } Q_T := (0, 1) \times [0, T], \\ u(0, t) = u(1, t) &= 0, \quad u(x, 0) = u^0, \quad \partial_t u(x, 0) = v^0,\end{aligned}\tag{6.0.1}$$

bzw. deren örtlich zweidimensionales Analogon

$$\begin{aligned}\partial_t^2 u - c^2 \Delta u &= f \quad \text{in } Q_T := \Omega \times [0, T], \\ u|_{\partial\Omega} &= 0, \quad u|_{t=0} = u^0, \quad \partial_t u|_{t=0} = v^0.\end{aligned}\tag{6.0.2}$$

Das Definitionsgebiet  $\Omega$  wird wieder als glatt berandet oder als konvexes Polygonegebiet vorausgesetzt. Die Problemdaten  $f, u^0, v^0$  sind ebenfalls glatt und kompatibel, so dass die Lösung ebenfalls als glatt angenommen werden kann. Unsere theoretischen Überlegungen haben gezeigt, dass bei hyperbolischen Problemen Irregularitäten in den Anfangsdaten oder der rechten Seite entlang der Charakteristiken fortgepflanzt werden. Im Gegensatz zu den elliptischen und parabolischen Problemen besitzen hyperbolische keinerlei „Glättungseigenschaft“. Lokale Störungen (bzw. Wellen) werden ungedämpft fortgepflanzt. Insbesondere gilt das Prinzip der „Energieerhaltung“, d.h.: Für  $f \equiv 0$  bleibt die Gesamtenergie (Summe aus kinetischer und elastischer Energie) in der Zeit erhalten:

$$\|\partial_t u(t)\|^2 + c^2 \|\nabla u(t)\|^2 = \|v^0\|^2 + c^2 \|\nabla u^0\|^2.\tag{6.0.3}$$

Diese charakteristische Eigenschaft sollten auch Diskretisierungen der Wellengleichung besitzen.

### 6.1 Differenzenverfahren für die Wellengleichung

Wir beginnen mit der örtlich eindimensionalen, homogenen Wellengleichung

$$\begin{aligned}\partial_t^2 u - c^2 \partial_x^2 u &= 0 \quad \text{in } Q_T := (0, 1) \times [0, T], \\ u(0, t) = u(1, t) &= 0, \quad u(x, 0) = u^0, \quad \partial_t u(x, 0) = v^0.\end{aligned}\tag{6.1.4}$$

Auf einem (äquidistanten) Orts-Zeit-Gitter  $\{x_n = nh, t_m = mk\}$  mit Ortsgitterweite  $h$  sowie Zeitschrittweite  $k$  lautet die zentrale Differenzenapproximation 2. Ordnung

$$k^{-2} \{U_n^m - 2U_n^{m-1} + U_n^{m-2}\} - c^2 h^{-2} \{U_{n-1}^{m-1} - U_n^{m-1} + U_{n+1}^{m-1}\} \tag{6.1.5}$$

zur Bestimmung der Approximationen  $U_n^m \sim u(x_n, t_m)$ . Dies ist eine (explizite) „Zweischrittformel“. Die Startwerte  $U_n^0$  und  $U_n^1$  werden aus den Anfangsbedingungen berechnet gemäß:

$$U_n^0 := u^0(x_n),$$

sowie unter Beachtung von

$$\begin{aligned} u(x_n, k) &= u(x_n, 0) + k \partial_t u(x_n, 0) + \frac{1}{2} k^2 \partial_t^2 u(x_n, 0) + \dots \\ &= u^0(x_n) + k v^0(x_n) + \frac{1}{2} c^2 k^2 \partial_x^2 u^0(x_n) + \dots \end{aligned}$$

durch

$$U_n^1 = u^0(x_n) + k v^0(x_n) + \frac{1}{2} c^2 k^2 h^{-2} \{ u^0(x_{n-1}) - 2u^0(x_n) + u^0(x_{n+1}) \}.$$

Mit dem Quotienten  $\sigma := k/h$  gilt für den zugehörigen Abschneidefehler:

$$\tau_{h,k} = \{ \partial_t^2 u - c^2 \partial_x^2 u \} + \frac{1}{12} h^2 (\sigma^2 - c^{-2}) \partial_t^4 u + \frac{1}{360} h^4 (\sigma^4 - c^{-4}) \partial_t^6 u + \dots$$

Offenbar ist  $\tau_{h,k} \equiv 0$  für  $\sigma = c^{-1}$ , d.h.: Die explizite Differenzenformel

$$U_n^m = U_{n-1}^{m-1} + U_{n+1}^{m-1} - U_n^{m-2}$$

ist eine „exakte“ Differenzendarstellung der Wellengleichung. Das Abhängigkeitsgebiet der Differenzenformel hängt offenbar von der Schrittweitenrelation  $\sigma = k/h$  ab.

1. Fall  $0 < \sigma \leq c^{-1}$ : Abhängigkeitsgebiet der Differenzenformel enthält das der Differentialgleichung;
2. Fall  $\sigma = c^{-1}$ : „Übereinstimmung“;
3. Fall  $\sigma > c^{-1}$ : Abhängigkeitsgebiet der Differenzenformel ist enthalten in dem der Differentialgleichung.

**Satz 6.1 (CFL-Bedingung):** *Notwendig für die Konvergenz*

$$U_n^m \rightarrow u(x_n, t^m) \quad (h, k \rightarrow 0), \tag{6.1.6}$$

für beliebige Anfangsdaten ist die Schrittweitenbedingung (sog. Bedingung von Courant<sup>1</sup>-Friedrichs<sup>2</sup>-Lewy<sup>3</sup>, kurz „CFL-Bedingung“)

$$\sigma := \frac{k}{h} \leq c^{-1}. \tag{6.1.7}$$

<sup>1</sup>Richard Courant (1888–1972): Deutscher Mathematiker; Prof. in Münster und Göttingen, nach Vertreibung durch die Nazis 1933 Prof. an der New York University, Gründer des berühmten, später nach ihm benannten „Courant-Instituts“; Beiträge zur Funktionentheorie und Mathematischen Physik, „Erfinder“ der Idee der Finite-Elemente-Methode (publiziert 1943 nach Vorarbeiten aus 1922).

<sup>2</sup>Otto Paul Friedrichs (1901–1982): Deutscher Mathematiker; Prof. in Braunschweig, emigrierte 1937 nach USA ans Courant-Institut in New York; wichtige Beiträge zu partiellen Differentialgleichungen der mathematischen Physik.

<sup>3</sup>Hans Lewy (1904–1988): Deutscher Mathematiker; 1927 Promotion in Göttingen bei F. Courant, dort auch Privatdozent; die u. a. nach ihm benannte „CFL-Stabilitätsbedingung“ ist in einer Arbeit von 1928 „Über die partiellen Differenzengleichungen der mathematischen Physik“ (Math. Annalen 100, 32–74) enthalten; 1933 Entlassung aus dem Staatsdienst und Emigration in die USA, ab 1935 Prof. in Berkeley; Beiträge zur Theorie partieller Differentialgleichungen.

**Beweis:** Sei  $\sigma > c^{-1}$ . Wenn in irgendeinem festen Gitterpunkt  $(x_n, t_m)$  für eine spezielle Anfangsbedingung  $U_n^m \rightarrow u(x_n, t_m)$  konvergiert für  $h, k \rightarrow 0$ , so können wir diese Anfangsdaten außerhalb des Abhängigkeitsintervalls von  $(x_n, t_m)$  bzgl. der Differenzenformel beliebig ändern, ohne dass  $U_n^m$  verändert wird. In diesem Fall kann also  $U_n^m$  nicht gegen die veränderte Lösung  $\tilde{u}(x_n, t_m)$  konvergieren. Q.E.D.

**Bemerkung 6.1:** Wir weisen darauf hin, dass für spezielle Anfangsdaten durchaus (theoretische) Konvergenz auch für  $\sigma > c^{-1}$  eintreten kann; in diesem Fall liegt aber numerische Instabilität vor. Die Schrittweitenbedingung (6.1.7) ist weniger restriktiv als die entsprechende Bedingung  $k \leq \frac{1}{2}ah^2$  bei der Wärmeleitungsgleichung.

Im folgenden werden wir die Konvergenz des Differenzenschemas (6.1.5) untersuchen. Dabei bedienen wir uns der Spektraltechnik, die wir bereits bei parabolischen Problemen kennengelernt haben. Der örtliche Differenzenoperator

$$A_h U_n^m := -\frac{c^2}{h^2} \{U_{n-1}^m - 2U_n^m + U_{n+1}^m\} \quad (6.1.8)$$

( $U_0^m = U_{N+1}^m = 0$ ) ist symmetrisch und positiv definit bzgl. des diskreten Skalarprodukts

$$(v, w)_h := h^2 \sum_{n=1}^N v_n w_n, \quad \|v\|_h := (v, v)_h^{1/2}.$$

Seine Eigenwerte seien  $0 < \lambda_1 \leq \dots \leq \lambda_N \sim 4c^2/h^2$  mit einem zugehörigen Orthonormalsystem  $\{w^{(n)}, n = 1, \dots, N\}$  von Eigenvektoren. Insbesondere gilt

$$(A_h v, v)_h \geq \lambda \|v\|_h^2 \quad (6.1.9)$$

mit einer von  $h$  unabhängigen Konstante  $\lambda > 0$ .

**Satz 6.2 (CFL-Bedingung):** Die explizite Differenzenformel (6.1.5) ist genau dann numerisch stabil, wenn die CFL-Bedingung  $k/h \leq c^{-1}$  erfüllt ist. Im Falle einer hinreichend glatten Lösung gilt dann die Konvergenzabschätzung

$$\max_{[0,T]} \|U_h^m - u(\cdot, t_m)\|_h \leq c(u)T^2\{k^2 + h^2\}. \quad (6.1.10)$$

**Beweis:** i) Für die Entwicklungskoeffizienten in

$$U_h^m = \sum_{n=1}^N a_n^m w^{(n)}$$

gilt

$$a_n^m - 2a_n^{m-1} + a_n^{m-2} + k^2 \lambda_n a_n^{m-1} = 0$$

bzw.

$$a_n^m + (k^2 \lambda_n - 2)a_n^{m-1} + a_n^{m-2} = 0. \quad (6.1.11)$$

Diese homogene Differenzengleichung hat die allgemeine Lösung

$$a_n^m = c_1 r_1^m + c_2 r_2^m$$

mit den Wurzeln  $r_i$  des charakteristischen Polynoms  $\rho(r) = r^2 + (k^2 \lambda_n - 2)r + 1$ :

$$r_{1,2} = \frac{2 - k^2 \lambda_n \pm \sqrt{(2 - k^2 \lambda_n)^2 - 4}}{2}.$$

Im Fall  $k^2 \lambda_n \leq 4$  ist  $|r_{1,2}| \leq 1$ , d. h.: Es liegt Stabilität vor. Im Fall  $k^2 \lambda_n > 4$  ist  $|r_2| > 1$ , d. h.: Es besteht Instabilität. Offenbar gilt für  $h \rightarrow 0$ :

$$k^2 \lambda_n \leq 4 \Leftrightarrow \sigma \leq c^{-1}. \quad (6.1.12)$$

ii) Sei nun  $k^2 \lambda_n \leq 4$  und die Lösung  $u$  glatt. Wir betrachten den Fehler der Ortsdiskretisierung getrennt von dem der Zeitdiskretisierung:

$$e_n^m := u(x_n, t_m) - U_n^m = u(x_n, t_m) - u_h(x_n, t_m) + u_h(x_n, t_m) - U_n^m =: \varepsilon_h(x_n, t_m) + E_n^m.$$

Für den Ortsdiskretisierungsfehler  $\varepsilon_h$  gilt  $\varepsilon_h^0 = \partial_t \varepsilon_h^0 = 0$  und

$$\partial_t^2 \varepsilon_h + A_h \varepsilon_h = \mathcal{O}(h^2).$$

Wir multiplizieren diese Identität mit  $\partial_t \varepsilon_h$ ,

$$\frac{1}{2} d_t \{ \| \partial_t \varepsilon_h \|_h^2 + (A_h \varepsilon_h, \varepsilon_h)_h \} = (\mathcal{O}(h^2), \partial_t \varepsilon_h)_h,$$

und integrieren über  $[0, t]$ :

$$\| \partial_t \varepsilon_h(t) \|_h^2 + (A_h \varepsilon_h(t), \varepsilon_h(t))_h = \| \varepsilon_h(0) \|_h^2 + (A_h \varepsilon_h(0), \varepsilon_h(0))_h + \int_0^t (\mathcal{O}(h^2), \partial_s \varepsilon_h)_h ds.$$

Dies impliziert

$$\max_{[0,t]} \{ \| \partial_t \varepsilon_h \|_h^2 + (A_h \varepsilon_h, \varepsilon_h)_h \} \leq t \mathcal{O}(h^2) \max_{[0,t]} \| \partial_t \varepsilon_h \|_h$$

bzw. nach Aufintegrieren bzgl. der Zeit:

$$\max_{[0,t]} \| \varepsilon_h \|_h \leq c t^2 h^2. \quad (6.1.13)$$

Der Faktor  $t^2$  lässt sich hier nicht vermeiden, da bei der Wellengleichung (im Gegensatz zur Wärmeleitungsgleichung) lokale Störungen *nicht* ausgedämpft werden.

iii) Für den Zeitdiskretisierungsfehler gilt

$$k^{-2}\{E^m - 2E^{m-1} + E^{m-2}\} + A_h E^{m-1} = \mathcal{O}(k^2). \quad (6.1.14)$$

Die Konstante in  $\mathcal{O}(k^2)$  hängt dabei von den Zeitableitungen von  $u_h(t)$  bis zur Ordnung 4 ab. Diese lassen sich durch die entsprechenden Zeitableitungen von  $u$  beschränken, was hier jedoch nicht ausgeführt wird. Für die  $n$ -te Fourier-Komponente  $E_n^m = (E^m, w^{(n)})_h$  gilt

$$E_n^m - 2E_n^{m-1} + E_n^{m-2} + k^2 \lambda_n E_n^{m-1} = (\mathcal{O}(k^4), w^{(n)})_h,$$

gleichmäßig bzgl.  $n$ . Das charakteristische Polynom dieser Differenzengleichung ist  $\rho(r) = r^2 + (k^2 \lambda_n - 2)r + 1$  mit Wurzeln  $|r_{1,2}| = 1$ . Die a priori Abschätzung für Lösungen inhomogener Differenzengleichungen aus Hilfssatz 6.1 liefert also:

$$|E_n^m| \leq c \left\{ \max_{\mu=0,1} |E_n^\mu| + m^2 \max_{\nu=2,\dots,m} \mathcal{O}(k^4) \right\}$$

bzw.

$$\|E^m\|_h^2 = \sum_{\nu=1}^m |E_n^\nu|^2 \leq c \{ \|E^0\|_h^2 + \|E^1\|_h^2 \} + ct_m^4 k^4.$$

Offenbar ist  $E^0 = 0$  und

$$\begin{aligned} E^1 &= u_h(t_1) - U^1 = u_h(t_1) - u^0 - ku^1 - \frac{1}{2}k^2 A_h u^0 \\ &= -\varepsilon(t_1) + u(t_1) - u^0 - ku^1 - \frac{1}{2}k^2 c^2 \partial_x^2 u(0) + \mathcal{O}(k^2 h^2) \\ &= -\varepsilon(t_1) + u(t_1) - u^0 - ku^1 - \frac{1}{2}k^2 \partial_t^2 u(0) + \mathcal{O}(k^2 h^2) \\ &= \mathcal{O}(k^3 + k^2 h^2) + \mathcal{O}(k^3). \end{aligned}$$

Dies führt auf

$$\|E^m\|_h^2 \leq c t_m^2 \{h^2 + k^2\}^2. \quad (6.1.15)$$

Kombination der Abschätzungen (6.1.13) und (6.1.15) vervollständigt schließlich den Beweis.  
Q.E.D.

**Hilfssatz 6.1 (Differenzengleichungen):** *Die Folge  $\{y_m\}_{m \in \mathbb{N}}$  genüge der linearen, inhomogenen Differenzengleichung*

$$\sum_{\nu=0}^R a_\nu y_{m+\nu} = g_m, \quad m \geq 0. \quad (6.1.16)$$

*Wenn alle Nullstellen  $\lambda_\nu$  des charakteristischen Polynoms*

$$\rho(z) := \sum_{\nu=0}^R a_\nu z^\nu$$

Betrag  $|\lambda_\nu| \leq 1$  haben, gilt die a priori Abschätzung

$$\max_{R \leq \nu \leq m} |y_\nu| \leq c_R \left\{ \max_{0 \leq \nu \leq R} |y_\nu| + m^2 \max_{0 \leq \nu \leq m} |g_\nu| \right\}, \quad m \geq R. \quad (6.1.17)$$

**Beweis:** Siehe das Kapitel über Mehrschrittmethoden im Band „Numerik 1 (Numerik gewöhnlicher Differentialgleichungen)“ [2]. Q.E.D.

**Bemerkung 6.2:** Die bisher erhaltenen Aussagen bleiben gültig, wenn man das explizite Differenzenschema auf das reine Anfangswertproblem („Cauchy-Problem“) der Wellengleichung anwendet. In diesem Fall ist man auf explizite Verfahren angewiesen, da zur Verwendung impliziter Formeln die notwendigen Randwerte fehlen.

Bei der ARWA der Wellengleichung kann man sich von der einschränkenden CFL-Bedingung durch Verwendung impliziter Differenzenformeln befreien, etwa der Art

$$k^{-2} \{U_h^m - 2U_h^{m-1} + U_h^{m-2}\} + \alpha A_h U_h^m + (1 - 2\alpha) A_h U_h^{m-1} + \alpha A_h^{m-2} = 0 \quad (6.1.18)$$

mit einem Parameter  $\alpha \in [0, 1]$ . Diese Formel hat den Abschneidefehler

$$\tau_h = h^2 \left\{ \left\{ \frac{1}{12} (\sigma^2 - 1) - \alpha \sigma^2 \right\} \partial_x^4 u + \mathcal{O}(h^2) \right\};$$

Sie ist also für beliebiges  $\alpha$  von zweiter Ordnung. Für ein implizites Differenzenschema enthält ihr Abhängigkeitsbereich offensichtlich den der Differentialgleichung.

**Satz 6.3 (Konvergenz impliziter Verfahren):** Die implizite Differenzenformel (6.1.18) ist im Falle  $\alpha \geq 1/4$  unbedingt stabil und im Fall  $0 < \alpha < 1/4$  stabil für

$$0 < \sigma \leq \frac{1}{c\sqrt{1 - 4\alpha}}; \quad (6.1.19)$$

Für andere  $\sigma$  ist sie instabil. Im stabilen Fall gilt die Konvergenzabschätzung

$$\max_{[0,T]} \|U_h^m - u(\cdot, t_m)\|_h \leq c(u) T \{k^2 + h^2\}. \quad (6.1.20)$$

**Beweis:** Für die Entwicklungskoeffizienten in

$$U_h^m = \sum_{n=1}^N a_n^m w^{(n)}$$

gilt

$$a_n^m - 2a_n^{m-1} + a_n^{m-2} + k^2 \lambda_n \{ \alpha a_n^m + (1 - 2\alpha) a_n^{m-1} + \alpha a_n^{m-2} \} = 0.$$

Das charakteristische Polynom dieser Differenzenformel

$$\rho(r) = (1 + \alpha k^2 \lambda_n) r^2 + (k^2 \lambda_n (1 - 2\alpha) - 2) r + (1 + k^2 \lambda_n \alpha)$$

hat die Wurzeln

$$r_{1,2} = \frac{2 - k^2 \lambda_n (1 - 2\alpha) \pm \sqrt{(k^2 \lambda_n (1 - 2\alpha) - 2)^2 - 4(1 + k^2 \lambda_n \alpha)^2}}{2(1 + \alpha k^2 \lambda_n)}.$$

Wir haben zwei Fälle zu unterscheiden:

a) Stabiler Fall:

$$(k^2 \lambda_n (1 - 2\alpha) - 2)^2 \leq 4(1 + k^2 \lambda_n \alpha)^2 \Rightarrow |r_{1,2}| = 1. \quad (6.1.21)$$

b) Instabiler Fall

$$(k^2 \lambda_n (1 - 2\alpha) - 2)^2 > 4(1 + k^2 \lambda_n \alpha)^2 \Rightarrow |r_2| > 1. \quad (6.1.22)$$

Der Kette äquivalenter Ungleichungen

$$\begin{aligned} k^2 \lambda_n (1 - 2\alpha) - 2 &\leq 2 + 2k^2 \lambda_n \alpha \\ k^2 \lambda_n - 2\alpha k^2 \lambda_n &\leq 4 + 2\alpha k^2 \lambda_n \\ k^2 \lambda_n - 4\alpha k^2 \lambda_n &\leq 4 \\ \frac{4k^2 c^2}{h^2} (1 - 4\alpha) &\leq 4 \\ \frac{k^2}{h^2} (1 - 4\alpha) &\leq c^{-2} \end{aligned}$$

entnehmen wir die Bedingungen  $\alpha \geq 1/4$  oder

$$0 < \alpha < 1/4, \quad \sigma \leq \frac{1}{c\sqrt{1 - 4\alpha}}. \quad (6.1.23)$$

Dagegen ist

$$\begin{aligned} k^2 \lambda_n (1 - 2\alpha) - 2 &\geq -2 - 2k^2 \lambda_n \alpha \\ k^2 \lambda_n - 2\alpha k^2 \lambda_n &\geq -2\alpha k^2 \lambda_n \\ k^2 \lambda_n &\geq 0, \end{aligned}$$

stets erfüllt. Dies beweist den die Stabilität betreffenden Teil des Satzes. Die Konvergenzabschätzung wird dann ähnlich wie im expliziten Fall gezeigt. Wir lassen die Details weg.

Q.E.D.

In zwei Raumdimensionen ist der Abhängigkeitsbereich der Wellengleichung kegelförmig (z. B. Kreiskegel bei kreisförmigem Grundgebiet). Die obigen Aussagen für den eindimensionalen Fall gelten sinngemäß auch in zwei Dimensionen. Bei Ortsdiskretisierung mit dem 5-Punkte-Operator  $A_h$  lautet das Analogon des expliziten Schemas (6.1.5)

$$k^{-2} \{U_h^m - 2U_h^{m-1} + U_h^{m-2}\} + c^2 A_h U_h^{m-1} = 0 \quad (6.1.24)$$

und hat die Stabilitätsbedingung

$$\sigma \leq \frac{1}{\sqrt{2}c}. \quad (6.1.25)$$

In drei Raumdimensionen verschärft sich diese Bedingung zu  $\sigma \leq (\sqrt{3}c)^{-1}$ .

## 6.2 Finite-Elemente-Verfahren für die Wellengleichung

Als Basis von Finite-Elemente-Diskretisierungen dient wieder die variationelle Formulierung von (6.0.2):

$$(\partial_t^2 u, \varphi) + (a \nabla u, \nabla \varphi) = (f, \varphi) \quad \forall \varphi \in V, \quad t > 0, \quad u|_{t=0} = 0, \quad (6.2.26)$$

mit dem üblichen Sobolew-Raum  $V := H_0^1(\Omega)$ . Im Sinne der „Rothe-Methode“ könnten diese Gleichung nun zunächst wieder mit einem Differenzenschema 2. Ordnung in der Zeit und anschließend die einzelnen Zeitschritte mit FE-Ansätzen im Ort diskretisiert werden. Dies führt wie bei den Differenzenverfahren zwangsläufig auf Zwei-Schritt-Schemata, mit denen die Energieerhaltung nicht zu bewerkstelligen ist. Wir wollen daher jetzt einen anderen Weg beschreiten, der etwas näher an der Vorgehensweise bei parabolischen Problemen ist. Durch Einführung der zusätzlichen Unbekannten  $v = \partial_t u$  geht die Wellengleichung über in das System

$$\partial_t u - v = 0, \quad (6.2.27)$$

$$\partial_t v - c^2 \Delta u = f, \quad (6.2.28)$$

mit den natürlichen Randbedingungen  $u|_{\partial\Omega} = v|_{\partial\Omega} = 0$  sowie den Anfangsbedingungen  $u|_{t=0} = u^0$  und  $v|_{t=0} = v^0$ . In variationeller Form schreibt sich dies wie

$$(\partial_t v, \varphi) - (v, \psi) = 0 \quad \forall \psi \in V, \quad t \in [0, T], \quad (6.2.29)$$

$$(\partial_t u, \psi) + c^2 (\nabla u, \nabla \varphi) = (f, \varphi), \quad \forall \varphi \in V, \quad (6.2.30)$$

mit den Anfangsbedingungen  $u|_{t=0} = u^0$  und  $v|_{t=0} = v^0$ . Dieses System von Differentialgleichungen ist wieder „hyperbolisch“, da alle Eigenwerte der Koeffizientenmatrix rein imaginär sind. Zur Diskretisierung wäre also ein Zeitschrittverfahren günstig, bei dem die imaginäre Achse gerade der Rand des Stabilitätsgebiets ist. Das Crank-Nicolson-Verfahren besitzt diese Eigenschaft.

Zur Diskretisierung dieses Systems verwenden wir das Rothe-Verfahren, d. h.: Zunächst wird bzgl. der Zeit diskretisiert. Dazu verwenden wir auf einem Zeitgitter

$$0 = t_0 < t_1 < \dots < t_m < \dots < t_M = T,$$

mit Schrittweiten  $k_m = t_m - t_{m-1}$  das Crank-Nicolson-Schema:

$$(u^m - u^{m-1}, \psi) - \frac{1}{2} k_m (v^m + v^{m-1}, \psi) = 0, \\ (v^m - v^{m-1}, \varphi) + \frac{1}{2} k_m (c^2 \nabla(u^m + u^{m-1}), \nabla \varphi) = 0,$$

für alle Testfunktionen  $\varphi$  und  $\psi$  mit Anfangswerten  $u^0$  und  $v^0$ . Bei diesem Zeitschritt-

verfahren bleibt die totale Energie in jedem einzelnen Zeitschritt erhalten. Dazu setzen wir im variationellen Schema  $\varphi := u^m - u^{m-1}$  und  $\psi := v^m - v^{m-1}$  und kombinieren die beiden resultierenden Gleichungen zu

$$\frac{1}{2} \|v^m\|^2 + \frac{1}{2} \|c\nabla u^m\|^2 = \frac{1}{2} \|v^{m-1}\|^2 + \frac{1}{2} \|c\nabla u^{m-1}\|^2. \quad (6.2.31)$$

Die einzelnen Probleme in jedem Zeitschritt  $t_{m-1} \rightarrow t_m$  werden nun mit Hilfe eines FE-Verfahrens diskretisiert. Dazu werden zu jedem Zeitlevel  $t_m$  FE-Ansatzräume  $V_h^m \subset V = H_0^1(\Omega)$  auf Gittern  $T_H^m$  der üblichen Art gewählt. Im folgenden betrachten wir zunächst den Fall, dass die Gitter und Ansatzräume zu allen Zeitpunkten dieselben sind. Die allgemeine Situation von mit der Zeit variierten Ortgittern ist in Abb. 6.1 skizziert.

Zu dem FE-Ansatz gehören wieder die Masse- und Steifigkeitsmatrizen

$$M_h = (m_{ij})_{ij} = ((\varphi_h^{(j)}, \varphi_h^{(i)}))_{ij}, \quad A_h = (a_{ij})_{ij} = ((c^2 \nabla \varphi_h^{(j)}, \nabla \varphi_h^{(i)}))_{ij},$$

wobei  $\{\varphi_h^{(i)}, i = 1, \dots, N\}$  die Knotenbasis von  $V_h$  ist. Dabei wird angenommen, dass beide Unbekannte  $u_h$  sowie  $v_h$  in demselben Ansatzraum  $V_h$  bestimmt werden. Dies ist wegen der physikalisch vorgegebenen Randbedingung  $v|_{\partial\Omega} = \partial_t u|_{\partial\Omega} = 0$  sinnvoll. Eigentlich bräuchte  $v_h$  im Hinblick auf die gewählte variationelle Formulierung aber nur in  $L^2(\Omega)$  gewählt zu werden.

