

Numerik der partiellen Differentialgleichungen

Skript zur Vorlesung
im
Herbstsemester 2022

Helmut Harbrecht

Stand: 6. Dezember 2024

Vorwort

Diese Mitschrift kann und soll nicht ganz den Wortlaut der Vorlesung wiedergeben. Sie soll einerseits das Nacharbeiten des Inhalts der Vorlesung erleichtern. Andererseits ermöglicht sie auch ein *Vorarbeiten* des Stoffes, so dass der interessierte Student allfällige Fragen bereits in die Vorlesung mitbringen kann. Inhalt, Aufbau und Darstellung der Vorlesung orientieren sich stark an dem unten genannten Buch von Dietrich Braess. Daher sei dieses Buch als weiterführende Lektüre wärmstens empfohlen.

Kapitel, die mit einem Stern markiert sind, beinhalten ergänzendes Material, wobei Kapitel 1 und 2 eine Wiederholung des vorausgesetzten Stoffes zu partiellen Differentialgleichungen aus der Vorlesung *Numerik der Differentialgleichungen* beinhalten.

Hilfreich, aber nicht notwendig, zum Verstehen der Vorlesung sind Kenntnisse aus der Numerischen Mathematik, wie man sie beispielsweise in folgenden Büchern findet:

- M. Hanke-Bourgeois: *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*, Teubner-Verlag
- R. Schaback und H. Wendland: *Numerische Mathematik*, Springer-Verlag
- J. Stoer und R. Bulirsch: *Numerische Mathematik I+II*, Springer-Verlag

Literatur zur Vorlesung:

- D. Braess: *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*, Springer-Verlag
- W. Hackbusch: *Theorie und Numerik elliptischer Differentialgleichungen*, Teubner-Verlag

Inhaltsverzeichnis

1	Partielle Differentialgleichungen*	5
1.1	Beispiele*	5
1.2	Charakterisierung*	8
1.3	Maximumprinzip*	9
2	Finite-Differenzen-Verfahren*	13
2.1	Poisson-Gleichung*	13
2.2	Beliebige Differentialoperatoren*	17
2.3	Diskretes Maximumprinzip*	18
2.4	Konvergenz*	19
3	Variationsformulierung	23
3.1	Sobolev-Räume	23
3.2	Variationsformulierung von Dirichlet-Problemen	28
3.3	Variationsformulierung von Neumann-Problemen	33
4	Galerkin-Verfahren	36
5	Finite Elemente	40
5.1	Vernetzung	40
5.2	Ansatzfunktionen auf Dreieckselementen	41
5.3	Ansatzfunktionen auf Viereckselementen	43
5.4	Dreidimensionaler Fall	44
5.5	Approximationseigenschaften	45
6	Fehleranalyse	54
7	Rechentechnische Betrachtungen	59
8	Mehrgitterverfahren	64
8.1	Glättungseigenschaft von Iterationsverfahren	64
8.2	Prolongation und Restriktion	68
8.3	Zweigitterverfahren	69
8.4	Mehrgitterverfahren	73
8.5	Konvergenz des V-Zyklus	75
8.6	Geschachtelte Iteration	80
9	Residuale Fehlerschätzer	82
9.1	Clément-Operator	82

9.2	A-posteriori-Fehlerschätzung	83
9.3	Untere Abschätzung	85
10	Nichtsymmetrische Bilinearformen	89
11	Parabolische Differentialgleichungen	92
11.1	Linienmethode	92
11.2	θ -Schema	93
11.3	Fehleranalyse	95
12	Lineare Elastizität*	98
12.1	Herleitung	98
12.2	Variationsformulierung	100
12.3	Elliptizitätsabschätzung	101
12.4	Starrkörperbewegungen	102
12.5	Lagrange-Multiplikatoren	104
12.6	Finite-Element-Approximation	105

1. Partielle Differentialgleichungen^{*}

1.1 Beispiele^{*}

Potentialgleichung: Es sei $\Omega \subset \mathbb{R}^2$ ein *Gebiet*, das ist eine offene, zusammenhängende Menge, und $\Gamma := \partial\Omega$ der Rand. Der Graph der Funktion $g : \Gamma \rightarrow \mathbb{R}$ beschreibe eine Drahtschlinge, die eine Seifenhaut aufspannt. Diese Seifenhaut lässt sich als Funktion $u : \bar{\Omega} \rightarrow \mathbb{R}$ beschreiben, deren Form minimale Oberfläche besitzt:

$$\int_{\Omega} \sqrt{1 + u_x^2 + u_y^2} \, dx \, dy \rightarrow \min .$$

Wegen $\sqrt{1+z} = 1 + \frac{z}{2} + \mathcal{O}(z^2)$ kann man den Integranden für kleine Werte von u_x und u_y ersetzen durch

$$F(u) := \frac{1}{2} \int_{\Omega} u_x^2 + u_y^2 \, dx \, dy \rightarrow \min .$$

Ist $u \in C^2(\Omega) \cap C(\bar{\Omega})$ mit $u|_{\Gamma} = g$ Lösung dieser Minimierungsaufgabe, dann folgt für beliebiges $v \in C^1(\Omega) \cap C(\bar{\Omega})$ mit $v|_{\Gamma} = 0$, dass

$$0 = \lim_{\varepsilon \rightarrow 0} \frac{F(u + \varepsilon v) - F(u)}{\varepsilon} = \int_{\Omega} u_x v_x + u_y v_y \, dx \, dy = \int_{\Omega} \langle \nabla u, \nabla v \rangle \, d\mathbf{x}. \quad (1.1)$$

Für $\mathbf{f} := \nabla uv$ liefert der Gaußsche Integralsatz die Identität

$$\int_{\Omega} \Delta uv \, d\mathbf{x} + \int_{\Omega} \langle \nabla u, \nabla v \rangle \, d\mathbf{x} = \int_{\Omega} \operatorname{div} \mathbf{f} \, d\mathbf{x} = \int_{\Gamma} \langle \mathbf{f}, \mathbf{n} \rangle \, d\sigma = \int_{\Gamma} \underbrace{v}_{=0} \frac{\partial u}{\partial \mathbf{n}} \, d\sigma = 0,$$

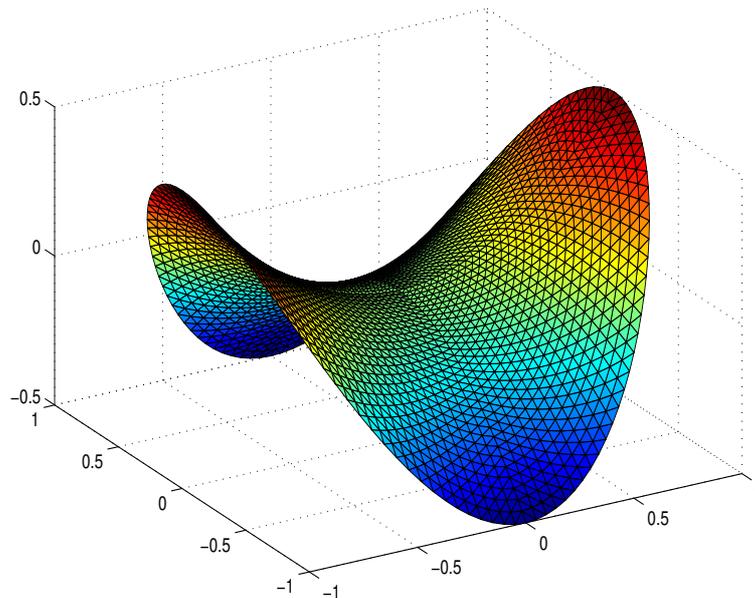
wobei $\Delta u = u_{xx} + u_{yy}$ den *Laplace-Operator* bezeichnet. Dies eingesetzt in (1.1) ergibt für u die Bedingung

$$0 = \int_{\Omega} \Delta uv \, d\mathbf{x}$$

für alle $v \in C^1(\Omega) \cap C(\bar{\Omega})$ mit $v|_{\Gamma} = 0$. Daher muss die Funktion u der *Potential-* oder *Laplace-Gleichung*

$$\Delta u(\mathbf{x}) = 0, \quad \mathbf{x} \in \Omega \quad (1.2)$$

genügen. Die Lösung zum Dirichletschen Problem der Laplace-Gleichung sieht wie folgt aus:



Eine einfache Lösungsformel für die Laplace-Gleichung gibt es im Fall des Kreises $\Omega = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_2 < 1\}$. Bei Einführung von Polarkoordinaten $x = r \cos \phi$, $y = r \sin \phi$ erkennt man, dass die Funktionen

$$r^k \cos(k\phi), \quad r^k \sin(k\phi), \quad k = 0, 1, \dots$$

der Potentialgleichung genügen. Entwickelt man die Randwerte in eine Fourier-Reihe

$$u(\cos \phi, \sin \phi) = a_0 + \sum_{k=0}^{\infty} \{a_k \cos(k\phi) + b_k \sin(k\phi)\},$$

so lässt sich die Lösung im Innern gemäß

$$u(x, y) = a_0 + \sum_{k=0}^{\infty} r^k \{a_k \cos(k\phi) + b_k \sin(k\phi)\}$$

darstellen.

Wärmeleitungsgleichung: In einem offenen, beschränkten Gebiet $\Omega \subset \mathbb{R}^d$ beschreibe die Funktion $u : \mathbb{R}_{\geq 0} \times \bar{\Omega} \rightarrow \mathbb{R}$ die Temperaturverteilung. Zum Zeitpunkt $t = 0$ liege die Anfangsverteilung $u(0, \mathbf{x}) = u_0(\mathbf{x}) \in C(\bar{\Omega})$ vor. Zusätzlich seien im Gebiet Ω die Wärmequelle $f \in C(\mathbb{R}_{>0} \times \Omega)$ und an dessen Rand $\Gamma = \partial\Omega$ die Temperaturverteilung $g \in C(\mathbb{R}_{>0} \times \Gamma)$ vorgegeben.

Aus dem Erhaltungssatz folgt nun für jedes Kontrollvolumen $V \subset \Omega$

$$\underbrace{\int_V \frac{\partial}{\partial t} u(t, \mathbf{x}) \, d\mathbf{x}}_{\text{Wärmegehalt in } V} = - \underbrace{\int_{\partial V} \langle \mathbf{q}(t, \mathbf{x}), \mathbf{n}(\mathbf{x}) \rangle \, d\sigma}_{\text{Wärmefluss von außen}} + \underbrace{\int_V f(t, \mathbf{x}) \, d\mathbf{x}}_{\text{Wärmequelle}}$$

Dem Materialgesetz gemäß genügt der Wärmefluss der Beziehung

$$\mathbf{q}(t, \mathbf{x}) = -c(\mathbf{x}) \nabla u(t, \mathbf{x})$$

mit der materialabhängigen Wärmeleitkonstante $c \geq c_0 > 0$. Eingesetzt in den Gaußschen Integralsatz folgt daher

$$-\int_{\partial V} \langle \mathbf{q}(t, \mathbf{x}), \mathbf{n}(\mathbf{x}) \rangle d\sigma = -\int_V \operatorname{div} \mathbf{q}(t, \mathbf{x}) d\mathbf{x} = \int_V \operatorname{div}(c(\mathbf{x})\nabla u(t, \mathbf{x})) d\mathbf{x}.$$

Für die Temperaturverteilung folgt somit für alle Kontrollvolumen V die Gleichung

$$\int_V \left\{ \frac{\partial}{\partial t} u(t, \mathbf{x}) - \operatorname{div}(c(\mathbf{x})\nabla u(t, \mathbf{x})) \right\} d\mathbf{x} = \int_V f(t, \mathbf{x}) d\mathbf{x},$$

dies bedeutet

$$\frac{\partial}{\partial t} u(t, \mathbf{x}) - \operatorname{div}(c(\mathbf{x})\nabla u(t, \mathbf{x})) = f(t, \mathbf{x}), \quad (t, \mathbf{x}) \in \mathbb{R}_{>0} \times \Omega.$$

Ist c konstant, etwa $c = 1$, so genügt die Temperaturverteilung $u \in C^2(\mathbb{R}_{>0} \times \Omega) \cap C(\mathbb{R}_{\geq 0} \times \bar{\Omega})$ der Gleichung

$$\frac{\partial}{\partial t} u(t, \mathbf{x}) - \Delta u(t, \mathbf{x}) = f(t, \mathbf{x}), \quad (t, \mathbf{x}) \in \mathbb{R}_{\geq 0} \times \Omega \quad (1.3)$$

mit dem d -dimensionalen Laplace-Operator $\Delta = \frac{\partial^2}{\partial x_1^2} + \dots + \frac{\partial^2}{\partial x_d^2}$.

Poisson-Gleichung: Sind die Daten f und g der Wärmeleitungsgleichung nicht zeitabhängig, dann stellt sich für $t \rightarrow \infty$ ein Gleichgewichtszustand ein. Dies bedeutet, es gilt $\partial u / \partial t = 0$ und (1.3) geht über in die *Poisson-Gleichung*

$$-\Delta u(\mathbf{x}) = f(\mathbf{x}), \quad \mathbf{x} \in \Omega. \quad (1.4)$$

Diese Gleichung hat in der Elektrostatik ebenfalls eine immense Bedeutung: ist in Ω die Ladungsdichte $f : \Omega \rightarrow \mathbb{R}$ bekannt, so genügt die Spannung u dort der Poisson-Gleichung.

Wellengleichung: Die Bewegung in einem idealen Gas wird durch drei Gesetze bestimmt. Wie üblich wird die Geschwindigkeit mit \mathbf{v} , die Dichte mit ρ und der Druck mit p bezeichnet.

1. *Kontinuitätsgleichung:*

$$\frac{\partial \rho}{\partial t} = -\rho_0 \operatorname{div} \mathbf{v}.$$

Wegen der Massenerhaltung ist die Änderung der Masse in einem Kontrollvolumen V gleich dem Fluss durch die Oberfläche, das ist $\int_{\partial V} \rho \langle \mathbf{v}, \mathbf{n} \rangle d\sigma$. Aus dem Gaußschen Integralsatz folgt daraus die Gleichung $\partial \rho / \partial t = -\operatorname{div}(\rho \mathbf{v})$. Die Approximation von ρ durch eine konstante, zeitlich unabhängige Dichte ρ_0 ergibt dann die obige Gleichung.

2. *Newtonsches Gesetz:*

$$\rho_0 \frac{\partial \mathbf{v}}{\partial t} = -\nabla p.$$

Der Druckgradient induziert ein Kraftfeld, das die Beschleunigung der Teilchen bewirkt.

3. *Zustandsgleichung:*

$$p = c^2 \rho.$$

In idealen Gasen ist der Druck bei konstanter Temperatur proportional zur Dichte. Aus den drei Gesetzen folgt

$$\frac{\partial^2 p}{\partial t^2} = c^2 \frac{\partial^2 \rho}{\partial t^2} = -c^2 \frac{\partial}{\partial t} (\rho_0 \operatorname{div} \mathbf{v}) = -c^2 \operatorname{div} \left(\rho_0 \frac{\partial \mathbf{v}}{\partial t} \right) = c^2 \operatorname{div}(\nabla p) = c^2 \Delta p.$$

Andere Beispiele für die *Wellengleichung*

$$\frac{\partial^2 p}{\partial t^2}(t, \mathbf{x}) = c^2 \Delta p(t, \mathbf{x}), \quad (t, \mathbf{x}) \in \mathbb{R}_{>0} \times \Omega \quad (1.5)$$

ergeben sich in zwei Raumdimensionen für eine schwingende Membran oder in einer Raumdimension für eine schwingende Saite.

1.2 Charakterisierung*

Sei $\Omega \subset \mathbb{R}^d$ ein Gebiet und $\mathcal{L} : C^2(\Omega) \rightarrow C(\Omega)$ ein allgemeiner linearer Differentialoperator zweiter Ordnung

$$(\mathcal{L}u)(\mathbf{x}) = - \sum_{i,j=1}^d a_{i,j}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j} u(\mathbf{x}) + \sum_{i=1}^d b_i(\mathbf{x}) \frac{\partial}{\partial x_i} u(\mathbf{x}) + c(\mathbf{x})u(\mathbf{x}), \quad (1.6)$$

wobei $\mathbf{A} = [a_{i,j}]_{i,j=1}^d \in [C(\Omega)]^{d \times d}$, $\mathbf{b} = [b_i]_{i=1}^d \in [C(\Omega)]^d$ und $c \in C(\Omega)$. Die zugehörige Differentialgleichung lautet dann

$$(\mathcal{L}u)(\mathbf{x}) = f(\mathbf{x}), \quad \mathbf{x} \in \Omega. \quad (1.7)$$

Da für $u \in C^2(\Omega)$ die zweiten Ableitungen symmetrisch sind, also $\partial^2 u / (\partial x_i \partial x_j) = \partial^2 u / (\partial x_j \partial x_i)$, kann ohne Beschränkung der Allgemeinheit $a_{i,j} = a_{j,i}$ angenommen werden. Demnach ist die Matrix \mathbf{A} symmetrisch und besitzt nur reelle Eigenwerte. Der Differentialoperator $-\sum_{i,j=1}^d a_{i,j} \partial^2 / (\partial x_i \partial x_j)$ ist der *Hauptteil* von \mathcal{L} .

Definition 1.1 Der Differentialoperator (1.6) heißt

- **elliptisch in \mathbf{x}** , falls die Eigenwerte von $\mathbf{A}(\mathbf{x})$ alle positiv sind,
- **parabolisch in \mathbf{x}** , falls $d-1$ Eigenwerte von $\mathbf{A}(\mathbf{x})$ alle positiv sind und ein Eigenwert verschwindet, aber $\operatorname{rang}([\mathbf{A}(\mathbf{x}), \mathbf{b}(\mathbf{x})]) = d$ ist,
- **hyperbolisch in \mathbf{x}** , falls $d-1$ Eigenwerte von $\mathbf{A}(\mathbf{x})$ alle positiv sind und ein Eigenwert negatives Vorzeichen besitzt.

Der Differentialoperator (1.6) heißt **elliptisch/parabolisch/hyperbolisch (in Ω)**, falls er elliptisch/parabolisch/hyperbolisch ist für alle $\mathbf{x} \in \Omega$. Entsprechend wird die Differentialgleichung (1.7) elliptisch/parabolisch/hyperbolisch genannt, wenn der zugehörige Differentialoperator diese Eigenschaft besitzt.

Beispiel 1.2 Die Potentialgleichung (1.2) und die Poisson-Gleichung (1.4) sind elliptisch, die Wärmeleitungsgleichung (1.3) ist parabolisch, während die Wellengleichung (1.5) hyperbolisch ist. \triangle

Zusätzlich zur Differentialgleichung (1.7) müssen noch geeignete Anfangs- oder Randbedingungen gefordert werden, um eine sachgemäße Aufgabenstellung zu ergeben.

Definition 1.3 Ein Problem heißt **sachgemäß gestellt**, wenn eine Lösung existiert, diese eindeutig ist und stetig von den vorgegebenen Daten abhängt. Andernfalls heißt das Problem **schlecht gestellt**.

Die Unterscheidung partieller Differentialgleichungen in verschiedene Typen ergäbe keinen Sinn, wenn nicht jeder Typ grundlegend andere Eigenschaften hätte.

1. *Elliptische Differentialgleichungen*: Bei elliptischen Problemen werden Randbedingungen vorgegeben: für gegebenes $f \in C(\Omega)$ und $g \in C(\Gamma)$ suche $u \in C^2(\Omega) \cap C(\overline{\Omega})$, so dass

$$\mathcal{L}u = f \text{ in } \Omega, \quad u = g \text{ auf } \Gamma. \quad (1.8)$$

Diese Randbedingungen heißen *Dirichlet-Randbedingungen*. In der Praxis treten oft auch *Neumann-Randbedingungen*, $\partial u / \partial \mathbf{n} = g$ auf Γ , auf. Lösungen elliptischer Differentialgleichungen erfüllen das Maximumprinzip (siehe nächster Abschnitt).

2. *Parabolische Differentialgleichungen*: Parabolische Differentialgleichungen beschreiben Diffusionsvorgänge. Die ausgezeichnete Koordinatenrichtung ist in der Regel die Zeit, so dass man oftmals die Differentialgleichung auf die Form $u_t + \mathcal{L}u = f$ bringen kann, wobei \mathcal{L} ein elliptischer Differentialoperator ist. Zusätzlich werden Anfangsrandwerte vorgegeben: für gegebenes $f \in C(\mathbb{R}_{>0} \times \Omega)$, $g \in C(\mathbb{R}_{>0} \times \Gamma)$ und $u_0 \in C(\overline{\Omega})$ suche $u \in C^2(\mathbb{R}_{>0} \times \Omega) \cap C(\mathbb{R}_{\geq 0} \times \overline{\Omega})$, so dass

$$\begin{aligned} u_t + \mathcal{L}u &= f \text{ in } \mathbb{R}_{>0} \times \Omega \\ u &= g \text{ auf } \mathbb{R}_{>0} \times \Gamma \quad (\text{Randbedingung}) \\ u(0, \cdot) &= u_0 \text{ auf } \overline{\Omega} \quad (\text{Anfangsbedingung}) \end{aligned}$$

3. *Hyperbolische Differentialgleichungen*: Hier ist ebenfalls eine Koordinate ausgezeichnet, die wieder als Zeit interpretiert werden kann. Daher lässt sich die Differentialgleichung oft schreiben als $u_{tt} + \mathcal{L}u = f$ mit einem elliptischen Differentialoperator \mathcal{L} . Hyperbolische Gleichungen beschreiben physikalisch gesehen Schwingungsvorgänge. Sinnvolle Probleme erhält man mit Anfangsbedingungen: für gegebenes $f \in C(\mathbb{R}_{>0} \times \Omega)$, $g \in C(\mathbb{R}_{>0} \times \Gamma)$ und $u_0, u_1 \in C(\overline{\Omega})$ suche $u \in C^2(\mathbb{R}_{>0} \times \Omega) \cap C(\mathbb{R}_{\geq 0} \times \overline{\Omega})$, so dass

$$\begin{aligned} u_{tt} + \mathcal{L}u &= f \text{ in } \mathbb{R}_{>0} \times \Omega \\ u &= g \text{ auf } \mathbb{R}_{>0} \times \Gamma \quad (\text{Randbedingung}) \\ u(0, \cdot) &= u_0, \quad u_t(0, \cdot) = u_1 \text{ auf } \overline{\Omega} \quad (\text{Anfangsbedingungen}) \end{aligned}$$

Wenn der Differentialoperator invariant gegenüber Bewegungen ist (also gegenüber Translation und Drehung), dann hat der elliptische Anteil \mathcal{L} die Form

$$\mathcal{L}u = -a\Delta u + cu.$$

1.3 Maximumprinzip*

Bei der Analyse von Differenzenverfahren spielt das diskrete Analogon des Maximumprinzips eine wichtige Rolle. Deshalb betrachten wir vorab eine einfache Fassung des Prinzips.

Dazu seien $\Omega \subset \mathbb{R}^d$ stets ein beschränktes Gebiet und der elliptische Differentialoperator von der Form

$$(\mathcal{L}u)(\mathbf{x}) = - \sum_{i,j=1}^d a_{i,j}(\mathbf{x}) u_{x_i, x_j}(\mathbf{x}). \quad (1.9)$$

Satz 1.4 (Maximumprinzip) Die Funktion $u \in C^2(\Omega) \cap C(\overline{\Omega})$ genüge der elliptischen Differentialgleichung $\mathcal{L}u = f \leq 0$ in Ω . Dann nimmt u sein Maximum auf dem Rand Γ an.

Beweis. (i) Wir führen den Beweis zunächst unter der stärkeren Voraussetzung $f < 0$. Angenommen, es sei $\mathbf{y} \in \Omega$ mit

$$u(\mathbf{y}) = \sup_{\mathbf{x} \in \Omega} u(\mathbf{x}) > \max_{\mathbf{x} \in \Gamma} u(\mathbf{x}).$$

Bei einer linearen Koordinatentransformation $\mathbf{x} \mapsto \boldsymbol{\xi} = \mathbf{U}\mathbf{x}$ lautet der Differentialoperator in den neuen Koordinaten

$$(\mathcal{L}u)(\mathbf{x}) = - \sum_{i,j=1}^d [\mathbf{U}\mathbf{A}(\mathbf{x})\mathbf{U}^T]_{i,j} u_{\xi_i, \xi_j}(\mathbf{x}),$$

wobei $\mathbf{A}(\mathbf{x}) = [a_{i,j}(\mathbf{x})]_{i,j=1}^d$ die Koeffizientenmatrix ist. Wegen der Symmetrie von $\mathbf{A}(\mathbf{x})$ können wir eine orthogonale Matrix \mathbf{U} wählen, mit der $\mathbf{U}\mathbf{A}(\mathbf{y})\mathbf{U}^T$ diagonal wird. Aus der positiven Definitheit schließen wir, dass die Diagonalelemente positiv sind. Weil \mathbf{y} Extrempunkt ist, gilt

$$\nabla u(\mathbf{y}) = \mathbf{0}, \quad u_{\xi_i, \xi_i}(\mathbf{y}) \leq 0.$$

Dies bedeutet

$$(\mathcal{L}u)(\mathbf{y}) = - \sum_{i,j=1}^d [\mathbf{U}\mathbf{A}(\mathbf{y})\mathbf{U}^T]_{i,j} u_{\xi_i, \xi_j}(\mathbf{y}) \geq 0$$

im Widerspruch zu $(\mathcal{L}u)(\mathbf{y}) = f(\mathbf{y}) < 0$.

(ii) Sei nun $f \leq 0$ angenommen und es gebe ein $\mathbf{y} \in \Omega$ mit $u(\mathbf{y}) > \max_{\mathbf{x} \in \Gamma} u(\mathbf{x})$. Die Hilfsfunktion

$$h(\mathbf{x}) := \|\mathbf{x} - \mathbf{y}\|_2^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_d - y_d)^2$$

ist auf Γ beschränkt. Wenn $\delta > 0$ hinreichend klein gewählt wird, nimmt also auch die Funktion

$$w = u + \delta h$$

ihr Maximum in einem Punkt \mathbf{z} im Innern an. Wegen $h_{x_i, x_j} = 2\delta_{i,j}$ ist

$$(\mathcal{L}w)(\mathbf{x}) = (\mathcal{L}u)(\mathbf{x}) + \delta(\mathcal{L}h)(\mathbf{x}) = f(\mathbf{x}) - 2\delta \sum_{i=1}^d a_{i,i}(\mathbf{x}) < 0$$

für alle $\mathbf{x} \in \Omega$. Wie im ersten Teil des Beweises ergibt sich daraus ein Widerspruch. \square

Folgerungen:

1. *Minimumprinzip:* Ist $\mathcal{L}u = f \geq 0$ in Ω , so nimmt u sein Minimum auf dem Rand Γ an.

Beweis. Man wende auf $v := -u$ das Maximumprinzip an. \square

2. *Vergleichsprinzip:* Wenn für $u, v \in C^2(\Omega) \cap C(\overline{\Omega})$ gilt

$$\mathcal{L}u \leq \mathcal{L}v \text{ in } \Omega, \quad u \leq v \text{ auf } \Gamma,$$

so folgt $u \leq v$ in Ω .

Beweis. Für $w := v - u$ ist nach Voraussetzung $\mathcal{L}w = \mathcal{L}v - \mathcal{L}u \geq 0$ und auf Γ auch $w \geq 0$. Nach dem Minimumprinzip folgt $\inf_{\mathbf{x} \in \Omega} w(\mathbf{x}) \geq 0$ und folglich $v(\mathbf{x}) \geq u(\mathbf{x})$ für alle $\mathbf{x} \in \Omega$. \square

3. *Eindeutigkeit der Lösung:* Die Lösung des Dirichlet-Problems (1.8) ist eindeutig.

Beweis. Seien u_1 und u_2 zwei Lösungen von (1.8), dann erfüllt $v = u_1 - u_2$ die Gleichung

$$\mathcal{L}v = 0 \text{ in } \Omega, \quad v = 0 \text{ auf } \Gamma.$$

Minimum- und Maximumprinzip implizieren

$$0 = \inf_{\mathbf{z} \in \Omega} v(\mathbf{z}) \leq v(\mathbf{x}) \leq \sup_{\mathbf{z} \in \Omega} v(\mathbf{z}) = 0, \quad \mathbf{x} \in \Omega.$$

\square

4. *Stetige Abhängigkeit von den Randdaten:* Die Lösung des Dirichlet-Problems (1.8) hängt stetig von den Randdaten ab. Sind u_1 und u_2 Lösungen zu verschiedenen Randwerten, so ist

$$\max_{\mathbf{x} \in \overline{\Omega}} |u_1(\mathbf{x}) - u_2(\mathbf{x})| = \max_{\mathbf{x} \in \Gamma} |u_1(\mathbf{x}) - u_2(\mathbf{x})|.$$

Beweis. Für $v := u_1 - u_2$ ist $\mathcal{L}v = 0$. Aus dem Maximumprinzip folgt

$$v(\mathbf{x}) \leq \max_{\mathbf{z} \in \Gamma} v(\mathbf{z}) \leq \max_{\mathbf{z} \in \Gamma} |v(\mathbf{z})|, \quad \mathbf{x} \in \Omega.$$

Ebenso liefert das Minimumprinzip die Aussage

$$v(\mathbf{x}) \geq \min_{\mathbf{z} \in \Gamma} v(\mathbf{z}) \geq -\max_{\mathbf{z} \in \Gamma} |v(\mathbf{z})|, \quad \mathbf{x} \in \Omega.$$

\square

Definition 1.5 Ein linearer Differentialoperator \mathcal{L} zweiter Ordnung heißt **gleichmäßig elliptisch**, wenn ein $\alpha > 0$ existiert, so dass die Koeffizientenmatrix $\mathbf{A}(\mathbf{x}) = [a_{i,j}(\mathbf{x})]_{i,j=1}^d$ der Abschätzung

$$\boldsymbol{\xi}^T \mathbf{A}(\mathbf{x}) \boldsymbol{\xi} \geq \alpha \|\boldsymbol{\xi}\|_2^2$$

für alle $\boldsymbol{\xi} \in \mathbb{R}^d$ und $\mathbf{x} \in \Omega$ genügt. Die Zahl α wird als **Elliptizitätskonstante** bezeichnet.

5. *Stetige Abhängigkeit von der rechten Seite:* Der Operator \mathcal{L} der Form (1.9) sei gleichmäßig elliptisch in Ω . Dann gibt es eine nur von Ω und der Elliptizitätskonstante α abhängige Zahl c , so dass für jedes $u \in C^2(\Omega) \cap C(\overline{\Omega})$ gilt

$$|u(\mathbf{x})| \leq \max_{\mathbf{z} \in \Gamma} |u(\mathbf{z})| + c \sup_{\mathbf{z} \in \Omega} |(\mathcal{L}u)(\mathbf{z})|, \quad \mathbf{x} \in \Omega.$$

Beweis. Sei $\Omega \subset \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 < R\}$ und setze

$$w(\mathbf{x}) = R^2 - \sum_{i=1}^d x_i^2.$$

Im Hinblick auf $w_{x_i, x_j} = -2\delta_{i,j}$ ist offensichtlich

$$\mathcal{L}w \geq 2\alpha \quad \text{und} \quad 0 \leq w \leq R^2 \quad \text{in } \Omega,$$

wobei α die Elliptizitätskonstante ist. Für

$$v(\mathbf{x}) := \max_{\mathbf{z} \in \Gamma} |u(\mathbf{z})| + w(\mathbf{x}) \frac{1}{2\alpha} \sup_{\mathbf{z} \in \Omega} |(\mathcal{L}u)(\mathbf{z})|$$

ist nach Konstruktion $\mathcal{L}v \geq |\mathcal{L}u|$ in Ω und $v \geq |u|$ auf Γ . Das Vergleichsprinzip liefert $-v(\mathbf{x}) \leq u(\mathbf{x}) \leq v(\mathbf{x})$ für alle $\mathbf{x} \in \Omega$ und wegen $w \leq R^2$ erhalten wir die gewünschte Abschätzung mit $c = R^2/(2\alpha)$. \square

6. *Elliptische Operatoren mit Term der Ordnung 0:* Für den allgemeineren Differentialoperator

$$(\mathcal{L}u)(\mathbf{x}) = c(\mathbf{x})u(\mathbf{x}) - \sum_{i,j=1}^d a_{i,j}(\mathbf{x})u_{x_i, x_j}(\mathbf{x}) \quad \text{mit} \quad c(\mathbf{x}) \geq 0$$

gilt ein abgeschwächtes Maximumprinzip. Aus $\mathcal{L}u \leq 0$ folgt

$$\max_{\mathbf{x} \in \overline{\Omega}} u(\mathbf{x}) \leq \max\{0, \max_{\mathbf{x} \in \Gamma} u(\mathbf{x})\}.$$

Beweis. Ein Beweis ist nur für $\mathbf{y} \in \Omega$ und $u(\mathbf{y}) = \sup_{\mathbf{x} \in \Omega} u(\mathbf{x}) > 0$ erforderlich. Dann ist $(\mathcal{L}u)(\mathbf{y}) - c(\mathbf{y})u(\mathbf{y}) \leq (\mathcal{L}u)(\mathbf{y}) \leq 0$. Außerdem ist durch den Hauptteil $\mathcal{L}u - cu$ ein elliptischer Operator der Form (1.9) definiert. Deshalb kann der Beweis wie für Satz 1.4 vollzogen werden. \square

2. Finite-Differenzen-Verfahren^{*}

2.1 Poisson-Gleichung^{*}

Im folgenden wollen wir uns auf die Poisson-Gleichung beschränken. Dazu seien $\Omega \subset \mathbb{R}^d$ ein beschränktes Gebiet, $f \in C(\Omega)$ und $g \in C(\Gamma)$. Gesucht ist $u \in C^2(\Omega) \cap C(\overline{\Omega})$, so dass

$$-\Delta u = f \text{ in } \Omega, \quad u = g \text{ auf } \Gamma.$$

Definition 2.1 Eine Lösung $u \in C^2(\Omega) \cap C(\overline{\Omega})$ der Poisson-Gleichung ist eine **klassische Lösung**. Gilt speziell $f = 0$, das heißt, ist $\Delta u = 0$ in Ω , so ist u **harmonisch**.