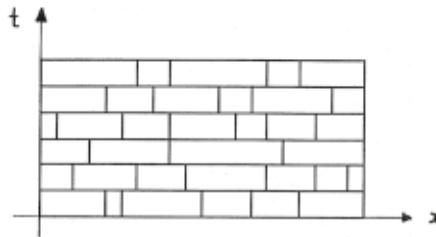


Abbildung 6.1: Raum-Zeit-Gitter mit „hängenden Knoten“

Bezeichnen wir nun die zugehörigen Knotenwertvektoren von  $u_h^m$  und  $v_h^m$  ebenfalls mit  $u_h^m$  und  $v_h^m$ , so erhält das Zeitschrittschema die Gestalt

$$\begin{aligned} M_h(U_h^m - U_h^{m-1}) + \frac{1}{2} k_m M_h(V_h^m - V_h^{m-1}) &= 0, \\ M_h(V_h^m - V_h^{m-1}) + \frac{1}{2} k_m A_h(U_h^m + U_h^{m-1}) &= 0. \end{aligned}$$

Dies kann nun so umgeformt werden, dass ein System von zwei sukzessive lösbarren Problemen entsteht:

$$\begin{aligned} (M_h + \frac{1}{4} k_m^2 A_h) U_h^m &= M_h U_h^{m-1} + k_m M_h V_h^{m-1} - \frac{1}{4} k_m^2 A_h U_h^{m-1}, \\ M_h V_h^m &= M_h V_h^{m-1} - \frac{1}{2} k_m A_h (U_h^m + U_h^{m-1}). \end{aligned}$$

In jedem Zeitschritt sind also eine (modifizierte) Ritz-Projektion sowie eine  $L^2$ -Projektion durchzuführen. Die Konvergenzanalyse dieses Verfahrens kann wieder mit Hilfe der oben schon beschriebenen „Energie-Technik“ erfolgen, was hier aber nicht ausgeführt wird.

### 6.3 Lösungsaspekte

Bei der Anwendung unbedingt stabiler, impliziter Differenzenschemata müssen in jedem Zeitschritt lineare Gleichungssysteme mit Koeffizientenmatrizen der Form  $I_h + \alpha k^2 A_h$  gelöst werden. Deren Kondition verhält sich im Falle  $k \sim h$  wie

$$\kappa_2(A_h) \sim 1. \quad (6.3.32)$$

Auf solchen gleichförmigen Gittern sind implizite Verfahren also verhältnismäßig kostengünstig. Dies ändert sich aber, wenn das Ortgitter zur Anpassung an irreguläre Lösungsstrukturen lokal verfeinert wird.

### 6.4 Übungen

**Übung 6.1:** Zum Abschluss noch ein paar Testfragen zum vorausgegangenen Stoff:

1. Wodurch unterscheiden sich das „Ritzsche Projektionsverfahren“ vom allgemeinen „Galerkin-Verfahren“, und was ist ein „Petrow-Galerkin-Verfahren“?
2. Was ist neben der Grobgitterkorrektur der wichtigste Bestandteil eines Mehrgitterverfahrens?
3. Welche ist die mit einem „quadratischen“ Finite-Elemente-Ansatz maximal erreichbare Konvergenzordnung  $\mathcal{O}(h^r)$  (bzgl. einer „minimalen“ Fehlernorm)?
4. Was versteht man unter der „Maximalwinkel-“ bzw. der „Minimalwinkel-Bedingung“ für Triangulierungen?
5. Was versteht man unter „Glättungseigenschaft“ eines Zeitschrittverfahrens, und besitzt das Crank-Nicolson-Verfahren diese Eigenschaft?
6. Was ist ein „isoparametrischer“ Finite-Elemente-Ansatz?
7. Was ist die Faustregel hinsichtlich der erforderlichen Ordnung von Quadraturformeln zur Berechnung der Steifigkeitsmatrix eines Finite-Elemente-Verfahrens?
8. Welcher Schrittweitenbedingung unterliegt das explizite Euler-Verfahren zur Lösung der mit dem 5-Punkte-Differenzenoperator im Ort diskretisierten Wärmeleitungsgleichung?
9. Welche Kondition in Abhängigkeit von der Gitterweite hat die Steifigkeitsmatrix einer Finite-Elemente-Diskretisierung der „Plattengleichung“ (biharmonischer Operator)?
10. Was ist eine „M-Matrix“, und welche Eigenschaften folgen daraus?

# A Lösungen der Übungsaufgaben

Im Folgenden sind Lösungen für die am Ende der einzelnen Kapitel formulierten Aufgaben zusammengestellt. Es handelt sich dabei nicht um „Musterlösungen“ mit vollständig ausformuliertem Lösungsweg, sondern nur um Lösungsansätze in knapper Form.

## A.1 Kapitel 1

**Lösung A.1.1:** Aus der Beschränktheit der ersten Ableitung von  $f$  folgt die Lipschitz-Stetigkeit von  $f$  bzgl.  $x$ , und es ist

$$\|f(t, x)\|_\infty \leq \|f(t, x) - f(t, 0)\|_\infty + \|f(t, 0)\|_\infty \leq K\|x\|_\infty + \|f(t, 0)\|_\infty$$

Somit existiert nach dem globalen Existenzsatz genau eine globale Lösung der AWA. Und diese ist nach dem Regularitätssatz aus  $C^\infty$ .

Sei nun  $u$  diese Lösung. So ist

$$u^{(k)} = f^{(k-1)}$$

Und damit hat die Taylor-Reihe von  $u$  im Entwicklungspunkt  $t_0$  gerade die angegebene Form. Es bleibt zu zeigen, dass die Taylorreihe für alle  $t > t_0$  konvergiert. Dazu betrachten wir das  $(n+1)$ -te Restglied der Taylorentwicklung von  $u$ .

$$R_{n+1} = \frac{f^{(n)}(\xi, u(\xi))}{(n+1)!} (t - t_0)^{n+1}$$

mit  $\xi \in (t_0, t)$ . Aus der gleichmäßigen Beschränktheit der Ableitungen von  $f$  folgt

$$R_{n+1} \rightarrow 0 \quad (n \rightarrow \infty).$$

und damit die Behauptung.

**Lösung A.1.2:** Wir setzen  $p := \partial_x u$ ,  $q := \partial_y u$ ,  $r := \partial_x^2 u$ ,  $s := \partial_x \partial_y u$ ,  $t := \partial_y^2 u$  und

$$\alpha := \partial_x^3 u, \beta := \partial_x^2 \partial_y u, \gamma = \partial_x \partial_y^2 u, \delta = \partial_y^3 u.$$

Differenzieren der Differentialgleichung ergibt

$$\begin{aligned} a_{11}\partial_x^3 u + 2a_{12}\partial_x^2 \partial_y u + a_{22}\partial_x \partial_y^2 u &= \partial_x f - a_{01}\partial_x^2 u - a_{02}\partial_x \partial_y u - a_{00}\partial_x u \\ a_{11}\partial_x^2 \partial_y u + 2a_{12}\partial_x \partial_y^2 u + a_{22}\partial_y^3 u &= \partial_y f - a_{01}\partial_x \partial_y u - a_{02}\partial_y^2 u - a_{00}\partial_y u \end{aligned}$$

Differenzieren von  $r, s, t$  entlang  $\Gamma$  ergibt

$$\begin{aligned} \partial_\tau r &= \partial_x r \partial_\tau x + \partial_y r \partial_\tau y = \alpha \partial_\tau x + \beta \partial_\tau y \\ \partial_\tau s &= \partial_x s \partial_\tau x + \partial_y s \partial_\tau y = \beta \partial_\tau x + \gamma \partial_\tau y \\ \partial_\tau t &= \partial_x t \partial_\tau x + \partial_y t \partial_\tau y = \gamma \partial_\tau x + \delta \partial_\tau y \end{aligned}$$

Zusammengenommen ergeben sich zwei  $3 \times 3$ -Gleichungssysteme für die gesuchten Ableitungen  $\alpha, \beta, \gamma, \delta$ :

$$\begin{aligned} a_{11}\alpha + 2a_{12}\beta + a_{22}\gamma &= \partial_x f - a_{01}r - a_{02}s - a_{00}p \\ \partial_\tau x\alpha + \partial_\tau y\beta &= \partial_\tau r \\ \partial_\tau x\beta + \partial_\tau y\gamma &= \partial_\tau s \end{aligned}$$

$$\begin{aligned} a_{11}\beta + 2a_{12}\gamma + a_{22}\delta &= \partial_y f - a_{01}s - a_{02}t - a_{00}q \\ \partial_\tau x\beta + \partial_\tau y\gamma &= \partial_\tau s \\ \partial_\tau x\gamma + \partial_\tau y\delta &= \partial_\tau t \end{aligned}$$

Beide haben dieselbe Koeffizientenmatrix  $B$  wie das entsprechende System zur Bestimmung der zweiten Ableitungen:

$$\begin{aligned} a_{11}r + 2a_{12}s + a_{22}t &= f - a_{01}p - a_{02}q - a_{00}u \\ \partial_\tau xr + \partial_\tau ys &= \partial_\tau p \\ \partial_\tau xs + \partial_\tau yt &= \partial_\tau q. \end{aligned}$$

Man überlegt sich leicht, dass im Falle  $\det B \neq 0$  die durch die beiden Gleichungssysteme bestimmten vier dritten Ableitungen eindeutig (und widerspruchsfrei) bestimmt sind.

**Lösung A.1.3:** Der Differentialoperator  $L = a_{11}\partial_x + 2a_{12}\partial_x\partial_y + a_{22}\partial_y^2 + \dots$  ist „elliptisch“ für  $a_{12}^2 - a_{11}a_{22} < 0$ , „parabolisch“ für  $a_{12}^2 - a_{11}a_{22} = 0$  und „hyperbolisch“ für  $a_{12}^2 - a_{11}a_{22} > 0$ .

- a) Der Operator  $L = \partial_x\partial_y - \partial_x$  ist wegen  $a_{12}^2 - a_{11}a_{22} = \frac{1}{4} > 0$  hyperbolisch.
- b) Der Operator  $L = \partial_x^2 + \partial_x\partial_y + y\partial_y^2 + 4$  ist wegen  $a_{12}^2 - a_{11}a_{22} = \frac{1}{4} - y$  hyperbolisch für  $y < \frac{1}{4}$ , parabolisch für  $y = \frac{1}{4}$  und elliptisch für  $y > \frac{1}{4}$ .
- c) Der Operator  $L = 2(\partial_x + \partial_y)^2 + \partial_y = 2\partial_x^2 + 4\partial_x\partial_y + 2\partial_y^2 + \partial_y$  ist wegen  $a_{12}^2 - a_{11}a_{22} = 4 - 4 = 0$  parabolisch.

**Lösung A.1.4:** Wir wollen im Folgenden die ersten partiellen Ableitungen der  $u_i$  ( $i = 1, 2$ ) bestimmen. Dazu führen wir folgende Abkürzungen ein:

$$r_i := \partial_x u_i \quad s_i := \partial_y u_i$$

Durch Ableiten in Tangentialrichtung erhalten wir

$$\partial_\tau u_i = r_i \partial_\tau x + s_i \partial_\tau y$$

Zusammen mit den 2 Differentialgleichungen ergibt sich das folgende LGS:

$$\begin{pmatrix} b_{11}^1 & b_{12}^1 & b_{21}^1 & b_{22}^1 \\ b_{11}^2 & b_{12}^2 & b_{21}^2 & b_{22}^2 \\ \partial_\tau x & 0 & \partial_\tau y & 0 \\ 0 & \partial_\tau x & 0 & \partial_\tau y \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \partial_\tau u_1 \\ \partial_\tau u_2 \end{pmatrix}$$

Wir erhalten für die Determinante der Matrix

$$\begin{vmatrix} b_{11}^1 & b_{12}^1 & b_{21}^1 & b_{22}^1 \\ b_{11}^2 & b_{12}^2 & b_{21}^2 & b_{22}^2 \\ \partial_\tau x & 0 & \partial_\tau y & 0 \\ 0 & \partial_\tau x & 0 & \partial_\tau y \end{vmatrix} = (\partial_\tau x)^2 (b_{21}^1 b_{22}^2 - b_{22}^1 b_{21}^2) \\ + (\partial_\tau y)^2 (b_{11}^1 b_{21}^2 - b_{11}^2 b_{21}^1) \\ + \partial_\tau x \partial_\tau y (b_{12}^1 b_{21}^2 - b_{21}^1 b_{12}^2 + b_{22}^1 b_{11}^2 - b_{11}^1 b_{22}^2).$$

Durch Setzen von

$$\hat{a}_{11} := b_{21}^1 b_{22}^2 - b_{22}^1 b_{21}^2 \\ \hat{a}_{22} := b_{11}^1 b_{21}^2 - b_{11}^2 b_{21}^1 \\ \hat{a}_{12} := \frac{1}{2} (b_{12}^1 b_{21}^2 - b_{21}^1 b_{12}^2 + b_{22}^1 b_{11}^2 - b_{11}^1 b_{22}^2)$$

sehen wir, dass die Lösbarkeits-Eigenschaften wieder von dem Vorzeichen des Terms

$$\hat{a}_{12}^2 - \hat{a}_{11} \hat{a}_{22}$$

abhängt. Im Vergleich dazu liefert die PDE 2. Ordnung

$$\hat{a}_{11} := a_{22} \\ \hat{a}_{22} := -a_{11} \\ \hat{a}_{12} := -\frac{a_{12} + a_{21}}{2}$$

und somit im Fall  $a_{12} = a_{21}$

$$\hat{a}_{12}^2 - \hat{a}_{11} \hat{a}_{22} = a_{12}^2 - a_{11} a_{22}.$$

Beide Vorgehensweisen liefern also die gleiche Typen-Einteilung.

**Lösung A.1.5:** (i) Wir verwenden den Beweisgang aus dem Text in leicht modifizierter Form. Für einen beliebigen Punkt  $x = (x_1, x_2) \in Q$  ist

$$v(x) = v(x_1, x_2) - v(x_1, 0) = \int_0^{x_2} \partial_2 v(x_1, \xi) d\xi.$$

Mit Hilfe der Hölderschen Ungleichung folgt

$$|v(x)|^2 \leq \left( \int_0^{x_2} \partial_2 v(x_1, \xi) d\xi \right)^2 \leq \int_0^1 |\partial_2 v(x_1, \xi)|^2 d\xi.$$

Wir integrieren diese Ungleichung unter Verwendung des Satzes von Fubini nacheinander bzgl. der Variablen  $x_1, x_2$ :

$$\begin{aligned} \int_Q |v(x)|^2 dx &\leq \int_0^1 \int_0^1 \left( \int_0^1 |\partial_2 v(x_1, \xi)|^2 d\xi \right) dx_1 dx_2 \\ &= \int_0^1 \left( \int_0^1 \int_0^1 |\partial_2 v(x_1, \xi)|^2 dx_1 d\xi \right) dx_2 = \int_Q |\partial_2 v(x)|^2 dx \end{aligned}$$

(ii) Für  $\Gamma := \{(0, 0)\}$  kann die Poincarésche Ungleichung *nicht* gelten. Zum Beweis konstruieren wir eine Folge von Funktionen  $u_k \in V_0(\Gamma; Q)$  mit den Eigenschaften

$$\liminf_{k \rightarrow \infty} \int_Q |u_k|^2 dx > 0, \quad \int_Q \|\nabla u_k\|^2 dx \rightarrow 0 \quad (k \rightarrow \infty).$$

Dazu setzen wir unter Verwendung von Polarkoordinaten  $(r, \theta)$ :

$$u_k(r, \theta) := r^{1/k}.$$

Für diese Funktionen ist  $\|\nabla u_k\| = |\partial_r u_k| = k^{-1} r^{-1+1/k}$  und folglich:

$$\begin{aligned} \int_Q |u_k|^2 dx &\geq \int_0^{\pi/2} \int_0^1 r^{2/k} r dr d\omega = \frac{\pi}{2} \int_0^1 r^{1+2/k} dr \\ &= \frac{\pi}{2} \frac{1}{2/k + 2} r^{2+2/k} \Big|_0^1 = \frac{\pi}{2} \frac{1}{2 + 2/k} \rightarrow \frac{\pi}{4} \quad (k \rightarrow \infty), \end{aligned}$$

sowie analog

$$\begin{aligned} \int_Q \|\nabla u_k\|^2 dx &= \frac{1}{k^2} \int_Q r^{-2+2/k} dx \leq \frac{\pi}{2k^2} \int_0^2 r^{-1+2/k} dr \\ &= \frac{\pi}{2k^2} \frac{k}{2} r^{2/k} \Big|_0^2 = \frac{\pi}{4} \frac{2^{2/k}}{k} \rightarrow 0 \quad (k \rightarrow \infty). \end{aligned}$$

Als Konsequenz dieses Resultats ist in diesem Fall die in dem Text verwendete „direkte Methode der Variationsrechnung“ (d. h. das „Minimalfolgenargument“) zum Nachweis der Existenz „schwacher“ Lösungen der zugehörigen 1. RWA des Laplace-Operators nicht anwendbar, da das Energiefunktional  $J(u)$  auf  $V_0(\Gamma; Q)$  nicht nach unten beschränkt ist. Tatsächlich bedeutet dies, dass in zwei (und höheren) Dimensionen die 1. RWA mit solchen einpunktigen Dirichlet-Randbedingungen *nicht* „wohl gestellt“ ist.

**Lösung A.1.6:** a) Seien  $u, v \in C^2(\Omega) \cap C^1(\bar{\Omega})$  Lösungen derselben 2. RWA. Dann erfüllt  $w := u - v$  die Gleichungen  $-\Delta w + aw = 0$  in  $\Omega$  und  $\partial_n w|_{\partial\Omega} = 0$ . Mit Hilfe partieller

Inegration folgt unter Aussnutzung der Randbedingung  $\partial_n w|_{\partial\Omega} = 0$ :

$$\int_{\Omega} \{ \|\nabla w\|^2 + a|w|^2 \} dx = \int_{\Omega} (-\Delta w + aw)w dx + \int_{\partial\Omega} \partial_n w w do = 0.$$

Dies impliziert  $w \equiv 0$ .

b) Mit denselben Bezeichnungen wie in (a) gilt nun  $(\partial_n w + \alpha w)|_{\partial\Omega} = 0$ . Damit ergibt sich dann wegen  $\alpha \geq 0$ :

$$\int_{\Omega} \{ \|\nabla w\|^2 + a|w|^2 \} dx = \int_{\Omega} (-\Delta w + aw)w dx + \int_{\partial\Omega} \partial_n w w do = - \int_{\partial\Omega} \alpha w^2 do \leq 0.$$

Dies impliziert wieder  $w \equiv 0$ .

Für  $a = 0$  kann in beiden Fällen nur auf  $\nabla w \equiv 0$  bzw.  $w \equiv \text{konst}$  geschlossen werden. Es fehlt aber eine zusätzliche Bedingung, um hieraus  $w \equiv 0$  folgern zu können. Eine solche Zusatzbedingung könnte z. B. die Forderung sein, daß nach Lösungen der RWAn mit verschwindendem Mittelwert gefragt ist:  $\int_{\Omega} u dx = 0$ .

**Lösung A.1.7:** Wir zeigen die beiden Ungleichungen einzeln:

i) Wir wollen zeigen  $u \geq 0$ . Es ist

$$-\Delta(-u) = -1 \leq 0 \quad \text{in } \Omega, \quad u = 0 \leq 0 \quad \text{auf } \partial\Omega$$

Nach dem Maximum-Prinzip ist also

$$-u \leq 0 \quad \text{oder} \quad \max_{\Omega} -u \leq \max_{\partial\Omega} -u = 0.$$

Also  $u \geq 0$  auf  $\overline{\Omega}$ .

ii) Nun zeigen wir  $u \leq \frac{1}{8}$ . Dazu betrachten wir die Funktion  $v = \frac{1}{4}(x(1-x) + y(1-y))$  mit

$$-\Delta v = \frac{1}{4}(2+2) = 1.$$

Wegen

$$\nabla v = \begin{pmatrix} 1-2x \\ 1-2y \end{pmatrix} = 0 \quad \Leftrightarrow \quad \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$$

sowie

$$v\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{1}{8}, \quad v = 0 \quad \text{auf } \partial Q_1$$

liegt in  $(x, y) = (1/2, 1/2)$  ein Maximum von  $v$ , und es ist  $v \geq 0$  in  $Q_1$ . Wegen  $\Omega \subset Q_1$  folgt nun

$$-\Delta(u-v) = 0 \leq 0 \quad \text{in } \Omega$$

sowie

$$u - v = -v \leq 0 \quad \text{auf } \partial\Omega.$$

Nach dem Maximumprinzip ist also

$$u - v \leq 0 \quad \text{oder} \quad \max_{\Omega} u - v \leq \max_{\partial\Omega} u - v \leq 0$$

und somit

$$u \leq v \leq \frac{1}{8}.$$

**Lösung A.1.8:** Zur Wiederholung: Wir setzen  $x_1 = r \cos \theta$ ,  $x_2 = r \sin \theta$  und  $u(x) = u(x_1, x_2) = u(r \cos \theta, r \sin \theta)$ . Mit Hilfe der Kettenregel gilt dann:

$$\begin{aligned}\partial_r^2 u(r, \theta) &= \partial_1^2 u(x) \cos^2 \theta + \partial_2 \partial_1 u(x) \sin \theta \cos \theta + \partial_1 \partial_2 u(x) \cos \theta \sin \theta + \partial_2^2 u(x) \sin^2 \theta, \\ \partial_\theta^2 u(r, \theta) &= \partial_1^2 u(x) r^2 \cos^2 \theta - \partial_2 \partial_1 u(x) r^2 \sin \theta \cos \theta - \partial_1 \partial_2 u(x) r^2 \cos \theta \sin \theta \\ &\quad - \partial_1 u(x) r \cos \theta - \partial_2 u(x) r \sin \theta + \partial_2^2 u(x) r^2 \cos^2 \theta.\end{aligned}$$

Also ist  $(\partial_r^2 + r^{-1} \partial_r + r^{-2} \partial_\theta^2)u(r, \theta) = (\partial_1^2 + \partial_2^2)u(x) = \Delta u(x)$ .

i) Die Randbedingungen ließt man direkt ab. Wir setzen  $\alpha := \pi/\omega$  und finden

$$\begin{aligned}\Delta s_\omega(r, \theta) &= (\partial_r^2 + r^{-1} \partial_r + r^{-2} \partial_\theta^2)(r^\alpha \sin \theta \alpha) \\ &= (\alpha - 1)\alpha r^{\alpha-2} \sin \theta \alpha + \alpha r^{\alpha-2} \sin \theta \alpha - r^{\alpha-2} \alpha^2 \sin \theta \alpha = 0.\end{aligned}$$

ii) Die Funktion  $s_\omega$  ist im Innern des Sektorabschnitts  $G$  beliebig oft differenzierbar. Ihre ersten und zweiten Ableitungen verhalten sich dort wie

$$|\partial_i s_\omega(r, \theta)| \leq c r^{\pi/\omega-1}, \quad |\partial_i \partial_j s_\omega(r, \theta)| \leq c r^{\pi/\omega-2}.$$

Zu überprüfen ist also die Existenz der (uneigentlichen) Integrale

$$\begin{aligned}J_1(\omega) &:= \int_G r^{2\pi/\omega-2} dx = \int_0^\omega \int_0^1 r^{2\pi/\omega-1} dr d\theta = \omega \int_0^1 r^{2\pi/\omega-1} dr, \\ J_2(\omega) &:= \int_G r^{2\pi/\omega-4} dx = \int_0^\omega \int_0^1 r^{2\pi/\omega-3} dr d\theta = \omega \int_0^1 r^{2\pi/\omega-3} dr.\end{aligned}$$

Für  $\pi < \omega \leq 2\pi$  ist  $2\pi/\omega - 1 \geq 0$  und folglich  $J_1(\omega)$  existent, aber  $2\pi/\omega - 3 < -1$  und folglich  $J_2(\omega)$  nicht existent. Im Fall  $\omega < \pi$  ist  $\nabla s_\omega$  beschränkt und somit (eigentlich) quadrat-integrierbar. Ferner ist  $2\pi/\omega - 3 > -1$  und somit auch  $\nabla^2 s_\omega$  wenigstens quadrat-integrierbar.

**Lösung A.1.9:** a) Es ist

$$\partial_x u(x, y) = \begin{cases} \frac{1}{2} (x - y)^{\frac{-1}{2}}, & x \geq y, \\ \frac{-1}{2} (y - x)^{\frac{-1}{2}}, & y < x, \end{cases}$$

und  $\partial_y u = -\partial_x u$ , so dass es genügt,  $\partial_x u$  zu betrachten. Wir betrachten dazu die

Ausschöpfung  $\Omega_\varepsilon := \{(x, y) \in \mathbb{R}^2 : x < y - \varepsilon \text{ oder } x > y + \varepsilon\}$  und wir erhalten

$$\int_{\Omega_\varepsilon} (\partial_x u)^2 \, d(x, y) = \int_0^1 \int_0^{y-\varepsilon} \frac{1}{4} (y-x)^{-1} \, dx \, dy + \int_0^1 \int_{y+\varepsilon}^1 \frac{1}{4} (x-y)^{-1} \, dx \, dy$$

Wir begnügen uns damit, das erste Integral auf der rechten Seite zu betrachten:

$$\begin{aligned} \int_0^1 \int_0^{y-\varepsilon} \frac{1}{4} (y-x)^{-1} \, dx \, dy &= \int_0^1 \left[ \frac{-\ln(y-x)}{4} \right]_0^{y-\varepsilon} \, dy \\ &= \frac{-\ln(\varepsilon)}{4} + \int_0^1 \frac{\ln(y)}{4} \, dy \rightarrow \infty \quad (\varepsilon \rightarrow 0) \end{aligned}$$

Also ist  $\partial_x u$  nicht in  $L^2(\Omega)$  und somit  $u \notin H^1(\Omega)$ .

b) Durch Anwendung der Kettenregel erhalten wir für die partiellen Ableitungen von  $u$ :

$$\partial_x u(x, y) = \cos\left(\ln\left(\frac{1}{r}\right)\right) \frac{-x}{r^2}, \quad \partial_y u(x, y) = \cos\left(\ln\left(\frac{1}{r}\right)\right) \frac{-y}{r^2}.$$

Somit erhalten wir

$$\nabla u \cdot \nabla u = (\partial_x u)^2 + (\partial_y u)^2 = \cos\left(\ln\left(\frac{1}{r}\right)\right) \frac{x^2 + y^2}{r^4} = \cos\left(\ln\left(\frac{1}{r}\right)\right)^2 \frac{1}{r^2}$$

Wir betrachten jetzt nur die Kreisbögen  $S_\varepsilon = \{(x, y) \in \mathbb{R}^2 : \varepsilon < x^2 + y^2 < 1\}$ . Und bestimmen das Integral

$$\begin{aligned} \int_{S_\varepsilon} \frac{\cos(\ln(\frac{1}{r}))^2}{r^2} \, d(x, y) &= \int_\varepsilon^1 \int_0^{\frac{\pi}{2}} \frac{\cos(\ln(\frac{1}{r}))^2}{r^2} r \, d\theta \, dr \\ &= \frac{\pi}{2} \int_\varepsilon^1 \frac{\cos(\ln(\frac{1}{r}))^2}{r} \, dr \\ &= \frac{\pi}{2} \int_{\ln(\varepsilon)}^0 \cos(x)^2 \, dx \\ &= \frac{\pi}{4} (\cos(x) \sin(x) + x) \Big|_{\ln(\varepsilon)}^0 \\ &= \frac{-\pi}{4} (\cos(\ln(\varepsilon)) \sin(\ln(\varepsilon)) + \ln(\varepsilon)) \rightarrow \infty \quad (\varepsilon \rightarrow 0) \end{aligned}$$

Also ist  $u \notin H^1(\Omega)$ .

**Lösung A.1.10:** Zunächst eine Feststellung, die Menge

$$M := \left\{ u \in H^1(\Omega) : \int_\Omega u(x) \, dx = 0 \right\}$$

ist konvex und abgeschlossen bzgl. der  $L^2$ -Topologie. Die Konvexität folgt sofort aus der

Linearität des Integrals, die Abgeschlossenheit, da

$$\left| \int_{\Omega} u(x) dx \right| \leq \int_{\Omega} |u(x)| dx \leq c \|u\|_{\Omega}.$$

Wir zeigen die Aussage durch ein Widerspruchssargument. Angenommen es gäbe keine Konstante  $c_{\Omega}$  mit der Eigenschaft. Dann gibt es eine Folge  $u_n$  aus  $M$ , so dass

$$\|u_n\|_{\Omega} > n \|\nabla u_n\|_{\Omega}$$

bzw.  $x_n := \|u_n\|_{\Omega}^{-1} \rightarrow 0$  ( $n \rightarrow \infty$ ). Wir können also o.B.d.A. annehmen, daß  $0 < x_n \leq 1$ ,  $n \in \mathbb{N}$ . Da  $0 \in M$  und  $M$  konvex, ist auch  $v_n := x_n u_n \in M$ , und es gilt  $\|v_n\|_{\Omega} = 1$ . Aus der Ungleichung für  $u_n$  erhalten wir

$$\|\nabla v_n\|_{\Omega} = x_n \|\nabla u_n\|_{\Omega} < \frac{x_n}{n} \|u_n\|_{\Omega} = \frac{\|v_n\|_{\Omega}}{n} \rightarrow 0 \quad (n \rightarrow \infty).$$

Also ist  $v_n$  in  $H^1(\Omega)$  beschränkt. Es gibt also eine Teilfolge, wir wollen o.B.d.A. annehmen dies wäre  $v_n$ , die schwach gegen  $v \in H^1(\Omega)$  konvergiert. Insbesonder konvergiert  $\nabla v_n$  schwach in  $L^2$  gegen  $\nabla v$ . Wegen obiger Ungleichung konvergiert aber  $\nabla v_n$  stark in  $L^2$  gegen 0, folglich ist  $\nabla v = 0$ . Da  $\Omega$  ein Gebiet ist (also insbesondere zusammenhängend) folgt daraus  $v = \text{konst}$  fast überall in  $\Omega$ .

Nach dem Rellichschen Auswahlssatz gibt es eine stark in  $L^2$  konvergente Teilfolge von  $v_n$ , die wir wieder mit  $v_n$  bezeichnen, wegen  $\|v_n\|_{\Omega} = 1$  folgt also aus der starken Konvergenz  $\|v\|_{\Omega} = 1$ , und somit  $v = 1$  fast überall auf  $\Omega$ . Andererseits ist  $M$  bezüglich der Norm  $\|\cdot\|_{\Omega}$  abgeschlossen, also folgt  $v \in M$  im Widerspruch zu  $v = 1$ . Damit ist die Behauptung bewiesen.

**Lösung A.1.11:** Wir können im allgemeinen nur Existenz und Eindeutigkeit einer Lösung  $u \in V$  der zugehörigen schwachen Formulierung, d. h. bezüglich

$$(\nabla u, \nabla \varphi) = (f, \varphi) \quad \forall \varphi \in V \tag{1.1.1}$$

erwarten. Wir zeigen zunächst die Existenz und Eindeutigkeit einer Lösung für den Raum

$$V := \left\{ u \in H^1(\Omega) : \int_{\Omega} u(x) dx = 0 \right\}.$$

Die Existenz einer solchen Lösung ist gegeben, wenn wir zeigen können, dass das Funktional

$$E(v) = \frac{1}{2} \|\nabla v\|_{\Omega}^2 - (f, v)_{\Omega}$$

in  $V$  ein Minimum annimmt. Zunächst zeigen wir, dass  $E$  in  $V$  nach unten beschränkt ist. Mithilfe der Poincaréschen Ungleichung aus der vorausgehenden Aufgabe

$$\int_{\Omega} f v dx \leq \|f\|_{\Omega} \|v\|_{\Omega} \leq c_{\Omega} \|f\|_{\Omega} \|\nabla v\|_{\Omega}.$$

Mithilfe der Youngschen Ungleichung folgt weiter

$$\int_{\Omega} fv \, dx \leq \frac{c_{\Omega}^2}{2} \|f\|_{\Omega}^2 + \frac{1}{2} \|\nabla v\|_{\Omega}^2$$

und damit die Beschränktheit von  $E(\cdot)$  nach unten:

$$E(v) \geq -\frac{c_{\Omega}^2}{2} \|f\|_{\Omega}^2.$$

Sei nun  $(v_k)_{k \in \mathbb{N}}, v_k \in V$  eine Minimalfolge bezüglich  $E(\cdot)$ :

$$E(v_k) \rightarrow d := \inf_{v \in V} E(v).$$

Wie im Skript zeigen wir mithilfe der Parallelogramm-Identität, dass  $(v_k)$  Cauchy-Folge bezüglich der Energienorm

$$\|v\|_E := \|\nabla v\|_{\Omega}$$

ist:

$$\begin{aligned} \|v_n - v_m\|_E^2 &= 2\|v_n\|_E^2 + 2\|v_m\|_E^2 - 4\|\frac{1}{2}(v_n + v_m)\|_E^2 \\ &\leq 4E(v_n) + 4(f, v_n) + 4E(v_m) + 4(f, v_m) - 8E(\frac{1}{2}(v_n + v_m)) - 8(f, \frac{1}{2}(v_n + v_m)) \\ &= 4E(v_n) + 4E(v_m) - 8E(\frac{1}{2}(v_n + v_m)) \end{aligned}$$

Unter Berücksichtigung von  $\lim_{k \rightarrow \infty} E(v_k) = d$  und  $E(\frac{1}{2}(v_n + v_m)) \geq d$  folgt

$$\lim \sup_{n, m \rightarrow \infty} \|v_n - v_m\|_E^2 \leq 0.$$

Aufgrund der Poincaréschen Ungleichung auf  $V$  gilt weiter

$$\|v_n - v_m\|_{H^1(\Omega)} \leq C \|v_n - v_m\|_E,$$

d. h.  $(v_k)_{k \in \mathbb{N}}$  ist Cauchy-Folge bezüglich der  $H^1$ -Norm und konvergiert folglich gegen einen Limes  $v \in H^1(\Omega)$ . Um  $v \in V$  zu garantieren, müssen wir noch zeigen, dass auch die Mittelwertbedingung im Limes erhalten bleibt

$$\left| \int_{\Omega} v dx \right| = \left| \int_{\Omega} v - v_k dx \right| \leq \|v - v_k\|_{\Omega} \rightarrow 0 (k \rightarrow \infty).$$

Wir haben also eine schwache Lösung  $v \in V$  gefunden. Aufgrund der Mittelwertbedingung an  $f$  bleibt (1.1.1) auch gültig, wenn man den Testraum auf  $H^1(\Omega)$  erweitert (Für  $\varphi \in H^1(\Omega)$  liegt  $\hat{\varphi} := \varphi - |\Omega|^{-1} \int_{\Omega} \varphi$  in  $M$  und

$$(\nabla u, \nabla \varphi) = (\nabla u, \nabla \hat{\varphi}), \quad (f, \varphi) = (f, \hat{\varphi}).$$

Wir haben also eine schwache Lösung  $v \in H^1(\Omega)$  für das Problem

$$(\nabla u, \nabla \varphi) = (f, \varphi) \quad \forall \varphi \in H^1(\Omega)$$

gefunden. Seien nun  $v_1, v_2$  zwei Lösungen, so gilt für  $w = v_1 - v_2$

$$(\nabla w, \nabla w) = 0$$

und daher  $v_1 = v_2 + const.$

**Lösung A.1.12:** a) Die RWA auf der gepunkteten Kreisscheibe ist *nicht* wohl gestellt, da das Energiefunktional

$$E(v) = \frac{1}{2} \|\nabla v\|_{\Omega}^2 - (f, v)_{\Omega}$$

auf  $H_0^1(\Omega)$  kein Minimum besitzt. Es existiert eine Folge  $(v_k)_{k \in \mathbb{N}} \subset H_0^1(\Omega)$ , welche bzgl. der  $H^1$ -Norm eine Cauchy-Folge ist, für deren Limes aber  $u(0) \neq 0$  ist. Eine reguläre schwache Lösung genügt also nicht der geforderten Randbedingung in  $x = 0$ .

b) Die RWA auf der geschlitzten Kreisscheibe ist wohl gestellt, denn zu jeder Cauchy-Folge  $(v_k)_{k \in \mathbb{N}} \subset H_0^1(\Omega)$  gehört aufgrund der Spurabschätzung

$$\|v\|_{L^2(\Gamma)} \leq c \|v\|_{H^1(\Omega)},$$

eine Cauchy-Folge  $(v_{k|\Gamma})_{k \in \mathbb{N}} \subset L^2(\Gamma)$  von Spuren. Der  $H^1$ -Limes  $v = \lim_{k \rightarrow \infty} v_k$  hat also automatisch die Spur  $v|_{\Gamma} \equiv 0$ . Jede reguläre schwache Lösung genügt also der geforderten homogenen Dirichlet-Randbedingung auf  $\Gamma$ .

**Lösung A.1.13:** Gesucht ist  $u \in H^1(\Omega)$  mit der Eigenschaft

$$(\nabla u, \nabla \varphi)_{\Omega} + (u, \varphi)_{\Omega} + (u, \varphi)_{\partial\Omega} = (f, \varphi)_{\Omega} + (g, \varphi)_{\partial\Omega} \quad \forall \varphi \in H^1(\Omega).$$

Für eine hinreichend reguläre Lösung folgt durch partielle Integration

$$(-\Delta u + u - f, \varphi)_{\Omega} + (\partial_n u + u - g, \varphi)_{\partial\Omega} = 0 \quad \forall \varphi \in H^1(\Omega),$$

und damit die Gültigkeit der gegebenen Differentialgleichung sowie der Randbedingung.

**Lösung A.1.14:** Für die Ungleichungen gilt:

- a)  $\|u\|_{L^\infty(\Omega)} \leq c \|u\|_{H^2(\Omega)}$ ,  $u \in H^2(\Omega)$ ,  $\Omega \subset \mathbb{R}^3$ ; (richtig)
- b)  $\|u\|_{L^\infty(\Omega)} \leq c \|u\|_{H^{1,1}(\Omega)}$ ,  $u \in H^{1,1}(\Omega)$ ,  $\Omega \subset \mathbb{R}^1$ ; (richtig)
- c)  $\|u\|_{L^\infty(\Omega)} \leq c \|u\|_{H^1(\Omega)}$ ,  $u \in H^1(\Omega)$ ,  $\Omega \subset \mathbb{R}^2$ ; (falsch)
- d)  $\|u\|_{L^1(\partial\Omega)} \leq c \|u\|_{H^{1,1}(\Omega)}$ ,  $u \in H^{1,1}(\Omega)$ ,  $\Omega \subset \mathbb{R}^2$  (richtig).

**Lösung A.1.15:** Wir betrachten zunächst die rechte Seite der Ungleichung. Sei  $u^0 \in H^1(\Omega)$ . Wir verwenden die  $L^2$ -Orthonormalbasis aus Eigenfunktionen des Laplace-Operators  $(v_k)_{k \in \mathbb{N}}$ . Für  $u^0$  gelte die Darstellung

$$u^0(x) = \sum_{j=0}^{\infty} u_j^0 v_j.$$

Die n-te Partialsumme

$$s_n := \sum_{j=0}^n u_j^0 v_j(x).$$

konvergiert in  $H^1(\Omega)$  gegen  $u^0$ . Insbesondere gilt

$$\|\nabla u^0\|_\Omega^2 = \lim_{n \rightarrow \infty} \|\nabla s_n\|_\Omega^2.$$

Für die Partialsumme können wir Integration und Differentiation mit der Summation vertauschen

$$(\nabla s_n, \nabla s_n)_\Omega = \sum_{j,k=0}^n u_j^0 u_k^0 (\nabla v_j(x), \nabla v_k(x))_\Omega$$

Wir nutzen aus, dass  $v_j$  auch schwache Lösungen des Eigenwertproblems des Laplace-Operators sind, d. h.:

$$(\nabla v_j, \nabla \varphi)_\Omega = \lambda_j (v_j, \varphi) \quad \forall \varphi \in H_0^1(\Omega).$$

Wir bekommen

$$(\nabla s_n, \nabla s_n)_\Omega = \sum_{j,k=0}^n u_j^0 u_k^0 \lambda_j (v_j(x), v_k(x))_\Omega.$$

Schließlich nutzen wir die Orthonormalitätseigenschaft der  $v_j$ . Insgesamt haben wir gezeigt

$$\|\nabla u^0\|_\Omega^2 = \sum_{j=0}^\infty (u_j^0)^2 \lambda_j. \quad (1.1.2)$$

Für die Lösung der Wärmeleitungsgleichung gilt die Darstellung

$$u(x, t) = \sum_{j=1}^\infty u_j^0 v_j(x) e^{-\lambda_j t}.$$

Differenzieren ergibt

$$\partial_t u(x, t) = \Delta u(x, t) = - \sum_{j=1}^\infty u_j^0 \lambda_j v_j(x) e^{-\lambda_j t}.$$

Mit der Parsevalschen Identität gilt

$$\|\partial_t u(x, t)\|_\Omega^2 = \|\Delta u(x, t)\|_\Omega^2 = \sum_{j=1}^\infty (u_j^0)^2 (\lambda_j)^2 e^{-2\lambda_j t}.$$

Die Funktion  $xe^{-2x}$  nimmt ihr Maximum auf  $[0, \infty)$  in  $x = 0.5$  an. Es gilt

$$xe^{-2x} \leq \frac{1}{2}e^{-1}.$$

Mit diesem Resultat können wir weiter abschätzen

$$\|\partial_t u(x, t)\|_{\Omega}^2 = \|\Delta u(x, t)\|_{\Omega}^2 = \frac{1}{2t} e^{-1} \sum_{j=1}^{\infty} (u_j^0)^2 \lambda_j.$$

Zusammen mit (1.1.2) haben wir gezeigt

$$\|\partial_t u(x, t)\|_{\Omega}^2 = \|\Delta u(x, t)\|_{\Omega}^2 \leq \frac{1}{2t} e^{-1} \|\nabla u^0\|_{\Omega}^2.$$

Wegen

$$\sqrt{\frac{1}{2}e^{-1}} = 0.42888\dots < 0.5$$

folgt daraus die Behauptung

**Lösung A.1.16:** Wir machen den Ansatz  $u(x, t) = \psi(t)v(x)$ . Damit ergibt sich

$$\partial_t^2 u - \Delta u = \psi''(t)v(x) - \psi(t)\Delta v(x) = 0$$

bzw.

$$\frac{\psi''(t)}{\psi(t)} = \frac{\Delta v(x)}{v(x)} = \text{const} = -\lambda.$$

Dies führt auf die Eigenwertprobleme

$$-\Delta v = \lambda v, \quad -\psi''(t) = \lambda \psi(t)$$

mit den in der Aufgabenstellung gegebenen Randwerten. Die Eigenfunktionen  $v_j$  für das Laplace-Problem mit Nullrandbedingungen definieren ein Orthonormalsystem in  $L^2(\Omega)$ . Wir wissen, dass die zugehörigen Eigenwerte  $\lambda_j$  positiv sind. Die Differentialgleichung in der Zeit hat dann die Fundamentallösung

$$\psi(t) = \sum_{j=0}^{\infty} a_j \sin(\sqrt{\lambda_j}t) + b_j \cos(\sqrt{\lambda_j}t).$$

Zusammen ergibt sich

$$u(x, t) = v(x)\psi(t) = \sum_{j=0}^{\infty} v_j(x) \left( a_j \sin(\sqrt{\lambda_j}t) + b_j \cos(\sqrt{\lambda_j}t) \right).$$

Machen wir für die Funktionen  $u^0(x), u^1(x)$  den Ansatz

$$u^0(x) = \sum_{j=0}^{\infty} u_j^0 v_j(x), \quad u^1(x) = \sum_{j=0}^{\infty} u_j^1 v_j(x)$$

erhalten wir durch Koeffizientenvergleich  $b_j = u_j^0$  sowie nach Ableiten  $a_j = \frac{u_j^1}{\sqrt{\lambda_j}}$ . Wir erhalten die Darstellung

$$u(x, t) = \sum_{j=0}^{\infty} v_j(x) \left( \frac{u_j^1}{\sqrt{\lambda_j}} \sin(\sqrt{\lambda_j}t) + u_j^0 \cos(\sqrt{\lambda_j}t) \right).$$

Es bleibt zu untersuchen, welche Regularitätsanforderungen wir an  $u^0$  und  $u^1$  stellen müssen, um

$$\partial_t^2 u, \Delta u \in L^2(\Omega \times (0, T))$$

garantieren zu können. Differenzieren führt zu

$$\partial_t^2 u = \Delta u = - \sum_{j=0}^{\infty} v_j(x) \left( u_j^1 \lambda_j^{\frac{1}{2}} \sin(\sqrt{\lambda_j}t) + u_j^0 \lambda_j \cos(\sqrt{\lambda_j}t) \right).$$

Mithilfe der Parsevalschen Identität erhalten wir

$$\begin{aligned} \|\partial_t^2 u\|_{L^2(\Omega \times (0, T))}^2 &= \int_0^T \int_{\Omega} \left( \sum_{j=0}^{\infty} v_j(x) \left( u_j^1 \lambda_j^{\frac{1}{2}} \sin(\sqrt{\lambda_j}t) + u_j^0 \lambda_j \cos(\sqrt{\lambda_j}t) \right) \right)^2 dx dt \\ &= \int_0^T \left( \sum_{j=0}^{\infty} \left( u_j^1 \lambda_j^{\frac{1}{2}} \sin(\sqrt{\lambda_j}t) + u_j^0 \lambda_j \cos(\sqrt{\lambda_j}t) \right)^2 \right)^2 dt. \end{aligned}$$

Wir ziehen Betragsstriche unter die Summe und schätzen die trigonometrischen Terme ab

$$\|\partial_t^2 u\|_{L^2(\Omega \times (0, T))}^2 \leq T \sum_{j=0}^{\infty} (\lambda_j u_j^0)^2 + \lambda_j (u_j^1)^2 + 2\lambda_j^{\frac{3}{2}} u_j^0 u_j^1.$$

Ausnutzen der Youngschen Ungleichung führt auf

$$\|\partial_t^2 u\|_{L^2(\Omega \times (0, T))}^2 \leq T \sum_{j=0}^{\infty} (\lambda_j u_j^0)^2 + \lambda_j (u_j^1)^2. \quad (1.1.3)$$

Wir wollen nun  $u^0 \in H^2(\Omega)$  annehmen. Wir definieren die n-te Partialsumme

$$s_n := \sum_{j=0}^n u_j^0 v_j(x).$$

Diese konvergiert dann in  $H^2(\Omega)$  gegen  $u^0$ . Damit folgt insbesondere

$$\infty > \|\Delta u^0\|_{\Omega}^2 = \lim_{n \rightarrow \infty} \|\Delta s_n\|_{\Omega}^2.$$

Für die Partialsumme können wir Integration und Differentiation mit der Summation

vertauschen

$$\begin{aligned} (\Delta s_n, \Delta s_n)_\Omega &= \sum_{j,k=0}^n u_j^0 u_k^0 (\Delta v_j(x), \Delta v_k(x))_\Omega \\ &= \sum_{j,k=0}^n u_j^0 u_k^0 (\lambda_j v_j(x), \lambda_k v_k(x))_\Omega. \end{aligned}$$

Schließlich folgt aufgrund der Orthonormalitätseigenschaft der  $v_n$

$$\infty > \lim_{n \rightarrow \infty} \sum_{j=0}^n (u_j^0 \lambda_j)^2 = \sum_{j=0}^{\infty} (u_j^0 \lambda_j)^2.$$

Dies ist genau der erste Teil der Summe in (1.1.3). Sei nun  $u^1 \in H^1(\Omega)$ . Die n-te Partiälsumme

$$r_n := \sum_{j=0}^n u_j^1 v_j(x).$$

konvergiert in  $H^1(\Omega)$  gegen  $u^1$  und wir haben

$$\infty > \|\nabla u^1\|_\Omega^2 = \lim_{n \rightarrow \infty} \|\nabla r_n\|_\Omega^2.$$

Nutzen wir aus, dass  $v_j$  auch schwache Lösung des Eigenwertproblems sind, erhalten wir mit derselben Argumentation wie oben

$$\begin{aligned} (\nabla r_n, \nabla r_n)_\Omega &= \sum_{j,k=0}^n u_j^1 u_k^1 (\nabla v_j(x), \nabla v_k(x))_\Omega \\ &= \sum_{j,k=0}^n u_j^1 u_k^1 \lambda_j (v_j(x), v_k(x))_\Omega. \end{aligned}$$

Schließlich folgt wieder mit der Orthonormalitätseigenschaft der  $v_n$

$$\infty > \sum_{j=0}^{\infty} (u_j^1)^2 \lambda_j.$$

Wir haben gezeigt, dass (1.1.3) und damit sowohl  $\partial_t^2 u$  als auch  $\Delta u$  in  $L^2(\Omega \times (0, T))$  wohldefiniert sind. Unter den getroffenen Regularitätsvoraussetzungen löst das oben konstruierte  $u(x, t)$  die Wellengleichung also in einem (starken)  $L^2$ -Sinne.

## A.2 Kapitel 2

**Lösung A.2.1:** Wir bestimmen den Abschneidefehler

$$\tau_h(x, y) := -\Delta_h^{(9)} u(x, y) - f_h(x, y) = -\Delta_h^{(9)} u_h - f - \frac{1}{12} h^2 \Delta f.$$

Taylor-Entwicklung von  $u(x \pm h, y)$ ,  $u(x, y \pm h)$  und  $u(x \pm h, y \pm h)$  ergibt:

$$\begin{aligned} u(x \pm h, y) &= (1 \pm h \partial_x + \frac{1}{2} h^2 \partial_x^2 \pm \frac{1}{6} h^3 \partial_x^3 + \frac{1}{24} h^4 \partial_x^4 \pm \frac{1}{120} h^5 \partial_x^5) u(x, y) \\ &\quad + \frac{1}{720} h^6 \partial_x^6 u(\xi, \eta), \\ u(x, y \pm h) &= (1 \pm h \partial_y + \frac{1}{2} h^2 \partial_y^2 \pm \frac{1}{6} h^3 \partial_y^3 + \frac{1}{24} h^4 \partial_y^4 \pm \frac{1}{120} h^5 \partial_y^5) u(x, y) \\ &\quad + \frac{1}{720} h^6 \partial_y^6 u(\xi, \eta), \\ u(x \pm h, y \pm h) &= (1 \pm h \partial_x \pm h \partial_y + \frac{1}{2} h^2 \partial_x^2 \pm h^2 \partial_x \partial_y + \frac{1}{2} h^2 \partial_y^2 \pm \frac{1}{6} h^3 \partial_x^3 \pm \frac{1}{2} h^3 \partial_x^2 \partial_y \\ &\quad \pm \frac{1}{2} h^3 \partial_x \partial_y^2 \pm \frac{1}{6} h^3 \partial_y^3 + \frac{1}{24} h^4 \partial_x^4 \pm \frac{1}{6} h^4 \partial_x^3 \partial_y + \frac{1}{4} h^4 \partial_x^2 \partial_y^2 \pm \frac{1}{6} h^4 \partial_x \partial_y^3 \\ &\quad + \frac{1}{24} h^4 \partial_y^4 \pm \frac{1}{120} h^5 \partial_x^5 \pm \frac{1}{24} h^5 \partial_x \partial_y^4 \pm \frac{1}{12} h^5 \partial_x^2 \partial_y^3 \pm \frac{1}{12} h^5 \partial_x^3 \partial_y^2 \pm \frac{1}{24} h^5 \partial_x^4 \partial_y \\ &\quad \pm \frac{1}{120} h^5 \partial_y^5) u(x, y) + (\frac{1}{720} h^6 \partial_x^6 \pm \frac{1}{120} h^6 \partial_x^5 \partial_y + \frac{1}{48} h^6 \partial_x^4 \partial_y^2 \pm \frac{1}{36} h^6 \partial_x^3 \partial_y^3 \\ &\quad + \frac{1}{48} h^6 \partial_x^2 \partial_y^4 \pm \frac{1}{120} h^6 \partial_x \partial_y^5 + \frac{1}{720} h^6 \partial_y^6) u(\xi, \eta). \end{aligned}$$

Damit folgt für den kompakten 9-Punkte-Operator:

$$\begin{aligned} \Delta_h^{(9)} u_h(x, y) &= \frac{1}{6h^2} \left\{ 4u(x \pm h, y) + 4u(x, y \pm h) + u(x \pm h, y \pm h) - 20u(x, y) \right\} \\ &= \frac{1}{6h^2} \left( (8 + 8 + 4 - 20) + h(0 + 0) \partial_x + h(0 + 0) \partial_y \right. \\ &\quad + \frac{1}{2} h^2 (8 + 4) \partial_x^2 + h^2 (0) \partial_x \partial_y + \frac{1}{2} h^2 (8 + 4) \partial_y^2 \\ &\quad + \frac{1}{6} h^3 (0 + 0) \partial_x^3 + \frac{1}{2} h^3 (0) \partial_x^2 \partial_y + \frac{1}{2} h^3 (0) \partial_x \partial_y^2 + \frac{1}{6} h^3 (0 + 0) \partial_y^3 \\ &\quad + \frac{1}{24} h^4 (8 + 4) \partial_x^4 + \frac{1}{6} h^4 (0) \partial_x^3 \partial_y + \frac{1}{4} h^4 (4) \partial_x^2 \partial_y^2 + \frac{1}{6} h^4 (0) \partial_x \partial_y^3 + \frac{1}{4} h^4 (8 + 4) \partial_y^4 \\ &\quad \left. + \frac{1}{120} h^5 (0 + 0 + 0 + 0 + 0 + 0) \partial_x^5 \right) u(x, y) + \mathcal{O}(M_6(u)h^4) \\ &= (\partial_x^2 + \partial_y^2) u(x, y) + \frac{1}{12} h^2 (\partial_x^4 + 2\partial_x^2 \partial_y^2 + \partial_y^4) u(x, y) + \mathcal{O}(M_6(u)h^4) \\ &= \Delta u(x, y) + \frac{1}{12} h^2 \Delta^2 u(x, y) + \mathcal{O}(M_6(u)h^4) \\ &= -f(x, y) - \frac{1}{12} h^2 \Delta f(x, y) + \mathcal{O}(M_6(u)h^4). \end{aligned}$$

Also ist  $\tau_h(x, y) = \mathcal{O}(h^4)$ , d. h.: Die Diskretisierung hat die Konsistenzordnung  $m = 4$ .