Wir werden mit Lösung stets die klassische Lösung meinen. Um diese zu berechnen, benötigen wir finite Differenzen:

Definition 2.2 Für $u \in C(\mathbb{R}^d)$ und eine Richtung $1 \leq j \leq d$ definieren wir die **Vorwärts-** oder **rechtsseitige Differenz** durch

$$(\partial_j^{+h}u)(\mathbf{x}) := \frac{u(\mathbf{x} + h\mathbf{e}_j) - u(\mathbf{x})}{h},$$

die **Rückwärts-** oder **linksseitige Differenz** durch

$$(\partial_j^{-h}u)(\mathbf{x}) := \frac{u(\mathbf{x}) - u(\mathbf{x} - h\mathbf{e}_j)}{h}$$

und die **symmetrische** oder **zentrale Differenz** durch

$$(\partial_j^h u)(\mathbf{x}) := \frac{u(\mathbf{x} + h\mathbf{e}_j) - u(\mathbf{x} - h\mathbf{e}_j)}{2h}.$$

Lemma 2.3 Ist $\{\mathbf{x} + t\mathbf{e}_j : |t| \leq 1\} \subset \overline{\Omega}$ und $u \in C^4(\overline{\Omega})$, dann gilt

$$\begin{aligned}\frac{\partial u}{\partial \mathbf{e}_j}(\mathbf{x}) &= (\partial_j^{\pm h} u)(\mathbf{x}) + R_1^{\pm}, & |R_1^{\pm}| &\leq \frac{h}{2} \|u\|_{C^2(\overline{\Omega})}, \\ \frac{\partial u}{\partial \mathbf{e}_j}(\mathbf{x}) &= (\partial_j^h u)(\mathbf{x}) + R_2, & |R_2| &\leq \frac{h^2}{6} \|u\|_{C^3(\overline{\Omega})},\end{aligned}$$

und

$$\begin{aligned}\frac{\partial^2 u}{\partial \mathbf{e}_j^2}(\mathbf{x}) &= (\partial_j^{-h} \partial_j^{+h} u)(\mathbf{x}) + R_3 \\ &= \frac{u(\mathbf{x} + h\mathbf{e}_j) - 2u(\mathbf{x}) + u(\mathbf{x} - h\mathbf{e}_j)}{h^2} + R_3, & |R_3| &\leq \frac{h^2}{12} \|u\|_{C^4(\overline{\Omega})}.\end{aligned}$$

Beweis. Es genügt, die Behauptung im Eindimensionalen zu beweisen. Taylor-Entwicklung von u liefert

$$u(x \pm h) = u(x) \pm hu'(x) + \frac{h^2}{2} u''(\xi), \quad \xi \in (x, x \pm h),$$

woraus sofort die erste Aussage folgt. Subtrahieren wir ferner

$$\begin{aligned}u(x - h) &= u(x) - hu'(x) + \frac{h^2}{2} u''(x) - \frac{h^3}{6} u'''(\xi_1), & \xi_1 &\in (x - h, x), \\ u(x + h) &= u(x) + hu'(x) + \frac{h^2}{2} u''(x) + \frac{h^3}{6} u'''(\xi_2), & \xi_2 &\in (x, x + h),\end{aligned}$$

so folgt die zweite Aussage

$$u(x + h) - u(x - h) = 2hu'(x) + \frac{h^3}{6} (u'''(\xi_2) + u'''(\xi_1)).$$

Schließlich folgt aus Addition der drei Gleichungen

$$\begin{aligned}u(x - h) &= u(x) - hu'(x) + \frac{h^2}{2} u''(x) - \frac{h^3}{6} u'''(x) + \frac{h^4}{24} u^{(4)}(\xi_1), & \xi_1 &\in (x - h, x), \\ -2u(x) &= -2u(x), \\ u(x + h) &= u(x) + hu'(x) + \frac{h^2}{2} u''(x) + \frac{h^3}{6} u'''(x) + \frac{h^4}{24} u^{(4)}(\xi_2), & \xi_2 &\in (x, x + h),\end{aligned}$$

dass

$$\frac{u(x + h) - 2u(x) + u(x - h)}{h^2} = u''(x) + \frac{h^2}{24} (u^{(4)}(\xi_1) + u^{(4)}(\xi_2)).$$

□

Zur Diskretisierung wird über das Gebiet Ω ein *Gitter* mit Maschenweite h gelegt

$$\begin{aligned}\Omega_h &:= \{\mathbf{x} \in \Omega : \mathbf{x} = h\mathbf{k} \text{ mit } \mathbf{k} \in \mathbb{Z}^d\}, \\ \Gamma_h &:= \{\mathbf{x} \in \Gamma : \exists 1 \leq i \leq d \text{ mit } x_i = hk, k \in \mathbb{Z}\}.\end{aligned}$$

In Anlehnung an $\overline{\Omega} = \Omega \cup \Gamma$ setzen wir $\overline{\Omega}_h := \Omega_h \cup \Gamma_h$. Punkte aus Γ_h werden *Randpunkte* genannt. Ein Gitterpunkt $\mathbf{x} \in \Omega_h$, der einen Nachbarn aus Γ_h besitzt, heißt *randnah*. Alle

anderen Punkte aus Ω_h sind *randfern*. Ist $\bar{\Omega}$ die Vereinigung von Würfeln der Kantenlänge h , so sprechen wir von einem *Würfelgebiet*. In diesem Fall besitzen dann auch alle Rand- und randnahe Punkte immer den Abstand h zu ihren Nachbarn.

In den Randpunkten \mathbf{x} aus Γ_h ist $u(\mathbf{x})$ durch die Randwerte $g(\mathbf{x})$ vorgegeben. Hingegen erhält man für jeden Punkt \mathbf{x} aus Ω_h eine Gleichung für $u(\mathbf{x})$, indem man die Poisson-Gleichung durch Differenzenquotienten approximiert. In jedem randfernen Gitterpunkt \mathbf{x} diskretisieren wir den Laplace-Operator durch $(\Delta_h u)(\mathbf{x}) := \sum_{i=1}^d (\partial_i^{-h} \partial_i^{+h} u)(\mathbf{x})$, wobei sich

$$(\Delta u)(\mathbf{x}) = \sum_{i=1}^d (\partial_i^{-h} \partial_i^{+h} u)(\mathbf{x}) + \mathcal{O}(h^2)$$

ergibt. Für $d = 2$ erhält man den sogenannten *5-Punkte-Differenzenstern*

$$\begin{bmatrix} \alpha_{NW} & \alpha_N & \alpha_{NO} \\ \alpha_W & \alpha_Z & \alpha_O \\ \alpha_{SW} & \alpha_S & \alpha_{SO} \end{bmatrix}_* = \frac{1}{h^2} \begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}_*$$

Dieser ist für Würfelgebiete ausreichend.

In beliebigen Gebieten muss der Differenzenquotient für randnahe Punkte entsprechend modifiziert werden. Für $u \in C^3(\bar{\Omega})$ erhält man durch Taylor-Entwicklung in einer Dimension

$$u_{xx} = \frac{2}{h_O(h_O + h_W)} u_O - \frac{2}{h_O h_W} u_Z + \frac{2}{h_W(h_O + h_W)} u_W + \mathcal{O}(h)$$

und in zwei Dimensionen

$$\begin{aligned} \Delta u = \Delta_h u + \mathcal{O}(h) &= \frac{2}{h_O(h_O + h_W)} u_O + \frac{2}{h_W(h_O + h_W)} u_W + \frac{2}{h_S(h_S + h_N)} u_S \\ &+ \frac{2}{h_N(h_S + h_N)} u_N - \left(\frac{2}{h_O h_W} + \frac{2}{h_S h_N} \right) u_Z + \mathcal{O}(h). \end{aligned}$$

Hierbei bezeichnet h die jeweils größte Schrittweite, das heißt, $h := \max\{h_W, h_O\}$ beziehungsweise $h := \max\{h_W, h_O, h_S, h_N\}$. Diese Diskretisierung des Laplace-Operators wird auch *Shortley-Weller-Approximation* genannt.

Beispiel 2.4 *Eindimensionaler Fall:* Sei

$$-u_{xx} = f \text{ in } (a, b), \quad u(a) = \alpha, \quad u(b) = \beta.$$

Für $h = (b - a)/n$ und $x_i = a + hi$, $i = 1, \dots, n - 1$, setzen wir $u_i = u(x_i)$ und $f_i = f(x_i)$. Dann erhalten wir das lineare Gleichungssystem

$$\frac{1}{h^2} \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-2} \\ u_{n-1} \end{bmatrix} = \begin{bmatrix} f_1 + \alpha/h^2 \\ f_2 \\ \vdots \\ f_{n-2} \\ f_{n-1} + \beta/h^2 \end{bmatrix}.$$

Zweidimensionales Würfelgebiet: Zur Lösung der Poisson-Gleichung im Einheitsquadrat

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

überziehen wir Ω mit einem Gitter der Maschenweite $h = 1/n$. Das entstehende Gleichungssystem wird übersichtlicher bei Benutzung von Doppelindizes $u_{i,j} = u(ih, jh)$, $1 \leq i, j < n$. Es ergibt sich das Gleichungssystem

$$4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} = f_{i,j}, \quad 1 \leq i, j < n$$

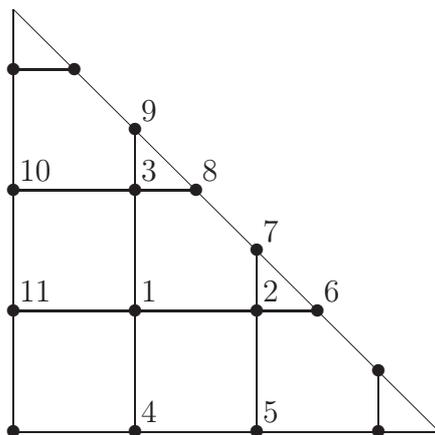
mit $f_{i,j} = h^2 f(ih, jh)$. Die Terme mit Indizes 0 oder n gelten als nicht geschrieben.

Im Fall von Würfelgebieten ist die Systemmatrix immer symmetrisch. Ordnet man die Indizes schachbrettartig an (zum Beispiel durch abwechselndes rotes und schwarzes Einfärben der zugehörigen Gitterpunkte), so erhält man ein Gleichungssystem der Form

$$\left[\begin{array}{c|c} 4 & \\ \cdots & \mathbf{A} \\ & 4 \end{array} \right] \mathbf{u} = \mathbf{f}.$$

$$\left[\begin{array}{c|c} \mathbf{A}^T & \\ & \cdots \\ & 4 \end{array} \right]$$

Beliebiges zweidimensionales Gebiet: Sei Ω ein rechtwinkliges gleichschenkliges Dreieck mit Katheten der Länge 7:



Zu lösen sei die Laplace-Gleichung mit Dirichlet-Randbedingungen. Für $h = 2$ enthält Ω_h drei Punkte. Es entsteht ein lineares Gleichungssystem für u_1 , u_2 und u_3 :

$$\begin{aligned} u_1 - \frac{u_2}{4} - \frac{u_3}{4} &= \frac{u_4}{4} + \frac{u_{11}}{4} \\ -\frac{u_1}{6} + u_2 &= \frac{u_5}{6} + \frac{u_6}{3} + \frac{u_7}{3} \\ -\frac{u_1}{6} + u_3 &= \frac{u_8}{3} + \frac{u_9}{3} + \frac{u_{10}}{6}. \end{aligned}$$

Man beachte, dass das System unsymmetrisch ist! △

Zusammengefasst erhalten wir demnach ein *Differenzenverfahren* für die näherungsweise Lösung des Poisson-Problems: suche eine *Gitterfunktion* $u_h : \bar{\Omega}_h \rightarrow \mathbb{R}$, so dass

$$\begin{aligned} -(\Delta_h u_h)(\mathbf{x}) &= f(\mathbf{x}) & \text{für alle } \mathbf{x} \in \Omega_h, \\ u_h(\mathbf{x}) &= g(\mathbf{x}) & \text{für alle } \mathbf{x} \in \Gamma_h. \end{aligned} \quad (2.1)$$

Sammelt man alle Unbekannten im Vektor \mathbf{u}_h , so führt (2.1) auf ein Gleichungssystem $\mathbf{A}_h \mathbf{u}_h = \mathbf{f}_h$.

2.2 Beliebige Differentialoperatoren*

Der allgemeine elliptische Differentialoperator

$$(\mathcal{L}u)(\mathbf{x}) = - \sum_{i,j=1}^d a_{i,j}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j} u(\mathbf{x}) + \sum_{i=1}^d b_i(\mathbf{x}) \frac{\partial}{\partial x_i} u(\mathbf{x}) + c(\mathbf{x})u(\mathbf{x})$$

wird diskretisiert durch

$$(\mathcal{L}_h u)(\mathbf{x}) = \left[- \sum_{i=1}^d a_{i,i}(\mathbf{x}) \partial_i^{-h} \partial_i^{+h} - \sum_{\substack{i,j=1 \\ i \neq j}}^d a_{i,j}(\mathbf{x}) \partial_i^h \partial_j^h + \sum_{i=1}^d b_i(\mathbf{x}) \partial_i^h + c(\mathbf{x}) \right] u(\mathbf{x}).$$

Falls $u \in C^4(\bar{\Omega})$ ist, dann gilt $|(\mathcal{L}u)(\mathbf{x}) - (\mathcal{L}_h u)(\mathbf{x})| = \mathcal{O}(h^2)$.

Beispiel 2.5 Im Zweidimensionalen ergibt sich

$$\begin{aligned} (\mathcal{L}_h u)(\mathbf{x}) &= \left[-a_{1,1}(\mathbf{x}) \partial_1^{-h} \partial_1^{+h} - 2a_{1,2}(\mathbf{x}) \partial_1^h \partial_2^h - a_{2,2}(\mathbf{x}) \partial_2^{-h} \partial_2^{+h} \right] u(\mathbf{x}) \\ &\quad + \left[b_1(\mathbf{x}) \partial_1^h + b_2(\mathbf{x}) \partial_2^h \right] u(\mathbf{x}) + c(\mathbf{x})u(\mathbf{x}) \\ &= \frac{1}{2h^2} \begin{bmatrix} a_{1,2}(\mathbf{x}) & -2a_{2,2}(\mathbf{x}) & -a_{1,2}(\mathbf{x}) \\ -2a_{1,1}(\mathbf{x}) & 4[a_{1,1}(\mathbf{x}) + a_{2,2}(\mathbf{x})] & -2a_{1,1}(\mathbf{x}) \\ -a_{1,2}(\mathbf{x}) & -2a_{2,2}(\mathbf{x}) & a_{1,2}(\mathbf{x}) \end{bmatrix}_* u(\mathbf{x}) \\ &\quad + \frac{1}{2h} \begin{bmatrix} 0 & b_2(\mathbf{x}) & 0 \\ -b_1(\mathbf{x}) & 0 & b_1(\mathbf{x}) \\ 0 & -b_2(\mathbf{x}) & 0 \end{bmatrix}_* u(\mathbf{x}) + \begin{bmatrix} 0 & 0 & 0 \\ 0 & c(\mathbf{x}) & 0 \\ 0 & 0 & 0 \end{bmatrix}_* u(\mathbf{x}). \end{aligned}$$

△

So schön dieser Stern auch ist, so lässt sich dennoch im allgemeinen keine Stabilität nachweisen. Dies liegt an der Diskretisierung der gemischten Ableitung $\partial^2/(\partial x_1 \partial x_2)$, die wir in Abhängigkeit vom Vorzeichen von $a_{1,2}(\mathbf{x})$ wie folgt modifizieren. Wir wählen

$$\frac{1}{2h^2} \begin{bmatrix} 0 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 0 \end{bmatrix}_* \text{ falls } a_{1,2}(\mathbf{x}) \geq 0 \text{ bzw. } \frac{1}{2h^2} \begin{bmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix}_* \text{ falls } a_{1,2}(\mathbf{x}) < 0.$$

Mit $a_{1,2}^+ := \max\{a_{1,2}, 0\}$ und $a_{1,2}^- := \min\{a_{1,2}, 0\}$ erhalten wir den Siebenpunktstern

$$\begin{aligned} (\mathcal{L}_h u)(\mathbf{x}) &= \frac{1}{h^2} \begin{bmatrix} a_{1,2}^-(\mathbf{x}) & |a_{1,2}(\mathbf{x})| - a_{2,2}(\mathbf{x}) & -a_{1,2}^+(\mathbf{x}) \\ |a_{1,2}(\mathbf{x})| - a_{1,1}(\mathbf{x}) & 2[|a_{1,2}(\mathbf{x})| - a_{2,2}(\mathbf{x})] & |a_{1,2}(\mathbf{x})| - a_{1,1}(\mathbf{x}) \\ -a_{1,2}^+(\mathbf{x}) & |a_{1,2}(\mathbf{x})| - a_{2,2}(\mathbf{x}) & a_{1,2}^-(\mathbf{x}) \end{bmatrix}_* u(\mathbf{x}) \\ &\quad + \frac{1}{2h} \begin{bmatrix} 0 & b_2(\mathbf{x}) & 0 \\ -b_1(\mathbf{x}) & 0 & b_1(\mathbf{x}) \\ 0 & -b_2(\mathbf{x}) & 0 \end{bmatrix}_* u(\mathbf{x}) + \begin{bmatrix} 0 & 0 & 0 \\ 0 & c(\mathbf{x}) & 0 \\ 0 & 0 & 0 \end{bmatrix}_* u(\mathbf{x}). \end{aligned}$$

Diese Diskretisierung ist ebenfalls konsistent von zweiter Ordnung, das heißt, es ist $|(\mathcal{L}u)(\mathbf{x}) - (\mathcal{L}_h u)(\mathbf{x})| = \mathcal{O}(h^2)$ falls $u \in C^4(\bar{\Omega})$. Unter der Bedingung

$$|a_{1,2}(\mathbf{x})| \leq \min\{a_{1,1}(\mathbf{x}), a_{2,2}(\mathbf{x})\},$$

lässt sich nun für den Hauptteil von \mathcal{L} das Sternlemma 2.6 anwenden.

Ist Ω ein Würfelgebiet, so führt das Randwertproblem

$$\begin{aligned} (\mathcal{L}u)(\mathbf{x}) &= f(\mathbf{x}) \quad \text{für alle } \mathbf{x} \in \Omega, \\ u(\mathbf{x}) &= g(\mathbf{x}) \quad \text{für alle } \mathbf{x} \in \Gamma, \end{aligned}$$

unter Verwendung der hier vorgestellten Diskretisierung auf das Differenzenverfahren

$$\begin{aligned} (\mathcal{L}_h u_h)(\mathbf{x}) &= f(\mathbf{x}) \quad \text{für alle } \mathbf{x} \in \Omega_h, \\ u_h(\mathbf{x}) &= g(\mathbf{x}) \quad \text{für alle } \mathbf{x} \in \Gamma_h. \end{aligned} \tag{2.2}$$

Dies ist gleichbedeutend mit einem linearen Gleichungssystem $\mathbf{A}_h \mathbf{u}_h = \mathbf{f}_h$ für die unbekannten Gitterwerte \mathbf{u}_h . Die Systemmatrix \mathbf{A}_h ist symmetrisch, falls $\mathbf{b} = \mathbf{0}$ gilt.

2.3 Diskretes Maximumprinzip*

Alle verwendeten Differenzensterne entsprechen einer gewichteten Mittelung von Nachbarwerten. Daher kann offensichtlich kein Wert größer sein als das Maximum über alle Nachbarwerte. Dies ist der Spezialfall der Theorie der Differenzensterne, deren Koeffizienten ein bestimmtes Vorzeichenverhalten aufweisen.

Lemma 2.6 (Sternlemma) Sei $k > 1$. Für die Zahlen α_ℓ und p_ℓ , $0 \leq \ell \leq k$, gelte $\alpha_\ell < 0$ für alle $\ell = 1, 2, \dots, k$ und

$$\sum_{\ell=0}^k \alpha_\ell \geq 0, \quad \sum_{\ell=0}^k \alpha_\ell p_\ell \leq 0.$$

Ferner sei $p_0 \geq 0$ oder $\sum_{\ell=0}^k \alpha_\ell = 0$. Dann folgt aus $p_0 \geq \max_{1 \leq \ell \leq k} \{p_\ell\}$ die Gleichheit

$$p_0 = p_1 = \dots = p_k.$$

Beweis. Aus den Voraussetzungen folgt

$$\sum_{\ell=1}^k \alpha_\ell (p_\ell - p_0) = \sum_{\ell=0}^k \alpha_\ell (p_\ell - p_0) = \sum_{\ell=0}^k \alpha_\ell p_\ell - p_0 \sum_{\ell=0}^k \alpha_\ell \leq 0.$$

In der links stehenden Summe sind alle Summanden wegen $\alpha_\ell < 0$ und $p_\ell - p_0 \leq 0$ nicht negativ. Also hat jeder Summand den Wert 0. Aus $\alpha_\ell \neq 0$ folgt die Behauptung. \square

Definition 2.7 Das Gebiet Ω_h heißt **(diskret) zusammenhängend**, wenn zu jedem Punktepaar $\mathbf{x}, \mathbf{y} \in \Omega_h$ auch ein Verbindungsweg existiert, der entlang der Gitterlinien und ganz in Ω_h verläuft.

Bemerkung Für genügend kleines h ist das Gebiet Ω_h diskret zusammenhängend. \triangle

Satz 2.8 (Diskretes Maximumprinzip) Sei u_h die Lösung der diskreten Differentialgleichung

$$(\mathcal{L}_h u_h)(\mathbf{x}) = f(\mathbf{x}) \leq 0 \quad \text{für alle } \mathbf{x} \in \Omega_h,$$

die von der Diskretisierung der elliptischen Differentialgleichung $\mathcal{L}u = f \leq 0$ in Ω herührt. Der Differenzenstern zu jedem Gitterpunkt in Ω_h genüge folgenden drei Bedingungen:

1. Alle Koeffizienten, abgesehen vom Zentrum, sind nicht positiv.
2. Der Koeffizient in Ostrichtung sei negativ: $\alpha_O < 0$.
3. Die Summe aller Koeffizienten ist nicht negativ.

Dann ist

$$\max_{\mathbf{x} \in \Omega_h} u_h(\mathbf{x}) \leq \max\{\max_{\mathbf{x} \in \Gamma_h} u_h(\mathbf{x}), 0\}.$$

Wenn das Maximum im Innern angenommen wird, die Koeffizienten in den Hauptrichtungen (also in zwei Dimensionen $\alpha_O, \alpha_W, \alpha_S, \alpha_N$) negativ sind und Ω_h zusammenhängend ist, dann ist u_h konstant.

Beweis. Wenn das Maximum im Punkt $\mathbf{z} \in \Omega_h$, also im Innern, angenommen wird, dann setze $p_0 := u_h(\mathbf{z}) > 0$ und identifiziere p_1, p_2, \dots, p_k mit den Werten in allen Nachbarpunkten, die im Differenzenstern auftreten. Wegen $\sum_{\ell=0}^k \alpha_\ell p_\ell = f(\mathbf{z}) \leq 0$ impliziert das Sternlemma $p_0 = p_1 = \dots = p_k$, das heißt, $u_h(\mathbf{z})$ stimmt mit allen Nachbarn überein.

Nun marschieren wir zum Rand: wir wiederholen dieses Argument solange im jeweils östlichen Nachbarn, bis wir am Rand angekommen sind.

Wenn Ω_h zusammenhängt, können wir gemäß der Voraussetzung das obige Argument solange in alle Hauptrichtungen anwenden, bis alle Punkte von $\overline{\Omega}_h$ erreicht sind. \square

Bemerkung Wenn man als dritte Bedingung sogar verlangt, dass die Summe aller Koeffizienten des Sterns 0 ergibt, dann folgt das strenge Maximumprinzip $\max_{\mathbf{x} \in \overline{\Omega}_h} u_h(\mathbf{x}) \leq \max_{\mathbf{x} \in \Gamma_h} u_h(\mathbf{x})$. \triangle

Aus dem diskreten Maximumprinzip kann man genau dieselben Folgerungen schließen wie aus dem kontinuierlichen Maximumprinzip. Insbesondere sei auf das Vergleichsprinzip und die stetige Abhängigkeit von den Daten f und g hingewiesen. Eine weitere wollen wir explizit benennen:

Proposition 2.9 Wenn das diskrete Maximumprinzip gilt, ist das Gleichungssystem $\mathbf{A}_h \mathbf{u}_h = \mathbf{f}_h$ eindeutig lösbar.

2.4 Konvergenz*

Auf Ω_h und $\overline{\Omega}_h$ definieren wir die Maximumnorm durch

$$\|v_h\|_{\Omega_h} := \max_{\mathbf{x} \in \Omega_h} |v_h(\mathbf{x})|, \quad \|v_h\|_{\overline{\Omega}_h} := \max_{\mathbf{x} \in \overline{\Omega}_h} |v_h(\mathbf{x})|.$$

Definition 2.10 Das Differenzenverfahren (2.2) heißt

- **konvergent** mit der Ordnung p , wenn

$$\|u - u_h\|_{\overline{\Omega}_h} = \mathcal{O}(h^p),$$

- **konsistent** mit der Ordnung p , wenn

$$\|\mathcal{L}_h u - \mathcal{L}u\|_{\Omega_h} = \mathcal{O}(h^p),$$

- **stabil** (bzgl. der rechten Seite), wenn eine Konstante $C_s > 0$ existiert, so dass für alle Gitterfunktionen v_h mit $v_h = 0$ am Rand gilt

$$\|v_h\|_{\overline{\Omega}_h} \leq C_s \|\mathcal{L}_h v_h\|_{\Omega_h}.$$

Beispiel 2.11 Auf Würfelgebieten sind wegen $\|\mathcal{L}_h u - \mathcal{L}u\|_{\Omega_h} = \mathcal{O}(h^2)$ die Differenzenverfahren (2.1) und (2.2) konsistent mit der Ordnung 2. Da die Shortley-Weller-Approximation nur von der Ordnung 1 ist, ist hingegen das Verfahren (2.1) für allgemeine Gebiete nur konsistent mit der Ordnung 1. \triangle

Bemerkung Stabilität bedeutet nichts anderes, als dass $\|\mathbf{A}_h^{-1}\|_{\infty} \leq C_s$ unabhängig von der Maschenweite h ist. Das sieht man wie folgt: Bezeichnen \mathbf{v}_h und \mathbf{w}_h die Vektoren der Werte der Gitterfunktion $v_h|_{\Omega_h}$ und $\mathcal{L}_h v_h$, dann folgt $\mathbf{w}_h = \mathbf{A}_h \mathbf{v}_h$. Die Stabilitätsbedingung kann nun übersetzt werden gemäß

$$\|v_h\|_{\overline{\Omega}_h} = \|\mathbf{v}_h\|_{\infty} = \|\mathbf{A}_h^{-1} \mathbf{w}_h\|_{\infty} \leq C_s \|\mathbf{w}_h\|_{\infty} = C_s \|\mathbf{A}_h \mathbf{v}_h\|_{\infty} = C_s \|\mathcal{L}_h v_h\|_{\Omega_h}.$$

Hieraus folgt das Behauptete, da diese Ungleichung für beliebige Vektoren \mathbf{w}_h gilt. \triangle

Satz 2.12 Ist ein Differenzenverfahren stabil und konsistent mit der Ordnung p , dann ist es auch konvergent mit der Ordnung p .

Beweis. Es gilt

$$\|u - u_h\|_{\overline{\Omega}_h} \leq C_s \|\mathcal{L}_h(u - u_h)\|_{\Omega_h} = C_s \|\mathcal{L}_h u - \mathcal{L}_h u_h\|_{\Omega_h}.$$

Wegen $(\mathcal{L}_h u_h)(\mathbf{x}) = f(\mathbf{x}) = (\mathcal{L}u)(\mathbf{x})$ für alle $\mathbf{x} \in \Omega_h$, folgt

$$\underbrace{\|u - u_h\|_{\overline{\Omega}_h}}_{\text{Diskretisierungsfehler}} \leq C_s \underbrace{\|\mathcal{L}_h u - \mathcal{L}u\|_{\Omega_h}}_{\text{Konsistenzfehler}} = \mathcal{O}(h^p).$$

□

Lemma 2.13 Sei Ω in der Menge $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 < R\}$ enthalten. Die Gitterfunktion v_h sei Lösung der Gleichung

$$\begin{aligned} -(\Delta_h v_h)(\mathbf{x}) &= 1 \quad \text{für alle } \mathbf{x} \in \Omega_h, \\ v_h(\mathbf{x}) &= 0 \quad \text{für alle } \mathbf{x} \in \Gamma_h. \end{aligned}$$

Dann gilt

$$0 \leq v_h(\mathbf{x}) \leq \frac{1}{2d}(R^2 - \|\mathbf{x}\|_2^2), \quad \mathbf{x} \in \overline{\Omega}_h.$$

Beweis. Man betrachte die Funktion $w(\mathbf{x}) = (R^2 - \|\mathbf{x}\|_2^2)/(2d)$. Da w ein Polynom zweiten Grades ist, verschwinden die bei der Bildung des Differenzensterns vernachlässigten Ableitungen, das heißt, es gilt $-\Delta_h w = -\Delta w = 1$ in Ω_h . Außerdem ist $w \geq 0$ auf Γ_h . Aus dem diskreten Vergleichsprinzip folgt daher $v_h(\mathbf{x}) \leq w(\mathbf{x})$ für alle $\mathbf{x} \in \overline{\Omega}_h$. \square

Dieses Lemma impliziert die Stabilität: Sei w_h mit $w_h = 0$ auf Γ_h beliebig, dann folgt

$$-\frac{(\Delta_h w_h)(\mathbf{x})}{\|\Delta_h w_h\|_{\Omega_h}} \leq 1 = -(\Delta_h v_h)(\mathbf{x}), \quad \mathbf{x} \in \Omega_h.$$

Das diskrete Vergleichsprinzip liefert sofort

$$\frac{w_h(\mathbf{x})}{\|\Delta_h w_h\|_{\Omega_h}} \leq v_h(\mathbf{x}) \leq \frac{1}{2d}(R^2 - \|\mathbf{x}\|_2^2), \quad \mathbf{x} \in \overline{\Omega}_h,$$

dies bedeutet

$$\|w_h\|_{\overline{\Omega}_h} \leq \frac{R^2}{2d} \|\Delta_h w_h\|_{\Omega_h}.$$

Korollar 2.14 Die Lösung der Poisson-Gleichung erfülle $u \in C^4(\overline{\Omega})$. Dann konvergiert das Differenzenverfahren (2.1) und es gilt

$$\|u - u_h\|_{\overline{\Omega}_h} = \mathcal{O}(h^p)$$

mit $p = 2$ im Falle von Würfelgebieten und $p = 1$ im Falle von allgemeinen Gebieten.