**Lösung A.2.2:** a) Zur Rechtfertigung der Formel schreiben wir

$$\frac{\|e_h\|_h}{\|e_h\|_{h/2}} = \frac{h^\alpha}{(h/2)^\alpha} = 2^\alpha$$

und erhalten durch Logarithmieren

$$\alpha = \frac{\log(\|e_h\|_h / \|e_h\|_{h/2})}{\log(2)}.$$

Wenn die Lösung  $u$  unbekannt ist, machen wir mit einer  $h$ -unabhängigen Verteilungsfunktion  $c(x)$  den (heuristischen) Ansatz

$$\begin{aligned} u_h - u_{h/2} &= u_h - u + u - u_{h/2} = e_{h/2} - e_h = c(h/2)^\alpha - ch^\alpha \\ &= ch^\alpha(2^{-\alpha} - 1) \\ u_{h/2} - u_{h/4} &= u_{h/2} - u + u - u_{h/4} = e_{h/4} - e_{h/2} = c(h/4)^\alpha - c(h/2)^\alpha \\ &= ch^\alpha(4^{-\alpha} - 2^{-\alpha}) = ch^\alpha 2^{-\alpha}(2^{-\alpha} - 1). \end{aligned}$$

Folglich gilt

$$\frac{\|u_h - u_{h/2}\|_h}{\|u_{h/2} - u_{h/4}\|_h} \approx 2^\alpha \quad \text{bzw.} \quad \alpha = \frac{\log(\|u_h - u_{h/2}\|_h / \|u_{h/2} - u_{h/4}\|_h)}{\log(2)}$$

b) Die inhärenten Konvergenzordnungen der gegebenen Folgen sind:

$$\begin{aligned} \alpha &= \frac{\log(\frac{33.627-30.318}{30.318-29.100})}{\log(2)} = \frac{\log(\frac{3.309}{1.218})}{\log(2)} \approx \frac{\log(2.716)}{\log(2)} \approx \frac{0.9994}{0.6931} \approx 1.44 \\ \alpha &= \frac{\log(\frac{30.318-29.100}{29.100-28.586})}{\log(2)} = \frac{\log(\frac{1.218}{0.514})}{\log(2)} \approx \frac{\log(2.369)}{\log(2)} \approx \frac{0.8624}{0.6931} \approx 1.24 \\ \alpha &= \frac{\log(\frac{29.100-28.586}{28.586-28.351})}{\log(2)} = \frac{\log(\frac{0.514}{0.235})}{\log(2)} \approx \frac{\log(2.187)}{\log(2)} \approx \frac{0.7826}{0.6931} \approx 1.12 \end{aligned}$$

und

$$\begin{aligned} \alpha &= \frac{\log(\frac{26.570-27.008}{27.008-27.883})}{\log(2)} = \frac{\log(\frac{-0.438}{-0.875})}{\log(2)} \approx \frac{\log(0.500)}{\log(2)} \approx \frac{-0.6920}{0.6931} \approx -0.998 \\ \alpha &= \frac{\log(\frac{27.008-27.883}{27.883-28.072})}{\log(2)} = \frac{\log(\frac{-0.875}{-0.189})}{\log(2)} \approx \frac{\log(4.629)}{\log(2)} \approx \frac{1.532}{0.6931} \approx 2.21 \\ \alpha &= \frac{\log(\frac{27.883-28.072}{28.072-28.117})}{\log(2)} = \frac{\log(\frac{-0.189}{-0.045})}{\log(2)} \approx \frac{\log(4.2)}{\log(2)} \approx \frac{1.435}{0.6931} \approx 2.07 \end{aligned}$$

**Lösung A.2.3:** a) Wir verwenden wieder die Greensche Identität

$$v_h(P) = h^2 \sum_{Q \in \Omega_h} G_h(P, Q) L_h v_h(Q) + \sum_{Q \in \partial \Omega_h} G_h(P, Q) v_h(Q)$$

mit der durch

$$L_h G_h(P, Q) = h^2 \delta(P, Q), \quad P \in \Omega_h, \quad G_h(P, Q) = \delta(P, Q), \quad P \in \partial \Omega_h, \quad Q \in \bar{\Omega}_h,$$

definierten diskreten Greenschen Funktion  $G_h : \bar{\Omega}_h \times \bar{\Omega}_h \rightarrow \mathbb{R}$ . Da der kompakte 9-Punkte-Operator konsistent ist und die Bedingungen (B1), (B2) und (B3) erfüllt, gilt für ihn das diskrete Maximumsprinzip sowie die Abschätzungen

$$0 \leq h^2 \sum_{Q \in \Omega_h} G_h(P, Q) \leq \frac{d_\Omega^2}{4}.$$

Anwendung der Greenschen Identität für die Fehlerfunktion  $e_h := u - u_h$  ergibt dann wegen  $e_h = 0$  auf  $\partial\Omega_h$ :

$$\begin{aligned} \max_{\bar{\Omega}_h} |e_h| &\leq \frac{d_\Omega^2}{4} \max_{\bar{\Omega}_h} |L_h e_h| = \frac{d_\Omega^2}{4} \max_{\bar{\Omega}_h} |L_h u - L_h u_h| \\ &= \frac{d_\Omega^2}{4} \max_{\bar{\Omega}_h} |L_h u - f - \frac{1}{12} h^2 \Delta f| \leq c M_6(u) h^4. \end{aligned}$$

c) Wir können denselben Ansatz wie in Teil (b) verwenden, da auch das modifizierte 9-Punkte-Schema konsistent ist und den Bedingungen (B1), (B2) und (B3) genügt. Ausgehend von der diskreten Greenschen Identität erhalten wir

$$e_h(P) = h^2 \sum_{Q \in \Omega_h^0} G_h(P, Q) L_h e_h(Q) + h^2 \sum_{Q \in \partial\Omega_h^*} G_h(P, Q) L_h e_h(Q)$$

Mit Hilfe der auch hier gültigen Abschätzungen

$$0 \leq h^2 \sum_{Q \in \Omega_h} G_h(P, Q) \leq \frac{d_\Omega^2}{4}, \quad \sum_{Q \in \partial\Omega_h^*} G_h(P, Q) \leq \frac{1}{2}$$

folgt

$$\begin{aligned} \max_{P \in \bar{\Omega}_h} |e_h| &\leq \frac{d_\Omega^2}{4} \max_{Q \in \Omega_h^0} |-\Delta_h^{(9)} u - f_h| + \frac{1}{2} \max_{Q \in \partial\Omega_h^*} |-\Delta_h^* - f| \\ &\leq c M_6(u) h^6 + c M_3(u) h^3. \end{aligned}$$

**Lösung A.2.4:** a) Wir prüfen zunächst die Formel für die Eigenwerte von  $A_h$  nach. Es gilt (nachrechnen):

$$A_h w^{\nu\mu} = \dots$$

b) Der Darstellung der Eigenwerte von  $A_h$  und der Taylor-Entwicklung des Cosinus

$$\cos(x) = 1 - \frac{x^2}{2} + \mathcal{O}(x^4)$$

entnehmen wir, dass

$$\begin{aligned}\lambda_{\max}(A_h) &= h^{-2}(4 - 4 \cos((1-h)\pi)) = h^{-2}(4 + 4 \cos(h\pi)) = 8h^{-2} + \mathcal{O}(1) \\ \lambda_{\min}(A_h) &= h^{-2}(4 - 4 \cos(h\pi)) = h^{-2}(4 - 4 + 2\pi^2h^2 + \mathcal{O}(h^4))\end{aligned}$$

und somit

$$\text{cond}_2(A_h) =: \frac{\lambda_{\max}(A_h)}{\lambda_{\min}(A_h)} = \frac{8 + \mathcal{O}(h^2)}{2\pi^2h^2 + \mathcal{O}(h^4)} = \frac{4}{\pi^2h^2} \frac{1 + \mathcal{O}(h^2)}{1 + \mathcal{O}(h^2)} = \frac{4}{\pi^2h^2} + \mathcal{O}(1).$$

**Lösung A.2.5:** a) Seien  $x, y \in \mathbb{R}^N$  mit der Eigenschaft  $x \geq y$  gegeben. Dann gilt für beliebige nicht-negative Zahlen  $a_{ij}^{(-1)}$ , dass:

$$\sum_j a_{ij}^{(-1)} x_i \geq \sum_j a_{ij}^{(-1)} y_i.$$

Angewendet auf die Multiplikation mit  $A^{-1}$  folgt:

$$Av \geq Aw \Rightarrow A^{-1}Av \geq A^{-1}Aw \Rightarrow v \geq w$$

und somit die Behauptung.

b) Für den Vektor  $w \in \mathbb{R}^N$  sei  $A_h w \geq (1, \dots, 1)^T$ . Die Matrix  $A_h$  ist invers-monoton, d.h.  $A_h^{-1} \geq 0$ . Also ist

$$w = A_h^{-1} A_h w \geq A_h^{-1} (1, \dots, 1)^T$$

d. h.: Der Vektor  $w$  ist komponentenweise größer oder gleich den jeweiligen (nicht-negativen) Zeilensummen von  $A_h^{-1}$ . Folglich gilt für die Maximale-Zeilensummen-Norm von  $A_h^{-1}$ :

$$\|A_h^{-1}\|_\infty \leq \|w\|_\infty.$$

c) Für die Maximale-Zeilensummen-Norm der Matrix  $A_h$  gilt offenbar  $\|A_h\|_\infty \leq 8h^{-2}$ . Für die Funktion  $w = x(1-x)/4 + y(1-y)/4$  gilt

$$-\Delta_h^{(5)} w = -\Delta w = 1.$$

Dies ist gleichbedeutend mit  $A_h w \geq (1, \dots, 1)^T$ . Nach Teil (b) folgt daraus

$$\|A_h^{-1}\|_\infty \leq \|w\|_\infty = \frac{1}{8}$$

Dies impliziert

$$\text{cond}_\infty(A_h) := \|A_h\|_\infty \|A_h^{-1}\|_\infty \leq h^{-2}.$$

**Lösung A.2.6:** Wir bezeichnen die Anzahl der inneren Gitterpunkt mit der  $y$ -Koordinate  $h$  mit  $n$ . Insgesamt gibt es dann gerade  $N = \frac{1}{2}(n^2 + n)$  innere Gitterpunkte.

a) Bei zeilenweiser Nummerierung, erhalten wir folgende Nummerierung

$$\begin{matrix} N \\ N-2 & N-1 \\ \vdots & \ddots \\ n+1 & n+2 & \cdots & 2n-1 \\ 1 & 2 & \cdots & n-1 & n \end{matrix}$$

dabei ist die Differenz der Nummer zwischen zwei benachbarten Punkten kleiner oder gleich  $n$ . Man beachte, das dieser Abstand für Punkte mit grösseren Nummern kleiner wird. Die Systemmatrix hat damit die Form

$$\begin{pmatrix} B_n & -I_{n-1} & 0 & \cdots & 0 \\ -I_{n-1} & B_{n-1} & -I_{n-2} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -I_2 & B_2 & -I_1 \\ 0 & \cdots & 0 & -I_1 & B_1 \end{pmatrix}.$$

Dabei sind

$$I_i = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \end{pmatrix} \in \mathbb{R}^{i \times i} \quad B_i = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix} \in \mathbb{R}^{i \times i}$$

und an der oberen linken Ecke sind entsprechen 0-en zu ergänzen. Damit ist der Speicher-aufwand  $O(Nn) = O(n^3)$  und der Rechenaufwand ist  $O(Nn^2) = O(N^2)$ .

b) Mit der Nummerierung

$$\begin{matrix} N-n+1 \\ \vdots & \ddots \\ 4 & & \\ 2 & 5 & N-1 \\ 1 & 3 & 6 & \cdots & N \end{matrix}$$

erhalten wir die folgende Systemmatrix:

$$\begin{pmatrix} 4I_1 & M_1 & & & \\ M_1^t & 4I_2 & \ddots & & \\ & M_2^t & \ddots & \ddots & \\ & & \ddots & 4I_{n-1} & M_{n-1} \\ & & & M_{n-1}^t & 4I_n \end{pmatrix}.$$

Wobei  $M_i$  durch

$$M_i = \begin{pmatrix} -1 & -1 & & \\ & \ddots & \ddots & \\ & & -1 & -1 \end{pmatrix} \in \mathbb{R}^{i \times (i+1)}$$

gegeben ist. Insgesamt liegen zwischen den Nummern der 4 benachbarten Felder höchstens die Nummern entlang einer Diagonalen, die Bandbreite ist also  $n + 1$ . Damit ist entsprechend zu Teil (a) der Speicherbedarf  $O(Nn) = O(n^3)$  und der Rechenaufwand ist durch  $O(Nn^2) = O(N^2)$  gegeben.

c) Im Falle der schachbrettartigen Nummerierung liegen zwischen den Indizes benachbarter Elemente höchstens  $N/2 + n/2$  Zahlen. Die entstehende Systemmatrix hat also die Bandbreite  $M/2 + n/2$ . Der Speicherbedarf ist demnach  $O(NN) = O(n^4)$  und der Rechenaufwand  $O(NN^2) = O(N^3)$ . Die Systemmatrix hat die Form

$$\begin{pmatrix} 4I_{\frac{N}{2}} & * \\ * & 4I_{\frac{N}{2}} \end{pmatrix}$$

wobei  $*$  eine Matrix mit höchstens vier von Null verschiedenen Einträgen pro Zeile ist. Wir erkennen, das die Zerlegung nur auf den mit  $*$  gekennzeichneten Einträgen operieren muss, so dass sich der Rechenaufwand auf  $O(n^5)$  verringert.

**Lösung A.2.7:** a) Die Eigenwerte der Iterationsmatrix  $B_\theta = I - \theta A$  sind  $\mu = 1 - \theta\lambda$  und folglich  $\mu_{\max} = 1 - \theta\lambda_{\min}$  sowie  $\mu_{\min} = 1 - \theta\lambda_{\max}$  bzw.

$$\rho(B_\theta) = \max\{|\mu|\} = \max\{|1 - \theta\lambda_{\max}|, |1 - \theta\lambda_{\min}|\}.$$

b) Im Falle  $0 < \lambda_{\min} \leq \lambda_{\max}$  gilt

$$0 < \theta < \frac{2}{\lambda_{\max}} \Leftrightarrow 0 < \theta\lambda_{\min} \leq \theta\lambda_{\max} < 2.$$

Dies wiederum ist äquivalent mit  $\max\{|1 - \theta\lambda_{\min}|, |1 - \theta\lambda_{\max}|\} < 1$ .

c) Ein einfaches geometrische Argument ergibt

$$\theta_{\text{opt}} = \frac{2}{\lambda_{\min} + \lambda_{\max}}.$$

**Lösung A.2.8:** Durch Nachrechnen erhalten wir das für  $1 \leq k, l \leq N$  die Funktionen

$$\varphi^{kl} := \sin\left(\frac{k\pi x}{N+1}\right) \sin\left(\frac{l\pi x}{N+1}\right)$$

Eigenfunktionen von  $A_x$  zum Eigenwert  $\lambda_x^{kl} = 2 - 2 \cos\left(\frac{k\pi}{N+1}\right) = 4 \sin^2\left(\frac{k\pi}{2(N+1)}\right)$ , und Eigenfunktionen von  $A_y$  zum Eigenwert  $\lambda_y^{kl} = 2 - 2 \cos\left(\frac{l\pi}{N+1}\right) = 4 \sin^2\left(\frac{l\pi}{2(N+1)}\right)$  sind. Wir haben also  $N^2$  Eigenfunktionen von  $A_x$  bzw.  $A_y$  aus Dimensionsgründen sind diese gerade ein vollständiges System von Eigenvektoren. Insbesondere ist  $A_x A_y = A_y A_x$ . Die zum ADI-Verfahren gehörige Iterationsmatrix hat gerade die Form

$$B_\sigma = (\sigma + A_y)^{-1} (\sigma - A_y) (\sigma + A_x)^{-1} (\sigma - A_x)$$

und hat die Eigenwerte

$$\lambda_\sigma^{kl} = \frac{\sigma - \lambda_y^{kl}}{\sigma + \lambda_y^{kl}} \frac{\sigma - \lambda_x^{kl}}{\sigma + \lambda_x^{kl}}.$$

Diese sind für  $\sigma > 0$  wohldefiniert, da die Eigenwerte von  $A_x$  und  $A_y$  positiv sind. Nun ist

$$\frac{|\sigma - c|}{|\sigma + c|} < 1$$

für jedes feste  $c > 0$  und beliebiges  $\sigma > 0$ , so dass für den Spektralradius von  $B_\sigma$  gilt:

$$\rho(B_\sigma) < 1.$$

Das ADI-Verfahren ist also unabhängig von  $\sigma > 0$  konvergent. Die Konvergenz ist am schnellsten falls  $\rho(B_\sigma)$  minimal ist. Wir suchen also

$$\min_{\sigma > 0} \max_{k,l} \left| \frac{\sigma - \lambda_y^{kl}}{\sigma + \lambda_y^{kl}} \frac{\sigma - \lambda_x^{kl}}{\sigma + \lambda_x^{kl}} \right|.$$

Da  $\lambda_x^{kl}$  von  $l$  unabhängig und  $\lambda_y^{kl}$  von  $k$  unabhängig ist, sowie  $\lambda_x^{kl} = \lambda_y^{lk}$  ist

$$\max_{k,l} \left| \frac{\sigma - \lambda_y^{kl}}{\sigma + \lambda_y^{kl}} \frac{\sigma - \lambda_x^{kl}}{\sigma + \lambda_x^{kl}} \right| = \max_k \frac{|\sigma - \lambda_x^{k1}|^2}{|\sigma + \lambda_x^{k1}|^2}$$

Wir betrachten jetzt den Graphen der Funktion  $x \mapsto \frac{|\sigma-x|}{|\sigma+x|}$ , und erkennen, dass das Maximum der Beziehung

$$\max_{\lambda_x^1 \leq x \leq \lambda_x^N} \frac{|\sigma - x|^2}{|\sigma + x|^2} = \max \left( \frac{|\sigma - \lambda_x^1|^2}{|\sigma + \lambda_x^1|^2}, \frac{|\sigma - \lambda_x^N|^2}{|\sigma + \lambda_x^N|^2} \right)$$

genügt. Damit erhalten wir, dass für den optimalen Parameter die Beziehung

$$\frac{|\sigma - \lambda_x^1|^2}{|\sigma + \lambda_x^1|^2} = \frac{|\sigma - \lambda_x^N|^2}{|\sigma + \lambda_x^N|^2}$$

gilt. Aus Monotoniegründen kann dies nur für  $\sigma \in [\lambda_x^1, \lambda_x^N]$  der Fall sein. Wir erhalten:

$$\begin{aligned} \frac{\sigma - \lambda_x^1}{\sigma + \lambda_x^1} &= \frac{\lambda_x^N - \sigma}{\sigma + \lambda_x^N} \\ \Leftrightarrow \quad \sigma^2 - \sigma \lambda_x^1 + \sigma \lambda_x^N - \lambda_x^1 \lambda_x^N &= -\sigma^2 - \sigma \lambda_x^1 + \sigma \lambda_x^N + \lambda_x^1 \lambda_x^N \\ \Leftrightarrow \quad \sigma &= \sqrt{\lambda_x^1 \lambda_x^N}. \end{aligned}$$

Im speziellen Fall erhalten wir also

$$\sigma = 4 \sin\left(\frac{\pi}{2(N+1)}\right) \sin\left(\frac{N\pi}{2(N+1)}\right)$$

und mit diesem  $\sigma$  ist

$$\rho(B_\sigma) = \frac{\cos^2\left(\frac{\pi}{N+1}\right)}{\left(1 + \sin\left(\frac{\pi}{N+1}\right)\right)^2}$$

### A.3 Kapitel 3

**Lösung A.3.1:** a) Wir suchen zunächst eine geeignete schwache Formulierung. Durch Multiplikation mit einer Funktion  $\varphi \in V = H^1$ , Integration über  $\Omega$ , und anschließende partielle Integration erhalten wir:

$$(\nabla u, \nabla \varphi)_\Omega - \langle \partial_n u, \varphi \rangle_{\partial\Omega} = (f, \varphi)_\Omega \quad \forall \varphi \in V.$$

Indem wir die Neumann-Randwerte einsetzen, folgt die schwache Formulierung:

$$(\nabla u, \nabla \varphi)_\Omega = (f, \varphi)_\Omega + \langle g, \varphi \rangle_{\partial\Omega} \quad \forall \varphi \in V.$$

Wir wissen bereits, dass obiges Problem nicht eindeutig lösbar ist, wir können also auch keine Bestapproximationseigenschaft erwarten. Fordern wir zusätzlich, dass  $\int_\Omega u = 0$  für alle  $u \in V$ , um eine eindeutige Lösung zu erhalten, so ist aufgrund der Poincaréschen Ungleichung für  $u \in V$

$$\|u\|_{H^1} \leq c \sqrt{(\nabla u, \nabla u)_\Omega}.$$

Also ist  $\sqrt{(\nabla u, \nabla u)_\Omega}$  eine zu  $\|u\|_{H^1}$  äquivalente Norm auf  $V$ . Sei nun  $V_h$  ein endlich-dimensionaler Teilraum von  $V$ , dann lautet das Ritz-Verfahren gegeben durch

$$(\nabla u_h, \nabla \varphi_h)_\Omega = (f, \varphi_h)_\Omega + \langle g, \varphi_h \rangle_{\partial\Omega} \quad \forall \varphi_h \in V_h.$$

Wir erhalten durch Galerkinorthogonalität für beliebiges  $v_h \in V_h$ :

$$\begin{aligned} (\nabla(u - u_h), \nabla(u - u_h))_\Omega &= (\nabla(u - u_h), \nabla(u - v_h))_\Omega + (\nabla(u - u_h), \nabla(v_h - u_h))_\Omega \\ &= (\nabla(u - u_h), \nabla(u - v_h))_\Omega \\ &\leq \|\nabla(u - u_h)\|_\Omega \|\nabla(u - v_h)\|_\Omega. \end{aligned}$$

Und somit, da  $v_h \in V_h$  beliebig war,

$$\|\nabla(u - u_h)\|_\Omega \leq \inf_{v_h \in V_h} \|\nabla(u - v_h)\|_\Omega.$$

b) Um eine variationelle Formulierung zu erhalten, betrachten wir den Rayleigh-Quotienten

$$R(u) = \frac{(\nabla u, \nabla u)_\Omega}{(u, u)_\Omega}$$

Es kann gezeigt werden, dass  $R(v) \geq \lambda$ , falls  $\lambda$  der kleinste Eigenwert des Laplace-Operators ist. Für einen Eigenvektor  $u$  zum Eigenwert  $\lambda$  ist  $R(u) = \lambda$ . Das Eigenwertproblem zum kleinsten Eigenwert lässt sich, mit  $V := H_0^1$ , schreiben als

$$\min_{u \in V} R(u) =: \lambda.$$

Entsprechend ist das Rayleigh-Ritz-Verfahren mit einem endlich-dimensionalen Teilraum  $V_h \subset V$

$$\min_{u_h \in V_h} R(u_h) =: \lambda_h.$$

Wir sehen hieraus sofort:

$$\lambda \leq \lambda_h$$

Für einen beliebigen Eigenwert  $\lambda_l$  gilt das min-max-Prinzip

$$\lambda_l = \min_{S_l \subset V, \dim(S_l)=l} \max_{v \in S_l} R(v)$$

sowie im Endlichdimensionalen

$$\lambda_l^h = \min_{S_l \subset V_h, \dim(S_l)=l} \max_{v_h \in S_l} R(v_h).$$

Mit Mitteln, die über den Stoff dieses Textes hinausgehen, zeigt man für einen konformen Finite-Elemente-Ansatz der Ordnung  $k$  die Fehlerabschätzung

$$|\lambda_l - \lambda_l^h| \leq Ch^{2k} \lambda_l^k.$$

c) Wir erhalten wieder durch Multiplikation mit einer Funktion  $v$  und partieller Integration

$$\begin{aligned} (\Delta^2 u, v) &= -(\nabla \Delta u, \nabla v) + (\partial_n \Delta u, v) \\ &= (\Delta u, \Delta v) + (\partial_n \Delta u, v) - (\Delta u, \partial_n v) \\ &= (\Delta u, \Delta v), \end{aligned}$$

und somit als schwache Formulierung

$$(\Delta u, \Delta v) = (-f, v) \quad \forall v \in H_0^2(\Omega).$$

Entsprechend ist das Ritzverfahren mit  $V_h \subset H_0^2(\Omega)$  endlichdimensional:

$$(\Delta u_h, \Delta v_h) = (-f, v_h) \quad \forall v_h \in V_h.$$

Um die Bestapproximationseigenschaft zu erhalten benötigen wir Stetigkeit und Koerzitivität von  $(\Delta \cdot, \Delta \cdot)$  auf  $H_0^2(\Omega)$ . Die Stetigkeit ist klar. Um die Koerzitivität zu zeigen, beachten wir, dass aufgrund der Poincaréschen Ungleichung

$$\|u\|_{H^2} \leq c |u|_{H^2}$$

gilt. Um nun die 2. Ableitungen durch den Laplace-Operator abzuschätzen betrachten wir das Vektorfeld

$$w = \begin{pmatrix} u_{11}u_{22} \\ -u_{11}u_{12} \end{pmatrix}.$$

Es gilt nach dem Gaußschen-Integralsatz unter Berücksichtigung der Randwerte

$$\int_{\Omega} u_{11}u_{22} - u_{12}^2 = \int_{\Omega} \nabla \cdot w = \int_{\partial\Omega} n \cdot w = 0.$$

Zusammengenommen ergibt sich

$$(\Delta u, \Delta u) = \int_{\Omega} |\Delta u|^2 = \int_{\Omega} (u_{11}u_{22} - u_{12}^2) = \int_{\Omega} u_{11}^2 + u_{22}^2 + u_{12}^2 = \|u\|_{H^2}^2$$

und somit die Koerzitivität.

**Lösung A.3.2:** a) Die Galerkin-Gleichungen entsprechen einem quadratischen linearen Gleichungssystem der Dimension  $N = \dim(V_h)$  für die Entwicklungskoeffizienten von  $u_h$  bzgl. einer beliebigen Basis von  $V_h$ . Zum Nachweis der eindeutigen Lösbarkeit genügt also der Nachweis der Eindeutigkeit. Für zwei Lösungen  $u_h^{(1)}, u_h^{(2)}$  erfüllt die Differenz  $w_h := u_h^{(1)} - u_h^{(2)}$  die Gleichung

$$a(w_h, \varphi_h) = 0 \quad \forall \varphi_h \in V_h.$$

Bei Wahl von  $\varphi_h = w_h$  folgt daher mit Hilfe der  $V$ -Elliptizität

$$0 = |a(w_h, w_h)| \geq \kappa \|w_h\|^2 \Rightarrow w_h = 0.$$

Zum Nachweis der Fehlerabschätzung verwenden wir zunächst die  $V$ -Elliptizität, dann die Galerkin-Orthogonalität mit einem beliebigen  $\varphi_h \in V_h$  und schließlich die Beschränktheit:

$$\|u - u_h\|^2 \leq \frac{1}{\kappa} |a(u - u_h, u - u_h)| = \frac{1}{\kappa} |a(u - u_h, u - \varphi_h)| \leq \frac{\alpha}{\kappa} \|u - u_h\| \|u - \varphi_h\|.$$

Dies impliziert offensichtlich die behauptete Ungleichung.

bi) Wir diskutieren zunächst die Lösbarkeit der Variationsaufgabe

$$a(u, \varphi) = l(\varphi) \quad \forall \varphi \in V,$$

unter der Voraussetzung, daß  $a(\cdot, \cdot)$  folgende Koerzitivitätseigenschaften besitzt:

$$\sup_{\varphi \in V} \frac{a(v, \varphi)}{\|\varphi\|} \geq \gamma \|v\|, \quad \sup_{\varphi \in V} \frac{a(\varphi, v)}{\|\varphi\|} \geq \gamma \|v\|, \quad v \in V.$$

Dies bedeutet, daß  $a(\cdot, \cdot)$  und die durch  $a^*(v, \varphi) := a(\varphi, v)$  definierte zugehörige „adjungierte“ Bilinearform koerzitiv sind. Die Bilinearformen  $a(\cdot, \cdot)$  und  $a^*(\cdot, \cdot)$  definieren mit Hilfe des Rieszschen Darstellungssatzes durch

$$(Av, \varphi) = a(v, \varphi), \quad (A^*v, \varphi) = a^*(v, \varphi) = (\varphi, v), \quad \varphi \in V,$$

lineare Operatoren  $A : V \rightarrow V$  und  $A^* : V \rightarrow V$ . Diese sind wegen der Koerzitivität der definierenden Bilinearformen injektiv:

$$Av = 0 \quad \Rightarrow \quad 0 = \sup_{\varphi \in V} \frac{a(v, \varphi)}{\|\varphi\|} \geq \gamma \|v\| \quad \Rightarrow \quad v = 0,$$

und analog für  $A^*$ . Wir definieren die symmetrische und positive Bilinearform

$$[v, \varphi] := (A^*v, A^*\varphi), \quad v, \varphi \in V.$$

Mit der zugehörigen Norm  $|v| := [v, v]^{1/2} = \|A^*v\|$  gilt dann:

$$\gamma \|v\| \leq \sup_{\varphi \in V} \frac{a(\varphi, v)}{\|\varphi\|} = \sup_{\varphi \in V} \frac{(\varphi, A^*v)}{\|\varphi\|} \leq \|A^*v\| = |v| \leq \gamma' \|v\|$$

mit

$$\gamma' := \sup_{\varphi \in V} \frac{\|A^*\varphi\|}{\|\varphi\|}.$$

Folglich definiert  $[\cdot, \cdot]$  ein Skalarprodukt auf  $V$ , welches zu dem auf  $V$  gegebenen Skalarprodukt  $(\cdot, \cdot)$  äquivalent ist. Nach dem Rieszschen Darstellungssatz existiert daher zu jedem linearen Funktional  $l \in V^*$  ein  $w \in V$ , so dass

$$l(\varphi) = [\varphi, v] = (A^*\varphi, A^*w) = (\varphi, AA^*w) = (\varphi, Av) \quad \forall \varphi \in V,$$

mit  $v := A^*w$ . Also ist  $v$  (eindeutige) Lösung der obigen Variationsaufgabe.

(bii) Die Koerzitivität von  $a(\cdot, \cdot)$  auf  $V$  reicht allein nicht aus, um die gewünschten Konvergenzaussagen zu erhalten. Setzt man aber voraus, daß  $a(\cdot, \cdot)$  auch auf den Teilmengen  $V_h \subset V$  koerzitiv ist und zwar gleichmäßig bzgl.  $h$ ,

$$\sup_{\psi_h \in V_h} \frac{a(v_h, \psi_h)}{\|\psi_h\|} \geq \gamma_h \|v_h\|, \quad v_h \in V_h, \quad \gamma_h \geq \gamma_0 > 0,$$

so lassen sich die obigen Aussagen beweisen. Mit beliebigem  $\varphi_h \in V_h$  gilt:

$$\begin{aligned}\|u - u_h\| &\leq \|u - \varphi_h\| + \|\varphi_h - u_h\| \\ &\leq \|u - \varphi_h\| + \frac{1}{\gamma_0} \sup_{\psi_h \in V_h} \frac{a(\varphi_h - u_h, \psi_h)}{\|\psi_h\|} \\ &\leq \|u - \varphi_h\| + \frac{1}{\gamma_0} \sup_{\psi_h \in V_h} \frac{a(\varphi_h - u, \psi_h)}{\|\psi_h\|} + \frac{1}{\gamma_0} \sup_{\psi_h \in V_h} \frac{a(u - u_h, \psi_h)}{\|\psi_h\|} \\ &\leq \left(1 + \frac{\alpha}{\gamma_0}\right) \|\varphi_h - u\|.\end{aligned}$$

Der Nachweis der „diskreten Koerzitivität“ erfordert spezielle Bedingungen an die Bilinearform  $a(\cdot, \cdot)$  und den Ansatzraum  $V_h$ . Zum Beispiel ist die Bilinearform

$$a(\cdot, \cdot) = (\nabla \cdot, \nabla \cdot) - \mu(\cdot, \cdot),$$

wenn  $\mu$  kein Eigenwert des Laplace-Operators ist, koerzitiv aber nicht  $V$ -elliptisch. Die Gültigkeit der (gleichmäßigen) diskreten Koerzitivität folgt dann über ein Widerspruchssargument mit Hilfe der Kompaktheit der Einbettung  $H_0^1(\Omega) \subset L^2(\Omega)$ . Angenommen, die Bilinearform  $a(\cdot, \cdot)$  ist nicht gleichmäßig koerzitiv auf  $V_h$ . Dann existieren eine Folge von Gitterweiten  $(h_k)_{k \in \mathbb{N}}$  und Funktionen  $v_k \in V_k := V_{h_k}$  mit den Eigenschaften

$$\|\nabla v_k\| = 1, \quad \sup_{\varphi_k \in V_k} \frac{a(v_k, \varphi_k)}{\|\nabla \varphi_k\|} < \frac{1}{k}, \quad k \in \mathbb{N}.$$

Wegen der Kompaktheit der Einbettung  $H_0^1(\Omega) \subset L^2(\Omega)$  existieren für die  $H^1$ -beschränkte Folge  $(v_k)_{k \in \mathbb{N}}$  eine Teilfolge  $(v_{k'})_{k' \in \mathbb{N}}$  und ein  $v \in L^2(\Omega)$  mit

$$\|v_{k'} - v\| \rightarrow 0 \quad (k' \rightarrow \infty).$$

Wegen der schwachen Kompaktheit der Einheitskugel in  $H_0^1(\Omega)$  kann o.B.d.A. erreicht werden, daß die Folge  $(v_{k'})_{k' \in \mathbb{N}}$  auch schwach in  $H_0^1(\Omega)$  gegen  $v$  konvergiert, d.h.:  $v \in H_0^1(\Omega)$  und

$$(\nabla v_{k'}, \nabla \varphi) \rightarrow (v, \varphi), \quad \varphi \in H_0^1(\Omega).$$

Dies impliziert zunächst für beliebiges  $\varphi \in H_0^1(\Omega)$ :

$$|(\nabla v, \nabla \varphi) - \mu(v, \varphi)| = \lim_{k' \rightarrow \infty} |(\nabla v_{k'}, \nabla \varphi) - \mu(v_{k'}, \varphi)| \leq 0,$$

und folglich  $v = 0$ , da  $\mu$  nach Voraussetzung  $\mu$  kein Eigenwert ist. Dies impliziert dann

$$\frac{(\nabla v_{k'}, \nabla v_{k'}) - \mu(v_{k'}, v_{k'})}{\|\nabla v_{k'}\|} \leq \sup_{\varphi_k' \in V_{k'}} \frac{(\nabla v_{k'}, \nabla \varphi_{k'}) - \mu(v_{k'}, \varphi_{k'})}{\|\nabla \varphi_{k'}\|} \rightarrow 0 \quad (k' \rightarrow \infty)$$

und folglich  $\|\nabla v_{k'}\| \rightarrow 0$  ( $k' \rightarrow \infty$ ) im Widerspruch zur Annahme  $\|\nabla v_k\| = 1$ .