Bemerkung Ist der Differentialoperator

$$(\mathcal{L}u)(\mathbf{x}) = - \sum_{i,j=1}^d a_{i,j}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j} u(\mathbf{x}) + c(\mathbf{x})u(\mathbf{x}), \quad c(\mathbf{x}) \geq 0$$

gleichmäßig elliptisch mit Elliptizitätskonstante $\alpha > 0$, dann folgt für $w(\mathbf{x}) = R^2 - \|\mathbf{x}\|_2^2$, dass $(\mathcal{L}_h w)(\mathbf{x}) = (\mathcal{L}w)(\mathbf{x}) \geq 2\alpha$ für alle $\mathbf{x} \in \Omega_h$ (vergleiche fünfte Folgerung des kontinuierlichen Maximumprinzips). Erfüllt der zugehörige Differenzenstern das diskrete Maximumprinzip (und damit das Vergleichsprinzip), so gilt daher Lemma 2.13 mit

$$0 \leq v_h(\mathbf{x}) \leq \frac{1}{2\alpha}(R^2 - \|\mathbf{x}\|_2^2), \quad \mathbf{x} \in \overline{\Omega}_h.$$

Enthält hingegen \mathcal{L} zusätzlich Terme erster Ordnung und ist $c > 0$, dann kann mittels einem Störargument für genügend kleine Maschenweite h Stabilität nachweisen werden. Auf Würfelgebieten erhalten wir auf diese Weise ebenfalls eine quadratische Konvergenzordnung für allgemeine Differentialoperatoren. \triangle

3. Variationsformulierung

3.1 Sobolev-Räume

Sei $\Omega \subset \mathbb{R}^d$ ein Gebiet mit stückweise glattem Rand. Der Funktionenraum $L^2(\Omega)$ besteht aus allen Funktionen, die über Ω quadratisch Lebesgue-integrierbar sind. Dabei werden zwei Funktionen miteinander identifiziert, wenn $u(\mathbf{x}) = v(\mathbf{x})$ für $\mathbf{x} \in \Omega$ abgesehen von einer Nullmenge gilt. Durch das Skalarprodukt

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

wird $L^2(\Omega)$ zu einem Hilbert-Raum mit der Norm

$$\|u\|_{L^2(\Omega)} := \sqrt{(u, u)_{L^2(\Omega)}} = \sqrt{\int_{\Omega} u^2(\mathbf{x}) \, d\mathbf{x}}.$$

Definition 3.1 Die Funktion $u \in L^2(\Omega)$ besitzt die (**schwache**) **Ableitung** $v = \partial^\alpha u$, falls $v \in L^2(\Omega)$ und

$$(v, \phi)_{L^2(\Omega)} = (-1)^{|\alpha|} (u, \partial^\alpha \phi)_{L^2(\Omega)} \quad \text{für alle } \phi \in C_0^\infty(\Omega)$$

gilt.

Hier bezeichnet $C^\infty(\Omega)$ den Raum der auf Ω beliebig oft stetig differenzierbaren Funktionen und $C_0^\infty(\Omega)$ den Unterraum der Funktionen, die nur auf einer kompakten Teilmenge von Ω von 0 verschiedene Werte annehmen.

Bemerkung Ist $u \in C^1(\Omega)$, dann liefert der Gaußsche Integralsatz

$$(\partial_{x_i} u, \phi)_{L^2(\Omega)} + (u, \partial_{x_i} \phi)_{L^2(\Omega)} = \int_{\Omega} \partial_{x_i} (u\phi) \, d\mathbf{x} = \int_{\partial\Omega} u\phi n_i \, d\sigma,$$

wobei $\mathbf{n} = (n_1, n_2, \dots, n_d)$ die nach außen gerichtete Normale an das Gebiet Ω bezeichnet. Folglich ist

$$(\partial_{x_i} u, \phi)_{L^2(\Omega)} = -(u, \partial_{x_i} \phi)_{L^2(\Omega)} \quad \text{für alle } \phi \in C_0^\infty(\Omega),$$

das heißt, die schwache Ableitung stimmt mit der üblichen überein. △

Definition 3.2 Für ganzzahliges $m \geq 0$ bezeichne der **Sobolev-Raum** $H^m(\Omega)$ die Menge aller Funktionen u in $L^2(\Omega)$, die schwache Ableitungen $\partial^\alpha u \in L^2(\Omega)$ für alle $|\alpha| \leq m$ besitzen.

Satz 3.3 Der Sobolev-Raum $H^m(\Omega)$, ausgestattet mit dem Skalarprodukt

$$(u, v)_{H^m(\Omega)} := \sum_{|\alpha| \leq m} (\partial^\alpha u, \partial^\alpha v)_{L^2(\Omega)}$$

und der zugehörigen Norm

$$\|u\|_{H^m(\Omega)} := \sqrt{(u, u)_{H^m(\Omega)}} = \sqrt{\sum_{|\alpha| \leq m} \|\partial^\alpha u\|_{L^2(\Omega)}^2},$$

ist ein Hilbert-Raum.

Beweis. Sei $\{v_n\}$ eine Cauchy-Folge in $H^m(\Omega)$, dann ist $\{\partial^\alpha v_n\}$ für alle $|\alpha| \leq m$ eine Cauchy-Folge in $L^2(\Omega)$. Aus der Vollständigkeit des $L^2(\Omega)$ folgt die Existenz von Funktionen $v^\alpha \in L^2(\Omega)$ mit

$$\|\partial^\alpha v_n - v^\alpha\|_{L^2(\Omega)} \xrightarrow{n \rightarrow \infty} 0.$$

Wir müssen nun nur noch die Identität $\partial^\alpha v = v^\alpha$ nachweisen. Ist $\{w_n\}$ eine Cauchy-Folge aus $L^2(\Omega)$ mit dem Grenzwert $w \in L^2(\Omega)$, dann folgt aus

$$(w - w_n, \phi)_{L^2(\Omega)} \leq \|w - w_n\|_{L^2(\Omega)} \|\phi\|_{L^2(\Omega)}$$

sofort $(w_n, \phi)_{L^2(\Omega)} \rightarrow (w, \phi)_{L^2(\Omega)}$ für jede Testfunktion $\phi \in C_0^\infty(\Omega)$. Somit ergibt sich

$$(v^\alpha, \phi)_{L^2(\Omega)} = \lim_{n \rightarrow \infty} (\partial^\alpha v_n, \phi)_{L^2(\Omega)} = \lim_{n \rightarrow \infty} (-1)^{|\alpha|} (v_n, \partial^\alpha \phi)_{L^2(\Omega)} = (-1)^{|\alpha|} (v, \partial^\alpha \phi)_{L^2(\Omega)},$$

dies bedeutet, es gilt tatsächlich $\partial^\alpha v = v^\alpha$. \square

Bemerkung Die Glattheit von H^m -Funktionen im Sinne klassischer C^k -Räume ist dimensionsabhängig. Für $d = 1$ sind alle Funktionen aus $H^1(\Omega)$ auch stetig, das heißt $H^1(\Omega) \subset C(\Omega)$. Ist $d = 2$, so enthält der Raum $H^1(\Omega)$ sogar Funktionen mit Punktsingularitäten. Beispielsweise gilt

$$u(r, \varphi) = \log \left(\log \frac{2}{r} \right) \in H^1(\{(r \cos \varphi, r \sin \varphi) : 0 \leq r < 1, 0 \leq \varphi < 2\pi\}).$$

Allgemein ist für $d \geq 3$ jede Funktion

$$u(\mathbf{x}) = r^{-\beta}, \quad \beta < (d-2)/2$$

eine H^1 -Funktion mit Punktsingularität im Nullpunkt. \triangle

Oftmals wichtig ist die Eigenschaft, dass $C^\infty(\Omega) \cap H^m(\Omega)$ dicht in $H^m(\Omega)$ liegt. Dieses Resultat wurde von Meyers und Serrin im Jahr 1964 bewiesen.

Definition 3.4 Die Vervollständigung von $C_0^\infty(\Omega)$ bezüglich der Sobolev-Norm $\|\cdot\|_{H^m(\Omega)}$ wird mit $H_0^m(\Omega)$ bezeichnet.

Offensichtlich ist der Hilbert-Raum $H_0^m(\Omega)$ ein abgeschlossener Unterraum von $H^m(\Omega)$. Außerdem ist $H_0^0(\Omega) = L^2(\Omega)$, so dass sich folgendes Schema ergibt:

$$\begin{array}{ccccccc} L^2(\Omega) & = & H^0(\Omega) & \supset & H^1(\Omega) & \supset & H^2(\Omega) & \supset & \dots \\ & & \parallel & & \cup & & \cup & & \\ & & H_0^0(\Omega) & \supset & H_0^1(\Omega) & \supset & H_0^2(\Omega) & \supset & \dots \end{array}$$

Im Sobolev-Raum $H^m(\Omega)$ wird durch

$$|u|_{H^m(\Omega)} := \sqrt{\sum_{|\alpha|=m} \|\partial^\alpha u\|_{L^2(\Omega)}^2}$$

die $H^m(\Omega)$ -Seminorm definiert. Ist $m > 0$, erfüllt sie alle Normeigenschaften bis auf die Definitheit. Denn es gilt beispielsweise $|u|_{H^m(\Omega)} = 0$ für jede konstante Funktion $u \in H^m(\Omega)$. Für Funktionen mit homogenen Randbedingungen ist die $H^m(\Omega)$ -Seminorm jedoch äquivalent zur $H^m(\Omega)$ -Norm.

Satz 3.5 (Poincaré-Friedrichssche Ungleichung) Sei Ω in einem d -dimensionalen Würfel der Kantenlänge s enthalten. Dann ist

$$\|v\|_{L^2(\Omega)} \leq s|v|_{H^1(\Omega)} \quad \text{für alle } v \in H_0^1(\Omega).$$

Beweis. Da $C_0^\infty(\Omega)$ dicht in $H_0^1(\Omega)$ ist, genügt es, die Ungleichung für $v \in C_0^\infty(\Omega)$ zu beweisen. Wir können $\Omega \subset \square := \{(x_1, x_2, \dots, x_d) : 0 \leq x_i \leq s\}$ annehmen und $v(\mathbf{x}) = 0$ für $\mathbf{x} \in \square \setminus \Omega$ setzen. Es folgt

$$v(x_1, x_2, \dots, x_d) = \underbrace{v(0, x_2, \dots, x_d)}_{=0} + \int_0^{x_1} \partial_{x_1} v(t, x_2, \dots, x_d) dt$$

und mit der Cauchy-Schwarzschen Ungleichung weiter

$$|v(\mathbf{x})|^2 \leq \left(\int_0^{x_1} 1^2 dt \right) \left(\int_0^{x_1} |\partial_{x_1} v(t, x_2, \dots, x_d)|^2 dt \right) \leq s \int_0^{x_1} |\partial_{x_1} v(t, x_2, \dots, x_d)|^2 dt.$$

Da die rechte Seite unabhängig von x_1 ist, ergibt sich

$$\int_0^s |v(\mathbf{x})|^2 dx_1 \leq s^2 \int_0^s |\partial_{x_1} v(\mathbf{x})|^2 dx_1.$$

Schließlich wird über die anderen Koordinaten integriert:

$$\int_{\square} |v(\mathbf{x})|^2 d\mathbf{x} \leq s^2 \int_{\square} |\partial_{x_1} v(\mathbf{x})|^2 d\mathbf{x} \leq s^2 |v|_{H^1(\Omega)}^2.$$

□

Bemerkung Die Poincaré-Friedrichssche Ungleichung gilt bereits, wenn homogene Randbedingungen lediglich auf einem Teil des Randes $\Gamma_D \subset \Gamma$ mit positivem $(d-1)$ -dimensionalem Maß vorgegeben sind. \triangle

Korollar 3.6 Wenn Ω beschränkt ist, sind in $H_0^m(\Omega)$ die Normen $\|\cdot\|_{H^m(\Omega)}$ und $|\cdot|_{H^m(\Omega)}$ äquivalent. Ist Ω in einem Würfel der Kantenlänge s enthalten, so ist

$$|v|_{H^m(\Omega)} \leq \|v\|_{H^m(\Omega)} \leq (1+s)^m |v|_{H^m(\Omega)} \quad \text{für } v \in H_0^m(\Omega).$$

Beweis. Wir zeigen die Aussage mit Hilfe von vollständiger Induktion. Für $m=0$ ist die Aussage offensichtlich richtig. Für den Induktionsschritt $m-1 \mapsto m$ sei ein $m > 0$ beliebig vorgegeben. Mit Hilfe der Induktionsannahme folgt

$$\begin{aligned} \|v\|_{H^m(\Omega)}^2 &= \|v\|_{H^{m-1}(\Omega)}^2 + |v|_{H^m(\Omega)}^2 \\ &\leq (1+s)^{2(m-1)} |v|_{H^{m-1}(\Omega)}^2 + |v|_{H^m(\Omega)}^2 \\ &= (1+s)^{2(m-1)} \left(\sum_{|\alpha|=m-1} \|\partial^\alpha v\|_{L^2(\Omega)}^2 \right) + |v|_{H^m(\Omega)}^2. \end{aligned}$$

Durch die Anwendung der Poincaré-Friedrichsschen Ungleichung auf Ableitungen erkennt man, dass $\|\partial^\alpha v\|_{L^2(\Omega)} \leq s \|\partial_{x_1} \partial^\alpha v\|_{L^2(\Omega)}$ für alle $|\alpha| \leq m-1$ und $v \in H_0^m(\Omega)$ gilt. Dies eingesetzt ergibt

$$\begin{aligned} \|v\|_{H^m(\Omega)}^2 &\leq s^2 (1+s)^{2(m-1)} \left(\sum_{|\alpha|=m-1} \|\partial_{x_1} \partial^\alpha v\|_{L^2(\Omega)}^2 \right) + |v|_{H^m(\Omega)}^2 \\ &\leq \left(s^2 (1+s)^{2(m-1)} + 1 \right) \sum_{|\alpha|=m} \|\partial^\alpha v\|_{L^2(\Omega)}^2 \\ &\leq (1+s)^{2m} |v|_{H^m(\Omega)}^2. \end{aligned}$$

□

Der Rand Γ eines Gebietes $\Omega \subset \mathbb{R}^d$ ist eine Menge vom Maß Null bezüglich des \mathbb{R}^d . Daher besitzen Funktionen $u \in L^2(\Omega)$ keine Randwerte. Allerdings zeigt der nachfolgende Satz, dass für Funktionen aus $H^1(\Omega)$ Randwerte vorgegeben werden können.

Satz 3.7 (Spursatz) Das Gebiet Ω sei beschränkt und besitze einen stückweise glatten Rand Γ . Ferner erfülle Ω eine Kegelbedingung, das heißt, die Innenwinkel an den Ecken seien positiv, so dass man einen Kegel mit positivem Scheitelwinkel derart in Ω verschieben kann, dass er die Ecken berührt. Dann gibt es eine beschränkte, lineare Abbildung

$$\gamma : H^1(\Omega) \rightarrow L^2(\Gamma), \quad \|\gamma(v)\|_{L^2(\Gamma)} \leq c \|v\|_{H^1(\Omega)},$$

so dass $\gamma(v) = v|_\Gamma$ für alle $v \in C^1(\overline{\Omega})$ gilt.

Beweis. Wir führen den Beweis der Übersichtlichkeit halber nur für Gebiete im \mathbb{R}^2 . Die Verallgemeinerung auf den \mathbb{R}^d ist offensichtlich und verbleibt dem Leser als Übung.

Der Rand ist stückweise glatt und an den (endlich vielen) Punkten, wo der Rand nicht glatt ist, gilt die Kegelbedingung. Daher können wir den Rand in endlich viele Randstücke $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \dots \cup \Gamma_n$ teilen, so dass auf jedem Teilstück Γ_m nach Drehung des Koordinatensystems folgendes gilt.

1. Für eine Funktion $\phi = \phi_m \in C^1([y_1, y_2])$ gilt

$$\Gamma_m = \{(x, y) \in \mathbb{R}^2 : x = \phi(y), y_1 \leq y \leq y_2\}.$$

2. Das Gebiet

$$\Omega_m := \{(x, y) \in \mathbb{R}^2 : \phi(y) < x < \phi(y) + \delta, y_1 < y < y_2\}$$

ist für ein $\delta > 0$ in Ω enthalten.

Für $v \in C^1(\overline{\Omega})$ und $(x, y) \in \Gamma_m$ gilt

$$v(\phi(y), y) = v(\phi(y) + t, y) - \int_0^t \partial_x v(\phi(y) + s, y) ds, \quad 0 \leq t \leq \delta.$$

Die Integration über t von 0 bis δ ergibt

$$\begin{aligned} \delta v(\phi(y), y) &= \int_0^\delta v(\phi(y) + t, y) dt - \int_0^\delta \int_0^t \partial_x v(\phi(y) + s, y) ds dt \\ &= \int_0^\delta v(\phi(y) + t, y) dt - \int_0^\delta \int_s^\delta \partial_x v(\phi(y) + s, y) dt ds \\ &= \int_0^\delta v(\phi(y) + t, y) dt - \int_0^\delta \partial_x v(\phi(y) + s, y) (\delta - s) ds. \end{aligned}$$

Wir quadrieren diese Gleichung und nutzen die Youngsche Ungleichung $(a+b)^2 \leq 2a^2 + 2b^2$ aus:

$$\delta^2 v^2(\phi(y), y) \leq 2 \left(\int_0^\delta v(\phi(y) + t, y) dt \right)^2 + 2 \left(\int_0^\delta \partial_x v(\phi(y) + t, y) (\delta - t) dt \right)^2.$$

Die Cauchy-Schwarzsche Ungleichung liefert nun

$$\begin{aligned} \delta^2 v^2(\phi(y), y) &\leq 2 \left(\int_0^\delta 1 dt \right) \left(\int_0^\delta v^2(\phi(y) + t, y) dt \right) \\ &\quad + 2 \left(\int_0^\delta (\delta - t)^2 dt \right) \left(\int_0^\delta |\partial_x v(\phi(y) + t, y)|^2 dt \right) \\ &= 2\delta \left(\int_0^\delta v^2(\phi(y) + t, y) dt \right) + \frac{2}{3} \delta^3 \left(\int_0^\delta |\partial_x v(\phi(y) + t, y)|^2 dt \right). \end{aligned}$$

Wir dividieren durch δ^2 und integrieren bezüglich y

$$\int_{y_1}^{y_2} v^2(\phi(y), y) dy \leq \frac{2}{\delta} \int_{\Omega_m} v^2 d(x, y) + \frac{2}{3} \delta \int_{\Omega_m} |\partial_x v|^2 d(x, y).$$

Das Kurvenelement auf Γ_m ist durch $do = \sqrt{1 + (\phi'(y))^2} dy$ gegeben. Deshalb haben wir mit $c_m := \max \{ \sqrt{1 + (\phi'(y))^2} : y_1 \leq y \leq y_2 \}$:

$$\int_{\Gamma_m} v^2 do \leq c_m \int_{y_1}^{y_2} v^2(\phi(y), y) dy \leq c_m \left\{ \frac{2}{\delta} \int_{\Omega_m} v^2 d(x, y) + \frac{2}{3} \delta \int_{\Omega_m} |\partial_x v|^2 d(x, y) \right\}.$$

Indem wir

$$c := \sqrt{\left(\frac{2}{\delta} + \frac{2}{3}\delta\right) \sum_{m=1}^n c_m}$$

setzen, erhalten wir schließlich

$$\|\gamma(v)\|_{L^2(\Gamma)} \leq c \|v\|_{H^1(\Omega)}.$$

Die Restriktion $\gamma : H^1(\Omega) \rightarrow L^2(\Gamma)$ ist folglich auf einer dichten Menge eine beschränkte Abbildung. Wegen der Vollständigkeit von $L^2(\Gamma)$ kann sie auf ganz $H^1(\Omega)$ erweitert werden, ohne die Schranke zu vergrößern. Denn für jedes $v \in H^1(\Omega)$ existiert eine Folge $\{v_k\} \subset C^1(\overline{\Omega})$ mit $\|v - v_k\|_{H^1(\Omega)} \rightarrow 0$ für $k \rightarrow \infty$. Aus

$$\|\gamma(v_k) - \gamma(v_\ell)\|_{L^2(\Gamma)} \leq c \|v_k - v_\ell\|_{H^1(\Omega)}$$

folgt, dass $\{\gamma(v_k)\}$ eine Cauchy-Folge in $L^2(\Gamma)$ ist. Für ihren Grenzwert $\gamma(v) := \lim_{k \rightarrow \infty} \gamma(v_k)$ gilt dann

$$\|\gamma(v)\|_{L^2(\Gamma)} = \lim_{k \rightarrow \infty} \|\gamma(v_k)\|_{L^2(\Gamma)} \leq c \lim_{k \rightarrow \infty} \|v_k\|_{H^1(\Omega)} = c \|v\|_{H^1(\Omega)}.$$

□

Bemerkung Ohne die Kegelbedingung ist die Aussage des Spursatzes im allgemeinen falsch. △

3.2 Variationsformulierung von Dirichlet-Problemen

Gegeben sei die partielle Differentialgleichung

$$-\sum_{i,j=1}^d \partial_{x_i} (a_{i,j}(\mathbf{x}) \partial_{x_j} u(\mathbf{x})) + c(\mathbf{x})u(\mathbf{x}) = f(\mathbf{x}) \quad \text{für } \mathbf{x} \in \Omega,$$

$$u(\mathbf{x}) = 0 \quad \text{für } \mathbf{x} \in \Gamma$$

mit $c(\mathbf{x}) \geq 0$ für alle $\mathbf{x} \in \Omega$. Mit Hilfe der Differentialoperatoren

$$\operatorname{div} \mathbf{f} := \partial_{x_1} f_1 + \partial_{x_2} f_2 + \cdots + \partial_{x_d} f_d, \quad \nabla g := \begin{bmatrix} \partial_{x_1} g \\ \partial_{x_2} g \\ \vdots \\ \partial_{x_d} g \end{bmatrix}$$

und $\mathbf{A} := [a_{i,j}]_{i,j=1}^d$ können wir diese auch verkürzt schreiben als

$$-\operatorname{div}(\mathbf{A} \nabla u) + cu = f \text{ in } \Omega, \quad u = 0 \text{ auf } \Gamma. \quad (3.1)$$

Eine Lösung $u \in C^2(\Omega) \cap C(\overline{\Omega})$ wird *klassische Lösung* genannt. Wie wir sehen werden, gibt es aber auch physikalisch sinnvolle *schwache Lösungen*, die aus der Variationsformulierung gewonnen werden.

Satz 3.8 (Charakterisierungssatz) Sei V ein linearer Raum und

$$a : V \times V \rightarrow \mathbb{R}$$

eine symmetrische, positive Bilinearform, das heißt, es sei $a(u, u) > 0$ für alle $u \in V \setminus \{0\}$. Ferner sei

$$\ell : V \rightarrow \mathbb{R}$$

ein lineares Funktional. Die Größe

$$J(v) := \frac{1}{2}a(v, v) - \ell(v)$$

nimmt in V ihr Minimum genau dann bei u an, wenn

$$a(u, v) = \ell(v) \quad \text{für alle } v \in V. \quad (3.2)$$

Außerdem gibt es höchstens eine Minimallösung.

Beweis. Für $u, v \in V$ und $t \in \mathbb{R}$ berechnen wir

$$\begin{aligned} J(u + tv) &= \frac{1}{2}a(u + tv, u + tv) - \ell(u + tv) \\ &= J(u) + t\{a(u, v) - \ell(v)\} + \frac{1}{2}t^2a(v, v). \end{aligned} \quad (3.3)$$

Wenn $u \in V$ die Bedingung (3.2) erfüllt, dann folgt mit $t = 1$

$$J(u + v) = J(u) + \frac{1}{2}a(v, v) > J(u), \quad \text{falls } v \neq 0 \text{ ist.}$$

Damit ist u also ein eindeutiger Minimalpunkt. Besitzt umgekehrt J bei u ein Minimum, dann muss für jedes $v \in V$ die Ableitung der Funktion $t \mapsto J(u + tv)$ bei $t = 0$ verschwinden. Nach (3.3) beträgt die Ableitung dort $a(u, v) - \ell(v)$, womit sich (3.2) ergibt. \square

Satz 3.9 Jede klassische Lösung der partiellen Differentialgleichung (3.1) ist Lösung des Variationsproblems

$$J(v) = \int_{\Omega} \left\{ \frac{1}{2}(\langle \mathbf{A} \nabla v, \nabla v \rangle + cv^2) - fv \right\} dx \rightarrow \inf$$

unter allen Funktionen in $C^2(\Omega) \cap C(\bar{\Omega})$ mit Nullrandwerten.

Beweis. Der Beweis erfolgt mit Hilfe des Gaußschen Integralsatzes:

$$\int_{\Omega} \operatorname{div}((\mathbf{A} \nabla u)v) dx = \int_{\Omega} \{ \operatorname{div}(\mathbf{A} \nabla u)v + \langle \mathbf{A} \nabla u, \nabla v \rangle \} dx = \int_{\partial \Omega} \langle \mathbf{A} \nabla u, \mathbf{n} \rangle v d\sigma.$$

Besitzt v homogene Randwerte, dann verschwindet der Randterm, und es ergibt sich

$$\int_{\Omega} \langle \mathbf{A} \nabla u, \nabla v \rangle dx = - \int_{\Omega} \operatorname{div}(\mathbf{A} \nabla u)v dx.$$

Wir setzen nun

$$a(u, v) := \int_{\Omega} \{ \langle \mathbf{A} \nabla u, \nabla v \rangle + cuv \} \, d\mathbf{x}, \quad \ell(v) = \int_{\Omega} f(\mathbf{x})v(\mathbf{x}) \, d\mathbf{x}.$$

Dann gilt für jedes $v \in C^1(\Omega) \cap C(\overline{\Omega})$ mit Nullrandbedingungen

$$a(u, v) - \ell(v) = \int_{\Omega} \{ \langle \mathbf{A} \nabla u, \nabla v \rangle + cuv - fv \} \, d\mathbf{x} = \int_{\Omega} \{ -\operatorname{div}(\mathbf{A} \nabla u) + cu - f \} v \, d\mathbf{x}.$$

Dieser Ausdruck wird 0 wenn $-\operatorname{div}(\mathbf{A} \nabla u) + cu = f$ ist, also u eine klassische Lösung von (3.1) ist. Aus dem Charakterisierungssatz 3.8 folgt nun die Minimaleigenschaft. \square

Mit derselben Schlussweise erkennt man, dass jede Lösung des Variationsproblems, sofern sie im Raum $C^2(\Omega) \cap C(\overline{\Omega})$ liegt, klassische Lösung von (3.1) ist. Um allerdings die Existenz einer Lösung zu zeigen, darf man sich nicht nur auf klassische Lösungen beschränken, sondern muss auch schwache Lösungen zulassen.

Definition 3.10 Sei H ein Hilbert-Raum. Eine Bilinearform $a : H \times H \rightarrow \mathbb{R}$ heißt **stetig**, wenn es ein $c_S > 0$ gibt, so dass

$$|a(u, v)| \leq c_S \|u\| \|v\| \quad \text{für alle } u, v \in H$$

ist. Eine Bilinearform a heißt **H-elliptisch**, kurz **elliptisch** oder **koerziv**, wenn mit einem $c_E > 0$ gilt

$$a(v, v) \geq c_E \|v\|^2 \quad \text{für alle } v \in H.$$

Mit einer stetigen, H -elliptischen Bilinearform a wird durch

$$\|v\|_a := \sqrt{a(v, v)}$$

offensichtlich eine Norm induziert, die zur Norm des Hilbert-Raums H äquivalent ist. Dies bedeutet, es existieren Konstanten $\underline{c}, \bar{c} > 0$ derart, dass gilt

$$\underline{c} \|v\| \leq \|v\|_a \leq \bar{c} \|v\| \quad \text{für alle } v \in H.$$

Die Norm $\|v\|_a$ wird *Energienorm* genannt.

Wie üblich wird der Raum der stetigen, linearen Funktionale auf einem normierten Raum V mit V' bezeichnet.

Satz 3.11 (Lax-Milgram) Sei V ein abgeschlossener Unterraum in einem Hilbert-Raum H und $a : H \times H \rightarrow \mathbb{R}$ eine stetige und V -elliptische Bilinearform. Für jedes $\ell \in V'$ hat das Variationsproblem

$$J(v) := \frac{1}{2} a(v, v) - \ell(v) \rightarrow \inf$$

genau eine Lösung in V .

Beweis. Wegen

$$J(v) \geq \frac{1}{2}c_E \|v\|^2 - \|\ell\| \|v\| = \frac{1}{2c_E} \underbrace{(c_E \|v\| - \|\ell\|)^2}_{\geq 0} - \frac{\|\ell\|^2}{2c_E} \geq -\frac{\|\ell\|^2}{2c_E}$$

ist J nach unten beschränkt. Setze $\underline{c} = \inf\{J(v) : v \in V\}$ und sei $\{v_n\}$ eine Minimalfolge. Dann ist

$$\begin{aligned} c_E \|v_n - v_m\|^2 &\leq a(v_n - v_m, v_n - v_m) \\ &= 2a(v_n, v_n) + 2a(v_m, v_m) - a(v_n + v_m, v_n + v_m) \\ &= 4J(v_n) + 4J(v_m) - 8J\left(\underbrace{\frac{v_n + v_m}{2}}_{\in V}\right) \\ &\leq 4J(v_n) + 4J(v_m) - 8\underline{c}. \end{aligned}$$

Wegen $J(v_n), J(v_m) \rightarrow \underline{c}$ folgt $\|v_n - v_m\| \rightarrow 0$ für $n, m \rightarrow \infty$. Also ist $\{v_n\}$ eine Cauchy-Folge in H und es existiert $u = \lim_{n \rightarrow \infty} v_n$. Da V abgeschlossen ist, gilt auch $u \in V$. Die Stetigkeit von J impliziert schließlich

$$J(u) = \lim_{n \rightarrow \infty} J(v_n) = \inf_{v \in V} J(v).$$

Die Lösung ist eindeutig, denn sind u_1 und u_2 zwei Lösungen, so ist $u_1, u_2, u_1, u_2, \dots$ offensichtlich eine Minimalfolge. Wie wir gesehen haben, ist jede Minimalfolge eine Cauchy-Folge, woraus sich $u_1 = u_2$ ergibt. \square

Bemerkung Im Spezialfall $V = H$ und $a(u, v) = (u, v)$, wobei (\cdot, \cdot) das definierende Skalarprodukt sei, ergibt sich aus der Kombination von Charakterisierungssatz 3.8 und Satz von Lax-Milgram 3.11 gerade der Rieszsche Darstellungssatz: Zu jedem $\ell \in H'$ gibt es ein Element $u \in H$ mit

$$(u, v) = \ell(v) \quad \text{für alle } v \in H.$$

\triangle

Nach diesen Vorbereitungen können wir nun den Lösungsbegriff präzisieren.

Definition 3.12 Eine Funktion $u \in H_0^1(\Omega)$ heißt **schwache Lösung** der partiellen Differentialgleichung (3.1), falls mit der zugehörigen Bilinearform

$$a(u, v) = \ell(v) \quad \text{für alle } v \in H_0^1(\Omega)$$

ist.

Satz 3.13 Es gelte $f \in L^2(\Omega)$ und

$$0 \leq c(\mathbf{x}) \leq \bar{c} < \infty, \quad 0 < \underline{\alpha} \|\boldsymbol{\xi}\|^2 \leq \boldsymbol{\xi}^T \mathbf{A}(\mathbf{x}) \boldsymbol{\xi} \leq \bar{\alpha} \|\boldsymbol{\xi}\|^2 < \infty$$

für alle $\mathbf{x} \in \Omega$ und $\boldsymbol{\xi} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$. Dann besitzt (3.1) genau eine schwache Lösung in $H_0^1(\Omega)$. Diese ist das Minimum des Variationsproblems

$$\frac{1}{2}a(v, v) - \ell(v) \rightarrow \inf$$

in $H_0^1(\Omega)$.

Beweis. Aufgrund der Abschätzung

$$\begin{aligned} a(u, v) &= \int_{\Omega} \{ \langle \mathbf{A} \nabla u, \nabla v \rangle + cuv \} \, d\mathbf{x} \\ &\leq \int_{\Omega} \{ \bar{\alpha} \|\nabla u\| \|\nabla v\| + \bar{c} |u| |v| \} \, d\mathbf{x} \\ &\leq \bar{\alpha} \sqrt{\sum_{i=1}^d \int_{\Omega} |\partial_{x_i} u|^2 \, d\mathbf{x}} \sqrt{\sum_{i=1}^d \int_{\Omega} |\partial_{x_i} v|^2 \, d\mathbf{x}} + \bar{c} \sqrt{\int_{\Omega} u^2 \, d\mathbf{x}} \sqrt{\int_{\Omega} v^2 \, d\mathbf{x}} \\ &\leq \max\{\bar{\alpha}, \bar{c}\} \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} \end{aligned}$$

ist die Bilinearform stetig auf $H^1(\Omega)$. Die Elliptizität folgt aus

$$\begin{aligned} a(v, v) &= \int_{\Omega} \{ \langle \mathbf{A} \nabla v, \nabla v \rangle + cv^2 \} \, d\mathbf{x} \\ &\geq \underline{\alpha} \sum_{i=1}^d \int_{\Omega} |\partial_{x_i} v|^2 \, d\mathbf{x} \\ &= \underline{\alpha} |v|_{H^1(\Omega)}^2. \end{aligned}$$

Wegen der Poincaré-Friedrichsschen Ungleichung sind $|\cdot|_{H^1(\Omega)}$ und $\|\cdot\|_{H_0^1(\Omega)}$ äquivalente Normen und damit ist $a(\cdot, \cdot)$ eine $H_0^1(\Omega)$ -elliptische Bilinearform. Weiterhin ist auch die Linearform $\ell : H^1(\Omega) \rightarrow \mathbb{R}$ stetig:

$$|\ell(v)| = \left| \int_{\Omega} f v \, d\mathbf{x} \right| \leq \|f\|_{L^2(\Omega)} \|v\|_{H^1(\Omega)}.$$

Gemäß des Satzes von Lax-Milgram 3.11 existiert folglich eine Lösung des Variationsproblems, welche aufgrund des Charakterisierungssatzes 3.8 zugleich schwache Lösung ist. \square

Bemerkungen

1. Das Funktional $\ell : H_0^1(\Omega) \rightarrow \mathbb{R}$ ist sogar stetig für alle $f \in H^{-1}(\Omega) := (H_0^1(\Omega))' \supset L^2(\Omega)$. Daher muss f in (3.1) noch nicht einmal quadratisch integrierbar sein.
2. Die Randwertaufgabe mit nichthomogenen Randwerten $u = g$ auf Γ lässt sich folgendermaßen auf die Form (3.1) zurückführen: Bestimme ein $u_g \in H^1(\Omega)$ derart, dass $u_g|_{\Gamma} = g$ ist im Sinne des Spursatzes 3.7. Der Ansatz $u = u_0 + u_g$ führt dann auf die folgende Variationsformulierung:

$$\text{suche } u_0 \in H_0^1(\Omega), \text{ so dass } a(u_0, v) = \ell(v) - a(u_g, v) \quad \text{für alle } v \in H_0^1(\Omega).$$

Eine alternative Betrachtungsweise ist die Suche nach einem $u \in H^1(\Omega)$ mit $u|_{\Gamma} = g$, so dass $a(u, v) = \ell(v)$ gilt für alle $v \in H_0^1(\Omega)$.

3. Dirichlet-Randbedingungen werden durch die Wahl des Raumes, in dem die Variationsformulierung gestellt ist, explizit gefordert. Sie heißen daher auch *wesentliche Randbedingungen*.

△

3.3 Variationsformulierung von Neumann-Problemen

Gegeben sei das Neumann-Problem

$$\begin{aligned} - \sum_{i,j=1}^d \partial_{x_i} (a_{i,j}(\mathbf{x}) \partial_{x_j} u(\mathbf{x})) + c(\mathbf{x})u(\mathbf{x}) &= f(\mathbf{x}) \quad \text{für } \mathbf{x} \in \Omega, \\ \sum_{i,j=1}^d a_{i,j}(\mathbf{x}) n_i(\mathbf{x}) \partial_{x_j} u(\mathbf{x}) &= g(\mathbf{x}) \quad \text{für } \mathbf{x} \in \Gamma \end{aligned}$$

mit

$$0 < \underline{c} \leq c(\mathbf{x}) \leq \bar{c} < \infty, \quad 0 < \underline{\alpha} \|\boldsymbol{\xi}\|^2 \leq \boldsymbol{\xi}^T \mathbf{A}(\mathbf{x}) \boldsymbol{\xi} \leq \bar{\alpha} \|\boldsymbol{\xi}\|^2 < \infty$$

für alle $\mathbf{x} \in \Omega$ und $\boldsymbol{\xi} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$. Diese Gleichung können wir mit Hilfe von Differentialoperatoren auch kurz schreiben als

$$-\operatorname{div}(\mathbf{A}\nabla u) + cu = f \text{ in } \Omega, \quad \langle \mathbf{A}\nabla u, \mathbf{n} \rangle = g \text{ auf } \Gamma. \quad (3.4)$$

Die Multiplikation der Differentialgleichung mit einer Testfunktion $\phi \in C^\infty(\Omega) \cap H^1(\Omega)$ führt auf

$$\int_{\Omega} \{-\operatorname{div}(\mathbf{A}\nabla u) + cu\} \phi \, d\mathbf{x} = \int_{\Omega} \{\langle \mathbf{A}\nabla u, \nabla \phi \rangle + cu\phi\} \, d\mathbf{x} - \int_{\Gamma} \underbrace{\langle \mathbf{A}\nabla u, \mathbf{n} \rangle}_{\stackrel{!}{=}g} \phi \, d\sigma \stackrel{!}{=} \int_{\Omega} f \phi \, d\mathbf{x}.$$

Wir erhalten demnach

$$a(u, v) = \int_{\Omega} \{\langle \mathbf{A}\nabla u, \nabla v \rangle + cuv\} \, d\mathbf{x}, \quad \ell(v) = \int_{\Omega} f v \, d\mathbf{x} + \int_{\Gamma} g v \, d\sigma.$$

Dabei ist die Bilinearform $a : H^1(\Omega) \times H^1(\Omega)$ offensichtlich stetig und wegen

$$a(u, u) \geq \underline{\alpha} \|u\|_{H^1(\Omega)}^2 + \underline{c} \|u\|_{L^2(\Omega)}^2 \geq \min\{\underline{\alpha}, \underline{c}\} \|u\|_{H^1(\Omega)}^2$$

elliptisch in ganz $H^1(\Omega)$. Das Funktional $\ell(v)$ ist für $f \in L^2(\Omega)$ und $g \in L^2(\Gamma)$ stetig aufgrund des Spursatzes:

$$|\ell(v)| \leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)} \|\gamma(v)\|_{L^2(\Gamma)} \leq C \|v\|_{H^1(\Omega)}.$$

Satz 3.14 Sei Ω ein beschränktes Gebiet mit stückweise glattem Rand, das der Kegelbedingung genügt, und seien $f \in L^2(\Omega)$ und $g \in L^2(\Gamma)$. Dann besitzt die Variationsaufgabe

$$J(v) = \frac{1}{2} a(v, v) - \ell(v) \rightarrow \inf$$

genau eine Lösung $u \in H^1(\Omega)$. Die Lösung der Variationsaufgabe ist genau dann in $C^2(\Omega) \cap C^1(\bar{\Omega})$ enthalten, wenn eine klassische Lösung der Randwertaufgabe (3.4) existiert. Beide Lösungen sind dann identisch.

Beweis. Da die Bilinearform $a : H^1(\Omega) \times H^1(\Omega) \rightarrow \mathbb{R}$ stetig und elliptisch ist, folgt die Existenz einer eindeutigen Lösung $u \in H^1(\Omega)$ aus dem Satz von Lax-Milgram. Insbesondere ist die Bilinearform auch symmetrisch und daher ist u durch

$$a(u, v) = \ell(v) \quad \text{für alle } v \in H^1(\Omega) \quad (3.5)$$

charakterisiert.

Sei nun speziell (3.5) für $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$ erfüllt. Für $v \in H_0^1(\Omega)$ ist $\gamma(v) = 0$ und es folgt

$$a(u, v) = (f, v)_{L^2(\Omega)} \quad \text{für alle } v \in H_0^1(\Omega).$$

Folglich ist u zugleich Lösung eines Dirichlet-Problems, wobei wir uns die Randwerte von u als vorgegeben vorstellen. Das heißt, es ist

$$-\operatorname{div}(\mathbf{A}\nabla u) + cu = f \text{ in } \Omega. \quad (3.6)$$

Für $v \in H^1(\Omega)$ liefert daher der Gaußsche Integralsatz

$$0 = a(u, v) - \ell(v) = \int_{\Omega} \{-\operatorname{div}(\mathbf{A}\nabla u) + cu - f\}v \, d\mathbf{x} + \int_{\Gamma} \{\langle \mathbf{A}\nabla u, \mathbf{n} \rangle - g\}v \, d\sigma.$$

Wegen (3.6) verschwindet das Gebietsintegral auf der rechten Seite. Angenommen, die Funktion $w := \langle \mathbf{A}\nabla u, \mathbf{n} \rangle - g \in C(\Gamma)$ verschwindet nicht. Dann ist $\int_{\Gamma} w^2 \, d\sigma > 0$. Weil $C^1(\overline{\Omega})$ dicht in $C(\overline{\Omega})$ ist, gibt es ein $v \in C^1(\overline{\Omega}) \subset H^1(\Omega)$ mit $\int_{\Gamma} vw \, d\sigma > 0$. Dies ist ein Widerspruch, und die Randbedingung ist erfüllt. \square

Bemerkung Im Gegensatz zu den Dirichlet-Randbedingungen ergeben sich die Neumann-Randbedingungen, ohne dass man sie explizit fordert. Daher spricht man von *natürlichen Randbedingungen*. \triangle

Ist $c(\mathbf{x}) \equiv 0$, so ist mit u offensichtlich für jedes $\eta \in \mathbb{R}$ auch $u + \eta$ eine Lösung von (3.4). Es liegt demnach keine Eindeutigkeit mehr vor und damit kann die Bilinearform a nicht mehr elliptisch sein. Indem man $v = 1$ in die Bilinearform einsetzt, erkennt man ferner, dass die *Kompatibilitätsbedingung*

$$\int_{\Omega} f \, d\mathbf{x} + \int_{\Gamma} g \, d\sigma = 0$$

erfüllt sein muss, damit überhaupt eine Lösung existiert.

Es bezeichne

$$V := \left\{ v \in H^1(\Omega) : \int_{\Omega} v \, d\mathbf{x} = 0 \right\} \subset H^1(\Omega)$$

den Unterraum aller H^1 -Funktionen, deren *Mittelwert*

$$\bar{v} := \frac{1}{|\Omega|} \int_{\Omega} v \, d\mathbf{x}$$

verschwindet. Da in $H^1(\Omega)$ eine Poincaré-Friedrichssche Ungleichung der Form

$$\|v\|_{L^2(\Omega)} \leq c(|\bar{v}| + |v|_{H^1(\Omega)})$$

gezeigt werden kann, folgt, dass die Bilinearform $a(\cdot, \cdot)$ V -elliptisch ist. Die Variationsformulierung

$$\text{suche } u \in V, \text{ so dass } a(u, v) = \ell(v) \quad \text{für alle } v \in V$$

liefert damit eine schwache Lösung des Neumann-Problems (3.4) im Falle $c(\mathbf{x}) \equiv 0$.

Bemerkung Man beachte, dass für jede Konstante $\eta \in \mathbb{R}$ gilt

$$(f + \eta, v)_{L^2(\Omega)} = \int_{\Omega} f v \, d\mathbf{x} + \underbrace{\eta \int_{\Omega} v \, d\mathbf{x}}_{=0} = (f, v)_{L^2(\Omega)} \quad \text{für alle } v \in V.$$

Dies bedeutet, in der Variationsformulierung darf die Funktion $f \in L^2(\Omega)$ um eine beliebige Konstante verschoben werden. Um die Kompatibilitätsbedingung zu erzwingen, macht man den Ansatz $\tilde{f} = f - \eta$ und erhält

$$0 \stackrel{!}{=} \int_{\Omega} \tilde{f} \, d\mathbf{x} + \int_{\Gamma} g \, d\sigma = \int_{\Omega} f \, d\mathbf{x} - \eta|\Omega| + \int_{\Gamma} g \, d\sigma,$$

das heißt,

$$\eta := \frac{1}{|\Omega|} \left\{ \int_{\Omega} f \, d\mathbf{x} + \int_{\Gamma} g \, d\sigma \right\}.$$

△

4. Galerkin-Verfahren

Es sei $a : V \times V \rightarrow \mathbb{R}$ eine stetige, elliptische und symmetrische Bilinearform und $\ell : V \rightarrow \mathbb{R}$ ein beschränktes lineares Funktional. Um die Lösung $u \in V$ des Variationsproblems

$$\text{suche } u \in V, \text{ so dass } a(u, v) = \ell(v) \quad \text{für alle } v \in V \quad (4.1)$$

numerisch zu approximieren, schränken wir uns auf einen endlichdimensionalen Teilraum $V_h \subset V$ ein. Dies führt auf das *Galerkin-Verfahren*:

$$\text{suche } u_h \in V_h, \text{ so dass } a(u_h, v_h) = \ell(v_h) \quad \text{für alle } v_h \in V_h. \quad (4.2)$$

Existenz und Eindeutigkeit der Lösung dieses endlichdimensionalen Variationsproblems ist durch den Satz von Lax-Milgram gesichert. Denn der Raum $V_h \subset V$ ist ein abgeschlossener Teilraum des Hilbert-Raums V , auf dem die Bilinearform elliptisch mit der gleichen Elliptizitätskonstante c_E ist. Insbesondere folgt aus

$$c_E \|u_h\|_V^2 \leq a(u_h, u_h) = \ell(u_h) \leq \|\ell\|_{V'} \|u_h\|_V$$

die Stabilität des Galerkin-Verfahrens:

$$\|u_h\|_V \leq \frac{1}{c_E} \|\ell\|_{V'}.$$

Bemerkung Für den Fall einer symmetrischen Bilinearform ist dem Charakterisierungssatz 3.8 gemäß das Variationsproblem (4.1) äquivalent zum Minimierungsproblem

$$J(v) = \frac{1}{2} a(v, v) - \ell(v) \rightarrow \inf_{v \in V}.$$

Die Ersetzung von V durch V_h , das heißt, der Übergang zu

$$J(v_h) = \frac{1}{2} a(v_h, v_h) - \ell(v_h) \rightarrow \inf_{v_h \in V_h},$$

wird aus historischen Gründen *Ritz-Galerkin-Verfahren* genannt und erschien bereits 1908 in einer Arbeit von W. Ritz. Die Lösung $u_h \in V_h$ dieses endlichdimensionalen Minimierungsproblems erhält man dann wieder nach dem Charakterisierungssatz aus (4.2). \triangle

Satz 4.1 (Céa-Lemma) Die Bilinearform $a : V \times V \rightarrow \mathbb{R}$ sei stetig und elliptisch, und $u \in V$ und $u_h \in V_h \subset V$ seien die Lösungen der Variationsprobleme (4.1) und (4.2). Dann gilt

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

Beweis. Nach Definition von u bzw. u_h gilt

$$\begin{aligned} a(u, v) &= \ell(v) \quad \text{für alle } v \in V, \\ a(u_h, v) &= \ell(v) \quad \text{für alle } v \in V_h. \end{aligned}$$

Wegen $V_h \subset V$ folgt durch Subtraktion

$$a(u - u_h, v) = 0 \quad \text{für alle } v \in V_h. \quad (4.3)$$

Sei $v_h \in V_h$. Mit $v := v_h - u_h \in V_h$ folgt aus (4.3) sofort $a(u - u_h, v_h - u_h) = 0$ und

$$\begin{aligned} c_E \|u - u_h\|_V^2 &\leq a(u - u_h, u - u_h) \\ &= a(u - u_h, u - v_h) + \underbrace{a(u - u_h, v_h - u_h)}_{=0} \\ &\leq c_S \|u - u_h\|_V \|u - v_h\|_V. \end{aligned}$$

Nach Kürzen erhalten wir $c_E \|u - u_h\|_V \leq c_S \|u - v_h\|_V$ und, da $v_h \in V_h$ beliebig war, die Behauptung. \square

Bemerkung Die Beziehung (4.3) wird Galerkin-Orthogonalität genannt. \triangle

Das Céa-Lemma zeigt, dass u_h *quasi-optimal* bezüglich der Minimierung des Fehlers $\|u - u_h\|_V$ ist, das heißt, dieser Ausdruck wird bis auf eine Konstante minimiert. Die Genauigkeit der Lösung hängt demnach wesentlich von der Approximationsgüte des Ansatzraums V_h ab.

Um die Lösung auszurechnen, benötigt man eine Basis $\{\varphi_1, \varphi_2, \dots, \varphi_N\}$ von V_h . Dann ist (4.2) äquivalent zu

$$\text{suche } u_h \in V_h, \text{ so dass } a(u_h, \varphi_i) = \ell(\varphi_i) \quad \text{für alle } i = 1, 2, \dots, N.$$

Der Ansatz

$$u_h = \sum_{j=1}^N z_j \varphi_j$$

führt zu dem linearen Gleichungssystem

$$\sum_{j=1}^N a(\varphi_j, \varphi_i) z_j = \ell(\varphi_i), \quad i = 1, 2, \dots, N,$$

das wir in Matrix-Vektor-Form schreiben:

$$\mathbf{A}_h \mathbf{z}_h = \mathbf{b}_h, \quad \mathbf{A}_h = [a(\varphi_j, \varphi_i)]_{i,j=1}^N, \quad \mathbf{z}_h = [z_j]_{j=1}^N, \quad \mathbf{b}_h = [\ell(\varphi_i)]_{i=1}^N.$$

Die Matrix \mathbf{A}_h wird *Steifigkeitsmatrix* genannt. Sie ist symmetrisch und positiv definit,

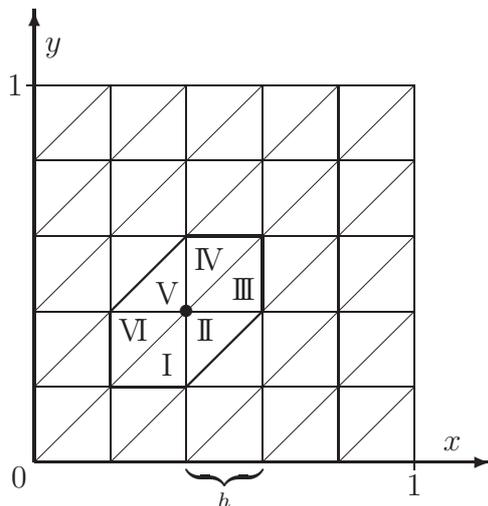
$$\mathbf{z}_h^T \mathbf{A}_h \mathbf{z}_h = \sum_{i,j=1}^N z_i a(\varphi_j, \varphi_i) z_j = a\left(\sum_{j=1}^N z_j \varphi_j, \sum_{i=1}^N z_i \varphi_i\right) \geq c_E \left\| \sum_{i=1}^N z_i \varphi_i \right\|_V^2,$$

da die Norm genau dann 0 ist, wenn $\mathbf{z}_h = \mathbf{0}$ gilt.

Beispiel 4.2 (Courant (1943)) Zu lösen sei die Poisson-Gleichung

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

im Einheitsquadrat $\Omega = (0, 1) \times (0, 1)$. Es werde $\bar{\Omega}$ wie folgt mit einem gleichmäßigen Dreiecksnetz der Maschenweite h überzogen:



Wir wählen

$$V_h = \{v \in C(\bar{\Omega}) : v \text{ ist in jedem Dreieck linear und } v = 0 \text{ auf } \Gamma\}.$$

In jedem Dreieck hat $v \in V_h$ die Form $v(x, y) = a + bx + cy$ und ist durch die Werte an den drei Eckpunkten des Dreiecks eindeutig bestimmt. Deshalb entspricht $\dim V_h = N$ der Anzahl der inneren Gitterpunkte. Ferner ist v global durch die Werte an diesen N Gitterpunkten \mathbf{x}_j gegeben. Zur Diskretisierung wählen wir die *nodale Basis*, die durch

$$\varphi_i(\mathbf{x}_j) = \delta_{i,j}$$

gegeben ist. Wir berechnen die Matrixelemente $[\mathbf{A}_h]_{i,j}$, wobei wir lokale Indizes wählen und Symmetrien ausnutzen. Für die Ableitungen gilt

	I	II	III	IV	V	VI	sonst
$\partial_x \varphi_Z$	0	$-h^{-1}$	$-h^{-1}$	0	h^{-1}	h^{-1}	0
$\partial_y \varphi_Z$	h^{-1}	h^{-1}	0	$-h^{-1}$	$-h^{-1}$	0	0

und daher folgt

$$\begin{aligned} a(\varphi_Z, \varphi_Z) &= \int_{\text{I-VI}} \|\nabla \varphi_Z\|^2 \, d\mathbf{x} \\ &= 2 \int_{\text{I-III}} \{|\partial_x \varphi_Z|^2 + |\partial_y \varphi_Z|^2\} \, d\mathbf{x} \\ &= 2 \int_{\text{II+III}} |\partial_x \varphi_Z|^2 \, d\mathbf{x} + 2 \int_{\text{I+II}} |\partial_y \varphi_Z|^2 \, d\mathbf{x} \\ &= 2h^{-2} \int_{\text{II+III}} d\mathbf{x} + 2h^{-2} \int_{\text{I+II}} d\mathbf{x} \\ &= 4. \end{aligned}$$

Weiter ergibt sich

$$\begin{aligned}
 a(\varphi_Z, \varphi_S) &= \int_{\text{I+II}} \langle \nabla \varphi_Z, \nabla \varphi_S \rangle \, d\mathbf{x} \\
 &= \int_{\text{I+II}} \partial_y \varphi_Z \partial_y \varphi_S \, d\mathbf{x} \\
 &= \int_{\text{I+II}} h^{-1}(-h^{-1}) \, d\mathbf{x} \\
 &= -1
 \end{aligned}$$

und aus Symmetriegründen

$$a(\varphi_Z, \varphi_S) = a(\varphi_Z, \varphi_N) = a(\varphi_Z, \varphi_O) = a(\varphi_Z, \varphi_W) = -1.$$

Schließlich ist

$$a(\varphi_Z, \varphi_{SW}) = \int_{\text{I+VI}} \underbrace{\langle \nabla \varphi_Z, \nabla \varphi_{SW} \rangle}_{=0} \, d\mathbf{x} = 0$$

und wegen der Symmetrie auch $a(\varphi_Z, \varphi_{NO}) = 0$. Demnach entsteht also ein Gleichungssystem mit genau derselben Matrix wie beim Differenzenverfahren mit dem Standard-5-Punkte-Stern:

$$\begin{bmatrix} \alpha_{NW} & \alpha_N & \alpha_{NO} \\ \alpha_W & \alpha_Z & \alpha_O \\ \alpha_{SW} & \alpha_S & \alpha_{SO} \end{bmatrix}_* = \begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}_* .$$

△

5. Finite Elemente

5.1 Vernetzung

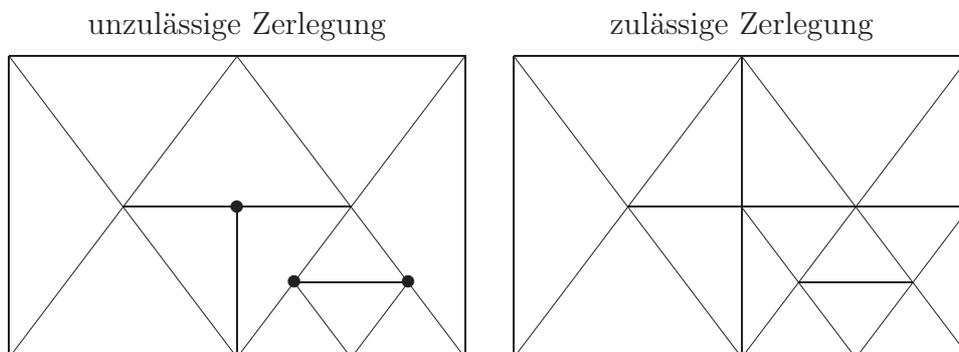
Zur numerischen Approximation der Lösung eines Variationsproblems benötigen wir eine systematische Vorgehensweise zur Konstruktion endlichdimensionaler Teilräume $V_h \subset V$. Dazu sei zunächst $\Omega \subset \mathbb{R}^2$ ein durch einen Polygonzug berandetes Gebiet.

Ein *Finite-Element-Raum* ist charakterisiert durch

1. die Art der Zerlegung: Am gebräuchlichsten sind Zerlegungen in Dreiecks- oder Viereckselemente.
2. die Wahl der Ansatzfunktionen: Die Ansatzfunktionen sind stückweise auf jedem Element definiert durch ein Polynom von vorgegebenem Grad.

Definition 5.1 Eine Zerlegung $\mathcal{T} = \{T_1, T_2, \dots, T_M\}$ von Ω in Dreiecks- oder Viereckselemente heißt **zulässig**, wenn folgende Eigenschaften erfüllt sind:

1. Es ist $\bar{\Omega} = \bigcup_{i=1}^M T_i$.
2. Besteht $T_i \cap T_j$, $i \neq j$, aus genau einem Punkt, so ist dieser ein Eckpunkt sowohl von T_i als auch von T_j .
3. Besteht $T_i \cap T_j$, $i \neq j$, aus mehr als einem Punkt, so ist $T_i \cap T_j$ eine Kante sowohl von T_i als auch von T_j .



Wir betrachten *Familien von Zerlegungen* $\{\mathcal{T}_h\}$, wobei jedes Element $T \in \mathcal{T}_h$ einen Durchmesser von höchstens $2h$ besitzen soll.

Als Ansatzfunktionen wählt man nun beispielsweise stückweise polynomiale Funktionen, das heißt $v|_T \in \mathcal{P}_m$ für alle $T \in \mathcal{T}$, wobei

$$\mathcal{P}_m := \left\{ v(x, y) = \sum_{0 \leq i+j \leq m} \alpha_{i,j} x^i y^j \right\}$$

die *Polynome* von vorgegebenen Grad $\leq m$ bezeichnet. Wie der nachfolgende Satz zeigt, muss eine stückweise polynomiale Funktion global stetig sein, damit sie in $H^1(\Omega)$ liegt.

Satz 5.2 Gegeben sei eine Zerlegung \mathcal{T} des Gebiets Ω und sei $k \geq 1$. Die Funktion $v : \overline{\Omega} \rightarrow \mathbb{R}$ erfülle $v|_T \in C^k(T)$ für jedes $T \in \mathcal{T}$. Dann ist $v \in H^k(\Omega)$ genau dann, wenn $v \in C^{k-1}(\overline{\Omega})$ gilt.

Beweis. Es genügt, den Beweis für $k = 1$ zu führen. Für $k > 1$ folgt die Aussage sofort aus einer Betrachtung der Ableitungen der Ordnung $k - 1$.

“ \Leftarrow ”: Sei $v \in C(\overline{\Omega})$. Für $i = 1, 2$ definieren wir $w_i : \Omega \rightarrow \mathbb{R}$ stückweise gemäß $w_i|_T := \partial_{x_i} v$ für jedes $T \in \mathcal{T}$. Es folgt für $\phi \in C_0^\infty(\Omega)$

$$\int_{\Omega} w_i \phi \, d\mathbf{x} = \sum_{T \in \mathcal{T}} \int_T \partial_{x_i} v \phi \, d\mathbf{x} = \sum_{T \in \mathcal{T}} \left\{ - \int_T v \partial_{x_i} \phi \, d\mathbf{x} + \int_{\partial T} v \phi n_i \, d\sigma \right\}.$$

Da v als stetig vorausgesetzt wurde, heben sich die Integrale über die inneren Kanten gegenseitig auf. Außerdem verschwindet ϕ auf Γ . Es bleibt also nur das Gebietsintegral übrig, dies bedeutet, es gilt

$$\int_{\Omega} w_i \phi \, d\mathbf{x} = - \int_{\Omega} v \partial_{x_i} \phi \, d\mathbf{x} \quad \text{für alle } \phi \in C_0^\infty(\Omega).$$

Folglich ist w_i nach Definition 3.1 die schwache Ableitung von v .

“ \Rightarrow ”: Sei $v \in H^1(\Omega)$. Wir betrachten v in der Umgebung einer Kante und drehen die Kante so, dass sie auf der y -Achse liegt. Sie umfasse speziell das Intervall $[\underline{y} - \delta, \overline{y} + \delta]$ mit $\underline{y} < \overline{y}$ und $\delta > 0$. Sei zunächst $v \in C^\infty(\Omega)$ angenommen. Für die Hilfsfunktion

$$\psi(x) := \int_{\underline{y}}^{\overline{y}} v(x, y) \, dy$$

folgt dann aus der Cauchy-Schwarzschen Ungleichung

$$\begin{aligned} |\psi(\overline{x}) - \psi(\underline{x})|^2 &= \left| \int_{\underline{x}}^{\overline{x}} \int_{\underline{y}}^{\overline{y}} \partial_x v(x, y) \, dy \, dx \right|^2 \\ &\leq \left| \int_{\underline{x}}^{\overline{x}} \int_{\underline{y}}^{\overline{y}} 1 \, dy \, dx \right| |v|_{H^1(\Omega)}^2 \\ &= |\overline{x} - \underline{x}| |\overline{y} - \underline{y}| |v|_{H^1(\Omega)}^2. \end{aligned}$$

Wegen der Dichtheit von $C^\infty(\Omega)$ in $H^1(\Omega)$ gilt diese Aussage auch für $v \in H^1(\Omega)$. Also ist $x \mapsto \psi(x)$ stetig, und zwar insbesondere bei $x = 0$. Da \underline{y} und \overline{y} abgesehen von $\underline{y} < \overline{y}$ beliebig waren, ist das nur möglich, wenn die stückweise stetige Funktion v auf der Kante stetig ist. \square

5.2 Ansatzfunktionen auf Dreieckselementen

Wir untersuchen im folgenden Dreieckselemente, bei denen auf jedem Element der Zerlegung Polynome vom Grad $\leq m$ zugelassen sind.

Lemma 5.3 Sei $m \geq 0$. In einem Dreieck T seien auf $m + 1$ parallelen Linien $\ell = 1 + 2 + \dots + (m + 1)$ Punkte $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_\ell$ angeordnet. Dann gibt es zu jedem $f \in C(T)$ genau ein Polynom p vom Grad m , das die Interpolationsaufgabe

$$p(\mathbf{z}_i) = f(\mathbf{z}_i), \quad i = 1, 2, \dots, \ell$$

löst.

Beweis. Für $m = 0$ ist nichts zu beweisen und wir nehmen an, der Beweis sei schon für $m - 1$ erbracht. Wegen der Invarianz unter affinen Transformationen können wir annehmen, dass die Punkte $\mathbf{z}_i = (x_i, y_i)$ für $i = 1, 2, \dots, m + 1$ auf der x -Achse liegen. Die eindimensionale Theorie liefert ein Polynom $p_0 = p_0(x)$ vom Grad m mit

$$p_0(x_i) = f(x_i, 0), \quad i = 1, 2, \dots, m + 1.$$

Nach Induktionsvoraussetzung gibt es ferner ein Polynom $q = q(x, y)$ vom Grad $m - 1$ mit

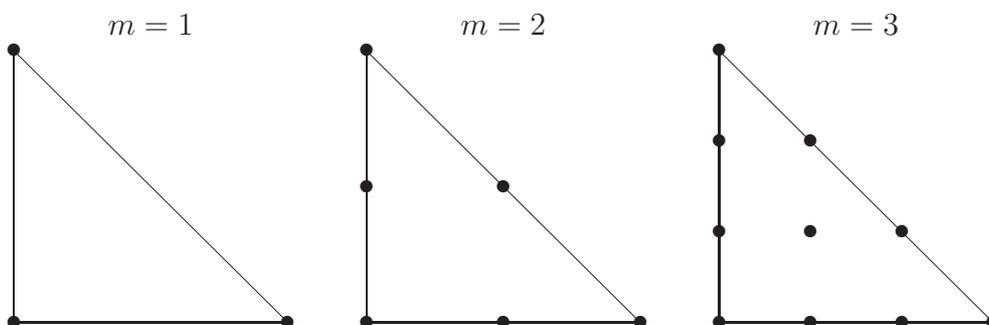
$$q(x_i, y_i) = \frac{1}{y_i} \{f(x_i, y_i) - p_0(x_i)\}, \quad i = m + 2, m + 3, \dots, \ell.$$

Offensichtlich löst $p(x, y) = p_0(x) + yq(x, y) \in \mathcal{P}_m$ die vorgegebene Interpolationsaufgabe. \square

Nach diesem Lemma werden die Ansatzfunktionen folglich eindeutig festgelegt, indem wir ihre Werte in $\ell = (m + 1)(m + 2)/2$ geeignet gewählten Interpolationspunkten vorgeben.

Definition 5.4 Die Polynome aus \mathcal{P}_m , die genau an einem der $\ell = (m + 1)(m + 2)/2$ Interpolationspunkte den Wert 1 annehmen und an allen anderen verschwinden, bilden die **nodale Basis**, auch **Lagrange-Basis** genannt.

Die Knoten der nodalen Basis für lineare, quadratische und kubische Dreieckselemente wählt man wie folgt:



Diese Wahl der Interpolationspunkte zur Konstruktion von Finite-Element-Räumen stellt sicher, dass die Ansatzfunktionen stetig sind. Denn die Restriktion einer Ansatzfunktion auf eine Kante ist ein eindimensionales Polynom, das durch die Festlegung von $m + 1$ Interpolationswerten eindeutig bestimmt ist. Da die gleichen Vorgaben im Nachbarelement gemacht werden, ist die Funktion stetig über die Kanten und damit global stetig.

5.3 Ansatzfunktionen auf Viereckselementen

Auf Rechtecksgittern werden die Polynomfamilien

$$\mathcal{Q}_m := \left\{ v(x, y) = \sum_{0 \leq i, j \leq m} \alpha_{i, j} x^i y^j \right\}$$

verwendet.

Bemerkung Es gilt $\mathcal{P}_m \subset \mathcal{Q}_m \subset \mathcal{P}_{2m}$. △

Lemma 5.5 Sei $m \geq 0$. In einem Rechteck $T = [0, a] \times [0, b]$ betrachte man die $(m+1)^2$ Punkte

$$\{(at_i, bt_j) : 0 \leq i, j \leq m\} \quad \text{mit} \quad 0 \leq t_0 < t_1 < \dots < t_m \leq 1.$$

Dann gibt es zu jedem $f \in C(T)$ genau ein Polynom $p \in \mathcal{Q}_m$, das die Interpolationsaufgabe

$$p(at_i, bt_j) = f(at_i, bt_j), \quad i, j = 0, 1, \dots, m$$

löst.