**Lösung A.3.3:** a) Exakte Integration ergibt die Werte

$$a_{ij} = (\nabla \varphi_h^i, \nabla \varphi_h^j) = \begin{cases} \frac{8}{3}, & j = i, \\ -\frac{1}{3}, & j \in \{i \pm 1, i \pm m, i \pm m \pm 1\}, \\ 0, & \text{sonst.} \end{cases}$$

b) Verwendung der 2-dimensionalen "Tensorprodukt-Trapezregel" ergibt:

$$a_{ij} = (\nabla \varphi_h^i, \nabla \varphi_h^j) = \begin{cases} 4, & j = i, \\ -1, & j \in \{i \pm 1, i \pm m\}, \\ 0, & \text{sonst.} \end{cases}$$

Der Finite-Elemente-Ansatz mit bilinearen Formfunktionen ergibt bei Verwendung der Trapezregel bis auf den Vorfaktor  $h^{-2}$  dieselbe Systemmatrix wie der 5-Punkte-Differenzenoperator.

**Lösung A.3.4:** Wir betrachten zunächst Dreiecke. Dann sind aufgrund des Zusammenhangs

$$\rho_T = h_T \sin(\alpha) \sin(\beta) \sin(\gamma)$$

die Bedingungen (a) und (b) äquivalent.

i) Sei nun a) erfüllt. Dann sind ferner alle Winckel gleichmässig von  $\pi$  wegbeschränkt. Somit gilt für jeden Winkel  $\alpha$

$$\sin(\alpha) \geq c > 0$$

und somit

$$\frac{\sin(\alpha)}{\sin(\beta)} \leq c \sin(\alpha) \leq c$$

Aus dem Sinussatz

$$\frac{\sin(\alpha)}{\sin(\beta)} = \frac{a}{b}$$

folgt also Bedingung (c). Die Umkehrung gilt nicht, wie man sich anhand der Abb. A.1 klar macht.

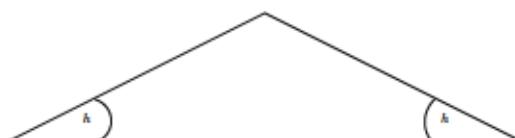


Abbildung A.1: Gegenbeispiel zur Äquivalenz der Formregularität bei Dreiecken

ii) Bei Vierecken sind zusätzlich die Implikationen  $(a) \Rightarrow (b), (c)$  sowie  $b \Rightarrow (c)$  falsch, wie man sich leicht anhand des ersten bzw. zweiten Rechtecks in Abb. A.2 überlegt.

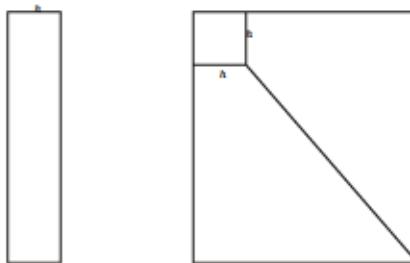


Abbildung A.2: Gegenbeispiele zur Äquivalenz der Formregularität bei Vierecken

**Lösung A.3.5:** a) Mit beliebigem  $\varphi_h \in V_h^{(1)}$  gilt

$$\|u - P_h u\|^2 = (u - P_h u, u - P_h u) = (u - P_h u, u - \varphi_h) \leq \|u - P_h u\| \|u - \varphi_h\|$$

und folglich

$$\|u - P_h u\| = \min_{\varphi_h \in V_h^{(1)}} \|u - \varphi_h\|.$$

Mit Hilfe der Abschätzung für die Knoteninterpolierende  $I_h u \in V_h^{(1)}$  für  $u \in H^2(\Omega)$ ,

$$\|u - I_h u\| \leq ch^2 \|\nabla^2 u\|,$$

folgt  $\|u - P_h u\| \leq ch^2 \|\nabla^2 u\|$ . Im Falle der Minimalregularität  $u \in L^2(\Omega)$  kann keine  $h$ -Potenz in der Fehlerabschätzung erwartet werden; es gilt aber:

$$\|u - P_h u\| = \min_{\varphi_h \in V_h^{(1)}} \|u - \varphi_h\| \leq \|u\|.$$

Für  $u \in H^1(\Omega)$  ist in mehr als einer Dimension die Knoteninterpolierende  $I_h u$  gar nicht definiert. In diesem Fall muß mit einer modifizierten „Quasi-Interpolierenden“  $\tilde{I}_h : H^1(\Omega) \rightarrow V_h^{(1)}$  gearbeitet werden, welche  $H^1$ -stabil ist. Dazu ordnen wir jedem Gitterknoten  $a$  den umgebenden Zellbereich  $S(a) = \cup\{T \in \mathcal{T}_h, a \in T\}$  zu und setzen

$$\tilde{I}_h u(a) := |S(a)|^{-1} \int_{S(a)} u(x) dx.$$

Für die so definierte Funktion  $\tilde{I}_h u \in V_h^{(1)}$  gilt dann die Fehlerabschätzung

$$\|u - \tilde{I}_h u\| \leq ch \|u\|_1.$$

Den recht komplizierten Beweis können wir hier nicht geben.

Alternativ kann man im Falle eines regulären Gebiets (glattberandeter oder konvexes Polygon/Polyeder) auch die Ritz-Projektion  $P_h u \in S_h^{(1)}$  verwenden. Für diese gilt mit der zugehörigen „dualen Lösung“  $z \in H_0^1(\Omega) \cap H^2(\Omega)$  von  $-\Delta z = \|u - R_h\|^{-1}(u - R_h u)$ :

$$\begin{aligned}
\|u - R_h u\| &= (\nabla(u - R_h u), \nabla z) = (\nabla(u - R_h u), \nabla(z - R_h z)) = (\nabla u, \nabla(z - R_h z)) \\
&\leq \|\nabla u\| \|\nabla(z - R_h z)\| \leq \|\nabla u\| \|\nabla(z - I_h z)\| \\
&\leq c_i h \|\nabla u\| \|\nabla^2 z\| \leq c_i c_s h \|\nabla u\|.
\end{aligned}$$

Damit ergibt sich dann

$$\|u - P_h u\| \leq \|u - R_h u\| \leq c_i c_s h \|\nabla u\|.$$

b) Für die negative Sobolew-Norm erhält man mit Hilfe der obigen Abschätzung für den Interpolationsfehler:

$$\begin{aligned}
\|u - P_h u\|_{-1} &= \sup_{\varphi \in H_0^1(\Omega)} \frac{(u - P_h, \varphi)}{\|\nabla \varphi\|} = \sup_{\varphi \in H_0^1(\Omega)} \frac{(u - P_h, \varphi - I_h \varphi)}{\|\nabla \varphi\|} \\
&\leq \|u - P_h u\| \sup_{\varphi \in H_0^1(\Omega)} \frac{\|\varphi - I_h \varphi\|}{\|\nabla \varphi\|} \leq c_2 c_1 h^3 \|\nabla^2 u\|.
\end{aligned}$$

**Lösung A.3.6:** In einer Dimension, d. h.  $\Omega = (0, 1)$ , gilt mit der Knoteninterpolierenden  $I_h u \in V_h^{(1)}$  für beliebiges  $\varphi_h \in V_h^{(1)}$ :

$$\begin{aligned}
(I_h u', \varphi'_h) &= \sum_{i=1}^{N+1} \int_{x_{i-1}}^{x_i} I_h u' \varphi'_h \, dx = \sum_{i=1}^{N+1} \left\{ - \int_{x_{i-1}}^{x_i} I_h u \varphi''_h \, dx + I_h u \varphi'_h \Big|_{x_{i-1}}^{x_i} \right\} \\
&= \sum_{i=1}^{N+1} \left\{ - \int_{x_{i-1}}^{x_i} u \varphi''_h \, dx + u \varphi'_h \Big|_{x_{i-1}}^{x_i} \right\} = \sum_{i=1}^{N+1} \int_{x_{i-1}}^{x_i} u' \varphi'_h \, dx = (u', \varphi'_h).
\end{aligned}$$

In diesem Fall stimmen also Ritz-Projektion  $R_h u$  und Knoteninterpolierende  $I_h u$  überein. Für die Interpolierende gilt nun i. Allg.:

$$\|u - I_h u\|_{-1} = \sup_{\varphi \in H_0^1(\Omega)} \frac{(u - I_h u, \varphi)}{\|\varphi'\|} \geq ch^2.$$

**Lösung A.3.7:** a) Mit Hilfe der Hölderschen Ungleichung folgt

$$\left| \int_{\Omega} (u - u_h) \omega \, dx \right| \leq \|\omega\| \|u - u_h\|,$$

d. h.: Die abzuschätzende Größe ist durch die  $L^2$ -Norm des Fehlers beschränkt. Also gilt nach einem Resultat der Vorlesung:

$$\left| \int_{\Omega} (u - u_h) \omega \, dx \right| \leq ch^2 \|\nabla^2 u\| \leq ch \|f\|.$$

b) Zunächst gilt:

$$\left| \int_{\Gamma} (u^2 - u_h^2) \, ds \right| = \left| \int_{\Gamma} (u - u_h)(u + u_h) \, ds \right| \leq \left( \int_{\Gamma} |u - u_h|^2 \, ds \right)^{1/2} \left( \int_{\Gamma} |u + u_h|^2 \, ds \right)^{1/2}.$$

Für den zweiten Faktor folgt mit Hilfe der  $L^2$ -Spurabschätzung aus der Vorlesung

$$\left( \int_{\Gamma} |u + u_h|^2 ds \right)^{1/2} \leq c \left( \int_{\Omega} |u + u_h|^2 ds \right)^{1/2} + \left( \int_{\Omega} |\nabla(u + u_h)|^2 ds \right)^{1/2} \leq c.$$

Für den ersten Faktor erhalten wir mit der  $L^1$ -Spurabschätzung angewendet für  $(u - u_h)^2$ :

$$\begin{aligned} \int_{\Gamma} |u - u_h|^2 ds &\leq c \int_{\Omega} |u - u_h|^2 dx + c \int_{\Omega} |\nabla(u - u_h)^2| dx \\ &\leq c \int_{\Omega} |u - u_h|^2 dx + c \int_{\Omega} |u - u_h| |\nabla(u - u_h)| dx \\ &\leq c \int_{\Omega} |u - u_h|^2 dx + c \left( \int_{\Omega} |u - u_h|^2 dx \right)^{1/2} \left( \int_{\Omega} |\nabla(u - u_h)|^2 dx \right)^{1/2} \end{aligned}$$

Die  $L^2$ -Fehlerabschätzungen aus der Vorlesung ergeben also

$$\left| \int_{\Gamma} (u^2 - u_h^2) ds \right| \leq ch^{3/2} \|\nabla^2 u\| \leq ch^{3/2} \|f\|.$$

**Lösung A.3.8:** ai)  $T$  kartesisches Einheitsdreieck:

$$P(T) = P_3(T), \quad p(a_i), \nabla p(a_i), p(z), \quad P(T) = P_3(T), \quad p(a_i), p(b_{ij}), p(z).$$

Sei  $p \in P(T)$ , welches bzgl. aller Knotenfunktionale verschwindet. Dann ist  $p|_{\Gamma_i} \equiv 0$ . Auf dem Einheitsdreieck ist folglich  $p(x, y) = cxy(x - y)$ , da jedes entlang von  $\partial T$  verschwindende Polynom die drei Faktoren  $x$ ,  $y$  und  $x - y$  enthalten muß. Wegen  $p(z) = 0$  folgt notwendig  $c = 0$ , d. h.  $p \equiv 0$ .

aiii) Ein Polynom  $p \in P_5(T)$  mit  $p(a_i) = 0$ ,  $\nabla p(a_i) = 0$ ,  $\nabla^2 p(a_i) = 0$ ,  $\partial_n p(m_i) = 0$  hat auf dem Einheitsdreieck notwenig die Gestalt  $p(x, y) = xy(1 - x - y)q(x, y)$  mit einem  $q \in P_2(T)$ . Wegen  $\nabla^2 p(0, 0) = 0$  gilt:

$$\begin{aligned} 0 &= \partial_x \partial_y p(0, 0) \\ &= (\partial_x \partial_y (xy(1 - x - y))q + \partial_y (xy(1 - x - y))\partial_x q \\ &\quad + \partial_x (xy(1 - x - y))\partial_y q + xy(1 - x - y)\partial_x \partial_y q)(0, 0) = q(0, 0). \end{aligned}$$

Dies impliziert  $q(0, 0) = 0$ . Analog folgt  $q(1, 0) = q(0, 1) = 0$ . Weiter gilt

$$\begin{aligned} 0 &= \partial_n p(m_1) = -\partial_y (xy(1 - x - y)q)(\frac{1}{2}, 0) \\ &= (-x(1 - x - y)q + xyq - xy(1 - x - y)\partial_y q)(\frac{1}{2}, 0) = -\frac{1}{4}q(m_1) \end{aligned}$$

und analog  $q(m_2) = q(m_3) = 0$ . Wegen der Unisolvenz von  $P_2(T)$  mit dem Satz  $\{q(a_i), q(m_i), i = 1, 2, 3\}$  von Knotenwerten folgt schließlich  $q \equiv 0$ . bzw.  $p \equiv 0$ .

aiii)  $T$  kartesisches Einheitsquadrat:

$$P(T) = \tilde{Q}_1(T) := P_1(T) \oplus \text{span}\{x^2 - y^2\}, \quad p(m_i), i = 1, \dots, 4.$$

Sei  $p \in P(T)$  mit  $p(m_i) = 0$ . Auf dem Einheitsquadrat ist  $p$  linear entlang der schrägen Verbindungslien zwischen den Mitten benachbarter Kanten und folglich gleich Null entlang dieser Linien. Damit ist auch  $\nabla p(m_i) = 0$ . Entlang jeder Linie durch eine Seitenmitte  $m_i$  verschwindet damit  $p$  in mindestens zwei Punkten und seine Richtungsableitung in mindestens einem Punkt. Da  $p$  entlang jeder Linie höchstens quadratisch ist, folgt  $p \equiv 0$ .

$$P(T) = \tilde{Q}_3(T) := P_3(T) \oplus \text{span}\{x^3y, xy^3\}, \quad p(a_i), \nabla p(a_i), i = 1, \dots, 4.$$

Sei  $p \in P(T)$ , welches bzgl. aller Knotenfunktionale verschwindet. Dann ist  $p|_{\partial T} \equiv 0$ . Folglich müßte  $p$  den Faktor  $cxy(1-x)(1-y)$  enthalten mit einer Konstante  $c \in \mathbb{R}$ . Da der Term  $x^2y^2$  aber nicht durch Elemente des Raumes  $P(T)$  erzeugt werden kann, muß  $c = 0$  sein. Dies bedeutet auch  $p \equiv 0$ .

b) Bei der Approximation der Laplace-Gleichung auf einem äquidistanten kartesischen Gitter gilt:

$$\dim V_h^{(3)} = \frac{9}{h^2}, \quad \dim \tilde{V}_h^{(3)} = \frac{5}{h^2}.$$

Die Anzahl der von Null verschiedenen Elemente pro Zeile der Systemmatrizen ist

$$V_h^{(3)} : N_z = 10, N_b = 16, N_a = 37, \quad \tilde{V}_h^{(3)} : N_z = 10, N_a = 27.$$

**Lösung A.3.9:** Wegen der Gültigkeit der Einbettung  $H^2(\Omega) \subset L^\infty(\Omega)$  in 2 und 3 Dimensionen erfüllt das Funktional  $F : H^2(\hat{T}) \rightarrow \mathbb{R}$ ,

$$F(v) = \max_{\hat{T}} |v - I_h v|$$

die Voraussetzungen des Bramble-Hilbert-Lemmas. Demnach gilt auf dem Referenzelement  $\hat{T}$

$$\max_{\hat{T}} |\hat{v} - I_h \hat{v}| \leq c \|\hat{\nabla}^2 \hat{v}\|_{\hat{T}}.$$

Die Transformationsargumente aus der Vorlesung ergeben

$$\max_T |v - I_h v| = \max_{\hat{T}} |\hat{v} - I_h \hat{v}| \leq c \|\hat{\nabla}^2 \hat{v}\|_{\hat{T}} \leq ch^2 h^{-\frac{d}{2}} \|\nabla^2 v\|_T.$$

Dies impliziert die behauptete Abschätzung in 2 Dimensionen. Für  $d = 3$  folgt

$$\max_{\Omega} |v - I_h v| \leq ch^{\frac{1}{2}} \|\nabla^2 v\|_{\Omega}.$$

**Lösung A.3.10:** Alle 4 Abschätzungen sind richtig:

i) Auf dem Referenzelement  $\hat{T}$  sind für den endlich dimensionalen Polynomraum  $P(\hat{T})$  alle Normen äquivalent, d. h. es gilt mit einer  $h$ -unabhängigen Konstante  $c > 0$ , so dass

$$\|\hat{\nabla}^2 \hat{v}_h\|_T \leq \|\hat{v}_h\|_{2,T} \leq c \|\hat{v}_h\|_T.$$

Die üblichen Transformationsargumente liefern nun

$$\|\nabla^2 v_h\|_T \leq ch^{-2} h \|\hat{\nabla}^2 \hat{v}_h\|_{\hat{T}} \leq ch^{-1} \|\hat{v}_h\|_{\hat{T}} \leq ch^{-2} \|v_h\|_T.$$

ii) Auf dem Quotientenraum

$$\frac{P(\hat{T})}{P_0}$$

sind die Normen  $\|\nabla \cdot\|_{1,\hat{T}}$  und  $\|\nabla \cdot\|_{\hat{T}}$  äquivalent. Mit dem Spurlemma gilt

$$\|\hat{\partial}_n \hat{v}_h\|_{\partial \hat{T}} \leq c \|\hat{\nabla} \hat{v}_h\|_{1,\hat{T}} \leq c \|\hat{\nabla} \hat{v}_h\|_{\hat{T}}.$$

Mit Transformationsargumenten folgt nun

$$\|\partial_n v_h\|_{\partial T} \leq ch^{-1} ch^{\frac{1}{2}} \|\hat{\partial}_n \hat{v}_h\|_{\partial \hat{T}} \leq ch^{-\frac{1}{2}} \|\hat{\nabla} \hat{v}_h\|_{\hat{T}} \leq ch^{-\frac{1}{2}} \|\nabla v_h\|_T.$$

iii) Wir nutzen die Äquivalenz der  $W^{1,\infty}$ - und der  $L^2$ -Norm auf  $P(\hat{T})$

$$\|\hat{\nabla} \hat{v}_h\|_{L^\infty(\hat{T})} \leq \|\hat{v}_h\|_{W^{1,\infty}(\hat{T})} \leq c \|\hat{v}_h\|_{\hat{T}}.$$

Nach Transformation folgt

$$\|\nabla v_h\|_{L^\infty(T)} \leq ch^{-1} \|\hat{\nabla} \hat{v}_h\|_{L^\infty(\hat{T})} \leq ch^{-1} \|\hat{v}_h\|_{\hat{T}} \leq ch^{-2} \|v_h\|_T.$$

iv) Ausnutzen der Äquivalenz von  $L^1$ - und  $L^2$ -Norm auf  $P(\hat{T})$  ergibt mit den Transformationsargumenten

$$\|v_h\|_{L^2(T)} \leq ch \|\hat{v}_h\|_{L^2(\hat{T})} \leq ch \|\hat{v}_h\|_{L^1(\hat{T})} \leq ch^{-1} \|\hat{v}_h\|_{L^1(T)}.$$

**Lösung A.3.11:** i)  $T$  kartesisches Einheitsdreieck:

$$P(T) = P_3(T), \quad p(a_i), \nabla p(a_i), p(z), \quad P(T) = P_3(T), \quad p(a_i), p(b_{ij}), p(z).$$

Sei  $p \in P(T)$ , welches bzgl. aller Knotenfunktionale verschwindet. Dann ist  $p|_{\Gamma_i} \equiv 0$ . Auf dem Einheitsdreieck ist folglich  $p(x, y) = cxy(1 - x - y)$ , da jedes entlang von  $\partial T$  verschwindende Polynom die drei Faktoren  $x$ ,  $y$  und  $1 - x - y$  enthalten muß. Wegen  $p(z) = 0$  folgt notwendig  $c = 0$ , d. h.  $p \equiv 0$ .

ii) Ein Polynom  $p \in P_5(T)$  mit  $p(a_i) = 0$ ,  $\nabla p(a_i) = 0$ ,  $\nabla^2 p(a_i) = 0$ ,  $\partial_n p(m_i) = 0$  hat auf dem Einheitsdreieck notwenig die Gestalt  $p(x, y) = xy(1 - x - y)q(x, y)$  mit einem  $q \in P_2(T)$ . Wegen  $\nabla^2 p(0, 0) = 0$  gilt:

$$\begin{aligned} 0 &= \partial_x \partial_y p(0, 0) \\ &= (\partial_x \partial_y (xy(1 - x - y))q + \partial_y (xy(1 - x - y))\partial_x q \\ &\quad + \partial_x (xy(1 - x - y))\partial_y q + xy(1 - x - y)\partial_x \partial_y q)(0, 0) = q(0, 0). \end{aligned}$$

Dies impliziert  $q(0,0) = 0$ . Analog folgt  $q(1,0) = \partial_y \partial_y p(1,0) = 0$  und  $q(0,1) = \partial_x \partial_x p(0,1) = 0$ . Weiter gilt

$$\begin{aligned} 0 &= \partial_n p(m_1) = -\partial_y(xy(1-x-y)q)(\frac{1}{2}, 0) \\ &= (-x(1-x-y)q + xyq - xy(1-x-y)\partial_y q)(\frac{1}{2}, 0) = -\frac{1}{4}q(m_1) \end{aligned}$$

und analog  $q(m_2) = q(m_3) = 0$ . Wegen der Unisolvenz von  $P_2(T)$  mit dem Satz  $\{q(a_i), q(m_i), i = 1, 2, 3\}$  von Knotenwerten folgt schließlich  $q \equiv 0$ . bzw.  $p \equiv 0$ .

iii)  $T$  kartesisches Einheitsquadrat:

$$P(T) = \tilde{Q}_1(T) := P_1(T) \oplus \text{span}\{x^2 - y^2\}, \quad p(m_i), \quad i = 1, \dots, 4.$$

Sei  $p \in P(T)$  mit  $p(m_i) = 0$ . Auf dem Einheitsquadrat ist  $p$  linear entlang der schrägen Verbindungslien zwischen den Mitten benachbarter Kanten und folglich gleich Null entlang dieser Linien. Damit ist auch  $\nabla p(m_i) = 0$ . Entlang jeder Linie durch eine Seitenmitte  $m_i$  verschwindet damit  $p$  in mindestens zwei Punkten und seine Richtungsableitung in mindestens einem Punkt. Da  $p$  entlang jeder Linie höchstens quadratisch ist, folgt  $p \equiv 0$ .

$$P(T) = \tilde{Q}_3(T) := P_3(T) \oplus \text{span}\{x^3y, xy^3\}, \quad p(a_i), \quad \nabla p(a_i), \quad i = 1, \dots, 4.$$

Sei  $p \in P(T)$ , welches bzgl. aller Knotenfunktionale verschwindet. Dann ist  $p|_{\partial T} \equiv 0$ . Folglich müßte  $p$  den Faktor  $cxy(1-x)(1-y)$  enthalten mit einer Konstante  $c \in \mathbb{R}$ . Da der Term  $x^2y^2$  aber nicht durch Elemente des Raumes  $P(T)$  erzeugt werden kann, muß  $c = 0$  sein. Dies bedeutet auch  $p \equiv 0$ .

**Lösung A.3.12:** a) Sei  $\mathbf{T}_h$  eine Triangulierung von  $\Omega$  durch Dreiecke. Dann wählen wir als Ansatz- und Testraum

$$V_h^{(1)} := \{\varphi \in C^0 : \varphi|_T \in P^1(T) \quad \forall T \in \mathbf{T}_h\}.$$

Die zugehörige Variationsgleichung lautet dann: Finde ein  $u_h \in V_h^{(1)}$ , so dass

$$(\nabla u_h, \nabla \varphi_h) + (u_h, \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h^{(1)}.$$

b) Zunächst gilt für die  $H^1$ -Norm und eine beliebige Funktion  $\varphi_h \in V_h^{(1)}$  aufgrund der Galerkin-Orthogonalität

$$\begin{aligned} \|u - u_h\|_1^2 &= (\nabla(u - u_h), \nabla(u - u_h)) + (u - u_h, u - u_h) \\ &= (\nabla(u - u_h), \nabla(u - \varphi_h)) + (u - u_h, u - \varphi_h) \\ &\leq \|\nabla(u - u_h)\| \|\nabla(u - \varphi_h)\| + \|u - u_h\| \|u - \varphi_h\| \\ &\leq \frac{1}{2} (\|\nabla(u - u_h)\|^2 + \|\nabla(u - \varphi_h)\|^2 + \|u - u_h\|^2 + \|u - \varphi_h\|^2) \\ &= \frac{1}{2} (\|u - u_h\|_1^2 + \|u - \varphi_h\|_1^2) \end{aligned}$$

und somit

$$\|u - u_h\|_1 \leq \inf_{\varphi_h \in V_h^{(1)}} \|u - \varphi_h\|_1.$$

Unter Verwendung der Standard-Approximationsabschätzungen folgt also

$$\|u - u_h\|_1 \leq ch\|\nabla^2 u\|$$

Für die  $L^2$ -Norm betrachten wir das duale Problem

$$-\Delta z + z = \frac{e}{\|e\|} \quad \text{in } \Omega, \quad \partial_n z = 0 \quad \text{auf } \partial\Omega,$$

wobei  $e = u - u_h$  ist. Es gilt hierfür die a priori Abschätzung  $\|\nabla^2 z\| \leq c$  und es folgt

$$\begin{aligned} \|e\| &= (\nabla e, \nabla(z - I_h z)) + (e, z - I_h z) \\ &\leq \|e\|_1 \|z - I_h z\|_1 \leq ch^2 \|\nabla^2 u\| \|\nabla^2 z\| \leq ch^2 \|\nabla^2 u\|. \end{aligned}$$

c) Wir erhalten einen konformen FE-Ansatzraum, indem wir zunächst für unsere Triangulierung  $T_h$  fordern, dass alle Rand/Eckpunkte der Triangulierung auf dem Rand von  $\Omega$  liegen. Dann definieren wir unseren Ansatzraum wie in (a), wobei wir im Falle das  $\partial\Omega$  ausserhalb von  $\bar{\Omega}_h$  liegt, die Testfunktionen linear bis zum Rand von  $\Omega$  fortsetzen. Im umgekehrten Fall, ist unsere Testfunktion lediglich auf  $\Omega \cap \Omega_h$  zu definieren. Im Falle nichthomogener Neumann-Bedingungen lautet die variationelle Formulierung

$$(\nabla u_h, \nabla \varphi_h) + (u_h, \varphi_h) = (f, \varphi_h) + (g, \varphi_h)_{\partial\Omega} \quad \forall \varphi_h \in V_h^{(1)}.$$

**Lösung A.3.13:** a) Als Ansatzraum verwenden wir

$$V_h^{(3)} := \{v_h \in H_0^1(\Omega_h) \mid v_{|T} \in P_3(T), T \in T_h\}$$

wobei die Dreiecke entlang des Randes  $\partial\Omega$  kubische Randkurven haben. Es sind die folgenden Fehlerabschätzungen zu erwarten:

$$\begin{aligned} \|\nabla(u - u_h)\|_{\Omega_h} &\leq ch^3 \|u\|_4, \\ \|u - u_h\|_{\Omega_h} &\leq ch^4 \|u\|_4, \end{aligned}$$

vorausgesetzt die Lösung ist  $u \in H^4(\Omega)$ .

b) Für den Fehler im Mittelwert ergibt sich mit Hilfe eines Dualitätsarguments:

$$\left| \int_{\Omega} u \, dx - \int_{\Omega} u_h \, dx \right| \leq ch^5 \|u\|_4.$$

**Lösung A.3.14:** a) Für das Referenzelement  $\tilde{T}$  definieren wir das Funktional  $F : H^2(\tilde{T}) \rightarrow P_1(\tilde{T})$  durch

$$F(v) := \|v - I_h v\|_{\partial\tilde{T}}.$$

Für dieses gilt dann aufgrund des Sparsatzes und der Interpolationsabschätzung bzgl. der  $L^2$ - und der  $H^1$ -Norm:

$$\begin{aligned}|F(v)| &= \|v - I_h v\|_{\partial \tilde{T}} \leq c \|v - I_h v\|_{2; \tilde{T}} \leq c \|v\|_{2; \tilde{T}}, \\ |F(v+w)| &\leq \|v + w - I_h(v+w)\|_{\partial \tilde{T}} \leq \|v - I_h v\|_{\partial \tilde{T}} + \|w - I_h w\|_{\partial \tilde{T}} = |F(v)| + |F(w)|, \\ F(q) &= 0, \quad q \in P_1(\tilde{T}).\end{aligned}$$

Nach dem Lemma von Bramble/Hilbert folgt die Abschätzung

$$|F(v)| \leq c \|\nabla^2 v\|_{\tilde{T}}$$

Das Transformationsargument aus dem Text ergibt dann

$$h_T^{-1/2} \|v - I_h v\|_{\partial T} \leq ch_T^2 h_T^{-1} \|\nabla^2 v\|_T,$$

woraus die behauptete Abschätzung folgt. Alternativ kann man auch wie folgt argumentieren: Mit Hilfe des Spursatzes auf der Zelle  $T$  mit Durchmesser  $h_T$  ergibt sich zunächst die Abschätzung

$$\|v - I_h v\|_{\partial T} \leq ch_T^{-1/2} \|v - I_h v\|_T + ch_T^{1/2} \|\nabla(v - I_h v)\|_T$$

und dann mit Hilfe der schon bekannten Fehlerabschätzungen über  $T$ :

$$\|v - I_h v\|_{\partial T} \leq ch_T^{-1/2} h_T^2 \|\nabla^2 v\|_T + ch_T^{1/2} h_T \|\nabla^2 v\|_T = ch_T^{3/2} \|\nabla^2 v\|_T.$$

b) Die Abschätzung

$$\|v - I_h v\|_{\partial T} \leq ch_T^{1/2} \|\nabla v\|_T$$

kann nicht gelten, da für Funktionen  $v \in H^1(\Omega)$  im Allg. die Knoteninterpolierende  $I_h u \in P_2(T)$  gar nicht definiert ist. Sie wäre aber gültig mit der „Quasi-Interpolierenden“  $\tilde{I}_h u$  (diskutiert in einer späteren Aufgabe) in der Form

$$\|v - I_h v\|_{\partial T} \leq ch_T^{1/2} \|\nabla v\|_{\tilde{T}}, \quad \tilde{T} = \cup\{\tau \in \mathbb{T}_h, \tau \cap T \neq \emptyset\}.$$

Die direkte Anwendung des Bramble/Hilbert-Lemmas ist hier aber nicht möglich, da die dort definierte Quasi-Interpolierende i. Allg. Polynome nicht reproduziert. Der Beweis dieser Abschätzung bedarf also noch einiger Arbeit.

**Lösung A.3.15:** Für die Lagrange-Interpolation in  $P(T) := P_2(T)$  gelten die Fehlerabschätzungen

- (i)  $\|\nabla^2(v - I_T v)\|_T \leq c_i h_T \|\nabla^3 v\|_T;$
- (ii)  $|(v - I_T v)(a)| \leq c_i h_T^2 \|\nabla^3 v\|_T;$
- (iii)  $\|\partial_n(v - I_T v)\|_{\partial T} \leq c_i h^{3/2} \|\nabla^3 v\|_T;$
- (iv)  $\|v - I_T v\|_T \leq c_i h_T^2 \|\nabla^2 v\|_T.$

**Lösung A.3.16:** a) Auf dem konvexen Polygonegebiet  $\Omega$  ist jede Lösung  $v \in H_0^1(\Omega)$  der Variationsgleichung

$$(\nabla v, \nabla \varphi) = (f, \varphi) \quad \forall \varphi \in H_0^1(\Omega),$$

auch in  $V = H^2(\Omega)$ , und es gilt die a priori Abschätzung

$$\|v\|_{H^2} \leq c_\Omega \|f\| = c_\Omega \|\Delta v\|.$$

Auf dem Teilraum  $H_0^2(\Omega) \subset H_0^1(\Omega)$  gilt also

$$a(v, v) := \|\Delta v\|^2 \geq c_\Omega^{-2} \|v\|_{H^2}^2,$$

d. h.: Die Bilinearform  $a(\cdot, \cdot)$  ist  $V$ -elliptisch. Folglich hat die Variationsgleichung

$$a(u, \varphi) = (f, \varphi) \quad \forall \varphi \in V,$$

für jedes  $f \in L^2(\Omega)$  eine eindeutige Lösung  $u \in V$ . Diese ist dann die sog. „schwache“ Lösung der biharmonischen Gleichung. Auf einem Rechteckgebiet ist  $u \in H^4(\Omega)$  und genügt der a priori Abschätzung

$$\|u\|_{H^4} \leq c_s \|\Delta^2 u\| = c_s \|f\|.$$

b) Sei  $V_h^{(5)} \subset V$  der Finite-Elemente-Raum basierend auf dem quintischen Argyris-Element. Die durch die Galerkin-Gleichungen

$$a(u_h, \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h^{(5)},$$

definierte Ritz-Approximation  $u_h \in V_h^{(5)}$  besitzt die Bestapproximationseigenschaft

$$\|\Delta(u - u_h)\| = \min_{\varphi_h \in V_h^{(5)}} \|\Delta(u - \varphi_h)\|.$$

Aufgrund der Sobolewschen Einbettung  $H^4(\Omega) \subset C^2(\bar{\Omega})$  ist die Knoteninterpolierende  $I_h u \in V_h^{(5)}$  von  $u$  wohl definiert, und es gilt (Bramble/Hilbert-Lemma):

$$\|\Delta(u - I_h u)\| \leq ch^2 \|u\|_{H^4}.$$

Dies impliziert die Energie-Fehlerabschätzung

$$\|u - u_h\|_{H^2} \leq c \|\Delta(u - u_h)\| \leq ch^2 \|u\|_{H^4}.$$

Zur Abschätzung des Fehlers bzgl. der  $L^2$ -Norm verwenden wir ein Dualitätsargument. Sei  $z \in V$  die schwache Lösung der Variationsgleichung

$$(\Delta\varphi, \Delta z) = (\varphi, u - u_h) \|u - u_h\|^{-1} \quad \forall \varphi \in V.$$

Dann ist  $z \in H^4(\Omega)$ , und es gilt  $\|z\|_{H^4} \leq c$ . Damit folgt mit Hilfe der Galerkin-Orthogonalität und der Interpolationsabschätzungen

$$\begin{aligned} \|u - u_h\| &= a(u - u_h, z) = a(u - u_h, z - I_h z) \leq \|\Delta(u - u_h)\| \|\Delta(z - I_h z)\| \\ &\leq ch^2 \|u\|_{H^4} h^2 \|z\|_{H^4} \leq ch^4 \|u\|_{H^4}. \end{aligned}$$

c) Die Spektral-Konditionen der zugehörigen Steifigkeitsmatrix  $A_h = (a(\varphi_h^j, \varphi_h^i))_{i,j}$  und Massematrix  $M_h = ((\varphi_h^j, \varphi_h^i))_{ij}$  sind gegeben durch

$$\kappa(A_h) = \frac{\lambda_{\max}(A_h)}{\lambda_{\min}(A_h)}, \quad \kappa(M_h) = \frac{\lambda_{\max}(M_h)}{\lambda_{\min}(M_h)} \leq c.$$

Mit Hilfe der inversen Beziehung folgt

$$\begin{aligned} \lambda_{\max}(A_h) &= \max_{x \in \mathbb{R}^N} \frac{(A_h x, x)}{|x|^2} \leq \max_{x \in \mathbb{R}^N} \frac{(A_h x, x)}{(M_h x, x)} \max_{x \in \mathbb{R}^N} \frac{(M_h x, x)}{|x|^2} \\ &= \max_{v_h \in V_h^{(5)}} \frac{a(v_h, v_h)}{\|v_h\|^2} \lambda_{\max}(M_h) \leq ch^{-4} \lambda_{\max}(M_h). \end{aligned}$$

Ferner

$$\begin{aligned} \lambda_{\min}(A_h) &= \min_{x \in \mathbb{R}^N} \frac{(A_h x, x)}{|x|^2} \geq \min_{x \in \mathbb{R}^N} \frac{(A_h x, x)}{(M_h x, x)} \min_{x \in \mathbb{R}^N} \frac{(M_h x, x)}{|x|^2} \\ &= \min_{v_h \in V_h^{(5)}} \frac{a(v_h, v_h)}{\|v_h\|^2} \lambda_{\min}(M_h) \geq \min_{v \in V} \frac{a(v, v)}{\|v\|^2} \lambda_{\min}(M_h) = \lambda_{\min}(\Delta^2) \lambda_{\min}(M_h). \end{aligned}$$

Dies impliziert

$$\kappa(A_h) \leq ch^{-4} \lambda_{\min}(\Delta^2)^{-1} \frac{\lambda_{\max}(M_h)}{\lambda_{\min}(M_h)} \leq ch^{-4} \lambda_{\min}(\Delta^2)^{-1} \kappa(M_h).$$

Da die Kondition der Massematrix  $\kappa(M_h) = \mathcal{O}(1)$  ist, folgt  $\kappa(A_h) = \mathcal{O}(h^{-4})$ .

**Lösung A.3.17:** a) Konvergenz in der Energie-Norm ist garantiert, für Quadraturordnung  $r \geq m - 2$ , mit der Ordnung  $m$  des FE-Ansatzes (Polynomgrad  $m - 1$ ) ist. Die optimale Konvergenzordnung wird erreicht für  $r \geq 2m - 3$ . Im Fall „kubische“ Elemente, d. h.  $m = 4$ , bedeutet dies  $r \geq 2$  für Konvergenz und  $r \geq 5$  für optimale Konvergenz. Geeignete, möglichst ökonomische Quadraturformeln wären für Konvergenz z. B. die Mittelpunktsregel und für optimale Konvergenz eine „Quasi“-Gauß-Formel 5-ter Ordnung.

b) Die 1. RWA des Laplace-Operators sei auf dem Einheitsquadrat mit bilinearen finiten Elementen auf einem äquidistanten, kartesischen Gitter diskretisiert. Die Elemente  $(\nabla \varphi_h^j, \nabla \varphi_h^i)$  der Systemmatrix werden mit der Mittelpunktregel berechnet:

Die Koordinaten  $(a_i, b_j)$  beschreiben einen Gitterpunkt auf einem äquidistanten kartesischen Gitter mit der Gitterweite  $h$ . Die quadratischen Zellen um  $(a_i, b_j)$  sollen der Einfachheit halber mit  $lo$ ,  $ro$ ,  $lu$  und  $ru$  für links oben etc. bezeichnet werden. Dann ergibt sich für die Basisfunktionen die Darstellung

$$\varphi_{i,j} := \begin{cases} -\frac{1}{h^2}(x - (a_i - h))(y - (b_j + h)) & (x, y) \in lo \\ \frac{1}{h^2}(x - (a_i + h))(y - (b_j + h)) & (x, y) \in ro \\ \frac{1}{h^2}(x - (a_i - h))(y - (b_j - h)) & (x, y) \in lu \\ -\frac{1}{h^2}(x - (a_i + h))(y - (b_j - h)) & (x, y) \in ru \end{cases}$$

mit dem Gradienten

$$\nabla \varphi_{i,j} = \begin{cases} \frac{1}{h^2} \begin{pmatrix} -y + b_j + h \\ -x + a_i - h \end{pmatrix} & (x, y) \in lo \\ \frac{1}{h^2} \begin{pmatrix} y - b_j - h \\ x - a_i - h \end{pmatrix} & (x, y) \in ro \\ \frac{1}{h^2} \begin{pmatrix} y - b_j + h \\ x - a_i + h \end{pmatrix} & (x, y) \in lu \\ \frac{1}{h^2} \begin{pmatrix} -y + b_j - h \\ -x + a_i + h \end{pmatrix} & (x, y) \in ru \end{cases}$$

Damit berechnen sich die Skalarprodukte mit Hilfe der Mittelpunktsregel zu

$$\begin{aligned} (\nabla \varphi_{i,j}, \nabla \varphi_{i,j}) &= \frac{4}{h^4} \int_{a_i-h}^{a_i} \int_{b_j}^{b_j+h} \begin{pmatrix} -y + b_j + h \\ -x + a_i - h \end{pmatrix}^2 dx dy \\ &\approx \frac{4}{h^2} \left( \frac{h^2}{4} + \frac{h^2}{4} \right) = 2 \end{aligned}$$

$$\begin{aligned} (\nabla \varphi_{i,j}, \nabla \varphi_{i+1,j}) &= \frac{2}{h^4} \int_{a_i}^{a_{i+1}} \int_{b_j}^{b_j+h} \begin{pmatrix} y - b_j - h \\ x - a_i - h \end{pmatrix} \begin{pmatrix} -y + b_{j+1} + h \\ -x + a_{i+1} - h \end{pmatrix} dx dy \\ &\approx \frac{2}{h^2} \left( -\frac{h^2}{4} + \frac{h^2}{4} \right) = 0 \end{aligned}$$

$$\begin{aligned} (\nabla \varphi_{i,j}, \nabla \varphi_{i+1,j+1}) &= \frac{1}{h^4} \int_{a_i}^{a_{i+1}} \int_{b_j}^{b_{j+1}} \begin{pmatrix} y - b_j - h \\ x - a_i - h \end{pmatrix} \begin{pmatrix} y - b_{j+1} + h \\ x - a_{i+1} + h \end{pmatrix} dx dy \\ &\approx \frac{1}{h^2} \left( -\frac{h^2}{2} - \frac{h^2}{2} \right) = -\frac{1}{2} \end{aligned}$$

Die Steifigkeitsmatrix hat damit die Blockform

$$\begin{pmatrix} D & N & 0 & 0 & \dots \\ N & D & N & 0 & \dots \\ 0 & N & D & N & \dots \\ \vdots & & \ddots & \ddots & \ddots \end{pmatrix},$$

mit den Untermatrizen

$$D = \begin{pmatrix} 2 & 0 & 0 & \dots \\ 0 & 2 & 0 & \dots \\ 0 & 0 & 2 & \dots \\ \vdots & & \ddots & \ddots \end{pmatrix}, \quad N = \begin{pmatrix} 0 & -1/2 & 0 & 0 & \dots \\ -1/2 & 0 & -1/2 & 0 & \dots \\ 0 & -1/2 & 0 & -1/2 & \dots \\ \vdots & & \ddots & \ddots & \ddots \end{pmatrix}$$

Dieses Verfahren entspricht dem 5-Punkte-Differenzenoperator, wobei hier die Richtungen zur Bestimmung des Differenzenquotienten diagonal zu den kartesischen Achsen verlaufen. Es handelt sich aber trotzdem um eine konsistente (Ordnung 2) Approximation des Laplace-Operators, da dieser invariant gegenüber Rotation des Koordinatensystems ist.

**Lösung A.3.18:** i) Eine Triangulierung  $T_h$  heißt „quasi-gleichförmig“ falls sie formregulär,

$$\sup_{h>0} \left( \max_{T \in T_h} \frac{h_T}{\rho_T} \right) \leq c,$$

und „größenregulär“,

$$\sup_{h>0} \left( \frac{\max h_T}{\min h_T} \right) \leq c,$$

ist.

ii) Die negative  $h$ -Potenz bei der Konditionsabschätzung stammt von der Benutzung einer inversen Beziehung zur Abschätzung von  $a(v_h, v_h)$ , diese und alle weiteren Schritte sind unabhängig von der konkreten Diskretisierung  $V_h \subset V$ , wie man sich beim Durchgehen des Beweises überzeugt:

$$\begin{aligned} \lambda_{\min}(A) &\geq \min_{\xi \in \mathbb{R}} \frac{\langle A\xi, \xi \rangle}{\langle M\xi, \xi \rangle} \min_{\xi \in \mathbb{R}} \frac{\langle M\xi, \xi \rangle}{\langle \xi, \xi \rangle} = \min_{v_h \in V_h} \frac{a(v_h, v_h)}{\|v_h\|^2} \lambda_{\min}(M) \\ &\geq \min_{v \in V} \frac{a(v, v)}{\|v\|^2} \lambda_{\min}(M) = \lambda_{\min}(-\Delta) \lambda_{\min}(M) \end{aligned}$$

und

$$\lambda_{\max}(A) \leq \max_{\xi \in \mathbb{R}} \frac{\langle A\xi, \xi \rangle}{\langle M\xi, \xi \rangle} \max_{\xi \in \mathbb{R}} \frac{\langle M\xi, \xi \rangle}{\langle \xi, \xi \rangle} = \max_{v_h \in V_h} \frac{a(v_h, v_h)}{\|v_h\|^2} \lambda_{\max}(M).$$

Mit der inversen Beziehung folgt:

$$a(v_h, v_h) = \sum_T \|\nabla v_h\|_T^2 \leq c \sum_T \rho_T^{-2} \|v_h\|_T^2 \leq c \left( \max_T \rho_T^{-2} \right) \|v_h\|^2.$$

Es gilt weiterhin  $\text{cond}_2(M) = O(1)$  und somit folgt zusammen mit der Formregularität:

$$\text{cond}_2(A) \leq c \max_T \rho_T^{-2} \text{cond}_2(M) = O\left(\max_T \rho_T^{-2}\right) = O(h^{-2}).$$

iii) Für den Fall, dass die Anzahl der an einer Ecke zusammenstoßenden Zellen beschränkt bleibt, gilt auch ohne Formregularität weiterhin  $\text{cond}_2(M) = O(1)$ . Die Argumentation in (ii) benötigte nur im letzten Schritt Formregularität, so dass folgt:

$$\text{cond}_2(A_h) = O(\rho^{-2}),$$

wobei  $\rho$  der minimale Innenkreisdurchmesser ist.

**Lösung A.3.19:** Es ist

$$\begin{aligned}\lambda_{\min}(A) &\geq \min_{\xi \in \mathbb{R}} \frac{\langle A\xi, \xi \rangle}{\langle M\xi, \xi \rangle} \min_{\xi \in \mathbb{R}} \frac{\langle M\xi, \xi \rangle}{\langle \xi, \xi \rangle} = \min_{v_h \in V_h} \frac{a(v_h, v_h)}{\|v_h\|^2} \lambda_{\min}(M) \\ &\geq \min_{v \in V} \frac{a(v, v)}{\|v\|^2} \lambda_{\min}(M) = \lambda_{\min}(-\Delta) \lambda_{\min}(M)\end{aligned}$$

und

$$\lambda_{\max}(A) \leq \max_{\xi \in \mathbb{R}} \frac{\langle A\xi, \xi \rangle}{\langle M\xi, \xi \rangle} \max_{\xi \in \mathbb{R}} \frac{\langle M\xi, \xi \rangle}{\langle \xi, \xi \rangle} = \max_{v_h \in V_h} \frac{a(v_h, v_h)}{\|v_h\|^2} \lambda_{\max}(M).$$

Mit der inversen Beziehung folgt:

$$a(v_h, v_h) = \sum_T \|\nabla v_h\|_T^2 \leq c \sum_T \rho_T^{-2} \|v_h\|_T^2 \leq c \max_T \rho_T^{-1} \|v_h\|^2.$$

Aufgrund der Formregulatitität ist  $\text{cond}_2(M) = O(1)$  und es folgt insgesamt

$$\text{cond}_2(A) \leq c \max_T \rho_T^{-2} \text{cond}_2(M) = O\left(\max_T \rho_T^{-2}\right).$$

**Lösung A.3.20:** a) Die variationelle Formulierung lautet:

$$\int_{\Omega} \alpha \nabla u \cdot \nabla \varphi \, dx + \int_{\Omega} \gamma u \varphi \, dx = \int_{\Omega} f \varphi \, dx + \int_{\partial\Omega} g \varphi \, do_x$$

b) Analog zum Vorgehen in der Vorlesung formuliert man das duale Problem: Finde  $z \in H^1$ , so dass

$$\int_{\Omega} \alpha \nabla \varphi \cdot \nabla z \, dx + \int_{\Omega} \gamma \varphi z \, dx = \frac{(\nabla e_h, \nabla \varphi)}{\|e_h\|_E}.$$

Setzen von  $\varphi = e_h$  liefert die Fehleridentität

$$\|e_h\|_E = \int_{\Omega} f(z + \psi_h) \, dx + \int_{\partial\Omega} g(z + \psi_h) \, dx - (\alpha \nabla U, \nabla z + \nabla \psi_h) - (\gamma U, z + \psi_h)$$

für beliebiges  $\psi_h \in V_h$ . Zellweises partielles Integrieren führt auf die Ungleichung

$$\|e_h\|_E \leq \sum_{T \in \mathcal{T}_h} \left\{ \|f + \Delta U - \gamma U\|_T \|z + \psi_h\|_T + \frac{1}{2} \|[\partial_n U]\|_{\partial T} \|z + \psi_h\|_{\partial T} \right\},$$

hierbei sei  $[\partial_n U]|_{\partial T \cap \partial\Omega} = 2(g - \partial_n U)$ , ansonsten der übliche Sprung über eine innere Kante. Man wählt nun wieder  $\psi_h = I_h z$  und folgert mit der Clément-Interpolationsabschätzung,

$$\|z - I_h z\|_T + h_T^{1/2} \|z - I_h z\|_{\partial T} \leq ch_T \|\nabla z\|_{\tilde{T}}$$

die Abschätzung:

$$\begin{aligned}
\|e_h\|_E &\leq \sum_{T \in \mathbb{T}_h} ch_T \left\{ \|f + \Delta U - \gamma U\|_T + \frac{1}{2} h_T^{-1/2} \|[\partial_n U]\|_{\partial T} \right\} \|\nabla z\|_{\bar{T}} \\
&\leq \left( \sum_{T \in \mathbb{T}_h} ch_T^2 \left\{ \|f + \Delta U - \gamma U\|_T^2 + \frac{1}{4} h_T^{-1} \|[\partial_n U]\|_{\partial T}^2 \right\} \right)^{1/2} \left( \sum_{T \in \mathbb{T}_h} \|\nabla z\|_z^2 \right)^{1/2} \\
&\leq \left( \sum_{T \in \mathbb{T}_h} ch_T^2 \left\{ \|f + \Delta U - \gamma U\|_T^2 + \frac{1}{4} h_T^{-1} \|[\partial_n U]\|_{\partial T}^2 \right\} \right)^{1/2}.
\end{aligned}$$

since  $\|\nabla z\|^2 \leq c$ .

c) Die Argumentation verläuft analog zu (b) mit dem modifizierten dualen Problem

$$\int_{\Omega} \alpha \nabla \varphi \cdot \nabla z \, dx + \int_{\Omega} \gamma \varphi z \, dx = \frac{(e_h, \varphi)}{\|e_h\|_E}.$$

Im Gegensatz zum dualen Problem in (b) ist hierbei aber die rechte Seite aus  $L^2$ , so dass  $z \in H^2(\Omega)$  erwartet werden kann. Dies ermöglicht daher die Nutzung der herkömmlichen Knoteninterpolierenden für  $\psi_h$  mit der Interpolationsfehlerabschätzung

$$\|z - I_h z\|_T + h_T^{1/2} \|z - I_h z\|_{\partial T} \leq ch_T^2 \|\nabla^2 z\|_T.$$

mit  $\|\nabla^2 z\| \leq c$ . Dies liefert

$$\|e_h\| \leq \left( \sum_{T \in \mathbb{T}_h} ch_T^4 \left\{ \|f + \Delta U - \gamma U\|_T^2 + \frac{1}{4} h_T^{-1} \|[\partial_n U]\|_{\partial T}^2 \right\} \right)^{1/2}.$$

**Lösung A.3.21:** Wir rekapitulieren die Spurabschätzung

$$\|\partial_n v\|_{\partial T} \leq c \{ h_T^{1/2} \|\Delta v\|_T + h_T^{-3/2} \|v\|_T \}.$$

Wegen  $[\partial_n u] = 0$  ergibt sich

$$\begin{aligned}
\|[\partial_n u_h]\|_{\partial T \setminus \partial \Omega}^2 &= \|[\partial_n e_h]\|_{\partial T \setminus \partial \Omega}^2 \leq c \sum_{T \subset \bar{T}} \|\partial_n e_h\|_{\partial T \setminus \partial \Omega}^2 \\
&\leq c \sum_{T \subset \bar{T}} \{ h_T \|\Delta e_h\|_T^2 + h_T^{-3} \|e_h\|_T^2 \} \\
&\leq c \sum_{T \subset \bar{T}} \{ h_T \|f + \Delta u_h\|_T^2 + h_T^{-3} \|e_h\|_T^2 \}
\end{aligned}$$

und damit

$$\begin{aligned}
\eta_{L^2}(u_h) &= \left( \sum_{T \in \mathbb{T}_h} h_T^4 \{ \|f + \Delta u_h\|_T^2 + \frac{1}{2} h_T^{-1} \|[\partial_n u_h]\|_{\partial T \setminus \partial \Omega}^2 \} \right)^{1/2} \\
&\leq c \|e_h\|_2 + \left( \sum_{T \in \mathbb{T}_h} h_T^4 \|f + \Delta u_h\|_T^2 \right)^{1/2}.
\end{aligned}$$

**Lösung A.3.22:** Für die Nebenbedingung  $\eta(h) \leq \text{TOL}$  wird im Optimum Gleichheit angenommen. Wir suchen also stationäre Punkte des Lagrangefunktionalen

$$\mathcal{L}(h, \lambda) = N(h) + \lambda(\eta(h) - \text{TOL}).$$

dies sind gerade solche Punkte in denen

$$\frac{d}{dt} \mathcal{L}(h + t\varphi, \lambda + t\mu)|_{t=0} = 0$$

für alle  $\varphi \in C(\bar{\Omega})$  und alle  $\mu \in \mathbb{R}$ . Indem wir zunächst  $\mu = 0$  und dann  $\varphi = 0$  wählen erhalten wir die beiden Variationsgleichungen

$$2 \int_{\Omega} \varphi (-h^{-3} + h\lambda A) dx = 0 \quad \forall \varphi \in C(\bar{\Omega})$$

und

$$\mu \left( \int_{\Omega} h^2 A dx - \text{TOL} \right) = 0 \quad \forall \mu \in \mathbb{R}.$$

Aus der ersten folgt

$$h = (\lambda A)^{-1/4}$$

durch dieser obiger Beziehung in die zweite Gleichung folgt

$$\lambda^{1/2} \text{TOL} = \int_{\Omega} A^{1/2} dx =: W.$$

Somit ist

$$\lambda = \left( \frac{W}{\text{TOL}} \right)^2$$

und es folgt:

$$h(x) = \left( \frac{\text{TOL}}{W} \right)^{1/2} A(x)^{-1/4}.$$

**Lösung A.3.23:** Mit  $u \in W^{2,\infty}(\Omega)$  liegt  $f$  in  $L^\infty(\Omega)$ . Da  $u_h$  stückweise linear ist, folgt:

$$\|f + \Delta u_h\|_T = \|f\|_T \leq \|f\|_\infty h_T.$$

Wir betrachten nun den zweiten Summanden:

$$\begin{aligned} \|[\partial_n u_h]\|_{\partial T} &= \|[\partial_n e]\|_{\partial T} \leq \|\partial_n e\|_{\partial T} + \|\partial_n \tilde{e}\|_{\partial T} \\ &\leq \|\nabla e\|_{\partial T} + \|\nabla \tilde{e}\|_{\partial T} \\ &\leq ch^{1/2} \left( \|\nabla e\|_{\infty, T} + \|\nabla^2 \tilde{e}\|_{\infty, T} \right) \\ &\leq ch^{3/2} \|\nabla^2 e\|_\infty = ch^{3/2} \|\nabla^2 u\|_\infty \end{aligned}$$

wobei  $\tilde{e}$  der Fehler auf den an  $\partial T$  angrenzenden Zellen ist. Insgesamt folgt somit die Behauptung.