Beweis. Wir lösen für jedes $j = 0, 1, \dots, m$ nacheinander die eindimensionalen Interpolationsprobleme

$$p_j(at_i) = f(at_i, bt_j), \quad i = 0, 1, \dots, m.$$

Ferner mögen

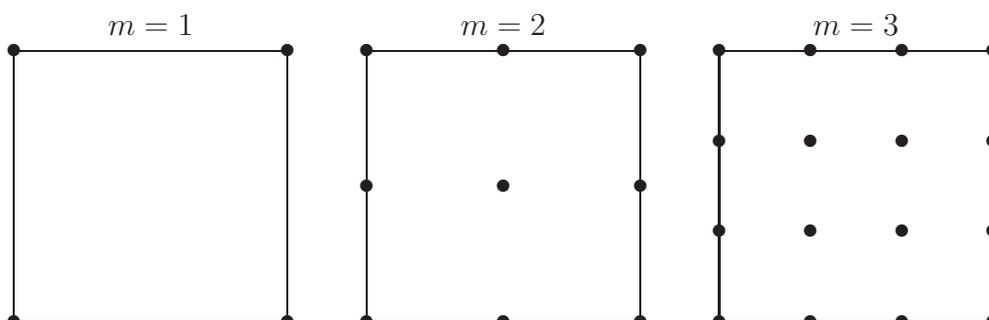
$$L_k(y) := \prod_{\substack{j=0 \\ j \neq k}}^m \frac{y - bt_j}{bt_k - bt_j}$$

die Lagrange-Polynome zu den Knoten $\{bt_0, bt_1, \dots, bt_m\}$ bezeichnen, das heißt, für jedes $k = 0, 1, \dots, m$ ist L_k ein Polynom vom Grad m mit $L_k(bt_j) = \delta_{k,j}$. Dann erfüllt das zusammengesetzte Polynom

$$p(x, y) = \sum_{k=0}^m p_k(x) L_k(y) \in \mathcal{Q}_m$$

offensichtlich die gewünschten Interpolationsbedingungen. □

Um globale Stetigkeit zu gewährleisten, werden auch hier wieder Knoten auf dem Rand des Rechtecks verteilt:



Mit Hilfe von affinen Transformationen lassen sich Rechtecke nur auf Parallelogramme abbilden. Um allgemeine Vierecke zu erzeugen, benötigt man bilineare Transformationen, unter denen dann aber der Polynomraum \mathcal{Q}_m nicht mehr invariant wäre.

Bemerkung Es gibt einen ganzen Zoo von verschiedenen Finiten Elementen, wir haben hier nur die gebräuchlichsten vorgestellt. Die recht populären Serendipity-Elemente sind beispielsweise auf jedem Element aus

$$\tilde{\mathcal{Q}}_2 := \mathcal{Q}_2 \cap \mathcal{P}_3 = \text{span}\{1, x, y, x^2, xy, y^2, x^2y, xy^2\}.$$

Die Interpolationsaufgabe ist durch Vorgabe der Interpolationsbedingungen auf den acht Punkten auf dem Rand des Vierecks eindeutig bestimmt.

Um Finite Elemente für den $H^2(\Omega)$ zu erhalten, benötigt man gemäß Satz 5.2 globale $C^1(\Omega)$ -Funktionen. Diese sind nicht einfach zu konstruieren und benötigen viel mehr Freiheitsgrade. \triangle

5.4 Dreidimensionaler Fall

Sei $\Omega \subset \mathbb{R}^3$ ein Polyeder. Eine Zerlegung $\mathcal{T} = \{T_1, T_2, \dots, T_M\}$ dieses Polyeders ist zulässig, wenn $\bar{\Omega} = \bigcup_{i=1}^M T_i$ und $T_i \cap T_j$, $i \neq j$, entweder leer oder eine gemeinsame Ecke, Kante oder Fläche ist. Man betrachtet Tetraeder-Elemente oder Quader-Elemente, bei denen die Ansatzfunktionen auf jedem Element $T \in \mathcal{T}$ jeweils

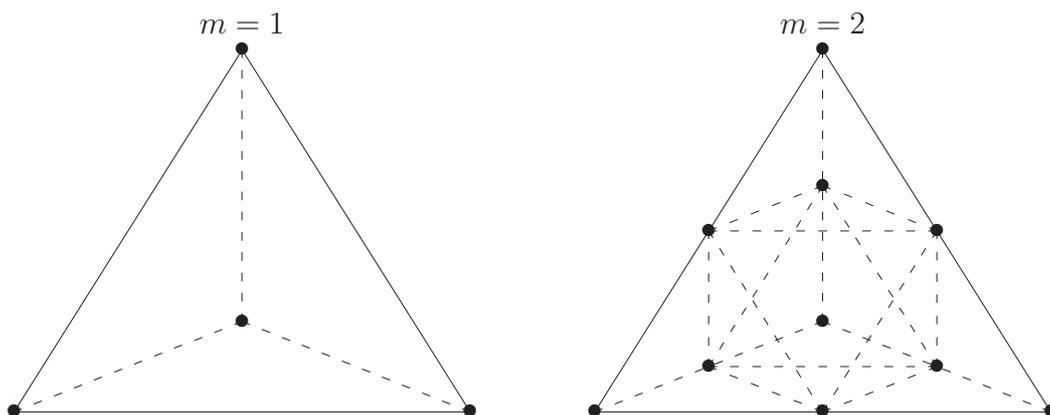
$$v|_T \in \mathcal{P}_m := \left\{ v(x, y, z) = \sum_{0 \leq i+j+k \leq m} \alpha_{i,j,k} x^i y^j z^k \right\}$$

oder

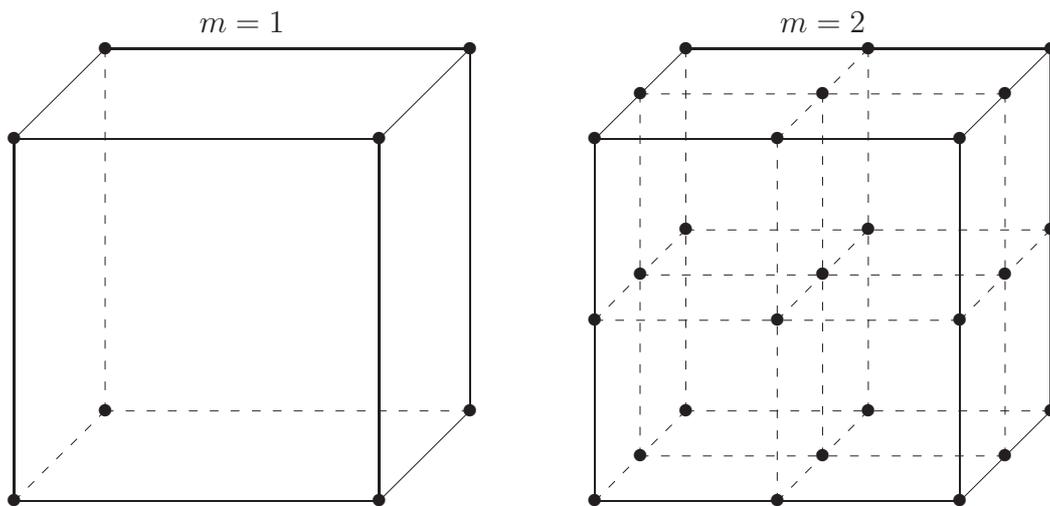
$$v|_T \in \mathcal{Q}_m := \left\{ v(x, y, z) = \sum_{0 \leq i,j,k \leq m} \alpha_{i,j,k} x^i y^j z^k \right\}$$

erfüllen. Satz 5.2 lässt sich völlig analog beweisen, die Ansatzfunktionen müssen also wieder global stetig sein, damit sie in $H^1(\Omega)$ enthalten sind.

Auch die beiden Lemmata 5.3 und 5.5 lassen sich entsprechend verallgemeinern. Auf Tetraedern werden die Ansatzfunktionen durch die Werte in den folgenden Interpolationenpunkten vorgegeben:



Auf Quadern wählt man die Interpolationenpunkte wie folgt:



Durch affine Abbildungen lassen sich beliebige Tetraeder beziehungsweise Spalte erzeugen. Beliebige Hexaeder-Elemente werden hingegen nur durch trilineare Transformationen erzeugt.

5.5 Approximationseigenschaften

Das C ea-Lemma besagt, dass die G ute der Galerkin-Approximation davon abhangt, wie gut sich die L osung im Finite-Element-Raum approximieren lasst. Daher m ussen wir die Approximationseigenschaften der Finite-Element-Raume untersuchen.

Definition 5.6 Eine Familie von Finite-Element-Raumen V_h mit Zerlegungen \mathcal{T}_h von $\Omega \subset \mathbb{R}^d$ heit **affine Familie**, wenn ein Tripel $(T_{\text{ref}}, \mathcal{P}_{\text{ref}}, \Sigma)$ mit den folgenden Eigenschaften existiert:

1. T_{ref} ist ein Polyeder im \mathbb{R}^d .
2. \mathcal{P}_{ref} ist ein Unterraum von $C(T_{\text{ref}})$ mit endlicher Dimension ℓ .
3. Σ ist eine Menge von ℓ linear unabhangigen Funktionalen auf \mathcal{P}_{ref} . Jedes $p \in \mathcal{P}_{\text{ref}}$ ist durch die Werte der ℓ Funktionalen aus Σ eindeutig bestimmt. Da sich die Funktionalen in der Regel auf Funktionswerte und Ableitungen an Punkten in T_{ref} beziehen, spricht man von (verallgemeinerten) Interpolationsbedingungen.
4. Zu jedem $T \in \mathcal{T}_h$ gibt es eine affine Abbildung $F_T : T_{\text{ref}} \rightarrow T$, so dass f ur jedes $v \in V_h$ gilt

$$v(\mathbf{x}) = p(F_T^{-1}\mathbf{x}) \text{ mit } p \in \mathcal{P}_{\text{ref}} \text{ f ur alle } \mathbf{x} \in T.$$

5. Es sei $V_h \subset C^k(\overline{\Omega})$. Die Restriktion von $p \in \mathcal{P}_{\text{ref}}$ auf jede Kante bzw. Flache von T_{ref} sowie die Ableitungen bis zur Ordnung k sind durch solche Interpolationsbedingungen aus Σ eindeutig bestimmt, die sich nur auf Gr oen an Punkten auf dieser Kante bzw. Flache beziehen.

Beispiel 5.7 St uckweise lineare Finite Elemente auf Dreiecken erhalt man durch die Wahl

$$T_{\text{ref}} := \Delta((0, 0), (1, 0), (0, 1)), \quad \mathcal{P}_{\text{ref}} := \mathcal{P}_1, \quad \Sigma := \{\delta_{(0,0)}, \delta_{(1,0)}, \delta_{(0,1)}\}.$$

Hierbei bezeichnet $\delta : C(\mathbb{R}^d) \rightarrow \mathbb{R}$ die *Delta-Distribution*

$$\delta_{\mathbf{y}}(f(\mathbf{x})) := \begin{cases} f(\mathbf{x}), & \text{falls } \mathbf{x} = \mathbf{y}, \\ 0, & \text{sonst.} \end{cases}$$

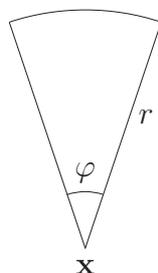
△

Wir haben gesehen, dass in zwei Raumdimensionen $H^1(\Omega)$ -Funktionen nicht stetig sein müssen. Hingegen sind Funktionen aus $H^2(\Omega)$ stetig, wie der folgende Einbettungssatz zeigt.

Satz 5.8 (Lemma von Sobolev) Es sei $\Omega \subset \mathbb{R}^2$ ein durch einen Polygonzug berandetes Gebiet und $m \geq 2$. Dann ist $H^m(\Omega) \subset C(\overline{\Omega})$, insbesondere gibt es eine Konstante $c > 0$, so dass gilt

$$\sup_{\mathbf{x} \in \overline{\Omega}} |v(\mathbf{x})| \leq c \|v\|_{H^2(\Omega)} \leq c \|v\|_{H^m(\Omega)} \quad \text{für alle } v \in H^m(\Omega).$$

Beweis. Wir zeigen zunächst $\sup_{\mathbf{x} \in \overline{\Omega}} |v(\mathbf{x})| \leq c \|v\|_{H^2(\Omega)}$ für $v \in C^\infty(\overline{\Omega}) \cap H^2(\Omega)$. Da ein polygonal berandetes Gebiet die Kegelbedingung erfüllt, gibt es zu jedem $\mathbf{x} \in \overline{\Omega}$ einen Kegel $K_{\mathbf{x}} \subset \overline{\Omega}$ mit Spitze in \mathbf{x} , der den Öffnungswinkel φ und den Radius r besitzt:



Betrachte die Hilfsfunktion $f_r \in C^2(\mathbb{R}^2)$ mit der Eigenschaft

$$f_r(\mathbf{y}) = \begin{cases} 1, & \text{falls } \|\mathbf{x} - \mathbf{y}\| < r/2, \\ 0, & \text{falls } \|\mathbf{x} - \mathbf{y}\| \geq r. \end{cases}$$

Dann ergibt sich für $\mathbf{x} + \rho \mathbf{e}_\theta \in K_{\mathbf{x}}$ mit $\|\mathbf{e}_\theta\| = 1$ mit Hilfe der Produktregel

$$\begin{aligned} v(\mathbf{x}) &= - \left[f_r(\mathbf{x} + \rho \mathbf{e}_\theta) v(\mathbf{x} + \rho \mathbf{e}_\theta) \right]_{\rho=0}^{\rho=r} \\ &= - \int_0^r \frac{\partial (f_r(\mathbf{x} + \rho \mathbf{e}_\theta) v(\mathbf{x} + \rho \mathbf{e}_\theta))}{\partial \rho} d\rho \\ &= - \underbrace{\left[\rho \frac{\partial (f_r(\mathbf{x} + \rho \mathbf{e}_\theta) v(\mathbf{x} + \rho \mathbf{e}_\theta))}{\partial \rho} \right]_{\rho=0}^{\rho=r}}_{=0} + \int_0^r \frac{\partial^2 (f_r(\mathbf{x} + \rho \mathbf{e}_\theta) v(\mathbf{x} + \rho \mathbf{e}_\theta))}{\partial \rho^2} \rho d\rho. \end{aligned}$$

Durch die Substitution $\mathbf{z} := \mathbf{x} + \rho \mathbf{e}_\theta$ folgt hieraus

$$\begin{aligned} \varphi |v(\mathbf{x})| &= \left| \int_0^\varphi \int_0^r \frac{\partial^2 (f_r(\mathbf{x} + \rho \mathbf{e}_\theta) v(\mathbf{x} + \rho \mathbf{e}_\theta))}{\partial \rho^2} \rho \, d\rho \, d\theta \right| \\ &= \left| \int_{K_{\mathbf{x}}} \frac{\partial^2 (f_r(\mathbf{z}) v(\mathbf{z}))}{\partial \rho^2} \, d\mathbf{z} \right| \\ &\leq \sqrt{\int_{K_{\mathbf{x}}} 1^2 \, d\mathbf{z}} \sqrt{\int_{K_{\mathbf{x}}} \left| \frac{\partial^2 (f_r(\mathbf{z}) v(\mathbf{z}))}{\partial \rho^2} \right|^2 \, d\mathbf{z}} \\ &\leq cr \sqrt{\varphi} \|v\|_{H^2(\Omega)}. \end{aligned}$$

Im letzten Schritt haben wir benutzt, dass gilt

$$\left| \frac{\partial^2 (f_r v)}{\partial \rho^2} \right| = \left| \frac{\partial^2 f_r}{\partial \rho^2} v + 2 \frac{\partial f_r}{\partial \rho} \frac{\partial v}{\partial \rho} + f_r \frac{\partial^2 v}{\partial \rho^2} \right| \leq \underbrace{\left| \frac{\partial^2 f_r}{\partial \rho^2} \right|}_{\leq c_f} |v| + 2 \underbrace{\left| \frac{\partial f_r}{\partial \rho} \right|}_{\leq c_f} \left| \frac{\partial v}{\partial \rho} \right| + \underbrace{|f_r|}_{\leq c_f} \left| \frac{\partial^2 v}{\partial \rho^2} \right|.$$

Da $C^\infty(\overline{\Omega}) \cap H^2(\Omega)$ dicht in $H^2(\Omega)$ liegt, ergibt sich schließlich

$$\sup_{\mathbf{x} \in \overline{\Omega}} |v(\mathbf{x})| \leq c \|v\|_{H^2(\Omega)} \quad \text{für alle } v \in H^2(\Omega).$$

Insbesondere lässt sich damit jedes $v \in H^2(\Omega)$ als gleichmäßiger Grenzwert von Funktionen aus $C^\infty(\overline{\Omega}) \cap H^2(\Omega)$ auffassen, weshalb v selbst stetig ist. \square

Satz 5.9 (Rellichscher Auswahlatz) Es sei $m \geq 0$ und Ω ein durch einen Polygonzug berandetes Gebiet. Dann ist die Einbettung $H^{m+1}(\Omega) \hookrightarrow H^m(\Omega)$ kompakt, das heißt, die Einheitskugel des $H^{m+1}(\Omega)$ ist kompakt bezüglich des $H^m(\Omega)$.

Beweis. Der interessierte Leser sei auf J. Wloka “Partielle Differentialgleichungen” verwiesen. \square

Lemma 5.10 Es sei $\Omega \subset \mathbb{R}^2$ ein durch einen Polygonzug berandetes Gebiet. Ferner sei $m \geq 2$ und in $\overline{\Omega}$ seien $\ell = m(m+1)/2$ Punkte $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_\ell$ vorgegeben, an denen die Interpolation $I : H^m(\Omega) \rightarrow \mathcal{P}_{m-1}$ durch Polynome vom Grad $m-1$ eindeutig bestimmt ist. Dann gibt es eine Konstante $c_I = c_I(\Omega, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_\ell)$, so dass gilt

$$\|v - Iv\|_{H^m(\Omega)} \leq c_I |v|_{H^m(\Omega)} \quad \text{für alle } v \in H^m(\Omega).$$

Beweis. Wir führen in $H^m(\Omega)$ die Norm

$$\|v\| := |v|_{H^m(\Omega)} + \sum_{k=1}^{\ell} |v(\mathbf{x}_k)|$$

ein und zeigen, dass die Normen $\|\cdot\|$ und $\|\cdot\|_{H^m(\Omega)}$ äquivalent sind. Dann folgt nämlich aus

$$\begin{aligned} \|v - Iv\|_{H^m(\Omega)} &\leq c\|v - Iv\| \\ &= c\left(|v - Iv|_{H^m(\Omega)} + \sum_{k=1}^{\ell} |(v - Iv)(\mathbf{x}_k)|\right) \\ &= c(|v - Iv|_{H^m(\Omega)}) \\ &= c|v|_{H^m(\Omega)} \end{aligned}$$

die Behauptung. Dabei haben wir ausgenutzt, dass Iv mit v in den Interpolationenpunkten übereinstimmt und dass $\partial_{\mathbf{x}}^{\alpha}(Iv) = 0$ ist für alle $|\alpha| = m$.

Beim Nachweis der Äquivalenz ist eine Richtung schnell erbracht. Nach Lemma 5.8 ist die Einbettung $H^m(\Omega) \hookrightarrow C(\bar{\Omega})$ stetig. Das bewirkt

$$|v(\mathbf{x}_k)| \leq c\|v\|_{H^m(\Omega)} \quad \text{für } k = 1, 2, \dots, \ell$$

und weiter $\|v\| \leq (1 + c\ell)\|v\|_{H^m(\Omega)}$.

Angenommen, die Umkehrung

$$\|v\|_{H^m(\Omega)} \leq c\|v\| \quad \text{für alle } v \in H^m(\Omega)$$

sei für jede positive Zahl c falsch. Dann gibt es eine Folge $\{v_i\}$ mit

$$\|v_i\|_{H^m(\Omega)} = 1, \quad \|v_i\| < \frac{1}{i}, \quad i = 1, 2, \dots$$

Nach dem Rellichschen Auswahlssatz (Satz 5.9) konvergiert eine Teilfolge in $H^{m-1}(\Omega)$. Ohne Beschränkung der Allgemeinheit können wir annehmen, dass es sich dabei um die ganze Folge handelt. Damit ist $\{v_i\}$ eine Cauchy-Folge in $H^{m-1}(\Omega)$. Aus $|v_i|_{H^m(\Omega)} \leq \|v_i\| \rightarrow 0$ und

$$\|v_i - v_j\|_{H^m(\Omega)}^2 \leq \|v_i - v_j\|_{H^{m-1}(\Omega)}^2 + 2\{|v_i|_{H^m(\Omega)}^2 + |v_j|_{H^m(\Omega)}^2\}$$

schließen wir, dass $\{v_i\}$ sogar eine Cauchy-Folge in $H^m(\Omega)$ ist. Wegen der Vollständigkeit des Raums existiert ein Grenzelement $v^* \in H^m(\Omega)$ mit $\|v_i - v^*\|_{H^m(\Omega)} \rightarrow 0$. Aus Stetigkeitsgründen folgt

$$\|v^*\|_{H^m(\Omega)} = 1, \quad \|v^*\| = 0.$$

Dies impliziert $|v^*|_{H^m(\Omega)} = 0$ und damit muss v^* ein Polynom aus \mathcal{P}_{m-1} sein. Wegen $v^*(\mathbf{x}_k) = 0$ für alle $k = 1, 2, \dots, \ell$ ist v^* das Nullpolynom, was im Widerspruch zu $\|v^*\|_{H^m(\Omega)} = 1$ steht. \square

Proposition 5.11 (Bramble-Hilbert-Lemma) Es sei $\Omega \subset \mathbb{R}^2$ ein durch einen Polygonzug berandetes Gebiet und $m \geq 2$. Ist $g : H^m(\Omega) \rightarrow \mathbb{R}$ ein beschränktes, lineares Funktional mit

$$g(p) = 0 \quad \text{für alle } p \in \mathcal{P}_{m-1},$$

dann gilt

$$|g(v)| \leq c|v|_{H^m(\Omega)} \quad \text{für alle } v \in H^m(\Omega).$$

Beweis. Es sei $I : H^m(\Omega) \rightarrow \mathcal{P}_{m-1}$ ein Interpolationsprojektor, der den Voraussetzungen von Lemma 5.10 genügt. Dann folgt

$$|g(v)| = |g(v - Iv)| \leq c_g \|v - Iv\|_{H^m(\Omega)} \leq c_g c_I |v|_{H^m(\Omega)}.$$

□

Definition 5.12 Eine Familie von Zerlegungen $\{\mathcal{T}_h\}$ heißt **nicht entartet**, wenn es eine Zahl $\kappa > 0$ gibt, so dass jedes $T \in \mathcal{T}_h$ einen Kreis vom Radius ρ_T enthält mit

$$\rho_T \geq \frac{h_T}{\kappa}.$$

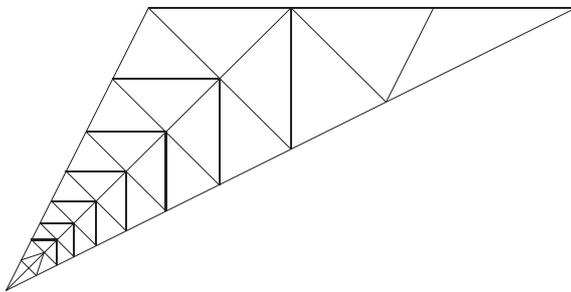
Hierbei bezeichnet $h_T \leq h := \max_{T \in \mathcal{T}_h} h_T$ den halben Durchmesser des Elements T . Ist die untere Schranke für den Inkreisradius ρ_T sogar unabhängig von T , gilt also

$$\rho_T \geq \frac{h}{\kappa},$$

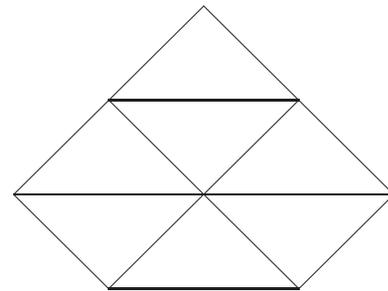
dann heißt die Familie $\{\mathcal{T}_h\}$ **quasi-uniform**.

Beispiel 5.13

nicht entartet:



quasi-uniform:



△

Bemerkung Eine quasi-uniforme Zerlegung ist offenbar auch nicht entartet. Nicht entartete Zerlegungen lassen jedoch im Gegensatz zu quasi-uniformen auch lokal verfeinerte Zerlegungen zu. △

Lemma 5.14 (Transformationsformel) Das Element T und das Referenzelement T_{ref} seien durch eine bijektive affine Abbildung

$$F_T : T_{\text{ref}} \rightarrow T, \quad \hat{\mathbf{x}} \mapsto \mathbf{x} = F_T(\hat{\mathbf{x}}) = \mathbf{B}\hat{\mathbf{x}} + \mathbf{b}$$

einander zugeordnet. Für $v \in H^m(T)$ ist dann die durch

$$v(\mathbf{x}) = v(F_T(\hat{\mathbf{x}})) = \hat{v}(\hat{\mathbf{x}})$$

transformierte Funktion \hat{v} aus $H^m(T_{\text{ref}})$ und es gibt eine Konstante $c = c(m)$, so dass

$$|\hat{v}|_{H^m(T_{\text{ref}})} \leq c \|\mathbf{B}\|_2^m |\det \mathbf{B}|^{-1/2} |v|_{H^m(T)}.$$

Beweis. Sei $v \in C^m(T)$ und damit auch $\hat{v} \in C^m(T_{\text{ref}})$, dann ergibt die Kettenregel

$$\frac{\partial \hat{v}(\hat{\mathbf{x}})}{\partial \hat{x}_i} = \frac{\partial v(\mathbf{x})}{\partial x_1} \frac{\partial x_1}{\partial \hat{x}_i} + \frac{\partial v(\mathbf{x})}{\partial x_2} \frac{\partial x_2}{\partial \hat{x}_i} = b_{1,i} \frac{\partial v(\mathbf{x})}{\partial x_1} + b_{2,i} \frac{\partial v(\mathbf{x})}{\partial x_2}.$$

Hieraus folgt

$$\left| \frac{\partial \hat{v}(\hat{\mathbf{x}})}{\partial \hat{x}_i} \right| \leq \{|b_{1,i}| + |b_{2,i}|\} \max_{j=1,2} \left| \frac{\partial v(\mathbf{x})}{\partial x_j} \right| \leq \|\mathbf{B}\|_1 \max_{j=1,2} \left| \frac{\partial v(\mathbf{x})}{\partial x_j} \right|$$

und rekursiv für beliebiges $\alpha \in \mathbb{N}_0^2$

$$|\partial_{\hat{\mathbf{x}}}^{\alpha} \hat{v}(\hat{\mathbf{x}})| \leq \|\mathbf{B}\|_1^{|\alpha|} \max_{\beta \text{ mit } |\beta|=|\alpha|} |\partial_{\mathbf{x}}^{\beta} v(\mathbf{x})|.$$

Deshalb ist

$$\sum_{|\alpha|=m} |\partial_{\hat{\mathbf{x}}}^{\alpha} \hat{v}(\hat{\mathbf{x}})|^2 \leq c \|\mathbf{B}\|_2^{2m} \sum_{|\beta|=m} |\partial_{\mathbf{x}}^{\beta} v(\mathbf{x})|^2$$

und somit

$$\begin{aligned} |\hat{v}|_{H^m(T_{\text{ref}})}^2 &= \sum_{|\alpha|=m} \int_{T_{\text{ref}}} |\partial_{\hat{\mathbf{x}}}^{\alpha} \hat{v}(\hat{\mathbf{x}})|^2 d\hat{\mathbf{x}} \\ &\leq c \|\mathbf{B}\|_2^{2m} \sum_{|\beta|=m} \int_{T_{\text{ref}}} |\partial_{\mathbf{x}}^{\beta} v(F_T(\hat{\mathbf{x}}))|^2 d\hat{\mathbf{x}} \\ &= c \|\mathbf{B}\|_2^{2m} \sum_{|\beta|=m} \int_T |\partial_{\mathbf{x}}^{\beta} v(\mathbf{x})|^2 |\det(\mathbf{B}^{-1})| d\mathbf{x} \\ &= c \|\mathbf{B}\|_2^{2m} |\det \mathbf{B}|^{-1} |v|_{H^m(T)}^2. \end{aligned}$$

Aufgrund der Dichtheit von $C^m(T) \cap H^m(T)$ in $H^m(T)$ folgt die Behauptung. \square

Das nachfolgende Lemma schätzt für ein beliebiges $T \in \mathcal{T}_h$ die Normen von \mathbf{B} und \mathbf{B}^{-1} explizit bezüglich des Parameters h ab. Die Konstanten sind unabhängig von T , vorausgesetzt, die Zerlegungen $\{\mathcal{T}_h\}$ sind quasi-uniform.

Lemma 5.15 Es sei T_{ref} ein festes, von der Zerlegung unabhängiges Referenzelement. Das Element T gehe aus T_{ref} durch eine bijektive affine Abbildung

$$F_T : T_{\text{ref}} \rightarrow T, \quad \hat{\mathbf{x}} \mapsto \mathbf{x} = F_T(\hat{\mathbf{x}}) = \mathbf{B}\hat{\mathbf{x}} + \mathbf{b}$$

hervor. Ferner enthalte T einen Kreis mit Radius h/κ und werde umschrieben von einem Kreis mit Radius h . Dann gilt

$$\|\mathbf{B}\|_2 \leq ch, \quad \|\mathbf{B}^{-1}\|_2 \leq c \frac{\kappa}{h}$$

mit einer Konstanten c für alle T mit dieser Eigenschaft.

Beweis. Es gibt Kreise mit Radius ρ_{ref} und h_{ref} , die in T_{ref} enthalten sind bzw. T_{ref} umschreiben. Demnach existiert auch ein $\hat{\mathbf{x}}_0 \in T_{\text{ref}}$, so dass $\hat{\mathbf{x}} \in T_{\text{ref}}$ für alle $\|\hat{\mathbf{x}} - \hat{\mathbf{x}}_0\|_2 \leq \rho_{\text{ref}}$.

Für $\mathbf{x}_0 := \mathbf{B}\hat{\mathbf{x}}_0 + \mathbf{b} \in T$ und $\mathbf{x} := \mathbf{B}\hat{\mathbf{x}} + \mathbf{b} \in T$ folgt $\mathbf{x} - \mathbf{x}_0 = \mathbf{B}(\hat{\mathbf{x}} - \hat{\mathbf{x}}_0)$ und $\|\mathbf{x} - \mathbf{x}_0\|_2 \leq 2h$. Hieraus schließen wir

$$\|\mathbf{B}\|_2 = \frac{1}{\rho_{\text{ref}}} \sup_{\|\hat{\mathbf{z}}\|_2 = \rho_{\text{ref}}} \|\mathbf{B}\hat{\mathbf{z}}\|_2 = \frac{1}{\rho_{\text{ref}}} \sup_{\substack{\|\hat{\mathbf{z}}\|_2 = \rho_{\text{ref}} \\ \mathbf{x} = \mathbf{B}(\hat{\mathbf{z}} + \hat{\mathbf{x}}_0) + \mathbf{b}}} \|\mathbf{x} - \mathbf{x}_0\|_2 \leq \frac{2h}{\rho_{\text{ref}}}.$$

Umgekehrt gibt es ein \mathbf{z}_0 mit $\mathbf{x} = \mathbf{z}_0 + \mathbf{z} \in T$ für alle $\|\mathbf{z}\|_2 \leq h/\kappa$ und folglich ist

$$\|\mathbf{B}^{-1}\|_2 = \frac{\kappa}{h} \sup_{\|\mathbf{z}\|_2 = h/\kappa} \|\mathbf{B}^{-1}\mathbf{z}\|_2 \leq 2h_{\text{ref}} \frac{\kappa}{h}.$$

□

Bemerkung Unter den Voraussetzungen von Lemma 5.15 kann man sofort folgende Schranken für die Determinanten von \mathbf{B} und \mathbf{B}^{-1} angeben. Wegen

$$\int_T 1 \, d\mathbf{x} = \int_{T_{\text{ref}}} |\det \mathbf{B}| \, d\hat{\mathbf{x}} \leq \pi h^2$$

folgt nämlich $|\det \mathbf{B}|^{1/2} \leq ch$ und wegen

$$|\det \mathbf{B}|^{-1} \leq \frac{\kappa^2}{\pi h^2} \int_T |\det \mathbf{B}|^{-1} \, d\mathbf{x} \leq \frac{\kappa^2}{\pi h^2} \int_{T_{\text{ref}}} 1 \, d\hat{\mathbf{x}}$$

folgt $|\det \mathbf{B}|^{-1/2} \leq c\kappa/h$. △

Wir können nun als Hauptresultat dieses Abschnittes die folgende Abschätzung für den Approximationsfehler in V_h beweisen. Den Approximationsfehler drücken wir in der gitterabhängigen Norm

$$\|v\|_{m,h} := \sqrt{\sum_{T \in \mathcal{T}_h} \|v\|_{H^m(T)}^2}$$

aus, da dann keine globale Glattheit von der Funktion v verlangt wird. Offensichtlich gilt jedoch

$$\|v\|_{m,h} = \|v\|_{H^m(\Omega)} \quad \text{für alle } v \in H^m(\Omega).$$

Satz 5.16 (Approximationsabschätzung) Es seien $\Omega \subset \mathbb{R}^2$ ein durch einen Polygonzug berandetes Gebiet, $k \geq 2$, $0 \leq m \leq k$, und $\{\mathcal{T}_h\}$ eine quasi-uniforme Familie von Zerlegungen. Der Operator I_h bezeichne die stückweise Interpolation durch Polynome vom Grad $k-1$. Dann gilt

$$\|v - I_h v\|_{m,h} \leq c_{\text{int}} h^{k-m} |v|_{H^k(\Omega)} \quad \text{für alle } v \in H^k(\Omega)$$

mit einer Konstante c_{int} , die nur von Ω , κ und k abhängt.