**Lösung A.3.24:** Für die Lagrange-Interpolation in  $P(T) := P_2(T)$  gelten die Fehlerabschätzungen

$$\begin{aligned} (i) \quad & \|\nabla^2(v - I_T v)\|_T \leq c_i h_T \|\nabla^3 v\|_T; \\ (ii) \quad & |(v - I_T v)(a)| \leq c_i h_T^2 \|\nabla^3 v\|_T; \\ (iii) \quad & \|\partial_n(v - I_T v)\|_{\partial T} \leq c_i h^{3/2} \|\nabla^3 v\|_T; \\ (iv) \quad & \|v - I_T v\|_T \leq c_i h_T^2 \|\nabla^2 v\|_T. \end{aligned}$$

**Lösung A.3.25:** a) Variationelle Formulierung im Raum  $V = H^1(\Omega)$ : Finde  $u \in V$  mit

$$a(u, \varphi) := (\alpha \nabla u, \nabla \varphi)_\Omega + (\gamma u, \varphi)_\Omega = (f, \varphi)_\Omega + (g, \varphi)_{\partial\Omega}, \quad \forall \varphi \in V.$$

Die Bilinearform  $a(\cdot, \cdot)$  ist offenbar symmetrisch,  $V$ -elliptisch und beschränkt auf  $V$ . Nach dem Darstellungssatz von Riesz (oder dem Lemma von Lax-Milgram) existiert also eine eindeutige Lösung  $u \in V$ . Ist diese hinreichend glatt, folgt durch partielle Integration

$$(-\nabla \cdot (\alpha \nabla u) + \gamma u - f, \varphi)_\Omega + (n \cdot (\alpha \nabla u) - g, \varphi)_{\partial\Omega} = 0, \quad \varphi \in V.$$

Nach dem Fundamentalsatz der Variationsrechnung impliziert dies, dass  $u$  klassische Lösung der RWA ist:

$$-\nabla \cdot (\alpha \nabla u) + \gamma u = f \quad \text{in } \Omega, \quad n \cdot (\alpha \nabla u)|_{\partial\Omega} = g.$$

b) Zur Herleitung der a posteriori Energienorm-Fehlerabschätzung schreiben wir mit Hilfe der Galerkin-Orthogonalität

$$\Sigma_h := (\alpha \nabla e_h, \nabla e_h)_\Omega + (\gamma e_h, e_h)_\Omega = (\alpha \nabla e_h, \nabla(e_h - i_h e_h))_\Omega + (\gamma e_h, e_h - i_h e_h)_\Omega.$$

Zellweise partielle Integration ergibt weiter

$$\begin{aligned} \Sigma_h &= \sum_{T \in \mathbf{T}_h} \left\{ (-\nabla \cdot (\alpha \nabla e_h), e_h - i_h e_h)_T + (\gamma e_h, e_h - i_h e_h)_T + (n \cdot (\alpha \nabla e_h), e_h - i_h e_h)_{\partial T} \right\} \\ &= \sum_{T \in \mathbf{T}_h} \left\{ (R(u_h), e_h - i_h e_h)_T + (r(u_h), e_h - i_h e_h)_{\partial T} \right\} \end{aligned}$$

mit den „Zell- und Kanten-Residuen“ (bzw. „Gleichungs- und Sprung-Residuen“)

$$R(u_h)|_T := f + \nabla \cdot (\alpha \nabla u_h) - \gamma u_h, \quad r_h|_\Gamma := \begin{cases} \frac{1}{2}[n \cdot (\alpha \nabla u_h)], & \text{für } \Gamma \subset \partial T \setminus \partial\Omega, \\ n \cdot (\alpha \nabla u_h) - g, & \text{für } \Gamma \subset \partial\Omega. \end{cases}$$

Mit Hilfe der Hölderschen Ungleichung folgt hieraus

$$|\Sigma_h| \leq \sum_{T \in \mathbf{T}_h} \left\{ \|R(u_h)\|_T \|e_h - i_h e_h\|_T + \|r(u_h)\|_{\partial T} \|e_h - i_h e_h\|_{\partial T} \right\}$$

Wir wählen die Approximation  $i_h e_h$  als verallgemeinerte Knoteninterpolierende mit der

Eigenschaft

$$\|e_h - i_h e_h\|_T + h_T^{1/2} \|e_h - i_h e_h\|_{\partial T} \leq \tilde{c}_i h_T \|\nabla e_h\|_{\tilde{T}},$$

wobei  $\tilde{T} := \cup\{T' \in \mathbb{T}_h \mid T' \text{ hat gemeinsame Kante mit } T\}$ . Damit ergibt sich weiter

$$\begin{aligned} |\Sigma_h| &\leq \tilde{c}_i \sum_{T \in \mathbb{T}_h} \{h_T \|R(u_h)\|_T \|\nabla e_h\|_{\tilde{T}} + h_T^{1/2} \|r(u_h)\|_{\partial T} \|\nabla e_h\|_{\tilde{T}}\} \\ &\leq \tilde{c}_i \left( \sum_{T \in \mathbb{T}_h} h_T^2 \{ \|R(u_h)\|_T^2 + h_T^{-1} \|r(u_h)\|_{\partial T}^2 \} \right)^{1/2} \left( \sum_{T \in \mathbb{T}_h} \|\nabla e_h\|_T^2 \right)^{1/2}. \end{aligned}$$

Wegen

$$\sum_{T \in \mathbb{T}_h} \|\nabla e_h\|_T^2 \leq c \|\nabla e_h\|_\Omega^2 \leq c(\alpha \nabla e_h, \nabla e_h)_\Omega + (\gamma e_h, e_h)_\Omega = c \Sigma_h.$$

ergibt sich die gewünschte a posteriori Fehlerabschätzung:

$$\|e_h\|_E \leq c \tilde{c}_i \left( \sum_{T \in \mathbb{T}_h} h_T^2 \{ \|R(u_h)\|_T^2 + h_T^{-1} \|r(u_h)\|_{\partial T}^2 \} \right)^{1/2}.$$

c) Zur Herleitung der  $L^2$ -Norm-Fehlerabschätzung verwenden wir wieder ein Dualitätsargument mit dem Funktional

$$J(\varphi) := \|e_h\|_\Omega^{-1} (\varphi, e_h)_\Omega, \quad J(e_h) = \|e_h\|_\Omega.$$

Sei  $z \in V \cap H^2(\Omega)$  die Lösung des dualen Problems

$$a(\varphi, z) = J(\varphi) \quad \forall \varphi \in V,$$

bzw.

$$-\nabla \cdot (\alpha \nabla z) + \gamma z = \|e_h\|_\Omega^{-1} e_h \quad \text{in } \Omega, \quad n \cdot (\alpha \nabla u)|_{\partial \Omega} = 0 *$$

Für diese gilt auf dem konvexen Polygonegebiet die a priori Abschätzung  $\|\nabla^2 z\|_\Omega \leq c_z$ . Damit erschließen wir wie zuvor:

$$\begin{aligned} \|e_h\| &= J(e_h) = a(e_h, z) = a(e_h, z - i_h z) \\ &= \sum_{T \in \mathbb{T}_h} \{R(u_h), z - i_h z\}_T + (r(u_h), z - i_h z)_{\partial T} \\ &\leq c_i \sum_{T \in \mathbb{T}_h} \{ \|R(u_h)\|_T h_T^2 \|\nabla^2 z\|_T + \|r(u_h)\|_{\partial T} h_T^{3/2} \|\nabla^2 z\|_T \} \\ &\leq \left( \sum_{T \in \mathbb{T}_h} h_T^4 \{ \|R(u_h)\|_T + h_T^{-1} \|r(u_h)\|_{\partial T}^2 \} \right)^{1/2} \left( \sum_{T \in \mathbb{T}_h} \|\nabla^2 z\|_T^2 \right)^{1/2} \end{aligned}$$

Dies ergibt die gewünschte a posteriori Fehlerabschätzung

$$\|e_h\| \leq \left( \sum_{T \in \mathbb{T}_h} h_T^4 \{ \|R(u_h)\|_T + h_T^{-1} \|r(u_h)\|_{\partial T}^2 \} \right)^{1/2}.$$

**Lösung A.3.26:** Ein Glättungsschritt des Richardson-Verfahren benötigt im Wesentlichen eine Matrix-Vektor-Multiplikation. Die Matrix-Vektor-Multiplikation  $A_h x_h$  kann mit  $9N_l$  a. Op. ausgeführt werden, da  $A_h$  maximal 9 Nicht-Nulleinträge pro Zeile hat. Ein Glättungsschritt braucht damit  $12N_l$  a. Op.

Die Berechnung des Defekts  $d_l = f_l - A_l x^l$  kostet  $10N_l$  a. Op. Für die  $L^2$ -Projektion auf das nächstfeinere Gitter müssen wir nun

$$\tilde{d}_l^{l-1} := r_l^{l-1} d_l.$$

berechnen. Innerhalb des Mehrgitter-Algorithmus kann die  $L^2$ -Projektion sehr effizient berechnet werden. Wir bezeichnen die Basisfunktionen auf Gitterlevel  $l$  mit  $\varphi_i^l$ . Nach Definition der  $L^2$ -Projektion ist die  $i$ -te Komponente von  $\tilde{d}_i^{l-1}$  durch

$$\tilde{d}_i^{l-1} = (r_l^{l-1} d_l, \varphi_i^{l-1}) = (d_l, \varphi_i^{l-1}).$$

gegeben. Da  $V_{l-1} \subset V_l$ , können wir  $\varphi_i^{l-1}$  darstellen als

$$\varphi_i^{l-1} = \sum_{j=1}^{N_l} \mu_{ij} \varphi_j^l,$$

wobei für einen Index  $i$  maximal 9 Werte  $\mu_{ij}$  nichttrivial sind. Daher reduziert sich die Berechnung der  $L^2$ -Projektion auf

$$\tilde{d}_i^{l-1} = \sum_{j=1}^{N_l} \mu_{ij} (d_l, \varphi_j^l) = \sum_{j=1}^{N_l} \mu_{ij} d_j^l$$

und benötigt  $9N_l$  Operationen.

Die Prolongation kostet 2 a.op. für die Interpolation in jedem neuen Knoten (bei weniger als  $N_l$  solcher Knoten). Zusätzlich kostet das Aufaddieren der Korrektur  $N_l$  Operationen. Zusammen sind das

$$(2 \cdot 12 + 10 + 9 + 2 + 1)N_l = 46N_l$$

Operationen auf Gitterlevel  $l$ . Die Dimension der diskreten Räume verhält sich wie

$$N_{l-k} \approx 2^{-2k} N_l$$

Innerhalb eines V-Zyklus, müssen wir die genannten Operationen genau einmal pro Gitterlevel ausführen. Für genügend großes  $l$  können wir die Kosten zum Lösen auf dem größten Gitter vernachlässigen. Daher kostet ein V-Zyklus insgesamt

$$\sum_{k=0}^l 46N_{l-k} = \sum_{k=0}^l \frac{46}{2^{2k}} N_l = \frac{4}{3} 46N_l (1 - 2^{-(2k+2)}) \leq \frac{4}{3} 46N_l.$$

Innerhalb eines W-Zyklus, führen wir auf Gitterlevel  $l - k$   $2^k$  Schritte mit oben gezählten Operationen durch. Zusammen kostet das:

$$\sum_{k=0}^l 2^k 46N_{l-k} = \sum_{k=0}^l \frac{46}{2^k} N_l = 2 \cdot 46N_l (1 - 2^{-k-1}) \leq 2 \cdot 46N_l$$

arithmetische Operationen.

### Lösung A.3.27:

- a) Man erhält den 5-Punkte-Differenzenoperator.
- b) „Unisolvenz“ bedeutet, dass für ein Polynom  $p \in P(T)$  aus  $\chi(p) = 0$  ( $r = 1, \dots, R$ ) notwendig  $p \equiv 0$  folgt.
- c) Die „Minimalwinkelbedingung“ besagt, dass alle Winkel der Dreiecke gleichmäßig von Null weg beschränkt sind; dagegen besagt die „Maximalwinkelbedingung“, dass alle diese Winkel gleichmäßig von  $\pi$  wegbeschränkt sind. Die „Minimalwinkelbedingung“ ist äquivalent zur „Formregularität“ (Quotient aus Umkreis- und Inkreisradius gleichmäßig beschränkt), während die „Größenregulärität“ damit gar nichts zu tun hat.
- d) Es ist  $\dim(P_2) = 6$ ,  $\dim(P_5) = 21$ ,  $\dim(Q_2) = 9$ .
- e) Die Spektralkondition der FE-Matrix wird durch die Ordnung des Differentialoperators bestimmt; im gegebenen Fall gilt  $\mathcal{O}(h^{-2})$ .

## A.4 Kapitel 4

**Lösung A.4.1:** Wir betrachten O.B.d.A. nur den ersten  $N$ -Zyklus. Das Gauss-Seidel Verfahren hat gerade die Form

$$\hat{x}_j^{(1)} = \frac{1}{a_{jj}} \left( b_j - \sum_{k < j} a_{jk} \hat{x}_k^{(1)} - \sum_{k > j} a_{jk} \hat{x}_k^{(0)} \right).$$

Aufgrund der Wahl der Abstiegsrichtung  $d^{(t)} = e_{t+1}$  gilt für die Iterierten des Koordinatenrelaxationsverfahrens  $x_j^{(t+1)} = x_j^{(t)}$  für  $j \neq t+1$ . Es genügt also zu zeigen, dass im Schritt  $t \rightarrow t+1$  die  $t+1$ -te Komponente von  $x^{(t+1)}$  auf den Richtigen Wert gesetzt wird, der Rest folgt per Induktion.

Hierzu beachten wir  $r^{(t)} = b - Ax^{(t)}$  und somit ist

$$\alpha_{t+1} = \frac{r_{t+1}^{(t)}}{a_{t+1,t+1}} = \frac{1}{a_{t+1,t+1}} \left( b_{t+1} - \sum_k a_{t+1,k} x_k^{(t)} \right).$$

Durch Einsetzen in die Verfahrensvorschrift folgt:

$$\begin{aligned} x_{t+1}^{(t+1)} &= x_{t+1}^{(t)} + \frac{b_{t+1}}{a_{t+1,t+1}} - \frac{1}{a_{t+1,t+1}} \sum_k a_{t+1,k} x_k^{(t)} \\ &= \frac{1}{a_{t+1,t+1}} \left( b_{t+1} - \sum_{j < t+1} a_{t+1,k} x_k^{(t)} - \sum_{j > t+1} a_{t+1,k} x_k^{(t)} \right). \end{aligned}$$

Verwendet man nun die Induktionsannahme  $x_k^{(t)} = \hat{x}_k^{(1)}$  für  $k < t + 1$  und  $x_k^{(t)} = \hat{x}_k^{(0)}$  für  $k > t + 1$ , so folgt die behauptete Äquivalenz.

**Lösung A.4.2:** a) Wir rekapitulieren aus einer früheren Aufgabe: Die Eigenwerte der Iterationsmatrix  $B_\theta = I - \theta A$  sind  $\mu = 1 - \theta\lambda$  und folglich  $\mu_{\max} = 1 - \theta\lambda_{\min}$  sowie  $\mu_{\min} = 1 - \theta\lambda_{\max}$  bzw.

$$\rho(B_\theta) = \max |\mu| = \max_{i=1,\dots,N} |1 - \theta\lambda_i|.$$

Im Falle  $0 < \lambda_{\min} \leq \lambda_{\max}$  gilt

$$0 < \theta < \frac{2}{\lambda_{\max}} \Leftrightarrow 0 < \theta\lambda_{\min} \leq \theta\lambda_{\max} < 2.$$

Dies wiederum ist äquivalent mit  $\max\{|1 - \theta\lambda_{\min}|, |1 - \theta\lambda_{\max}|\} < 1$ . Ein einfaches geometrische Argument ergibt den Wert

$$\theta_{\text{opt}} = \frac{2}{\lambda_{\min} + \lambda_{\max}}$$

für minimales  $\rho(B_\theta)$ , d. h. für schnellste Konvergenz.

b) Die beste Glättungseigenschaft ist aber charakterisiert durch

$$|1 - \theta\lambda_{\max}| = \min_{i=1,\dots,N} |1 - \theta\lambda_i|,$$

da dann die hochfrequenten Fehleranteile in der Darstellung

$$|e_h^{(t)}|^2 = \sum_{i=1}^{N_h} \varepsilon_i^2 (1 - \theta\lambda_i)^{2t}$$

am schnellsten gedämpft werden. Dies ist der Fall für

$$\theta_{\text{opt}} = \frac{1}{\lambda_{\max}}.$$

Dass in diesem Fall die niederfrequenten Fehleranteile nur sehr langsam gedämpft werden, spielt keine Rolle, da man ja nur an dem Glättungseffekt interessiert ist.

**Lösung A.4.3:** Es gibt Probleme beim Beweis der Approximationseigenschaft. Im Beweis aus dem Text ist

$$v_{L-1} = A_{L-1}^{-1} r_L^{L-1} f_L, \quad v_L = A_L^{-1} f_L.$$

Aufgrund der Definition der Operatoren  $A_l$  folgt

$$a(v_L, \varphi_L) = (f_L, \varphi_L) \quad \forall \varphi_L \in V_L$$

sowie

$$a(v_{L-1}, \varphi_{L-1}) = (r_L^{L-1} f_L, \varphi_{L-1}) \quad \forall \varphi_{L-1} \in V_{L-1}.$$

Da  $r_L^{L-1}$  nicht die  $L^2$ -Projektion ist, ist i. Allg.

$$(r_L^{L-1} f_L, \varphi_{L-1}) \neq (f_L, \varphi_{L-1})$$

somit sind  $v_L$  und  $v_{L-1}$  nicht die Ritz-Projektion der selben kontinuierlichen Funktion.  
Was bleibt ist, dass  $v_L$  die Ritz-Projektion von

$$a(v, \varphi) = (f_L, \varphi) \quad \forall \varphi \in V$$

sowie  $v_{L-1}$  die Ritz-Projektion von

$$a(\tilde{v}, \varphi) = (r_L^{L-1} f_L, \varphi) \quad \forall \varphi \in V$$

ist. Damit ist

$$\|v_L - v\| \leq ch_L^2 \|\nabla^2 v\| \leq ch_L^2 \|f_L\|$$

sowie

$$\|v_{L-1} - \tilde{v}\| \leq ch_{L-1}^2 \|\nabla^2 \tilde{v}\| \leq ch_L^2 \|r_L^{L-1} f_L\|.$$

Damit bleibt noch zu zeigen, dass

$$\|r_L^{L-1} f_L\| \leq c \|f_L\|$$

sowie

$$\|v - \tilde{v}\| \leq ch_L^2 \|f_L\|.$$

Während man die erste Ungleichung unter Ausnutzung der endlichen Dimension des Raumes  $V_L$  zeigen kann, ist die 2. Ungleichung i. Allg. nicht gültig.

**Lösung A.4.4:** Hier tritt das Problem bei der Glättungseigenschaft auf. Mithilfe einer inversen Ungleichung können wir zwar zeigen

$$\|A_L\| \leq ch^{-2}.$$

Es bleibt zu zeigen, dass

$$\|S_L\| \leq c < 1$$

für  $S_L = I_L - \theta A_L$  mit einer  $L$ -unabhängigen Konstante  $c$ . Da  $A_L$  nicht symmetrisch ist, können wir die Spektralargumente aus dem Text hier nicht direkt übertragen.  $A_L$  ist jedoch positiv definit. Um das zu sehen, wenden wir den Gaußschen Integralsatz an:

$$(\partial_1 u, u) = \frac{1}{2} \int_{\Omega} \partial_1(u^2) dx = \frac{1}{2} \int_{\partial\Omega} n_1 u^2 ds = 0$$

für  $u \in H_0^1(\Omega)$ . Es folgt

$$a(u, u) = \|\nabla u\|^2.$$

Damit sind die Realteile der (komplexen) Eigenwerte von  $A_L$  positiv

$$\Re(\lambda_i) > 0 \quad i = 1 \dots N_L.$$

Die Eigenwerte von  $S_L = I_L - \theta A_L$  sind  $1 - \theta\lambda_i, i = 1 \dots N_L$ . Es gilt

$$\begin{aligned} |1 - \theta\lambda| &= |1 - \theta\Re(\lambda) - \theta\Im(\lambda)| = \{(1 - \theta\Re(\lambda))^2 + \theta^2(\Im(\lambda))^2\}^{\frac{1}{2}} \\ &= \{1 - 2\theta\Re(\lambda) + \theta^2(\Im(\lambda)^2 + \Re(\lambda)^2)\}^{\frac{1}{2}} \end{aligned}$$

Wählen wir

$$\theta < \max_{i=1 \dots N_L} \frac{2\Re(\lambda_i)}{|\lambda_i|^2},$$

so ist

$$\text{spr}(S_L) = \max_{i=1 \dots N_L} |1 - \theta\lambda_i| < c < 1.$$

gleichmäßig in  $L$ . Somit gibt es zu jedem  $\epsilon > 0$  eine Norm  $\|\cdot\|_*$  mit

$$\|S_L\|_* \leq c + \epsilon.$$

Es bleibt noch die Frage offen, inwieweit die  $L$ -unabhängige Konvergenzrate in der Norm  $\|\cdot\|$  erhalten bleibt.

**Lösung A.4.5:** Ein Glättungsschritt des Richardson-Verfahren benötigt im Wesentlichen eine Matrix-Vektor-Multiplikation. Die Matrix-Vektor-Multiplikation  $A_h x_h$  kann mit  $9N_l$  a. Op. ausgeführt werden, da  $A_h$  maximal 9 Nicht-Nulleinträge pro Zeile hat. Ein Glättungsschritt braucht damit  $12N_l$  a. Op.

Die Berechnung des Defekts  $d_l = f_l - A_l x^l$  kostet  $10N_l$  a. Op. Für die  $L^2$ -Projektion auf das nächstfeinere Gitter müssen wir nun

$$\tilde{d}^{l-1} := r_l^{l-1} d_l$$

berechnen. Innerhalb des Mehrgitter-Algorithmus kann die  $L^2$ -Projektion sehr effizient berechnet werden. Wir bezeichnen die Basisfunktionen auf Gitterlevel  $l$  mit  $\varphi_i^l$ . Nach Definition der  $L^2$ -Projektion ist die  $i$ -te Komponente von  $\tilde{d}^{l-1}$  durch

$$\tilde{d}_i^{l-1} = (r_l^{l-1} d^l, \varphi_i^{l-1}) = (d^l, \varphi_i^{l-1}).$$

gegeben. Da  $V_{l-1} \subset V_l$ , können wir  $\varphi_i^{l-1}$  darstellen als

$$\varphi_i^{l-1} = \sum_{j=1}^{N_l} \mu_{ij} \varphi_j^l,$$

wobei für einen Index  $iv$  maximal 9 Werte  $\mu_{ij}$  nichttrivial sind. Daher reduziert sich die Berechnung der  $L^2$ -Projektion auf

$$\tilde{d}_i^{l-1} = \sum_{j=1}^{N_l} \mu_{ij}(d^l, \varphi_i^l) = \sum_{j=1}^{N_l} \mu_{ij} d_i^l$$

und benötigt  $9N_l$  Operationen.

Die Prolongation kostet 2 a. Op. für die Interpolation in jedem neuen Knoten (bei wenigstens  $N_l$  solcher Knoten). Zusätzlich kostet das Aufaddieren der Korrektur  $N_l$  Operationen. Zusammen sind das

$$(2 \cdot 12 + 10 + 9 + 2 + 1)N_l = 46N_l$$

Operationen auf Gitterlevel  $l$ . Die Dimension der diskreten Räume verhält sich wie

$$N_{l-k} \approx 2^{-2k} N_l$$

Innerhalb eines V-Zyklus, müssen wir die genannten Operationen genau einmal pro Gitterlevel ausführen. Für genügend großes  $l$  können wir die Kosten zum Lösen auf dem gröbstens Gitter vernachlässigen. Daher kostet ein V-Zyklus insgesamt

$$\sum_{k=0}^l 46N_{l-k} = \sum_{k=0}^l \frac{46}{2^{2k}} N_l = \frac{4}{3} 46N_l (1 - 2^{-(2k+2)}) \leq \frac{4}{3} 46N_l.$$

Innerhalb eines W-Zyklus, führen wir auf Gitterlevel  $l-k$   $2^k$  Schritte mit oben gezählten Operationen durch. Zusammen kostet das die folgenden a. Op.:

$$\sum_{k=0}^l 2^k 46N_{l-k} = \sum_{k=0}^l \frac{46}{2^k} N_l = 2 \cdot 46N_l (1 - 2^{-k-1}) \leq 2 \cdot 46N_l.$$

**Lösung A.4.6:** a) Die Eigenwerte der Iterationsmatrix  $B_\theta = I - \theta A$  sind  $\mu = 1 - \theta\lambda$  und folglich  $\mu_{\max} = 1 - \theta\lambda_{\min}$  sowie  $\mu_{\min} = 1 - \theta\lambda_{\max}$  bzw.

$$\rho(B_\theta) = \max\{|\mu|\} = \max\{|1 - \theta\lambda_{\max}|, |1 - \theta\lambda_{\min}|\}.$$

b) Im Falle  $0 < \lambda_{\min} \leq \lambda_{\max}$  gilt

$$0 < \theta < \frac{2}{\lambda_{\max}} \Leftrightarrow 0 < \theta\lambda_{\min} \leq \theta\lambda_{\max} < 2.$$

Dies wiederum ist äquivalent mit  $\max\{|1 - \theta\lambda_{\min}|, |1 - \theta\lambda_{\max}|\} < 1$ .

c) Ein einfaches geometrische Argument ergibt

$$\theta_{\text{opt}} = \frac{2}{\lambda_{\min} + \lambda_{\max}}.$$

**Lösung A.4.7:** a) Die schwache Formulierung lautet

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

$$(a\nabla u, \nabla \varphi) + (bu, \varphi) = (f, \varphi) \quad \forall \varphi \in H_0^1(\Omega).$$

Wir definieren die Bilinearform

$$a(u, \varphi) = (a\nabla u, \nabla \varphi) + (bu, \varphi)$$

Um Existenz und Eindeutigkeit einer Lösung zu gewährleisten (mit dem Lax-Milgram-Lemma oder dem Rieszschen Darstellungssatz) brauchen wir die Abschätzung

$$a(u, u) \geq \alpha \|u\|_{H^1(\Omega)}^2$$

für eine Konstante  $\alpha > 0$ . Sei

$$a_0 := \min_{x \in \Omega} a(x), \quad b_0 := \min_{x \in \Omega} b(x).$$

Es gilt

$$a(u, u) \geq a_0 \|\nabla u\|^2 + b_0 \|u\|^2.$$

Daher sind die Bedingungen  $a_0 > 0$  und  $b_0 \geq 0$  hinreichend, denn es folgt mit der Poincaréschen Ungleichung

$$a(u, u) \geq a_0 \|\nabla u\|^2 \geq \frac{a_0}{c_P + 1} \|u\|_{H^1(\Omega)}^2.$$

b) Wir definieren eine reguläre Triangulierung  $T_h$  des Polyeders  $\Omega$  in abgeschlossene Tetraeder  $T$  sowie einen Satz von linearen Funktionalen  $\mathcal{X}$  auf  $P(T)$ , so dass ein Polynom  $p \in P_2(T)$  eindeutig durch die Werte von  $\mathcal{X}(p)$  festgelegt ist. Für den Fall quadratischer Polynome ist das einzige übliche konforme Element das Lagrange-Element, welches Punktwerte in Knotenpunkten und Seitenmittelpunkten verwendet. Der dazugehörige Ansatzraum ist

$$V_h^{(2)} := \{v_h \in C(\bar{\Omega}) \mid v_h \in P_2(T) \ \forall T \in \mathbb{T}_h, v_h = 0 \text{ auf } \partial\Omega\}.$$