Beweis. Es gilt für alle $0 \leq \ell \leq m$ gemäß Lemmata 5.14 und 5.15

$$|v - I_h v|_{H^\ell(T)} \leq c \underbrace{\|\mathbf{B}^{-1}\|_2^\ell}_{\leq (c\kappa/h)^\ell} |\det \mathbf{B}|^{1/2} |\hat{v} - I\hat{v}|_{H^\ell(T_{\text{ref}})} \leq ch^{-\ell} |\det \mathbf{B}|^{1/2} |\hat{v} - I\hat{v}|_{H^\ell(T_{\text{ref}})}.$$

Aus Lemma 5.10 folgt

$$|\hat{v} - I\hat{v}|_{H^\ell(T_{\text{ref}})} \leq \|\hat{v} - I\hat{v}\|_{H^k(T_{\text{ref}})} \leq c_I |\hat{v}|_{H^k(T_{\text{ref}})},$$

und damit die Abschätzung

$$|v - I_h v|_{H^\ell(T)} \leq ch^{-\ell} |\det \mathbf{B}|^{1/2} |\hat{v}|_{H^k(T_{\text{ref}})}.$$

Der Rücktransport auf das Element T liefert wieder mit Hilfe der Lemmata 5.14 und 5.15

$$|v - I_h v|_{H^\ell(T)} \leq ch^{-\ell} |\det \mathbf{B}|^{1/2} \underbrace{\|\mathbf{B}\|_2^k}_{\leq (ch)^k} |\det \mathbf{B}|^{-1/2} |v|_{H^k(T)} \leq ch^{k-\ell} |v|_{H^k(T)}.$$

Durch Aufsummation ergibt schließlich

$$\|v - I_h v\|_{H^m(T)}^2 = \sum_{\ell=0}^m |v - I_h v|_{H^\ell(T)}^2 \leq c |v|_{H^k(T)}^2 \sum_{\ell=0}^m h^{2(k-\ell)} \leq ch^{2(k-m)} |v|_{H^k(T)}^2$$

und daher die Behauptung. \square

Bemerkung Für $v \in H^2(\Omega)$ und stetige, stückweise lineare Ansatzfunktionen auf Dreiecken gelten demnach die Abschätzungen

$$\|v - I_h v\|_{H^1(\Omega)} \leq c_{int} h |v|_{H^2(\Omega)}, \quad \|v - I_h v\|_{L^2(\Omega)} \leq c_{int} h^2 |v|_{H^2(\Omega)}.$$

Diese Abschätzungen gelten auch für stetige, stückweise bilineare Ansatzfunktionen auf Parallelogrammen, da hier die Elementabbildungen ebenfalls affin sind. Im Fall beliebiger Vierecke benötigt man noch eine zusätzliche Bedingung an die Vierecke, um sicherzustellen, dass diese nicht degenerieren. Dann gelten beide Abschätzungen auch auf beliebigen Vierecken. \triangle

Für Funktionen $v_h \in V_h$ kann man stärkere Sobolev-Normen durch schwächere abschätzen, wenn man entsprechende h -Potenzen opfert. Dies ist die Aussage der folgenden *inversen Abschätzung*.

Satz 5.17 (inverse Abschätzung) Sei $\{\mathcal{T}_h\}$ eine quasi-uniforme Familie von Zerlegungen des Gebiets Ω . Der Ansatzraum V_h bestehe aus durch stückweise Polynome vom Grad s gegebenen Funktionen. Dann gibt es eine Konstante c_{inv} , welche nur von s , t und κ abhängt, so dass für $0 \leq m \leq t$ gilt

$$\|v_h\|_{t,h} \leq c_{inv} h^{m-t} \|v_h\|_{m,h} \quad \text{für alle } v_h \in V_h.$$

Beweis. Es genügt,

$$|v|_{H^t(T_{\text{ref}})} \leq c |v|_{H^m(T_{\text{ref}})} \quad \text{für alle } v \in \mathcal{P}_{\text{ref}} \quad (5.1)$$

zu zeigen. Mit der Transformationsformel aus Lemma 5.14 folgt dann genau wie im Beweis von Satz 5.16 die Umrechnung auf die einzelnen Elemente. Dabei kommt der Faktor ch^{m-t} in die Abschätzung. Die Summation der quadrierten Ausdrücke über alle Elemente liefert dann die Behauptung.

Zum Nachweis von (5.1) seien $\ell := m(m+1)/2$ Punkte $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_\ell \in T_{\text{ref}}$ so gewählt, dass die Interpolation I in \mathcal{P}_{m-1} stets eine eindeutige Lösung besitzt. Dann sind

$$\|v\| := |v|_{H^m(T_{\text{ref}})} + \sum_{j=1}^{\ell} |v(\mathbf{x}_j)|$$

und $\|\cdot\|_{H^m(T_{\text{ref}})}$ äquivalente Normen (vgl. Beweis von Lemma 5.10). Außerdem sind auf $\mathcal{P}_{\text{ref}} \oplus \mathcal{P}_{m-1}$ wegen der endlichen Dimension $\|\cdot\|$ und $\|\cdot\|_{H^t(T_{\text{ref}})}$ ebenfalls äquivalent. Schließlich ist wegen $Iv \in \mathcal{P}_{m-1}$ stets $|Iv|_{H^t(T_{\text{ref}})} = |Iv|_{H^m(T_{\text{ref}})} = 0$. Zusammen folgt nun (5.1):

$$\begin{aligned} |v|_{H^t(T_{\text{ref}})} &= |v - Iv|_{H^t(T_{\text{ref}})} \leq \|v - Iv\|_{H^t(T_{\text{ref}})} \leq c\|v - Iv\| \\ &= c \left\{ |v - Iv|_{H^m(T_{\text{ref}})} + \sum_{j=1}^{\ell} \underbrace{|(v - Iv)(\mathbf{x}_j)|}_{=0} \right\} \\ &= c|v|_{H^m(T_{\text{ref}})}. \end{aligned}$$

□

Bemerkung Für stetige, stückweise lineare Ansatzfunktionen auf Dreiecken haben wir demnach

$$\|v_h\|_{H^1(\Omega)} \leq c_{inv} h^{-1} \|v_h\|_{L^2(\Omega)} \quad \text{für alle } v_h \in V_h.$$

Diese Abschätzung gilt ebenfalls für stetige, stückweise bilineare Ansatzfunktionen auf Vierecken. △

6. Fehleranalyse

Aus Satz 5.16 haben wir gefolgert, dass für eine quasi-uniforme Familie von Zerlegungen, basierend auf stückweise linearen Ansatzfunktionen auf Dreiecken oder stückweise bilinearen Ansatzfunktionen auf Vierecken, gilt

$$\|v - I_h v\|_{H^1(\Omega)} \leq c_{int} h |v|_{H^2(\Omega)} \quad \text{für alle } v \in H^2(\Omega).$$

Nach dem Céa-Lemma erhalten wir daraus

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{c_S}{c_E} c_{int} h |u|_{H^2(\Omega)},$$

vorausgesetzt, wir können $u \in H^2(\Omega)$ zeigen.

Definition 6.1 Es sei $H_0^1(\Omega) \subset V \subset H^1(\Omega)$ und $a(\cdot, \cdot)$ eine V -elliptische Bilinearform. Das Variationsproblem

$$\text{suche } u \in V, \text{ so dass } a(u, v) = (f, v)_{L^2(\Omega)} \quad \text{für alle } v \in V$$

heißt **$H^s(\Omega)$ -regulär**, wenn es eine Konstante c_R gibt, so dass zu jedem $f \in H^{s-2}(\Omega)$ eine Lösung $u \in H^s(\Omega)$ existiert mit

$$\|u\|_{H^s(\Omega)} \leq c_R \|f\|_{H^{s-2}(\Omega)}.$$

Diese Definition wird zunächst nur für $s \geq 2$ herangezogen. Diese Einschränkung entfällt, wenn negative Normen erklärt sind.

Beispiel 6.2 Auf dem Gebiet

$$\Omega := \{(r \cos \varphi, r \sin \varphi) : 0 < r < 1, 0 < \varphi < \omega\}$$

mit den Randkomponenten

$$\begin{aligned} \Gamma_1 &= \{(r, 0) : 0 \leq r \leq 1\}, \\ \Gamma_2 &= \{(r \cos \omega, r \sin \omega) : 0 \leq r \leq 1\}, \\ \Gamma_3 &= \{(\cos \varphi, \sin \varphi) : 0 \leq \varphi \leq \omega\} \end{aligned}$$

betrachten wir die Funktion

$$u(x, y) = \hat{u}(r, \varphi) = (r^2 - r^{\frac{\pi}{\omega}}) \sin\left(\frac{\pi}{\omega} \varphi\right).$$

Mit $x = r \cos \varphi$ und $y = r \sin \varphi$ ergibt sich

$$\frac{1}{r} \frac{\partial \hat{u}}{\partial r} + \frac{\partial^2 \hat{u}}{\partial r^2} + \frac{1}{r^2} \frac{\partial^2 \hat{u}}{\partial \varphi^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

Demnach erhalten wir

$$\begin{aligned} u_{xx} + u_{yy} &= \left(2 - \frac{\pi}{\omega} r^{\frac{\pi}{\omega}-2}\right) \sin\left(\frac{\pi}{\omega} \varphi\right) + \left[2 - \frac{\pi}{\omega} \left(\frac{\pi}{\omega} - 1\right) r^{\frac{\pi}{\omega}-2}\right] \sin\left(\frac{\pi}{\omega} \varphi\right) \\ &\quad - (1 - r^{\frac{\pi}{\omega}-2}) \frac{\pi^2}{\omega^2} \sin\left(\frac{\pi}{\omega} \varphi\right) \\ &= \left(4 - \frac{\pi^2}{\omega^2}\right) \sin\left(\frac{\pi}{\omega} \varphi\right). \end{aligned}$$

Dies bedeutet, die Funktion u ist eindeutige Lösung des Dirichlet-Problems

$$-\Delta u = \left(\frac{\pi^2}{\omega^2} - 4\right) \sin\left(\frac{\pi}{\omega} \varphi\right) \text{ in } \Omega, \quad u = 0 \text{ auf } \Gamma_1 \cup \Gamma_2 \cup \Gamma_3.$$

Es liegt $r^{\pi/\omega} \sin(\pi\varphi/\omega)$ und damit auch u in $H^2(\Omega)$ genau dann, wenn $\pi/\omega \geq 1$, also wenn $\omega \leq \pi$ gilt. Da aber die rechte Seite wegen $\sin(\pi\varphi/\omega)$ in $L^2(\Omega)$ liegt, ist das Problem für $\omega > \pi$ nicht $H^2(\Omega)$ -regulär. \triangle

Satz 6.3 ($H^2(\Omega)$ -Regulariät) Es sei $\Omega \subset \mathbb{R}^2$ ein durch einen Polygonzug berandetes konvexes Gebiet und

$$a(u, v) = \int_{\Omega} \langle \mathbf{A} \nabla u, \nabla v \rangle \, d\mathbf{x}, \quad u, v \in H_0^1(\Omega)$$

eine elliptische Bilinearform mit Lipschitz-stetigen Koeffizienten $a_{i,j}$. Dann ist das Variationsproblem

$$\text{suche } u \in H_0^1(\Omega), \text{ so dass } a(u, v) = (f, v)_{L^2(\Omega)} \text{ für alle } v \in H_0^1(\Omega)$$

$H^2(\Omega)$ -regulär.

Beweis. Einen Beweis dieses Satzes findet der interessierte Leser zum Beispiel in P. Grisvard "Elliptic Problems in Nonsmooth Domains". \square

Satz 6.4 (Konvergenz) Es sei $\Omega \subset \mathbb{R}^2$ ein durch einen Polygonzug berandetes konvexes Gebiet und $\{\mathcal{T}_h\}$ eine quasi-uniforme Familie von Zerlegungen von Ω . Ist $f \in L^2(\Omega)$, dann erfüllt die mit dem Galerkin-Verfahren berechnete Näherungslösung $u_h \in V_h$ die Abschätzung

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{c_S}{c_E} c_{int} c_R h \|f\|_{L^2(\Omega)}.$$

Beweis. Nach Satz 6.3 ist das zugrundeliegende Variationsproblem $H^2(\Omega)$ -regulär, das heißt, seine Lösung u erfüllt $\|u\|_{H^2(\Omega)} \leq c_R \|f\|_{L^2(\Omega)}$. Nach Satz 5.16 gibt es daher ein $v_h \in V_h$ mit

$$\|u - v_h\|_{H^1(\Omega)} \leq c_{int} h \|u\|_{H^2(\Omega)} \leq c_{int} h \|u\|_{H^2(\Omega)}.$$

Mit Hilfe des Céa-Lemmas (Satz 4.1) folgt schließlich

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{c_S}{c_E} \|u - v_h\|_{H^1(\Omega)} \leq \frac{c_S}{c_E} c_{int} c_R h \|f\|_{L^2(\Omega)}.$$

□

Bemerkung Bei quadratischen Finiten Elementen erhält man nach Satz 5.16 eine höhere Fehlerordnung, sofern $H^3(\Omega)$ -Regularität vorliegt. Um jedoch $H^3(\Omega)$ -reguläre Lösungen zu erhalten, muss im allgemeinen ein glattes und damit krumm berandetes Gebiet zugrundeliegen, welches dann nicht mehr in Dreiecke zerlegt werden kann. \triangle

Satz 6.5 (Aubin-Nitsche-Lemma) Sei H ein Hilbert-Raum mit der Norm $\|\cdot\|_H$ und dem Skalarprodukt (\cdot, \cdot) . Es sei V ein Unterraum, der durch die Norm $\|\cdot\|_V$ zum Hilbert-Raum wird. Ferner sei die Einbettung $V \hookrightarrow H$ stetig, das heißt, es ist $\|v\|_H \leq c\|v\|_V$ für alle $v \in V$.

Vorgelegt sei das Variationsproblem

$$\text{suche } u \in V, \text{ so dass } a(u, v) = (f, v) \quad \text{für alle } v \in V, \quad (6.1)$$

wobei die Bilinearform $a : V \times V \rightarrow \mathbb{R}$ stetig und V -elliptisch sei. Dann gilt für die Finite-Element-Lösung u_h in $V_h \subset V$

$$\|u - u_h\|_H \leq c_S \|u - u_h\|_V \sup_{g \in H \setminus \{0\}} \left\{ \frac{1}{\|g\|_H} \inf_{v_h \in V_h} \|\varphi_g - v_h\|_V \right\},$$

wenn jedem $g \in H$ die eindeutige (schwache) Lösung $\varphi_g \in V$ des Variationsproblems

$$\text{suche } \varphi_g \in V, \text{ so dass } a(w, \varphi_g) = (g, w) \quad \text{für alle } w \in V \quad (6.2)$$

zugeordnet wird.

Beweis. Die Norm eines Elements in einem Hilbert-Raum lässt sich mittels eines Dualitätsarguments bestimmen:

$$\|w\|_H = \sup_{g \in H \setminus \{0\}} \frac{(g, w)}{\|g\|_H}. \quad (6.3)$$

Wir erinnern, dass u und u_h durch

$$\begin{aligned} a(u, v) &= (f, v) \quad \text{für alle } v \in V, \\ a(u_h, v_h) &= (f, v_h) \quad \text{für alle } v_h \in V_h \end{aligned}$$

gegeben sind. Deshalb ist $a(u - u_h, v_h) = 0$ für alle $v_h \in V_h$. Weiter folgt, wenn wir in (6.2) $w := u - u_h$ setzen, dass

$$(g, u - u_h) = a(u - u_h, \varphi_g) = a(u - u_h, \varphi_g - v_h) \leq c_S \|u - u_h\|_V \|\varphi_g - v_h\|_V.$$

Das Dualitätsargument (6.3) liefert nun

$$\|u - u_h\|_H = \sup_{g \in H \setminus \{0\}} \frac{(g, u - u_h)}{\|g\|_H} \leq c_S \|u - u_h\|_V \sup_{g \in H \setminus \{0\}} \frac{\|\varphi_g - v_h\|_V}{\|g\|_H}.$$

□

Bemerkung Das Variationsproblem (6.2) heißt das zu (6.1) *duale Problem*. \triangle

Proposition 6.6 (L^2 -Fehlerabschätzung) Unter den Voraussetzungen von Satz 6.4 gilt

$$\|u - u_h\|_{L^2(\Omega)} \leq c_S c_{int} c_R h \|u - u_h\|_{H^1(\Omega)}.$$

Gilt außerdem $f \in L^2(\Omega)$ und damit $u \in H^2(\Omega)$, dann folgt

$$\|u - u_h\|_{L^2(\Omega)} \leq \frac{c_S^2}{c_E} c_{int}^2 c_R^2 h^2 \|f\|_{L^2(\Omega)}.$$

Beweis. Wegen $H_0^1(\Omega) \subset L^2(\Omega)$ und $\|v\|_{L^2(\Omega)} \leq \|v\|_{H^1(\Omega)}$ für alle $v \in H_0^1(\Omega)$ ist die Einbettung $H_0^1(\Omega) \hookrightarrow L^2(\Omega)$ stetig. Damit ist das Aubin-Nitsche-Lemma mit

$$\begin{aligned} H &:= L^2(\Omega), & \|\cdot\|_H &:= \|\cdot\|_{L^2(\Omega)}, \\ V &:= H_0^1(\Omega), & \|\cdot\|_V &:= \|\cdot\|_{H^1(\Omega)}, \end{aligned}$$

anwendbar.

Für gegebenes $g \in L^2(\Omega)$ sei $\varphi_g \in H^2(\Omega)$ die Lösung des dualen Problems (6.2). Die Approximationseigenschaft zusammen mit der $H^2(\Omega)$ -Regularität liefern dann

$$\|\varphi_g - I_h \varphi_g\|_{H^1(\Omega)} \leq c_{int} h \|\varphi_g\|_{H^2(\Omega)} \leq c_{int} h \|\varphi_g\|_{H^2(\Omega)} \leq c_{int} c_R h \|g\|_{L^2(\Omega)}.$$

Zusammen mit dem Aubin-Nitsche-Lemma folgt

$$\begin{aligned} \|u - u_h\|_{L^2(\Omega)} &\leq c_S \|u - u_h\|_{H^1(\Omega)} \sup_{g \in L^2(\Omega) \setminus \{0\}} \inf_{v_h \in V_h} \frac{\|\varphi_g - v_h\|_{H^1(\Omega)}}{\|g\|_{L^2(\Omega)}} \\ &\leq c_S c_{int} c_R h \|u - u_h\|_{H^1(\Omega)}. \end{aligned}$$

Anwendung von Satz 6.4 liefert schließlich auch den zweiten Teil der Behauptung. \square

Satz 6.7 (L^∞ -Fehlerabschätzung) Unter den Voraussetzungen von Satz 6.4 gilt

$$\sup_{\mathbf{x} \in \Omega} |u(\mathbf{x}) - u_h(\mathbf{x})| \leq ch \|f\|_{L^2(\Omega)}.$$

Beweis. Zu $u \in H^2(\Omega)$ sei $I_h u$ die Interpolierende. Wir zeigen zunächst

$$\sup_{\mathbf{x} \in T} |u(\mathbf{x}) - I_h u(\mathbf{x})| \leq ch \|f\|_{L^2(\Omega)} \quad \text{für alle } T \in \mathcal{T}_h,$$

was dann sofort

$$\sup_{\mathbf{x} \in \Omega} |u(\mathbf{x}) - I_h u(\mathbf{x})| \leq ch \|f\|_{L^2(\Omega)} \tag{6.4}$$

impliziert.

Sei $T \in \mathcal{T}_h$ beliebig aber fest. Es bezeichne \hat{u} die durch die affine Transformation resultierende Funktion auf T_{ref} und $I\hat{u}$ die Interpolierende in \mathcal{P}_{ref} . Wegen der Einbettung $H^2(T_{\text{ref}}) \hookrightarrow C(T_{\text{ref}})$ folgt

$$\sup_{\mathbf{x} \in T} |u(\mathbf{x}) - I_h u(\mathbf{x})| = \sup_{\hat{\mathbf{x}} \in T_{\text{ref}}} |\hat{u}(\hat{\mathbf{x}}) - I\hat{u}(\hat{\mathbf{x}})| \leq c \|\hat{u} - I\hat{u}\|_{H^2(T_{\text{ref}})}.$$

Lemmata 5.10 und 5.14 liefern

$$\sup_{\mathbf{x} \in T} |u(\mathbf{x}) - I_h u(\mathbf{x})| \leq c |\hat{u}|_{H^2(T_{\text{ref}})} \leq ch |u|_{H^2(T)},$$

und somit die gewünschte Abschätzung

$$\sup_{\mathbf{x} \in T} |u(\mathbf{x}) - I_h u(\mathbf{x})| \leq ch \|u\|_{H^2(\Omega)} \leq ch \|f\|_{L^2(\Omega)}.$$

Es seien $v \in V_h$ und \hat{v} die affin auf T_{ref} transformierte Funktion $v|_T$. Da $\hat{v} \in \mathcal{P}_{\text{ref}}$ ist, gilt

$$\sup_{\hat{\mathbf{x}} \in T_{\text{ref}}} |\hat{v}(\hat{\mathbf{x}})| \leq c \|\hat{v}\|_{L^2(T_{\text{ref}})}.$$

Dies bedeutet gemäß Lemma 5.14, dass

$$\sup_{\mathbf{x} \in T} |v(\mathbf{x})| \leq \frac{c}{h} \|v\|_{L^2(T)},$$

und folglich ist

$$\sup_{\mathbf{x} \in \Omega} |v(\mathbf{x})| \leq \frac{c}{h} \|v\|_{L^2(\Omega)} \quad \text{für alle } v \in V_h. \quad (6.5)$$

Wir kombinieren nun (6.4) und (6.5):

$$\begin{aligned} \sup_{\mathbf{x} \in \Omega} |u(\mathbf{x}) - u_h(\mathbf{x})| &\leq \sup_{\mathbf{x} \in \Omega} |u(\mathbf{x}) - I_h u(\mathbf{x})| + \sup_{\mathbf{x} \in \Omega} |u_h(\mathbf{x}) - I_h u(\mathbf{x})| \\ &\leq c \left\{ h \|f\|_{L^2(\Omega)} + \frac{1}{h} \|u_h - I_h u\|_{L^2(\Omega)} \right\}. \end{aligned}$$

Hieraus folgt wegen

$$\|u_h - I_h u\|_{L^2(\Omega)} \leq \underbrace{\|u - u_h\|_{L^2(\Omega)}}_{\substack{\leq ch^2 \|f\|_{L^2(\Omega)} \\ \text{nach Proposition 6.6}}} + \underbrace{\|u - I_h u\|_{L^2(\Omega)}}_{\substack{\leq ch^2 \|f\|_{L^2(\Omega)} \\ \text{nach Satz 5.16 und} \\ H^2(\Omega)\text{-Regularität}}} \leq ch^2 \|f\|_{L^2(\Omega)}$$

die Behauptung. □

Bemerkung Diese L^∞ -Fehlerabschätzung ist nicht scharf. Man kann

$$\sup_{\mathbf{x} \in \Omega} |u(\mathbf{x}) - u_h(\mathbf{x})| \leq ch^2 |\log h|^{3/2} \|u\|_{C^2(\Omega)}$$

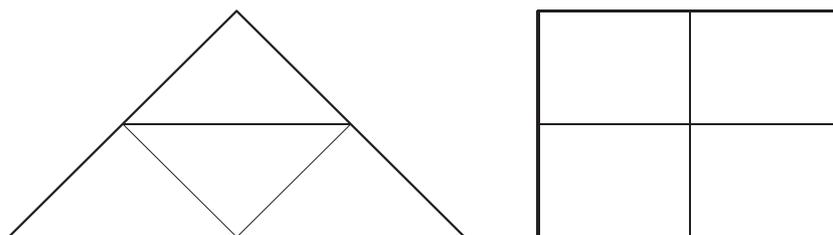
beweisen. Der logarithmische Term verschwindet sogar im Fall $d = 3$. △

7. Rechentechnische Betrachtungen

Die Umsetzung der Finite-Elemente-Methode am Computer lässt sich in folgende Einzelschritte zerlegen:

1. Netzgenerierung
2. Assemblierung, das ist das Aufstellen der Steifigkeitsmatrix und der diskreten rechten Seite
3. Lösung des linearen Gleichungssystems
4. a-posteriori Fehleranalyse: falls die Lösung nicht zufriedenstellend ist, markiere zu verfeinernde Elemente und gehe zu Schritt 1 zurück
5. Visualisierung, das heißt, die graphische Darstellung der Lösung

Netzgenerierung: Für die Erzeugung der ursprünglichen Triangulierung gibt es zahlreiche Möglichkeiten, beispielsweise per Hand durch den erfahrenen Anwender oder vollautomatisch durch Meshing Tools. Ausgehend von einer groben Triangulierung des Gebiets kann man durch uniformes Unterteilen jedes Dreiecks bzw. Vierecks in vier neue Dreiecke bzw. Vierecke beliebig feine quasi-uniforme Gitter generieren:



Wir wollen nun eine Methode vorstellen, um einige wenige Elemente einer bestehenden Triangulierung zu unterteilen. Dies wird nötig bei einer schlechten Approximationsgüte des Gitters, welche a-priori zu erwarten ist, etwa in der Nähe einer einspringenden Ecke, oder die während der Rechnung durch einen Fehlerschätzer gemeldet wird.

Algorithmus 7.1 (Netzverfeinerung)

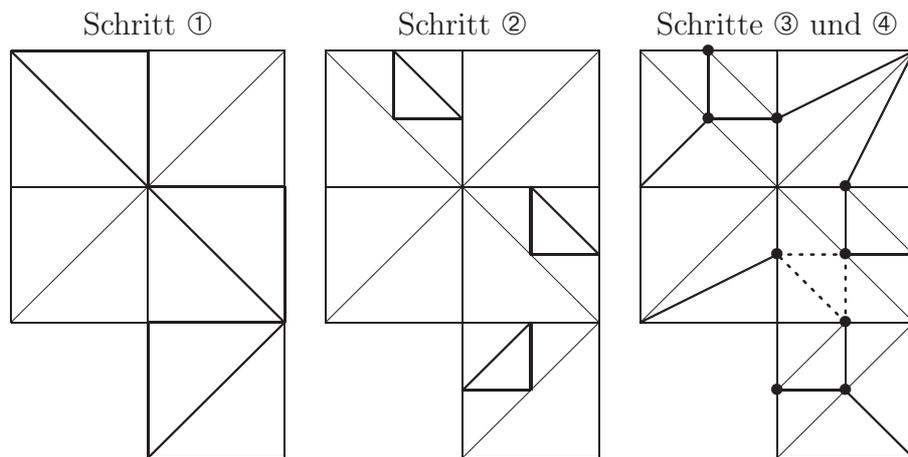
input: zulässige Triangulierung mit als zu verfeinern markierten Elementen

output: zulässige, verfeinerte Triangulierung

- ① verfeinere alle markierten Dreiecke in vier Dreiecke
- ② liegen auf den Kanten eines Dreiecks T (ohne die eigenen Ecken) mehr als ein Eckpunkt eines anderen Dreiecks, so verfeinere T ebenfalls 1:4
- ③ wiederhole ② bis kein weiteres Dreieck dazukommt

- ④ jedes Dreieck mit vier Ecken auf seinen Kanten wird halbiert, wobei die neue Kante "grün" markiert wird

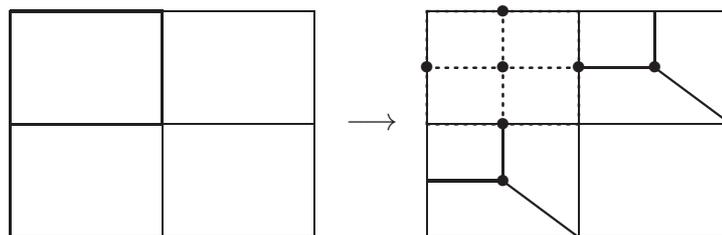
Beispiel 7.2 (Netzverfeinerung)



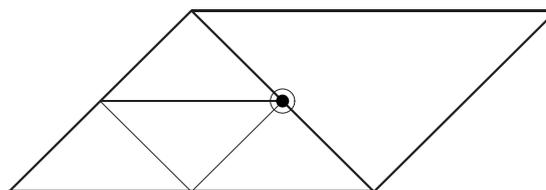
△

Bemerkungen

1. Grüne Kanten heißen *Transitionskanten* und die Dreiecke entsprechend *Transitions-elemente*. Sollen später weitere Verfeinerungen in Transitionselementen vorgenommen werden, sind die Transitionskanten zu eliminieren.
2. Nicht entartete Triangulierungen bleiben nicht entartet, weil ursprüngliche Winkel höchstens halbiert werden.
3. Ein zu Algorithmus 7.1 analoges Vorgehen ist auch bei Vierecken möglich:



4. Eine Verfeinerung ist auch ohne Transitionselemente möglich, allerdings stellen die dann entstehenden *hängenden Knoten* keine echten Freiheitsgrade dar, sondern sind durch Werte auf der zugrundeliegenden Kante bestimmt.



5. Natürlich ist auch eine Neuvernetzung ohne die Berücksichtigung des alten Gitters möglich.

△

Assemblierung: Bei Finiten Elementen mit einer nodalen Basis $\{\varphi_i\}_{i=1}^N$, wie beispielsweise bei linearen oder quadratischen Dreieckselementen, stellt man die Steifigkeitsmatrix am besten *element-orientiert* auf. Ist etwa

$$a(u, v) = \int_{\Omega} \langle \mathbf{A} \nabla u, \nabla v \rangle \, d\mathbf{x},$$

dann folgt

$$a(\varphi_i, \varphi_j) = \int_{\Omega} \langle \mathbf{A} \nabla \varphi_i, \nabla \varphi_j \rangle \, d\mathbf{x} = \sum_{T \in \mathcal{T}_h} \int_T \langle \mathbf{A} \nabla \varphi_i, \nabla \varphi_j \rangle \, d\mathbf{x} =: \sum_{T \in \mathcal{T}_h} a_T(\varphi_i, \varphi_j). \quad (7.1)$$

Hierbei muss nur über diejenigen Dreiecke summiert werden, die gleichzeitig zum Träger von φ_i und φ_j gehören.