Eine Basis von  $V_h^{(2)}$  wird durch

$$\varphi_i(a_j) = \delta_{ij}, \quad i, j = 1 \dots N,$$

definiert, wobei die Menge  $\{a_i, i = 1 \dots N\}$  aus Knoten- und Seitenmittelpunkten bestehe. Die Steifigkeitsmatrix  $A_h = (a_{ij})_{i,j=1}^N$  und der Lastvektor  $b_h = (b_i)_{i=1}^N$  berechnen sich zu

$$a_{ji} = a(\varphi_i, \varphi_j), \quad b_i = (f, \varphi_i).$$

Das diskrete System

$$a(u_h, \varphi_i) = (f, \varphi_i) \quad \forall \varphi_i \in V_h^{(2)}$$

ist nun äquivalent zum linearen Gleichungssystem

$$A_h x_h = b_h,$$

wobei  $x_i$  die Komponenten von  $u_h$  bezüglich der oben definierten Basis bezeichne:

$$u_h = \sum_{i=1}^N x_i \varphi_i.$$

c) Unter der Annahme  $u \in H^3(\Omega)$ , können wir mit Standardargumenten (Galerkin-Orthogonalität, Bestapproximationseigenschaft, Interpolationsabschätzungen) zeigen, dass

$$\|u - u_h\|_E := a(u, u)^{\frac{1}{2}} \leq ch^2 \|\nabla^3 u\|.$$

Mithilfe eines Dualitätsarguments bekommen wir in der  $L^2$ -Norm

$$\|u - u_h\| \leq ch^3 \|\nabla^3 u\|.$$

Die Spektralkondition hängt von der Ordnung des Differentialoperators ab. Für den gegebenen elliptischen Differentialoperator zweiter Ordnung gilt

$$\kappa_2(A_h) = \mathcal{O}(h^{-2}).$$

di) Von einem Startwert  $x^{(0)}$  definiert das Gauß-Seidel-Verfahren Iteriert durch

$$x_i^{(k+1)} = b_i - \sum_{j < k} a_{ij} x_j^{(k)} - \sum_{j > k} a_{ij} x_j^{(k+1)}, \quad i = 1 \dots N.$$

Ein Gauß-Seidel-Schritt reduziert den Gesamtfehler um den Faktor

$$\rho = 1 - \mathcal{O}(h^2).$$

Die Anzahl an Iterationen  $T$ , die man benötigt, um den Anfangsfehler um den Faktor  $\epsilon = 10^{-3}$  zu reduzieren, ist daher

$$\rho^T = \epsilon \quad \Leftrightarrow \quad T = \frac{\ln(\epsilon)}{\ln(\rho)} = -3 \frac{\ln(10)}{\ln(1 - ch^2)} \simeq ch^{-2},$$

da  $\ln(1 - ch^r) \simeq ch^r + \mathcal{O}(h^{2r})$ . Eine Iteration besteht aus  $2N$  Subtraktionen und  $N - 1$  Matrix-Vektor-Multiplikationen. Da die Matrix  $A_h$  dünn besetzt ist, kostet jede Iteration  $\mathcal{O}(N)$  a. Op.. Da  $N \approx h^{-3}$ , brauchen wir insgesamt  $\mathcal{O}(N^{\frac{5}{3}})$  a. Op..

dii) Das Gradientenverfahren ist ausgehend von einem Startwert  $x^{(0)}$  definiert durch

$$x^{(t+1)} = x^{(t)} + \alpha_t(r^{(t)}), \quad \alpha_t = \frac{\|r^{(t)}\|^2}{(A_h r^{(t)}, r^{(t)})},$$

wobei  $r^{(t)} := b_h - A_h x^{(t)}$  das Residuum in Schritt  $t$  bezeichnet. Ein Schritt des Gradien-

tenverfahrens reduziert den Fehler um den Faktor

$$\rho = 1 - \frac{1 - \kappa(A_h)^{-1}}{1 + \kappa(A_h)^{-1}} \approx 1 - 2\kappa(A_h)^{-1} = 1 - \mathcal{O}(h^2).$$

Hier haben wir benutzt, dass die Spektralkondition  $\kappa(A_h)$  sich wie  $\mathcal{O}(h^{-2})$  verhält. Eine Iteration besteht im Wesentlichen wieder aus einer konstanten Zahl an Matrix-Vektor-Multiplikationen und kostet daher  $\mathcal{O}(N)$  a. Op. Die Iterationszahl ist wieder  $\mathcal{O}(N^{\frac{5}{3}})$ . Ein besseres Ergebnis liefert das CG-Verfahren. Ein Schritt des CG-Verfahrens reduziert den Fehler um einen Faktor

$$\rho = 1 - \frac{1 - \kappa(A_h)^{-\frac{1}{2}}}{1 + \kappa(A_h)^{-\frac{1}{2}}} \approx 1 - 2\kappa(A_h)^{-\frac{1}{2}} = 1 - \mathcal{O}(h).$$

Die Anzahl Iterationen  $T$ , um den Anfangsfehler um den Faktor  $\epsilon$  zu reduzieren, kann daher mit

$$T \leq \frac{1}{2} \sqrt{\kappa(A_h)} \ln\left(\frac{2}{\epsilon}\right) + 1.$$

abgeschätzt werden, d. h.

$$T \in \mathcal{O}(N^{\frac{1}{3}}).$$

Die Anzahl a. Op. ist damit  $\mathcal{O}(N^{\frac{4}{3}})$ .

## A.5 Kapitel 5

**Lösung A.5.1:** a) Es genügt zu zeigen, dass bei numerischer Integration über dem Einheitsdreieck  $\hat{T}$  mit Hilfe der Dreiecks-Trapezregel für die drei Knotenbasisfunktionen gilt:

$$Q_T(\hat{\varphi}_i \hat{\varphi}_j) = \delta_{ij}.$$

Paarweises Einsetzen der Basisfunktionen

$$\hat{\varphi}_1 = x, \quad \hat{\varphi}_2 = y, \quad \hat{\varphi}_3 = 1 - x - y,$$

in die Quadraturformel liefert aber sofort:

$$\hat{Q}_T(\hat{\varphi}_i \hat{\varphi}_j) = \frac{1}{6} \{ \hat{\varphi}_i(0,0) \hat{\varphi}_j(0,0) + \hat{\varphi}_i(1,0) \hat{\varphi}_j(1,0) + \hat{\varphi}_i(0,1) \hat{\varphi}_j(0,1) \} = \frac{1}{6} \delta_{ij}.$$

Durch Rücktransformation auf die Gitterzellen und „Assemblierung“ erhält man:

$$(\tilde{M}_h)_{ij} = \delta_{ij} \sum_{\hat{T} \ni x_{ii}} \frac{|T|}{3}.$$

Durch Masse-Lumping erhält man also eine positive Diagonalmatrix.

b) Nach dem diskreten Maximumsprinzip für finite Elemente ist im Fall, wenn alle Innenwinkel der Triangulierung kleiner oder gleich  $\pi/2$  sind, die Steifigkeitsmatrix  $A_h$  eine

M-Matrix Es muss also nur gezeigt werden, dass beim Skalieren mit  $k$  und Addieren der „gelumpten“ Massematrix  $M_h$  die M-Matrixeigenschaft nicht verloren geht:

1. Die Diagonaldominanz,

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

bleibt unter Skalierung und Addition positiver Beiträge auf der Diagonalen erhalten.

2. Die Eigenschaft „von nicht-negativen Typ“,  $a_{ii} > 0$ ,  $a_{ij} \leq 0$ ,  $\forall i, j \neq i$ , ebenso.
3. Da durch die Addition positiver Diagonalelemente keine Einträge der Steifigkeitsmatrix ausgelöscht werden können, ist die resultierende Systemmatrix wieder irreduzibel.

- c) Nicht ausgeführt.

**Lösung A.5.2:** Nicht ausgeführt.

**Lösung A.5.3:** Nicht ausgeführt.

## A.6 Kapitel 6

**Lösung A.6.1:**

1. Das Ritzsche Projektionsverfahren ist für variationelle Probleme mit *symmetrischer* Bilinearform  $a(.,.)$  definiert. Ausgangspunkt ist die Formulierung über das Optimierungsproblem

$$\min_{u_h \in V_h} E(u_h); \quad E(\varphi) = \frac{1}{2}a(\varphi, \varphi) - (f, \varphi).$$

Dem gegenüber ist das Galerkinverfahren auch für nicht-symmetrische Bilinearformen definiert. Es bedient sich direkt der variationellen Formulierung: Finde  $u_h \in V_h$ , so dass:

$$a(u_h, \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h.$$

Die Petrov-Galerkin-Verfahren verwenden unterschiedliche Ansatz- und Testräume für die variationelle Formulierung:

$$u_h \in V_h^{\text{Ansatz}}, \quad \varphi_h \in V_h^{\text{Test}}.$$

2. Der Glätter auf den feineren Gitterlevel.

3. Für einen quadratischen Ansatz erhält man in der Energienorm  $\|\nabla e_h\| \leq h^2 \|\nabla^3 u\|$  und der  $L^2$ -Norm  $\|e_h\| \leq h^3 \|\nabla^3 u\|$ . Es ist aber sogar möglich die Norm noch eine Stufe auf eine sog. „negative Sobolevnorm“ abzuschwächen und damit 4te Ordnung 4 zu erhalten: Zu beliebigen  $\psi \in H^1$  sei  $z$  Lösung des dualen Problems

$$(\nabla \varphi, \nabla z) = \frac{(\varphi, \psi)}{\|\psi\|_1}.$$

Es gilt die a priori-Abschätzung  $\|z\|_3 \leq \left\| \psi / \|\psi\|_1 \right\|_1 \leq c$  unabhängig von  $\psi$ . Einsetzen von  $e_h$  in die Gleichung:

$$\frac{(e_h, \psi)}{\|\psi\|_1} \leq \|\nabla e_h\| \|\nabla z\| \leq ch^2 \|\nabla^3 u\| ch^2 \|\nabla^3 z\| \leq ch^4 \|\nabla^3 u\|.$$

4. Die Maximalwinkelbedingung fordert, dass alle in einer Gitterhierarchie auftretenden Innenwinkel gleichmäßig nach oben von  $180^\circ$  weg beschränkt sind; analog fordert die Minimalwinkelbedingung eine Abschätzung nach unten von  $0^\circ$  weg.  
 5. Unter der Glättungseigenschaft versteht man die Gültigkeit einer Abschätzung der Form

$$\|U_h^M - u_h^m\| \leq c \frac{k^r}{t_m^r} \|u_h^0\|,$$

wobei  $r$  die Ordnung des Verfahrens, und  $U^M$  die FE-Näherung an die kontinuierliche Lösung  $u_h^m$  zum Zeitpunkt  $t_m$  ist.

Das Crank-Nicolson-Schema ist zwar A-stabil aber nicht *stark* A-stabil. Eine Glättungseigenschaft ist deshalb nicht zu erwarten und tatsächlich zeigen numerische Ergebnisse, dass das Crank-Nicolson-Schema keine Glättungseigenschaft besitzt.

6. Bei einem isoparametrischen Ansatz ist die Transformation einer lokalen Zelle (Dreieck, Rechteck) auf die Referenzzelle vom selben polynomialen Ansatzraum wie der FE-Ansatz.  
 7. Die Ordnung  $r$  der Quadraturformel sollte so gewählt werden, dass er zum einen zulässig ist und zum anderen  $r \geq 2m - 3$ .  
 8. Es ergibt sich eine Schrittweitenbedingung der Form

$$k \leq c \frac{1}{h^2}.$$

9. Der biharmonische Operator ist von 4ter Ordnung, die Kondition ist also

$$\text{cond}_2(A_h) = \mathcal{O}(h^{-4}).$$

10. Eine M-Matrix ist eine quadratische Matrix mit starker Diagonaldominanz:

$$\sum_{j \neq i} a_{ij} \leq a_{ii} \quad \forall i, \text{ und } \exists k \text{ s.d. } \sum_{j \neq k} a_{kj} < a_{kk},$$

und von nicht-negativen Typs:

$$a_{ii} > 0, \quad a_{ij} \leq 0, \quad \forall i, j \neq i.$$

Eine M-Matrix ist regulär und ihre Inverse ist elementweise nicht negativ,  $A_h^{-1} \geq 0$ .

# Index

- 5-Punkte-Operator, 52, 63, 70, 91  
7-Punkte-Operator, 55  
9-Punkte-Operator  
    gestreckter, 52  
    kompakter, 53, 91  
 $L^2$ -Norm-Fehler, 143  
 $L^2$ -Norm-Fehlerschätzer, 150  
 $L^2$ -Projektion, 119, 164, 251  
 $L^\infty$ -Norm-Fehler, 121, 143  
 $M$ -Matrix, 90  
 $V$ -elliptisch, 79  
 $dG(0)$ -Verfahren, 231  
 $dG(r)$ -Verfahren, 231  
 $h/p$ -Finite-Elemente-Methode, 84  
A-orthogonal, 176  
A-Stabilität, 212  
    starke, 213, 219  
    strenge, 213  
Abhängigkeitsbereich, 42, 249  
Ableitung  
    partielle, 1  
    schwache, 26  
    verallgemeinerte, 26  
Abschneidefehler, 50, 208, 241, 248  
Abstiegsrichtung, 174  
Abstiegsverfahren, 174  
Adaptionsstrategie, 156  
ADI-Verfahren, 69, 239, 241  
Adini (????–), 104  
Adini-Plattenelement, 104  
Allgemeiner Störungssatz, 130  
Anfangs-Randwert-Aufgabe, 35  
Approximationseigenschaft, 193  
Argyris (1913–2004), 102  
Argyris-Plattenelement, 102  
Assemblierung, 128  
Aubin (1939–), 87  
Aubin-Nitsche-Trick, 87  
Bachwalow (1934–2005), 185  
Banach (1892–1945), 2  
Bandmatrix, 56, 66  
Baryzentrische Koordinaten, 135  
BDF-Verfahren, 208, 209  
Bestapproximationseigenschaft, 81, 119  
Bestimmtheitsbereich, 42  
Biharmonischer Operator, 7, 23, 101, 132, 167  
Bramble (1932–), 110  
Bramble-Hilbert-Lemma, 111, 137, 234  
Brandt (1938–), 185  
Cauchy (1789–1857), 12  
Cauchy-Problem, 12, 248  
CFL-Bedingung, 244, 245  
CG-Verfahren, 176, 178, 200, 239  
    quadrirtes, 181  
Charakteristik, 12  
Charakteristisches Polynom, 247  
Cholesky (1975–1918), 69  
Cholesky-Zerlegung, 69, 239  
Courant (1888–1972), 244  
Crank (1916–2006), 208  
Crank-Nicolson-Verfahren, 209, 213, 223, 231, 236, 238, 240, 250  
    gedämpftes, 222  
    Schnentrapezform, 236  
    Tangententrapezform, 237  
Dämpfung, 191  
Defekt, 174  
Defektgleichung, 189  
Defektkorrektur-Iteration, 67  
Diagonaldominanz  
    erweiterte , 57  
    irreduzible, 58  
Differentialgleichung  
    2. Ordnung, 10  
    gewöhnliche, 10  
    lineare, 9  
    nichtlineare, 9, 15  
Differentialoperator  
    elliptischer, 13  
    hyperbolischer, 13  
    parabolischer, 13  
Differenzenapproximation, 49  
Differenzengleichung, 247  
Differenzenoperator, 49  
Differenzenverfahren  
    A(0)-stabil, 207  
    explizit, 204

- implizit, 204
- Diffusions-Transport-Gleichung, 78
- Dirac (1902–1984), 123
- Dirac-Funktion, 123
- Direkte Methode, 19
- Dirichlet (1805–1859), 16
- Divergenz, 1
- Du Fort (????–????), 218
- Du Fort-Frankel-Verfahren, 218
- Duale Lösung, 145
- Duales Problem, 145
- Dualitätsargument, 87, 232
- Eckensingularität, 24, 45, 161
- Effektivitätsindex, 146, 153
- Eigenwerte, 74
- Eigenwertproblem, 34
- Einschrittverfahren, 204
- Element-(Last)-Vektor, 128
- Element-(Steifigkeits)-Matrix, 128
- Energie, 43
- Energie-Form, 187, 203
- Energie-Methode, 40, 223
- Energie-Norm, 19, 81, 174
- Energieerhaltung, 243
- Energiefunktional, 77
- Energienorm-Fehler, 87, 143
- Energienorm-Fehlerschätzer, 147
- Euklid (ca. 355–290 v. Chr.), 1
- Euklidische Vektornorm, 1
- Euklidisches Produkt, 1
- Euler (1707–1783), 205
- Explizites Euler-Verfahren, 207, 215
- F-Zyklus, 190, 239
- Fedorenko (1930–2009), 185
- Fehler
  - lokaler, 127
- Fehlerfunktional, 144
- Fehlerindikator, 155
- Finite Elemente, 84
  - bi-kubisch, 104, 105
  - bi-linear, 103
  - bi-quadratisch, 103
  - isoparametrisch, 108
  - konform, 99
- konstant, 99
- kubisch, 101, 166
- linear, 85, 99, 104
- nicht-konform, 99
- parametrisch, 108
- quadratisch, 100, 104
- quartisch, 102
- quintisch, 102
- tri-linear, 105
- tri-quadratisch, 105
- Finite-Elemente-Interpolierende, 99
- Finite-Elemente-Raum, 230
- Fixpunktiteration, 68
- Formfunktion, 84
- Formregularität, 85
- Fourier (1768–1830), 34
- Fourier-Entwicklung, 34
- Frankel (1919–1978), 218
- Friedrichs (1901–1982), 244
- Galerkin (1871–1945), 81
- Galerkin-Gleichungen, 176
- Galerkin-Orthogonalität, 232
- Galerkin-Verfahren, 81
- Gauß-Seidel-Verfahren, 69
- Gaußsches Eliminationsverfahren, 67
- Gauß (1777–1855), 2
- geschachteltes MG-Verfahren, 190
- Gitterfeinheit, 49
- Gitternummerierung
  - diagonale, 66
  - Schachbrett, 67
  - zeilenweise, 66
- Gittersteuerung, 155
- Gittertransfer, 191
- Glätter, 190
- Glättungseigenschaft, 40, 193, 213, 219, 222, 237
- Glättungsiteration, 186
- Größenregularität, 85
- Gradient, 1
- Gradientenverfahren, 175
- Gram (1850–1916), 177
- Gram-Schmidt-Algorithmus, 177
- Green (1793–1841), 2

- Greensche Funktion, 17, 123, 124  
Greensche Identität, 61  
Grobgitteroperatoren, 191  
Grobgitterproblem, 188  
  
hängender Knoten, 251  
Hölder (1859–1937), 3  
Höldersche Ungleichung, 3, 29  
Hackbusch (1948–), 185  
Hadamard (1865–1963), 9  
Halbbandbreite, 56  
Halbgruppen-Methode, 40  
Hauptteil, 12, 14  
Hermite (1822–1901), 98  
Hermite-Element, 98  
Hestenes (1906–1991), 176  
Hilbert (1862–1943), 2  
Hilbert (????–), 110  
  
ILU-Verfahren, 69  
Implizites Euler-Verfahren, 207, 209, 225, 231  
Integralformel von Green, 2  
Integralsatz von Gauß, 2  
Interpolationsabschätzung, 112, 114  
Interpolationskonstante, 232, 234  
Interpolierende, 86  
inverse Beziehung, 117  
inverse Monotonie, 59, 74, 216  
Iterationsmatrix, 68  
  
Jacobi (1804–1851), 69  
Jacobi-Verfahren, 69  
Jordan (1838–1922), 10  
  
Kantenresiduum, 145  
Kegelschnitt, 12  
Knoten, 85  
Knotenbasis, 85  
Knotenbasisfunktion, 86  
Knotenfunktional, 98  
Knotenwert, 85, 98  
Koerzitivität, 79  
Konditionierung, 130, 167  
Konditionszahl, 74  
Konformität, 99  
  
Konsistenz, 50, 51, 208  
Konsistenzordnung, 50  
Kontinuitätsgleichung, 5  
Konvergenz, 62  
Konvergenzordnung, 50, 73  
Konvergenzrate, 68  
Koordinatenrelaxation, 175  
Kovalevskaya (1850–1891), 15  
Kronecker (1823–1891), 61  
Kronecker-Symbol, 61  
Krylov (1879–1955), 173  
Krylow-Raum, 176  
Krylow-Raum-Methode, 173  
Kutta (1867–1944), 231  
  
Lösung  
  klassische, 16, 35  
  schwache, 20, 46  
Lösungskomplexität, 238  
Lagrange (1736–1813), 85  
Lagrange-Basis, 85  
Lagrange-Element, 98  
Lagrange-Hermite-Interpolation, 114  
Lagrange-Interpolation, 85  
Laplace (1749–1827), 2  
Laplace-Operator, 2, 14, 33, 45, 52, 65, 173  
Lastvektor, 128, 225  
Lax (1926–), 79  
Lax-Milgram-Lemma, 79  
Lebesgue (1875–1941), 25  
Lebesgue-Raum, 29  
Legendre (1752–1833), 177  
Legendre-Polynom, 177  
Lewy (1904–1988), 244  
Line Search, 174  
Linienmethode, 203, 207  
  
M-Matrix, 60, 74, 216, 229  
Mass-Lumping, 229  
Massematrix, 128, 204, 225, 251  
Matrix  
  diagonal-dominant, 90  
  von nicht-negativem Typ, 57, 90  
Matrizennorm, 1  
Maximumnorm, 2  
Maximumprinzip, 22, 124

- diskretes, 58
- elliptisches, 22
- für finite Elemente, 90
- parabolisches, 37
- Maximumprinzipmethode, 215
- Mehrgitter-Zyklus, 188
- Mehrgitterverfahren, 184, 239
- Mehrstellenformel, 72
- Milgram (1912–1961), 79
- Minimalfolge, 19
- Mittelpunktregel, 135
- Mittelpunkts-Verfahren, 208
- Monombasis, 83
- Morley (1924–2011), 101
- Morley-Plattenelement, 101
- Multiindex, 109
- Nabla-Operator, 1
- Nachglättung, 189
- Navier (1785–1836), 5
- Navier-Stokes-Gleichungen, 5
- Neumann 1832–1925, 16
- Newton (1643–1727), 5
- Nicolson (1917–1968), 208
- Nitsche (1926–1996), 87
- Normaleneinheitsvektor, 1
- Numerische Dissipativität, 213
- Numerische Integration, 134, 138, 168
- Operator**
  - kompakt, 34
  - positiv-definit, 33
  - symmetrisch, 33
- Orthonormalsystem, 193
- Orts-Zeit-Diskretisierung, 206
- Padé (1785–1836), 211
- Padé-Schema, 211
- Padé-Tafel, 211
- Parallelogrammidentität, 20
- Parseval (1755–1836), 40
- Parsevalsche Identität, 40
- PCG-Verfahren, 182
- Petrow (1912–1987), 82
- Petrow-Galerkin-Verfahren, 82, 231
- Plattengleichung, 7
- Poincaré (1854–1912), 3
- Poincarésche Ungleichung, 3, 29, 45, 46, 111
- Poisson (1781–1840), 4
- Poisson-Gleichung, 4, 6, 14, 15, 77, 120
- Pollution-Effekt, 162
- Polygonzugmethode, 205
- Projektionsmethode, 81
- Prolongation, 188, 189
- Punktfehler-Schätzer, 151
- Punktgitter, 49
- Quadraturfehler, 137
- Quadraturformel, 135
- Rückwärtsdifferenzenformel, 208
- Randbedingungen**
  - Dirichletsche, 16, 35, 77
  - natürliche, 78
  - Neumannsche, 16, 35, 78
  - Robinsche, 16, 35, 78
- Randwertaufgabe, 16, 45
- Referenzelement, 106, 113, 129
- Referenztransformation, 106
- Rellich (1906–1955), 32
- Rellichscher Auswahlsatz, 32
- Residuum, 145
- Restriktion, 188, 189
- Reynolds (1842–1912), 4
- Richardson (1881–1953), 68
- Richardson-Verfahren, 68, 185, 187, 200, 219
- Riemann (1826–1866), 17
- Riesz (1880–1956), 79
- Rieszscher Darstellungssatz, 79
- Ritz (1878–1909), 80
- Ritz-Projektion, 119, 226, 251
- Ritzsches Projektions-Verfahren, 80
- Robin (1855–1897), 16
- Rothe (1895–1988), 206
- Rothe-Methode, 206, 224, 250
- Runge (1856–1927), 231
- Runge-Kutta-Verfahren, 231
- Satz**
  - Mehrgitterkomplexität, 196
  - Mehrgitterkonvergenz, 195
  - von Cauchy-Kovalevskaya, 15

- von Lax-Milgram, 162  
Schlitzgebiet, 24  
Schmidt (1876–1959), 177  
Schnelle Fourier-Transformation, 67  
Schrittweitensteuerung, 230  
Seidel, von (1821–1896), 69  
Separabilität, 240  
Shortley (1910–????), 54  
Shortley-Weller-Approximation, 54, 57, 63, 65, 73, 88  
Simplex, 106  
Simpson-Verfahren, 208  
Sobolew (1908–1989), 20  
Sobolew-Raum, 20, 26, 29, 77  
Sobolewsche Ungleichung, 32, 47  
Sobolow-Raum, 46  
SOR-Verfahren, 69  
Spektral-Elemente-Methode, 84  
Spektral-Galerkin-Verfahren, 84  
Spektral-Methode, 40, 219  
Spektral-Theorie, 33  
Spektralkondition, 179, 239  
Spektralnorm, 217  
Spektralverfahren, 84  
Spezieller Störungssatz, 132  
Splitting-Methode, 239  
Spur-Abschätzung, 30  
Spur-Lemma, 30  
Stabilität, 50  
Startwert, 237  
Steifigkeitsmatrix, 128, 204, 225, 251  
Stiefel (1909–1978), 176  
Stokes (1819–1903), 5  
Strukturregularität, 84  
  
Taylor (1685–1731), 10  
Taylor-Reihe, 10  
Teilschritt- $\theta$ -Verfahren, 214  
Tensorprodukt-Basis, 83  
Tensorprodukt-Gauß-Formel, 137  
Tensorprodukt-Simpson-Regel, 137  
Tensorprodukt-Trapezregel, 163  
Tensorproduktformel, 136  
Transfermatrix, 225  
Trapezregel, 135  
  
Triangulierung, 84  
Tschebyscheff (1821–1894), 180  
Tschebyscheff-Approximation, 180  
Typeneinteilung, 10, 44  
  
Unisolvenz, 98, 109, 164, 170  
Unstetiges Galerkin-Verfahren, 231  
  
V-Zyklus, 189  
Variableneparation, 36  
Variationelle Formulierung, 78, 203, 250  
Variationsgleichung, 6, 78, 81  
Verfeinerungsstrategie  
    Fehlerbalancierung, 156  
    Fest-Raten, 157  
    Gitteroptimierung, 157  
Verträglichkeit, 50  
Vieta (1540–1603), 13  
Vietascher Wurzelsatz, 13  
von Neumann (1903–1957), 217  
Von Neumannsche Stabilitätsanalyse, 218  
Vorglättung, 188  
Vorkonditionierung, 68, 182  
    ADI, 184  
    Diagonal, 183  
    ICCG, 183  
    SSOR, 183  
  
W-Zyklus, 189  
Wärmeleitungsgleichung, 4, 5, 14, 34, 203, 235  
Wellengleichung, 3, 7, 14, 41, 243, 250  
Weller (????–????), 54  
Wilson (1931–), 104  
Wilson-Plattenelement, 104  
Wohlgestelltheit, 9, 15  
  
Zeilensummenkriterium  
    schwaches, 58  
Zeitschrittverfahren, 207  
Zeitschrittweite, 204, 230, 243  
Zelle, 84  
Zellgewicht, 146  
Zellresiduum, 145  
Zusammenhang, 57  
Zweigittermethode, 189

Zweigitteroperator, 192  
Zweischrittformel, 243  
Zweischrittverfahren, 204





## **Über dieses Buch**

Dieser einführende Text basiert auf Vorlesungen innerhalb eines mehrsemestrigen Zyklus "Numerische Mathematik", den der Autor über einen Zeitraum von 25 Jahren an der Universität Heidelberg gehalten hat. Im vorliegenden dritten Teil werden numerische Verfahren zur approximativen Lösung partieller Differentialgleichungen behandelt. Dabei finden wieder sowohl theoretisch-mathematische als auch praktische Aspekte Berücksichtigung.

Das Verständnis der Inhalte erfordert neben dem Stoff der ersten beiden Bände „Numerik 0 (Einführung in die Numerische Mathematik)“ und „Numerik 1 (Numerik gewöhnlicher Differentialgleichungen)“ nur solche Vorkenntnisse, wie sie üblicherweise in den Grundvorlesungen über Analysis und Lineare Algebra vermittelt werden. Zur Erleichterung des Selbststudiums dienen wieder theoretische und praktische Übungsaufgaben mit Lösungen.

## **Über den Autor**

Rolf Rannacher, Prof. i. R. für Numerische Mathematik an der Universität Heidelberg; Studium der Mathematik an der Universität Frankfurt am Main – Promotion 1974; Habilitation 1978 in Bonn; 1979/1980 Vis. Assoc. Prof. an der University of Michigan (Ann Arbor, USA), dann Professor in Erlangen und Saarbrücken – in Heidelberg seit 1988; Spezialgebiet „Numerik partieller Differentialgleichungen“, insbesondere „Methode der finiten Elemente“ mit Anwendungen in Natur- und Ingenieurwissenschaften; hierzu über 160 publizierte wissenschaftliche Arbeiten.



**UNIVERSITÄT  
HEIDELBERG**  
ZUKUNFT  
SEIT 1386

ISBN 978-3-946054-38-2

A standard linear barcode representing the ISBN number 978-3-946054-38-2.