Die Steifigkeitsmatrix berechnet man nun, indem man für jedes $T \in \mathcal{T}_h$ den durch (7.1) gegebenen additiven Beitrag ermittelt. Wenn jedes Element n Knoten enthält, hat man die Element-Steifigkeitsmatrix

$$\mathbf{A}_T = \begin{bmatrix} a_T(\varphi_{i_1}, \varphi_{i_1}) & a_T(\varphi_{i_2}, \varphi_{i_1}) & \cdots & a_T(\varphi_{i_n}, \varphi_{i_1}) \\ a_T(\varphi_{i_1}, \varphi_{i_2}) & a_T(\varphi_{i_2}, \varphi_{i_2}) & \cdots & a_T(\varphi_{i_n}, \varphi_{i_2}) \\ \vdots & \vdots & \ddots & \vdots \\ a_T(\varphi_{i_1}, \varphi_{i_n}) & a_T(\varphi_{i_2}, \varphi_{i_n}) & \cdots & a_T(\varphi_{i_n}, \varphi_{i_n}) \end{bmatrix} \in \mathbb{R}^{n \times n}$$

zu bilden. Außerdem transformiert man das Element T auf das Referenzelement T_{ref} . Sei $F_T : T_{\text{ref}} \rightarrow T$, $\hat{\mathbf{x}} \mapsto \mathbf{x} = \mathbf{B}\hat{\mathbf{x}} + \mathbf{x}_0$ die zugehörige affine Abbildung. Dann ist der Beitrag von T gegeben durch

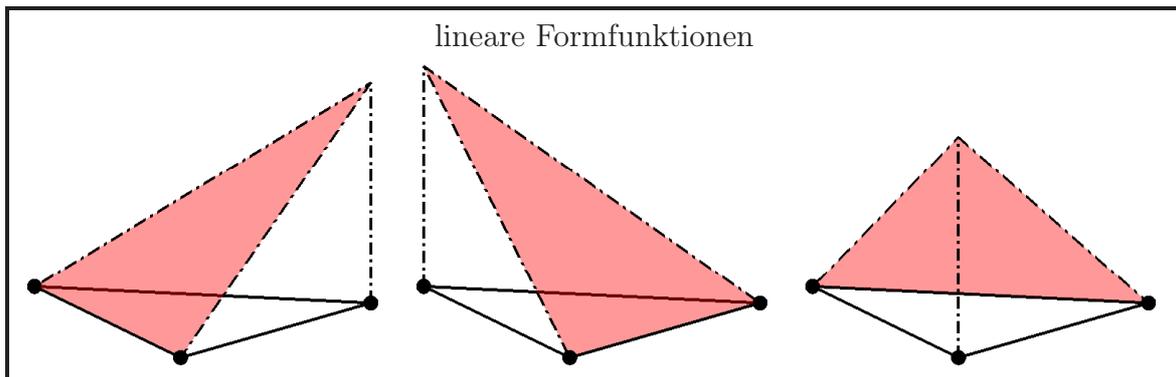
$$a_T(\varphi_{i_k}, \varphi_{i_\ell}) = \frac{|T|}{|T_{\text{ref}}|} \int_{T_{\text{ref}}} \langle \mathbf{A}\mathbf{B}^{-T} \nabla_{\hat{\mathbf{x}}} \psi_k, \mathbf{B}^{-T} \nabla_{\hat{\mathbf{x}}} \psi_\ell \rangle \, d\hat{\mathbf{x}}. \quad (7.2)$$

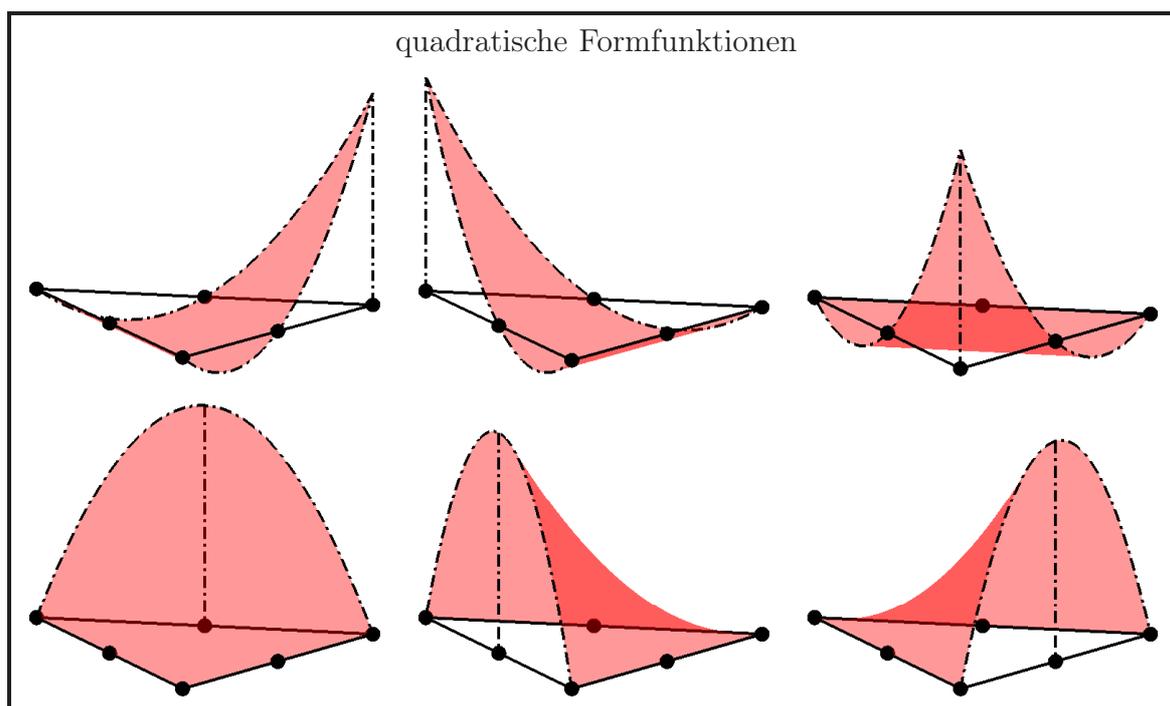
Jede Funktion aus der nodalen Basis fällt nach der Transformation auf das Referenzdreieck mit einer der normierten *Formfunktionen* $\psi_1, \psi_2, \dots, \psi_n$ zusammen. Wählt man als Referenzelement das Dreieck $\Delta((0, 0), (1, 0), (0, 1))$, so ist bei linearen Elementen

$$\psi_1(\hat{x}, \hat{y}) = 1 - \hat{x} - \hat{y}, \quad \psi_2(\hat{x}, \hat{y}) = \hat{x}, \quad \psi_3(\hat{x}, \hat{y}) = \hat{y}$$

und bei quadratischen

$$\begin{aligned} \psi_1(\hat{x}, \hat{y}) &= (1 - \hat{x} - \hat{y})(1 - 2\hat{x} - 2\hat{y}), & \psi_2(\hat{x}, \hat{y}) &= \hat{x}(2\hat{x} - 1), & \psi_3(\hat{x}, \hat{y}) &= \hat{y}(2\hat{y} - 1), \\ \psi_4(\hat{x}, \hat{y}) &= 4\hat{x}(1 - \hat{x} - \hat{y}), & \psi_5(\hat{x}, \hat{y}) &= 4\hat{x}\hat{y}, & \psi_6(\hat{x}, \hat{y}) &= 4\hat{y}(1 - \hat{x} - \hat{y}). \end{aligned}$$





Um (7.2) auszurechnen, ist im allgemeinen eine Quadraturformel notwendig, wie sie in folgender Tabelle zu finden ist:

Quadraturpunkt	Gewicht	exakt in
$\left(\frac{1}{3}, \frac{1}{3}\right)$	$\frac{1}{2}$	\mathcal{P}_1
$\left(\frac{1}{2}, \frac{1}{2}\right), \left(\frac{1}{2}, 0\right), \left(0, \frac{1}{2}\right)$	$\frac{1}{6}$	\mathcal{P}_2
$\left(\frac{1}{5}, \frac{1}{5}\right), \left(\frac{1}{5}, \frac{3}{5}\right), \left(\frac{3}{5}, \frac{1}{5}\right)$	$-\frac{27}{96}$ $\frac{25}{96}$	\mathcal{P}_3
$\left(\frac{1}{3}, \frac{1}{3}\right)$ $\left(\frac{6+\sqrt{15}}{21}, \frac{6+\sqrt{15}}{21}\right), \left(\frac{9-2\sqrt{15}}{21}, \frac{6+\sqrt{15}}{21}\right), \left(\frac{6+\sqrt{15}}{21}, \frac{9-2\sqrt{15}}{21}\right)$ $\left(\frac{6-\sqrt{15}}{21}, \frac{6-\sqrt{15}}{21}\right), \left(\frac{9+2\sqrt{15}}{21}, \frac{6-\sqrt{15}}{21}\right), \left(\frac{6-\sqrt{15}}{21}, \frac{9+2\sqrt{15}}{21}\right)$	$\frac{9}{80}$ $\frac{155+\sqrt{15}}{2400}$ $\frac{155-\sqrt{15}}{2400}$	\mathcal{P}_5

Ist der Integrand ein Polynom, so kann man das Integral auch analytisch mit Hilfe der Formel

$$\int_{T_{\text{ref}}} \hat{x}^p \hat{y}^q (1 - \hat{x} - \hat{y})^r d(\hat{x}, \hat{y}) = \frac{p!q!r!}{(p+q+r+2)!}$$

ausrechnen. Da nur n^2 Integrale pro Element $T \in \mathcal{T}_h$ zu bilden sind und $N \sim |\mathcal{T}_h|$ gilt, lässt sich die gesamte Steifigkeitsmatrix in linearer Komplexität aufstellen.

Das Aufstellen der diskreten rechten Seite \mathbf{b}_h wird analog ebenfalls element-orientiert durchgeführt.

Innere Kondensation: Obwohl sich die Steifigkeitsmatrix additiv aus $(n \times n)$ -Untermatrizen zusammensetzt, ist die Bandbreite durchweg größer als n (vgl. Beispiel 4.2). Eine Sonderrolle spielen die Knotenvariablen, die zu Knoten im Innern der Elemente gehören. So hat zum Beispiel das quadratische Viereckselement oder das kubische Dreieckselement jeweils einen derartigen Knoten. Die Elimination der zugehörigen Variablen ändert nur die Matrixelemente für die Knoten des betroffenen Elements, wobei der Aufwand hierzu dem eines Cholesky-Verfahrens für eine $(n \times n)$ -Matrix entspricht. Man spricht von *innerer* oder auch *statischer Kondensation*.

Dirichlet-Randdaten: Bei einem Dirichlet-Problem mit homogenen Randwerten wird $V_h \subset H_0^1(\Omega)$ einfach als die lineare Hülle aller nodalen Basisfunktionen gewählt, welche einem inneren Knoten zugeordnet sind. Die den Randknoten zugeordneten Ansatzfunktionen werden also auf Null gesetzt.

Auf ähnliche Weise kann man dann auch inhomogene Dirichlet-Randwerte g umsetzen: Man wählt die Koeffizienten der den Randknoten zugeordneten Ansatzfunktionen so, dass

$$\sum_{i=1}^N g_i \varphi_i|_{\Gamma} \approx g \quad \text{auf } \Gamma$$

gilt. Am einfachsten interpoliert man g in den Randknoten. Die Koeffizienten der Basisfunktionen zu Knoten im Gebietsinneren werden auf Null gesetzt. Die so konstruierte Funktion $g_h = \sum_{i=1}^N g_i \varphi_i$ erfüllt $g_h \in H^1(\Omega)$. Man benötigt nun nur noch eine Funktion $u_h \in V_h \subset H_0^1(\Omega)$, so dass das homogene Dirichlet-Problem

$$a(u_h, v_h) = \ell(v_h) - a(g_h, v_h)$$

für alle $v_h \in V_h$ erfüllt ist.

Man beachte aber, dass diese Konstruktion einer Fortsetzung der Randwerte g in das Gebietsinnere nicht gleichmäßig $H^1(\Omega)$ -stabil ist. Im Fall von nichtlinearen Differentialgleichungen konvergieren die Lösungsverfahren daher immer schlechter.

Übrige Schritte: Das Lösen des Gleichungssystems $\mathbf{A}_h \mathbf{z}_h = \mathbf{b}_h$ ist Gegenstand des nächsten Kapitels. Danach werden residuale Fehlerschätzer eingeführt. Hingegen ist die Visualisierung der Lösung kein Gegenstand dieser Vorlesung.

8. Mehrgitterverfahren

8.1 Glättungseigenschaft von Iterationsverfahren

Um klassische Iterationsverfahren zur Lösung des linearen Gleichungssystems

$$\mathbf{A}_j \mathbf{u}_j = \mathbf{f}_j$$

zu konstruieren, zerlegen wir die Systemmatrix gemäß $\mathbf{A}_j = \mathbf{D}_j - \mathbf{L}_j - \mathbf{U}_j$ in die Diagonalmatrix \mathbf{D}_j , die echte obere Dreiecksmatrix \mathbf{U}_j und die echte untere Dreiecksmatrix \mathbf{L}_j . Damit erhalten wir das

- *Jacobi-Verfahren* oder *Gesamtschrittverfahren* mit der Iterationsvorschrift

$$\mathbf{D}_j \mathbf{u}_j^{(k+1)} = (\mathbf{L}_j + \mathbf{U}_j) \mathbf{u}_j^{(k)} + \mathbf{f}_j, \quad k = 0, 1, 2, \dots,$$

- *Gauß-Seidel-Verfahren* oder *Einzel-schrittverfahren* mit der Iterationsvorschrift

$$(\mathbf{D}_j - \mathbf{L}_j) \mathbf{u}_j^{(k+1)} = \mathbf{U}_j \mathbf{u}_j^{(k)} + \mathbf{f}_j, \quad k = 0, 1, 2, \dots,$$

- *Richardson-Verfahren* mit der Iterationsvorschrift

$$\mathbf{u}_j^{(k+1)} = \mathbf{u}_j^{(k)} + \alpha_j (\mathbf{f}_j - \mathbf{A}_j \mathbf{u}_j^{(k)}), \quad k = 0, 1, 2, \dots$$

Mit Hilfe der Lösung $\mathbf{u}_j = \mathbf{A}_j^{-1} \mathbf{f}_j$ lässt sich das Jacobi-Verfahren auch in der Form

$$\mathbf{u}_j - \mathbf{u}_j^{(k+1)} = \mathbf{D}_j^{-1} (\mathbf{L}_j + \mathbf{U}_j) (\mathbf{u}_j - \mathbf{u}_j^{(k)}), \quad k = 0, 1, 2, \dots$$

und das Gauß-Seidel-Verfahren in der Form

$$\mathbf{u}_j - \mathbf{u}_j^{(k+1)} = (\mathbf{D}_j - \mathbf{L}_j)^{-1} \mathbf{U}_j (\mathbf{u}_j - \mathbf{u}_j^{(k)}), \quad k = 0, 1, 2, \dots$$

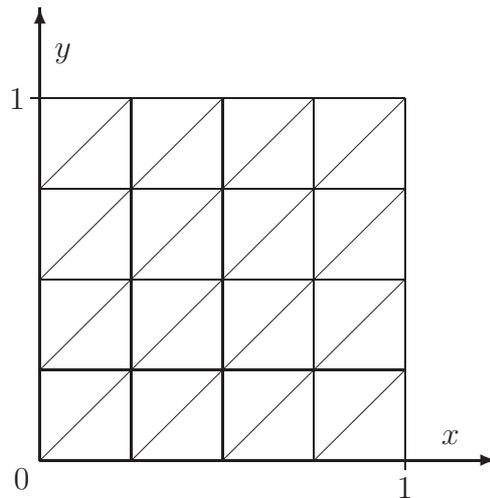
schreiben. Besitzt \mathbf{A}_j den konstanten Eintrag d_j auf der Diagonalen, so ist das Richardson-Verfahren mit $\alpha_j := 1/d_j$ äquivalent zum Jacobi-Verfahren.

Beobachtung. Stellt man den Fehlervektor $\mathbf{u}_j - \mathbf{u}_j^{(k)}$ bezüglich einer Basis aus Eigenvektoren dar, so sieht man, dass die zu großen Eigenwerten gehörigen Komponenten (*hohe Frequenzen*) schnell gedämpft werden, die zu kleinen Eigenwerten gehörigen Anteile (*niedrige Frequenzen*) jedoch nicht. Der Fehler wird also geglättet.

Beispiel 8.1 Wir betrachten die Poisson-Gleichung

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

im Einheitsquadrat $\Omega = (0, 1) \times (0, 1)$. Es werde $\bar{\Omega}$ wie folgt mit einem gleichmäßigen Dreiecksnetz der Maschenweite $h_j := 2^{-(j+1)}$ überzogen:



Die Diskretisierung der Bilinearform

$$a(u, v) = \int_{\Omega} \langle \nabla u, \nabla v \rangle \, d\mathbf{x}$$

durch stückweise lineare Finite Elemente führt auf ein lineares Gleichungssystem $\mathbf{A}_j \mathbf{u}_j = \mathbf{f}_j$ mit dem Standard-5-Punkte-Stern

$$\begin{bmatrix} \alpha_{NW} & \alpha_N & \alpha_{NO} \\ \alpha_W & \alpha_Z & \alpha_O \\ \alpha_{SW} & \alpha_S & \alpha_{SO} \end{bmatrix}_{\star} = \begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}_{\star},$$

vergleiche Beispiel 4.2.

Die Eigenwerte von \mathbf{A}_j lauten

$$\lambda_{k,\ell} = 4 \left(\sin^2 \frac{k\pi h_j}{2} + \sin^2 \frac{\ell\pi h_j}{2} \right), \quad 1 \leq k, \ell \leq N_j := 2^{j+1} - 1.$$

Die zugehörigen Eigenfunktionen $\mathbf{v}_{k,\ell}$ sind gegeben durch die Knotenwerte der kontinuierlichen Funktionen

$$v_{k,\ell}(x, y) = \sin(k\pi x) \sin(\ell\pi y), \quad 1 \leq k, \ell \leq N_j.$$

Die Eigenwerte der Iterationsmatrix $\mathbf{I} - \alpha_j \mathbf{A}_j$ des Richardson-Verfahrens für $\alpha_j = 1/8$ sind folglich

$$0 \leq \mu_{k,\ell} = 1 - \frac{1}{2} \left(\sin^2 \frac{k\pi h_j}{2} + \sin^2 \frac{\ell\pi h_j}{2} \right), \quad 1 \leq k, \ell \leq N_j.$$

Für den betragsgrößten Eigenwert $\mu_{1,1}$ folgt

$$\rho(\mathbf{I} - \mathbf{A}_j/8) = |\mu_{1,1}| \approx 1 - \frac{\pi^2 h_j^2}{4},$$

das heißt, die Konvergenz der Richardson-Iteration verlangsamt sich für $h_j \rightarrow 0$.

Die Eigenwerte mit einem hochfrequenten Anteil in mindestens einer Richtung sind gerade $\{\mu_{k,\ell} : k \geq (N_j + 1)/2 \text{ oder } \ell \geq (N_j + 1)/2\}$. Für diese Eigenwerte gilt

$$|\mu_{k,\ell}| \leq 1 - \frac{1}{2} \left(\sin^2 \frac{\pi}{4} + 0 \right) = \frac{3}{4},$$

das heißt, die hochfrequenten Anteile des Fehlers $\mathbf{u}_j - \mathbf{u}_j^{(k)}$ werden pro Iterationsschritt um mindestens den Faktor $3/4$ gedämpft. \triangle

Mehrgitterprinzip. Es sei

$$V_0 \subset V_1 \subset V_2 \subset \dots \subset H_0^1(\Omega)$$

eine geschachtelte Folge von Finite-Element-Räumen, die durch uniformes Verfeinern des Grobgitterraums V_0 entstehen. Zur Bestimmung der Lösung $\mathbf{u}_j = \mathbf{A}_j^{-1}\mathbf{f}_j$ im Finite-Element-Raum V_j führt man zunächst einige Schritte eines Iterationsverfahrens durch, um die hochfrequenten Fehleranteile zu dämpfen. Die niederfrequenten Anteile lassen sich dann auf der nächstgröberen Zerlegung V_{j-1} in den Griff bekommen.

Um die Glättungseigenschaft von Iterationsverfahren mathematisch präzise formulieren zu können, benötigen wir spezielle diskrete Normen:

Definition 8.2 Sei $\mathbf{A}_j = \mathbf{X}_j \mathbf{D}_j \mathbf{X}_j^T \in \mathbb{R}^{N_j \times N_j}$ die beim Galerkin-Verfahren in V_j erhaltene Systemmatrix. Die diskreten Normen $\|\cdot\|_{s,j}$ auf \mathbb{R}^{N_j} sind für $s \in \mathbb{R}$ gegeben durch

$$\|\mathbf{v}_j\|_{s,j} = \sqrt{\mathbf{v}_j^T \mathbf{A}_j^s \mathbf{v}_j} \quad \text{mit} \quad \mathbf{A}_j^s = \mathbf{X}_j \mathbf{D}_j^s \mathbf{X}_j^T.$$

Hierbei bezeichnen $\mathbf{X}_j = [\mathbf{x}_{j,1}, \mathbf{x}_{j,2}, \dots, \mathbf{x}_{j,N_j}]$ die aus den Eigenvektoren $\{\mathbf{x}_{j,k}\}_{k=1}^{N_j}$ von \mathbf{A}_j zusammengesetzte orthogonale Matrix und $\mathbf{D}_j = \text{diag}(\lambda_{j,1}, \lambda_{j,2}, \dots, \lambda_{j,N_j})$ die Diagonalmatrix der zugehörigen Eigenwerte $\{\lambda_{j,k}\}_{k=1}^{N_j}$.

Offensichtlich gibt es von j unabhängige Konstanten $\underline{c}, \bar{c} > 0$, so dass für alle $v_j \in V_j$ mit zugehörigem Koeffizientenvektor \mathbf{v}_j gilt

$$\begin{aligned} \frac{\underline{c}}{h_j} \|v_j\|_{L^2(\Omega)} &\leq \|\mathbf{v}_j\|_{0,j} \leq \frac{\bar{c}}{h_j} \|v_j\|_{L^2(\Omega)}, \\ \underline{c} \|v_j\|_{H^1(\Omega)} &\leq \|\mathbf{v}_j\|_{1,j} \leq \bar{c} \|v_j\|_{H^1(\Omega)}. \end{aligned} \tag{8.1}$$

Weiterhin erfüllt die diskrete Norm $\|\cdot\|_{s,j}$ eine verallgemeinerte Cauchy-Schwarzsche Ungleichung.

Lemma 8.3 Es gilt die verallgemeinerte Cauchy-Schwarzsche Ungleichung

$$|a(v_j, w_j)| \leq \|\mathbf{v}_j\|_{1+t,j} \|\mathbf{w}_j\|_{1-t,j}$$

für alle $v_j, w_j \in V_j$ und $t \in \mathbb{R}$.

Beweis. Mit $\{(\lambda_k, \mathbf{x}_k)\}_{k=1}^{N_j}$ bezeichnen wir die Eigenpaare von \mathbf{A}_j , wobei die Eigenvektoren normiert seien, also $\mathbf{x}_k^T \mathbf{x}_\ell = \delta_{k,\ell}$. Stellt man die Koeffizientenvektoren \mathbf{v}_j und \mathbf{w}_j in der Eigenbasis dar

$$\mathbf{v}_j = \sum_{k=1}^{N_j} \sigma_k \mathbf{x}_k, \quad \mathbf{w}_j = \sum_{k=1}^{N_j} \tau_k \mathbf{x}_k,$$

so folgt

$$\begin{aligned} |a(v_j, w_j)| &= |\mathbf{w}_j^T \mathbf{A}_j \mathbf{v}_j| = \left| \sum_{k=1}^{N_j} \lambda_k \sigma_k \tau_k \right| = \left| \sum_{k=1}^{N_j} \sigma_k \lambda_k^{(1+t)/2} \tau_k \lambda_k^{(1-t)/2} \right| \\ &\leq \sqrt{\sum_{k=1}^{N_j} \sigma_k^2 \lambda_k^{1+t}} \sqrt{\sum_{k=1}^{N_j} \tau_k^2 \lambda_k^{1-t}} = \|\mathbf{v}_j\|_{1+t,j} \|\mathbf{w}_j\|_{1-t,j}, \end{aligned}$$

das heißt, die Behauptung. \square

Satz 8.4 (Glättungseigenschaft des Richardson-Verfahrens) Es seien $0 < \underline{\alpha} \leq \alpha_j \leq 1/\lambda_{\max}(\mathbf{A}_j)$ und $\mathbf{S}_j := \mathbf{I} - \alpha_j \mathbf{A}_j$ die zugehörige Iterationsmatrix des Richardson-Verfahrens. Dann gilt mit einer von j unabhängigen Konstanten $c > 0$

$$\|\mathbf{S}_j^\ell \mathbf{v}_j\|_{2,j} \leq \frac{c}{\sqrt{\ell}} \|\mathbf{v}_j\|_{1,j} \quad \text{für alle } \mathbf{v}_j \in \mathbb{R}^{N_j}.$$

Beweis. Es bezeichne $\{\lambda_k\}_{k=1}^{N_j}$ die Eigenwerte von \mathbf{A}_j und $\{\mathbf{x}_k\}_{k=1}^{N_j}$ die zugehörigen ortho-normierten Eigenvektoren. Für $\mathbf{v}_j = \sum_{k=1}^{N_j} \sigma_k \mathbf{x}_k$ erhalten wir

$$\mathbf{S}_j^\ell \mathbf{v}_j = \sum_{k=1}^{N_j} \sigma_k (1 - \alpha_j \lambda_k)^\ell \mathbf{x}_k.$$

Folglich ist

$$\begin{aligned} \|\mathbf{S}_j^\ell \mathbf{v}_j\|_{2,j}^2 &= \|\mathbf{A}_j \mathbf{S}_j^\ell \mathbf{v}_j\|_2^2 \\ &= \sum_{k=1}^{N_j} (1 - \alpha_j \lambda_k)^{2\ell} \lambda_k^2 \sigma_k^2 \\ &= \frac{1}{\alpha_j} \sum_{k=1}^{N_j} (1 - \alpha_j \lambda_k)^{2\ell} (\alpha_j \lambda_k) (\lambda_k \sigma_k^2) \\ &\leq \underbrace{\frac{1}{\alpha_j} \max_{k=1}^{N_j} \{(1 - \alpha_j \lambda_k)^{2\ell} (\alpha_j \lambda_k)\}}_{\leq c} \underbrace{\sum_{k=1}^{N_j} \lambda_k \sigma_k^2}_{=\|\mathbf{v}_j\|_{1,j}^2}. \end{aligned}$$

Den mittleren Term bekommen wir wie folgt in den Griff: Aufgrund unserer Voraussetzung ist $0 \leq \alpha_j \lambda_k \leq 1$ für alle $k = 1, 2, \dots, N_j$. Daher schätzen wir ab

$$\max_{k=1}^{N_j} \{(1 - \alpha_j \lambda_k)^{2\ell} (\alpha_j \lambda_k)\} \leq \max_{0 \leq \xi \leq 1} \{(1 - \xi)^{2\ell} \xi\}.$$

Auf $[0, 1]$ nimmt die Funktion $g(\xi) := (1 - \xi)^{2\ell} \xi$ ihr Maximum wegen

$$g'(\xi) = (1 - \xi)^{2\ell-1} (1 - (2\ell + 1)\xi)$$

in $\xi = 1/(2\ell + 1)$ an. Daher folgt

$$\max_{k=1}^{N_j} \{(1 - \alpha_j \lambda_k)^{2\ell} (\alpha_j \lambda_k)\} \leq g \left(\frac{1}{2\ell + 1} \right) = \underbrace{\left(\frac{2\ell}{2\ell + 1} \right)^{2\ell}}_{\leq 1} \frac{1}{2\ell + 1} \leq \frac{1}{2\ell}.$$

□

8.2 Prolongation und Restriktion

Wir betrachten die geschachtelte Folge von Finite-Element-Räumen

$$V_0 \subset V_1 \subset V_2 \subset \dots \subset H_0^1(\Omega),$$

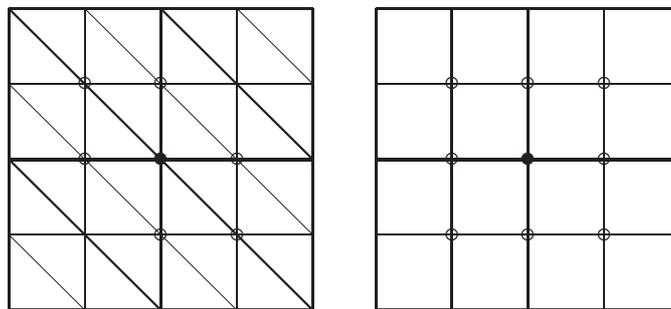
die durch uniformes Verfeinern erzeugt wird. Das Variationsproblem

$$\text{suche } u_j \in V_j, \text{ so dass } a(u_j, v_j) = \ell(v_j) \quad \text{für alle } v_j \in V_j$$

führt auf das lineare Gleichungssystem $\mathbf{A}_j \mathbf{u}_j = \mathbf{f}_j$. Dabei stellt sich die Frage, wie $\mathbf{A}_j \mathbf{u}_j = \mathbf{f}_j$ und $\mathbf{A}_{j-1} \mathbf{u}_{j-1} = \mathbf{f}_{j-1}$ zusammenhängen. Dazu seien $\{\mathbf{x}_{j,k}\}$ die Knoten der nodalen Basis $\{\varphi_{j,k}\}$ aus V_j . Für $v_j \in V_j$ ist dann das diskrete Analogon gegeben durch $\mathbf{v}_j = [v_j(\mathbf{x}_{j,k})]_{k=1}^{N_j}$.

Restriktion: Wegen $V_{j-1} \subset V_j$ können wir jede Basisfunktion $\varphi_{j-1,k}$ aus V_{j-1} durch die nodalen Basisfunktionen $\{\varphi_{j,\ell}\}$ aus V_j darstellen: $\varphi_{j-1,k} = \sum_{\ell} \varphi_{j-1,k}(\mathbf{x}_{j,\ell}) \varphi_{j,\ell}$. Daher kann der Vektor $\mathbf{f}_{j-1} = [\ell(\varphi_{j-1,k})]_{k=1}^{N_{j-1}}$ berechnet werden aus den Komponenten des Vektors $\mathbf{f}_j = [\ell(\varphi_{j,k})]_{k=1}^{N_j}$. Dies entspricht der *Restriktion* $\mathbf{f}_{j-1} = \mathbf{I}_j^{j-1} \mathbf{f}_j$. Im Fall linearer Finite Elemente auf Dreiecken bzw. bilinearer Finite Elemente auf Vierecken erhalten wir

$$\mathbf{I}_j^{j-1} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1 & 1/2 \\ 0 & 1/2 & 1/2 \end{bmatrix}_* \quad \text{bzw.} \quad \mathbf{I}_j^{j-1} = \begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 1/4 \end{bmatrix}_*$$



Beispiel 8.5 Im eindimensionalen Fall gilt

$$\begin{bmatrix} f_{j-1,1} \\ f_{j-1,2} \\ \vdots \\ f_{j-1,N_{j-1}} \end{bmatrix} = \begin{bmatrix} 1/2 & 1 & 1/2 & & & \\ & 1/2 & 1 & 1/2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & & 1/2 & 1 & 1/2 \end{bmatrix} \begin{bmatrix} f_{j,1} \\ f_{j,2} \\ \vdots \\ f_{j,N_j} \end{bmatrix}$$

△

Prolongation: Die *Prolongation* ist die Übersetzung einer Darstellung von $v_{j-1} \in V_{j-1}$ bezüglich der nodalen Basis in die Darstellung bezüglich der nodalen Basis in V_j . Dies entspricht der Interpolation. Für lineare Finite Elemente auf Dreiecken bzw. bilineare Finite Elemente auf Vierecken kann sie ebenfalls schematisch dargestellt werden als

$$\mathbf{I}_{j-1}^j = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1 & 1/2 \\ 0 & 1/2 & 1/2 \end{bmatrix}_* \quad \text{bzw.} \quad \mathbf{I}_{j-1}^j = \begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 1/4 \end{bmatrix}_* .$$

Achtung: Vorsicht bei der obigen Schreibweise! Bei der Restriktion werden Daten zusammengefasst, bei der Prolongation entsprechend verteilt.

Für $\mathbf{f}_j := [\ell(\varphi_{j,k})]_{k=1}^{N_j}$ folgt

$$\begin{aligned} \mathbf{v}_{j-1}^T \mathbf{I}_j^{j-1} \mathbf{f}_j &= \sum_{k=1}^{N_{j-1}} v_{j-1}(\mathbf{x}_{j-1,k}) \ell(\varphi_{j-1,k}) \\ &= \ell \left(\sum_{k=1}^{N_{j-1}} v_{j-1}(\mathbf{x}_{j-1,k}) \varphi_{j-1,k} \right) \\ &= \ell(v_{j-1}). \end{aligned}$$

Andererseits gilt

$$\begin{aligned} (\mathbf{I}_{j-1}^j \mathbf{v}_{j-1})^T \mathbf{f}_j &= \sum_{k=1}^{N_j} v_{j-1}(\mathbf{x}_{j,k}) \ell(\varphi_{j,k}) \\ &= \ell \left(\sum_{k=1}^{N_j} v_{j-1}(\mathbf{x}_{j,k}) \varphi_{j,k} \right) \\ &= \ell(v_{j-1}). \end{aligned}$$

Dies zeigt $\mathbf{I}_j^{j-1} = (\mathbf{I}_{j-1}^j)^T$, das heißt, Prolongation und Restriktion sind zueinander adjungiert.

Wählt man ein $w_{j-1} \in V_{j-1}$ und setzt $\mathbf{w}_{j-1} := [w_{j-1}(\mathbf{x}_{j-1,k})]_{k=1}^{N_{j-1}}$, dann folgt

$$\begin{aligned} \mathbf{v}_{j-1}^T \mathbf{A}_{j-1} \mathbf{w}_{j-1} &= a(w_{j-1}, v_{j-1}) \\ &= (\mathbf{I}_{j-1}^j \mathbf{v}_{j-1})^T \mathbf{A}_j (\mathbf{I}_{j-1}^j \mathbf{w}_{j-1}) \\ &= \mathbf{v}_{j-1}^T \mathbf{I}_j^{j-1} \mathbf{A}_j \mathbf{I}_{j-1}^j \mathbf{w}_{j-1}. \end{aligned}$$

Daher gilt $\mathbf{A}_{j-1} = \mathbf{I}_j^{j-1} \mathbf{A}_j \mathbf{I}_{j-1}^j$. Setzt man für $j \leq J$ schließlich $\mathbf{I}_j^J := \mathbf{I}_{j-1}^j \mathbf{I}_{j-2}^{j-1} \cdots \mathbf{I}_j^{j+1}$ und $\mathbf{I}_J^j := \mathbf{I}_{j+1}^j \mathbf{I}_{j+2}^{j+1} \cdots \mathbf{I}_J^{j-1}$, so ergibt sich $\mathbf{A}_j = \mathbf{I}_J^j \mathbf{A}_J \mathbf{I}_j^J$. Die Matrix \mathbf{A}_j der groben Zerlegung lässt sich also durch die Matrix \mathbf{A}_J der feinen Zerlegung darstellen. Analog gilt für die rechte Seite offensichtlich die Beziehung $\mathbf{f}_j = \mathbf{I}_j^J \mathbf{f}_J$.

8.3 Zweigitterverfahren

Für das Zweigitterverfahren betrachten wir zwei Finite-Element-Räume $V_{j-1} \subset V_j$, wobei V_j durch uniformes Verfeinern aus V_{j-1} hervorgehe. Ein Iterationsschritt des Zweigitterverfahrens setzt sich dann wie folgt zusammen:

1. *A-priori-Glättung.* Setze

$$\mathbf{u}_j^{(\text{pre},0)} = \mathbf{u}_j^{\text{alt}}$$

und führe K Schritte eines Iterationsverfahrens (zum Beispiel des Richardson-Verfahrens) durch

$$\mathbf{u}_j^{(\text{pre},k)} = \mathbf{u}_j^{(\text{pre},k-1)} + \alpha_j(\mathbf{f}_j - \mathbf{A}_j \mathbf{u}_j^{(\text{pre},k-1)}), \quad k = 1, 2, \dots, K.$$

2. *Grobitterkorrektur.* Löse die Defektgleichung auf der nächstgrößeren Triangulierung

$$\mathbf{A}_{j-1} \mathbf{e}_{j-1} = \mathbf{I}_j^{j-1}(\mathbf{f}_j - \mathbf{A}_j \mathbf{u}_j^{(\text{pre},K)})$$

und addiere die erhaltene Näherung

$$\mathbf{u}_j^{(\text{post},0)} = \mathbf{u}_j^{(\text{pre},K)} + \mathbf{I}_{j-1}^j \mathbf{e}_{j-1}.$$

3. *A-posteriori-Glättung.* Führe L Schritte eines Iterationsverfahrens durch

$$\mathbf{u}_j^{(\text{post},\ell)} = \mathbf{u}_j^{(\text{post},\ell-1)} + \alpha_j(\mathbf{f}_j - \mathbf{A}_j \mathbf{u}_j^{(\text{post},\ell-1)}), \quad \ell = 1, 2, \dots, L$$

und setze

$$\mathbf{u}_j^{\text{neu}} = \mathbf{u}_j^{(\text{post},L)}.$$

Bezeichnet \mathbf{u}_j die exakte Lösung des Gleichungssystems $\mathbf{A}_j \mathbf{u}_j = \mathbf{f}_j$, dann ist die Iterationsmatrix des Zweigitterverfahrens gegeben durch

$$\mathbf{u}_j - \mathbf{u}_j^{\text{neu}} = \mathbf{S}_j^L (\mathbf{I} - \mathbf{I}_{j-1}^j \mathbf{A}_{j-1}^{-1} \mathbf{I}_j^{j-1} \mathbf{A}_j) \mathbf{S}_j^K (\mathbf{u}_j - \mathbf{u}_j^{\text{alt}}). \quad (8.2)$$

Neben der Glättungseigenschaft des Iterationsverfahrens brauchen wir noch eine Aussage über die Approximationsgüte der Grobitterkorrektur.

Lemma 8.6 Es seien $u_j, u_j^{(\text{pre},K)} \in V_j$ und $e_{j-1} \in V_{j-1}$ die zu $\mathbf{u}_j, \mathbf{u}_j^{(\text{pre},K)} \in \mathbb{R}^{N_j}$ und $\mathbf{e}_{j-1} \in \mathbb{R}^{N_{j-1}}$ gehörigen Funktionen. Dann ist

$$e_{j-1} = P_{j-1}(u_j - u_j^{(\text{pre},K)})$$

die Galerkin-Projektion des nach der A-priori-Glättung erhaltenen Fehlers, das heißt, es gilt

$$a(e_{j-1}, v_{j-1}) = a(u_j - u_j^{(\text{pre},K)}, v_{j-1}) \quad \text{für alle } v_{j-1} \in V_{j-1}.$$

Beweis. Für alle $v_{j-1} \in V_{j-1}$ gilt

$$\begin{aligned} a(e_{j-1}, v_{j-1}) &= \mathbf{v}_{j-1}^T \mathbf{A}_{j-1} \mathbf{e}_{j-1} \\ &= \mathbf{v}_{j-1}^T \mathbf{I}_j^{j-1} (\mathbf{f}_j - \mathbf{A}_j \mathbf{u}_j^{(\text{pre},K)}) \\ &= (\mathbf{I}_{j-1}^j \mathbf{v}_{j-1})^T (\mathbf{f}_j - \mathbf{A}_j \mathbf{u}_j^{(\text{pre},K)}) \\ &= (\mathbf{I}_{j-1}^j \mathbf{v}_{j-1})^T \mathbf{A}_j (\mathbf{u}_j - \mathbf{u}_j^{(\text{pre},K)}) \\ &= a(u_j - u_j^{(\text{pre},K)}, v_{j-1}). \end{aligned}$$

□

Bezeichnet $v_{j-1} = P_{j-1}v_j \in V_{j-1}$ die Galerkin-Projektion von $v_j \in V_j$, dann gilt

$$a(v_j, w_{j-1}) = a(P_{j-1}v_j, w_{j-1}) = a(v_{j-1}, w_{j-1})$$

für alle $w_{j-1} \in V_{j-1}$. Daher folgt

$$\mathbf{w}_{j-1}^T \mathbf{A}_{j-1} \mathbf{v}_{j-1} = (\mathbf{I}_{j-1}^j \mathbf{w}_{j-1})^T \mathbf{A}_j \mathbf{v}_j = \mathbf{w}_{j-1}^T \mathbf{I}_j^{j-1} \mathbf{A}_j \mathbf{v}_j$$

für alle $\mathbf{w}_{j-1} \in \mathbb{R}^{N_{j-1}}$. Dies bedeutet

$$\mathbf{A}_{j-1} \mathbf{v}_{j-1} = \mathbf{I}_j^{j-1} \mathbf{A}_j \mathbf{v}_j,$$

beziehungsweise

$$\mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1} = (\mathbf{I} - \mathbf{I}_{j-1}^j \mathbf{A}_{j-1}^{-1} \mathbf{I}_j^{j-1} \mathbf{A}_j) \mathbf{v}_j.$$

Lemma 8.7 Es seien $v_j \in V_j$ und $v_{j-1} = P_{j-1}v_j \in V_{j-1}$ die zu $\mathbf{v}_j \in \mathbb{R}^{N_j}$ und $\mathbf{v}_{j-1} \in \mathbb{R}^{N_{j-1}}$ gehörigen Funktionen. Ferner sei $\Omega \subset \mathbb{R}^d$ ein konvexes Polygonebiet. Dann ist

$$\|\mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1}\|_{0,j} \leq c \|\mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1}\|_{1,j}$$

mit einer von j unabhängigen Konstanten c .

Beweis. Das Aubin-Nitsche-Lemma (Satz 6.5) liefert

$$\|v_j - P_{j-1}v_j\|_{L^2(\Omega)} \leq c_S \|v_j - P_{j-1}v_j\|_{H^1(\Omega)} \sup_{g \in L^2(\Omega) \setminus \{0\}} \left\{ \frac{1}{\|g\|_{L^2(\Omega)}} \inf_{w_{j-1} \in V_{j-1}} \|\varphi_g - w_{j-1}\|_{H^1(\Omega)} \right\},$$

wobei $\varphi_g \in H_0^1(\Omega)$ die Lösung des dualen Problems

$$a(w, \varphi_g) = (g, w)_{L^2(\Omega)} \quad \text{für alle } w \in H_0^1(\Omega)$$

bezeichnet. Aufgrund der $H^2(\Omega)$ -Regularität ist $\varphi_g \in H^2(\Omega)$ und es folgt

$$\|v_j - P_{j-1}v_j\|_{L^2(\Omega)} \leq c_A h_{j-1} \|v_j - P_{j-1}v_j\|_{H^1(\Omega)}.$$

Wegen $h_{j-1} = 2h_j$ und (8.1) ergibt sich schließlich

$$\begin{aligned} \|\mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1}\|_{0,j} &\leq \frac{\bar{c}}{h_j} \|v_j - P_{j-1}v_j\|_{L^2(\Omega)} \\ &\leq 2c_A \bar{c} \|v_j - P_{j-1}v_j\|_{H^1(\Omega)} \\ &\leq 2\underline{c} c_A \bar{c} \|\mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1}\|_{1,j}. \end{aligned}$$

□

Satz 8.8 (Approximationseigenschaft) Unter den Voraussetzungen von Lemma 8.7 gilt

$$\|\mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1}\|_{1,j} \leq c \|\mathbf{v}_j\|_{2,j}.$$

Beweis. Die Galerkin-Orthogonalität führt auf

$$\|\mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1}\|_{1,j}^2 = a(v_j - v_{j-1}, v_j - v_{j-1}) = a(v_j - v_{j-1}, v_j).$$

Die verallgemeinerte Cauchy-Schwarzsche Ungleichung (Lemma 8.3) liefert daher

$$\|\mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1}\|_{1,j}^2 \leq \|\mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1}\|_{0,j} \|\mathbf{v}_j\|_{2,j}.$$

Lemma 8.7 impliziert

$$\|\mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1}\|_{1,j}^2 \leq c \|\mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1}\|_{1,j} \|\mathbf{v}_j\|_{2,j},$$

woraus die Behauptung folgt. \square

Satz 8.9 (Konvergenz des Zweigitterverfahrens) Für das Zweigitterverfahren mit K A-priori-Glättungsschritten und ohne A-posteriori-Glättung gilt

$$\|\mathbf{u}_j - \mathbf{u}_j^{\text{neu}}\|_{1,j} \leq \frac{c}{\sqrt{K}} \|\mathbf{u}_j - \mathbf{u}_j^{\text{alt}}\|_{1,j}$$

mit einer von j unabhängigen Konstanten c .

Beweis. Bezeichnet $\mathbf{v}_j := \mathbf{S}_j^K(\mathbf{u}_j - \mathbf{u}_j^{\text{alt}})$, dann ist $\mathbf{v}_{j-1} = \mathbf{A}_{j-1}^{-1} \mathbf{I}_j^{j-1} \mathbf{A}_j \mathbf{v}_j$ die zugehörige Vektordarstellung der Galerkin-Projektion. Wegen

$$\mathbf{u}_j - \mathbf{u}_j^{\text{neu}} = (\mathbf{I} - \mathbf{I}_{j-1}^j \mathbf{A}_{j-1}^{-1} \mathbf{I}_j^{j-1} \mathbf{A}_j) \mathbf{v}_j = \mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1}$$

folgt aus Satz 8.8

$$\|\mathbf{u}_j - \mathbf{u}_j^{\text{neu}}\|_{1,j} = \|\mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1}\|_{1,j} \leq c \|\mathbf{v}_j\|_{2,j} = c \|\mathbf{S}_j^K(\mathbf{u}_j - \mathbf{u}_j^{\text{alt}})\|_{2,j}.$$

Mit der Glättungseigenschaft (Satz 8.4) ergibt sich schließlich

$$\|\mathbf{u}_j - \mathbf{u}_j^{\text{neu}}\|_{1,j} \leq \frac{c}{\sqrt{K}} \|\mathbf{u}_j - \mathbf{u}_j^{\text{alt}}\|_{1,j}.$$

\square

Bemerkungen

1. Satz 8.9 besagt, dass das Zweigitterverfahren bei einer genügend großen Anzahl von A-priori-Glättungsschritten konvergiert, und zwar mit einer von j unabhängigen Konvergenzrate.
2. Die Aussage gilt auch, wenn zusätzlich zur A-priori-Glättung auch eine A-posteriori-Glättung durchgeführt wird.
3. In der Praxis ist das Gauß-Seidel-Verfahren der beste Glätter.

\triangle

8.4 Mehrgitterverfahren

Es sei

$$V_0 \subset V_1 \subset V_2 \subset \dots \subset H_0^1(\Omega)$$

eine geschachtelte Folge von Finite-Element-Räumen, die durch uniformes Verfeinern des Grobgitterraums V_0 entstehen. Das allgemeine Mehrgitterverfahren lässt sich dann wie folgt beschreiben:

1. *A-priori-Glättung.* Setze

$$\mathbf{u}_j^{(\text{pre},0)} = \mathbf{u}_j^{\text{alt}}$$

und führe K Glättungsschritte durch

$$\mathbf{u}_j^{(\text{pre},k)} = \mathbf{u}_j^{(\text{pre},k-1)} + \alpha_j(\mathbf{f}_j - \mathbf{A}_j \mathbf{u}_j^{(\text{pre},k-1)}), \quad k = 1, 2, \dots, K.$$

2. *Restriktion.* Restringiere das Residuum auf das nächstgrößere Gitter

$$\mathbf{r}_{j-1} = \mathbf{I}_j^{j-1}(\mathbf{f}_j - \mathbf{A}_j \mathbf{u}_j^{(\text{pre},K)}).$$

3. *Grobgitterkorrektur.* Falls $j = 1$ ist, dann löse das Gleichungssystem $\mathbf{A}_0 \mathbf{e}_0 = \mathbf{r}_0$ exakt, andernfalls wende P Schritte des Mehrgitteralgorithmus auf $\mathbf{A}_{j-1} \mathbf{e}_{j-1} = \mathbf{r}_{j-1}$ an mit Startnäherung $\mathbf{0}$.

4. *Prolongation.* Addiere die prolongierte Grobgitterkorrektur

$$\mathbf{u}_j^{(\text{post},0)} = \mathbf{u}_j^{(\text{pre},K)} + \mathbf{I}_{j-1}^j \mathbf{e}_{j-1}.$$

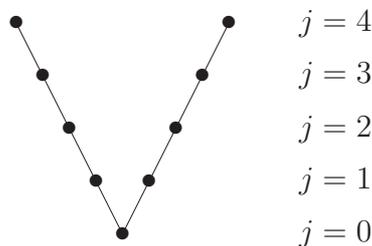
5. *A-posteriori-Glättung.* Führe L Glättungsschritte durch

$$\mathbf{u}_j^{(\text{post},\ell)} = \mathbf{u}_j^{(\text{post},\ell-1)} + \alpha_j(\mathbf{f}_j - \mathbf{A}_j \mathbf{u}_j^{(\text{post},\ell-1)}), \quad \ell = 1, 2, \dots, L$$

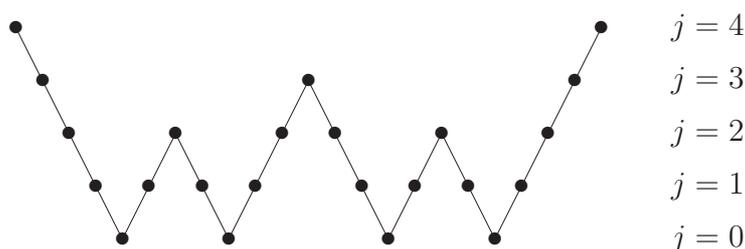
und setze

$$\mathbf{u}_j^{\text{neu}} = \mathbf{u}_j^{(\text{post},L)}.$$

Zur Implementierung des Mehrgitteralgorithmus ist folglich ein rekursiver Aufruf der eigenen Prozedur notwendig. Für $P = 1$ erhält man den *V-Zyklus*:



Für $P = 2$ ergibt sich hingegen der *W-Zyklus*:



Lemma 8.10 Die Iterationsmatrix $\mathbf{S}_j = \mathbf{I} - \alpha_j \mathbf{A}_j$ des Richardson-Verfahrens mit $0 \leq \alpha_j \leq 1/\lambda_{\max}(\mathbf{A}_j)$ erfüllt

$$\|\mathbf{S}_j^\ell \mathbf{v}_j\|_{1,j} \leq \|\mathbf{v}_j\|_{1,j}, \quad \ell = 1, 2, \dots$$

für alle $\mathbf{v}_j \in \mathbb{R}^{N_j}$.

Beweis. Seien $\{(\lambda_k, \mathbf{x}_k)\}$ die Eigenpaare der Systemmatrix \mathbf{A}_j . Zerlegen wir $\mathbf{v}_j = \sum_{k=1}^{N_j} \sigma_k \mathbf{x}_k \in \mathbb{R}^{N_j}$ in die Eigenbasis von \mathbf{A}_j , dann folgt

$$\|\mathbf{S}_j^\ell \mathbf{v}_j\|_{1,j}^2 = (\mathbf{S}_j^\ell \mathbf{v}_j)^T \mathbf{A}_j \mathbf{S}_j^\ell \mathbf{v}_j = \sum_{k=1}^{N_j} \sigma_k^2 (1 - \alpha_j \lambda_k)^{2\ell} \lambda_k \leq \sum_{k=1}^{N_j} \sigma_k^2 \lambda_k = \mathbf{v}_j^T \mathbf{A}_j \mathbf{v}_j = \|\mathbf{v}_j\|_{1,j}^2.$$

□

Satz 8.11 (Konvergenz des W-Zyklus) Zu jedem $\rho \in (0, 1)$ gibt es ein K_0 derart, dass bei der Verwendung von $K \geq K_0$ A-priori-Glättungsschritten für die mit dem W-Zyklus erhaltenen Iterierten gilt

$$\|\mathbf{u}_j - \mathbf{u}_j^{\text{neu}}\|_{1,j} \leq \rho \|\mathbf{u}_j - \mathbf{u}_j^{\text{alt}}\|_{1,j}$$

für alle $j = 1, 2, \dots$

Beweis. Wir beweisen die Aussage mit Hilfe von vollständiger Induktion. Für $j = 1$ stimmt der W-Zyklus mit dem Zweigitterverfahren überein und das Behauptete folgt sofort aus Satz 8.9.

Für den Induktionsschritt $j - 1 \mapsto j$ sei $\hat{\mathbf{e}}_{j-1}$ die exakte und \mathbf{e}_{j-1} die durch den W-Zyklus berechnete Grobgitterkorrektur. Aus

$$\mathbf{u}_j - \mathbf{u}_j^{\text{neu}} = \mathbf{u}_j - \mathbf{u}_j^{(\text{pre},K)} - \mathbf{I}_{j-1}^j \hat{\mathbf{e}}_{j-1} + \mathbf{I}_{j-1}^j (\hat{\mathbf{e}}_{j-1} - \mathbf{e}_{j-1})$$

folgt

$$\|\mathbf{u}_j - \mathbf{u}_j^{\text{neu}}\|_{1,j}^2 = \|\mathbf{u}_j - \mathbf{u}_j^{(\text{pre},K)} - \mathbf{I}_{j-1}^j \hat{\mathbf{e}}_{j-1}\|_{1,j}^2 + \|\mathbf{I}_{j-1}^j (\hat{\mathbf{e}}_{j-1} - \mathbf{e}_{j-1})\|_{1,j}^2.$$

Dabei fällt der gemischte Term wegen Lemma 8.6 weg:

$$\begin{aligned} & (\mathbf{u}_j - \mathbf{u}_j^{(\text{pre},K)} - \mathbf{I}_{j-1}^j \hat{\mathbf{e}}_{j-1})^T \mathbf{A}_j \mathbf{I}_{j-1}^j (\hat{\mathbf{e}}_{j-1} - \mathbf{e}_{j-1}) \\ &= a(u_j - u_j^{(\text{pre},K)} - \hat{e}_{j-1}, \hat{e}_{j-1} - e_{j-1}) \\ &= a(u_j - u_j^{(\text{pre},K)} - P_{j-1}(u_j - u_j^{(\text{pre},K)}), \underbrace{\hat{e}_{j-1} - e_{j-1}}_{\in V_{j-1}}) \\ &= 0. \end{aligned}$$

Mit

$$\|\mathbf{I}_{j-1}^j (\hat{\mathbf{e}}_{j-1} - \mathbf{e}_{j-1})\|_{1,j}^2 = (\hat{\mathbf{e}}_{j-1} - \mathbf{e}_{j-1})^T \underbrace{\mathbf{I}_{j-1}^{j-1} \mathbf{A}_j \mathbf{I}_{j-1}^j}_{=\mathbf{A}_{j-1}} (\hat{\mathbf{e}}_{j-1} - \mathbf{e}_{j-1}) = \|\hat{\mathbf{e}}_{j-1} - \mathbf{e}_{j-1}\|_{1,j-1}^2,$$

Satz 8.9 und der Induktionsvoraussetzung ergibt sich

$$\begin{aligned} \|\mathbf{u}_j - \mathbf{u}_j^{\text{neu}}\|_{1,j}^2 &= \|\mathbf{u}_j - \mathbf{u}_j^{(\text{pre},K)} - \mathbf{I}_{j-1}^j \widehat{\mathbf{e}}_{j-1}\|_{1,j}^2 + \|\widehat{\mathbf{e}}_{j-1} - \mathbf{e}_{j-1}\|_{1,j-1}^2 \\ &\leq \frac{c^2}{K} \|\mathbf{u}_j - \mathbf{u}_j^{\text{alt}}\|_{1,j}^2 + \rho^4 \|\widehat{\mathbf{e}}_{j-1} - \underbrace{\mathbf{e}_{j-1}^{\text{alt}}}_{=0}\|_{1,j-1}^2, \end{aligned}$$

wobei wir hier verwendet haben, dass zwei W-Zyklen aufgerufen werden. Aus

$$\begin{aligned} 0 &\leq a(v_j - P_{j-1}v_j, v_j - P_{j-1}v_j) \\ &= a(v_j, v_j) - 2 \underbrace{a(v_j, P_{j-1}v_j)}_{=a(P_{j-1}v_j, P_{j-1}v_j)} + a(P_{j-1}v_j, P_{j-1}v_j) \\ &= a(v_j, v_j) - a(P_{j-1}v_j, P_{j-1}v_j) \end{aligned}$$

für alle $v_j \in V_j$ folgt, da $\widehat{\mathbf{e}}_{j-1} = P_{j-1}(u_j - u_j^{(\text{pre},K)})$ gemäß Lemma 8.6 gilt, dass

$$a(\widehat{\mathbf{e}}_{j-1}, \widehat{\mathbf{e}}_{j-1}) \leq a(u_j - u_j^{(\text{pre},K)}, u_j - u_j^{(\text{pre},K)}).$$

Dies bedeutet aber

$$\|\widehat{\mathbf{e}}_{j-1}\|_{1,j-1} \leq \|\mathbf{u}_j - \mathbf{u}_j^{(\text{pre},K)}\|_{1,j} = \|\mathbf{S}_j^K(\mathbf{u}_j - \mathbf{u}_j^{\text{alt}})\|_{1,j}.$$

Lemma 8.10 liefert dann

$$\|\widehat{\mathbf{e}}_{j-1}\|_{1,j-1} \leq \|\mathbf{u}_j - \mathbf{u}_j^{\text{alt}}\|_{1,j}$$

und wir erhalten schließlich

$$\|\mathbf{u}_j - \mathbf{u}_j^{\text{neu}}\|_{1,j}^2 \leq \left(\frac{c^2}{K} + \rho^4\right) \|\mathbf{u}_j - \mathbf{u}_j^{\text{alt}}\|_{1,j}^2.$$

Wählen wir $K \geq c^2/(\rho^2(1-\rho^2))$, dann folgt $c^2/K + \rho^4 \leq \rho^2$ und damit die Behauptung. \square

8.5 Konvergenz des V-Zyklus

Wir beschränken unsere Analyse auf den symmetrischen V-Zyklus, das heißt, es gelte im folgenden stets $K = L$ und $P = 1$. Ferner seien $A_j, S_j : V_j \rightarrow V_j$ die zu $\mathbf{A}_j, \mathbf{S}_j : \mathbb{R}^{N_j} \rightarrow \mathbb{R}^{N_j}$ gehörigen Operatoren, das heißt, $A_j v_j, S_j v_j \in V_j$ sind die $\mathbf{A}_j \mathbf{v}_j, \mathbf{S}_j \mathbf{v}_j \in \mathbb{R}^{N_j}$ entsprechenden Funktionen. Den Iterationsoperator des symmetrischen V-Zyklus bezeichnen wir mit $E_j : V_j \rightarrow V_j$ beziehungsweise $\mathbf{E}_j : \mathbb{R}^{N_j} \rightarrow \mathbb{R}^{N_j}$, dies bedeutet, es gilt

$$u_j - u_j^{\text{neu}} = E_j(u_j - u_j^{\text{alt}}) \quad \text{bzw.} \quad \mathbf{u}_j - \mathbf{u}_j^{\text{neu}} = \mathbf{E}_j(\mathbf{u}_j - \mathbf{u}_j^{\text{alt}}).$$

Lemma 8.12 Der symmetrische V-Zyklus erfüllt die Rekursionsformel

$$E_0 = 0, \quad E_j = S_j^K (I - (I - E_{j-1})P_{j-1})S_j^K, \quad j = 1, 2, \dots$$

Beweis. Wir beweisen die Aussage mittels vollständiger Induktion. Ist $j = 1$, dann ist die Grobgitterkorrektur exakt und die Behauptung folgt aus Lemma 8.6:

$$E_1 = S_1^K(I - P_0)S_1^K = S_1^K(I - (I - E_0)P_0)S_1^K.$$

Für den Induktionsschritt $j - 1 \mapsto j$ sei \widehat{e}_{j-1} die exakte und e_{j-1} die durch den V-Zyklus berechnete Grobgitterkorrektur. Aus Lemma 8.6 folgt dann $\widehat{e}_{j-1} = P_{j-1}(u_j - u_j^{(\text{pre},K)})$ und aus der Induktionsvoraussetzung $\widehat{e}_{j-1} - e_{j-1} = E_{j-1}(\widehat{e}_{j-1} - 0)$, also

$$e_{j-1} = (I - E_{j-1})\widehat{e}_{j-1} = (I - E_{j-1})P_{j-1}(u_j - u_j^{(\text{pre},K)}).$$

Daraus ergibt sich schließlich die Behauptung:

$$\begin{aligned} u_j - u_j^{\text{neu}} &= S_j^K(u_j - u_j^{(\text{post},0)}) \\ &= S_j^K(u_j - u_j^{(\text{pre},K)} - e_{j-1}) \\ &= S_j^K(I - (I - E_{j-1})P_{j-1})(u_j - u_j^{(\text{pre},K)}) \\ &= S_j^K(I - (I - E_{j-1})P_{j-1})S_j^K(u_j - u_j^{\text{alt}}) \\ &= E_j(u_j - u_j^{\text{alt}}). \end{aligned}$$

□

Lemma 8.13 Der Iterationsoperator E_j ist symmetrisch und positiv semidefinit bezüglich des durch die Bilinearform $a(\cdot, \cdot)$ induzierten Innenprodukts, das heißt, es gilt

$$a(E_j v_j, w_j) = a(v_j, E_j w_j) \quad \text{und} \quad a(E_j v_j, v_j) \geq 0 \quad \text{für alle } v_j, w_j \in V_j.$$

Beweis. Wir beweisen die Aussage mittels vollständiger Induktion. Ist $j = 0$, so ist die Behauptung klar. Um den Induktionsschritt $j - 1 \mapsto j$ zu zeigen, beachten wir, dass für alle $v_j, w_j \in V_j$ gilt

$$\begin{aligned} a(S_j v_j, w_j) &= (\mathbf{S}_j \mathbf{v}_j)^T \mathbf{A}_j \mathbf{w} \\ &= \mathbf{v}_j^T (\mathbf{I} - \alpha_j \mathbf{A}_j) \mathbf{A}_j \mathbf{w} \\ &= \mathbf{v}_j^T \mathbf{A}_j (\mathbf{I} - \alpha_j \mathbf{A}_j) \mathbf{w} \\ &= \mathbf{v}_j^T \mathbf{A}_j \mathbf{S}_j \mathbf{w} \\ &= a(v_j, S_j w_j). \end{aligned}$$

Galerkin-Orthogonalität und Induktionsannahme liefern die Gleichung

$$\begin{aligned} a(\underbrace{E_{j-1} P_{j-1} S_j^K v_j}_{\in V_{j-1}}, S_j^K w_j) &= a(E_{j-1} P_{j-1} S_j^K v_j, P_{j-1} S_j^K w_j) \\ &= a(P_{j-1} S_j^K v_j, E_{j-1} P_{j-1} S_j^K w_j) = a(S_j^K v_j, \underbrace{E_{j-1} P_{j-1} S_j^K w_j}_{\in V_{j-1}}). \end{aligned}$$

Ferner ergibt sich aus der Definition der Galerkin-Projektion, dass

$$a(P_{j-1} v_j, w_j) = a(v_j, P_{j-1} w_j), \quad a((I - P_{j-1})v_j, w_j) = a(v_j, (I - P_{j-1})w_j).$$

Zusammen folgt daher die Symmetrie:

$$\begin{aligned}
a(E_j v_j, w_j) &= a\left(S_j^K (I - (I - E_{j-1})P_{j-1})S_j^K v_j, w_j\right) \\
&= a\left((I - P_{j-1})S_j^K v_j, S_j^K w_j\right) + a(E_{j-1}P_{j-1}S_j^K v_j, S_j^K w_j) \\
&= a\left(S_j^K v_j, (I - P_{j-1})S_j^K w_j\right) + a\left(S_j^K v_j, E_{j-1}P_{j-1}S_j^K w_j\right) \\
&= a\left(v_j, S_j^K (I - (I - E_{j-1})P_{j-1})S_j^K w_j\right) \\
&= a(v_j, E_j w_j).
\end{aligned}$$

Schließlich ergibt sich die Nichtnegativität gemäß

$$\begin{aligned}
a(E_j v_j, v_j) &= a\left(S_j^K (I - (I - E_{j-1})P_{j-1})S_j^K v_j, v_j\right) \\
&= a\left((I - P_{j-1})S_j^K v_j, S_j^K v_j\right) + a(E_{j-1}P_{j-1}S_j^K v_j, S_j^K v_j) \\
&= a\left(\underbrace{(I - P_{j-1})S_j^K v_j, (I - P_{j-1})S_j^K v_j}_{\geq 0 \text{ da } a \text{ elliptisch}}\right) + a\left(\underbrace{E_{j-1}P_{j-1}S_j^K v_j, P_{j-1}S_j^K v_j}_{\geq 0 \text{ nach Induktionsannahme}}\right) \\
&\geq 0.
\end{aligned}$$

□

Lemma 8.14 Der Iterationsoperator E_j des symmetrischen V-Zyklus mit jeweils K A-priori- und K A-posteriori-Glättungsschritten genügt der Abschätzung

$$a(E_j v_j, v_j) \leq \frac{c^*}{K + c^*} a(v_j, v_j) \quad \text{für alle } v_j \in V_j$$

mit einer von j unabhängigen Konstanten c^* .

Beweis. Wir führen den Beweis in drei Schritten.

(i.) Wir zeigen zunächst, dass

$$a((I - S_j)S_j^{2K} v_j, v_j) \leq \frac{1}{2K} a((I - S_j^{2K})v_j, v_j) \quad (8.3)$$

für alle $v_j \in V_j$ ist. Dazu seien $\{(\lambda_k, \mathbf{x}_k)\}$ wieder die Eigenpaare der Systemmatrix \mathbf{A}_j . Zerlegen wir $\mathbf{v}_j = \sum_{k=1}^{N_j} \sigma_k \mathbf{x}_k$ in die Eigenbasis von \mathbf{A}_j , dann folgt für alle $\ell \in \mathbb{N}$ die Abschätzung

$$\begin{aligned}
a((I - S_j)S_j^\ell v_j, v_j) &= \mathbf{v}_j^T \mathbf{A}_j \underbrace{(\mathbf{I} - \mathbf{S}_j)}_{=\alpha_j \mathbf{A}_j} \mathbf{S}_j^\ell \mathbf{v}_j \\
&= \alpha_j \mathbf{v}_j^T \mathbf{A}_j^2 \mathbf{S}_j^\ell \mathbf{v}_j \\
&= \alpha_j \sum_{k=1}^{N_j} \sigma_k^2 \lambda_k^2 \underbrace{(1 - \alpha_j \lambda_k)}_{\leq 1}^\ell \\
&\leq \alpha_j \sum_{k=1}^{N_j} \sigma_k^2 \lambda_k^2 (1 - \alpha_j \lambda_k)^{\ell-1} \\
&= a((I - S_j)S_j^{\ell-1} v_j, v_j).
\end{aligned}$$

Daraus ergibt sich dann mit Hilfe einer Teleskopsumme die Zwischenbehauptung (8.3):

$$\begin{aligned} a((I - S_j)S_j^{2K}v_j, v_j) &= \frac{1}{2K} \left\{ \underbrace{a((I - S_j)S_j^{2K}v_j, v_j) + \cdots + a((I - S_j)S_j^{2K}v_j, v_j)}_{2K\text{-mal}} \right\} \\ &\leq \frac{1}{2K} \sum_{\ell=0}^{2K-1} a((I - S_j)S_j^\ell v_j, v_j) \\ &= \frac{1}{2K} a((I - S_j^{2K})v_j, v_j). \end{aligned}$$

(ii.) Die Stetigkeit der Bilinearform $a(\cdot, \cdot)$ liefert zusammen mit der Approximations-eigenschaft (Satz 8.8) und (8.1), dass

$$\begin{aligned} a((I - P_{j-1})S_j^K v_j, (I - P_{j-1})S_j^K v_j) &\leq c_S \|(I - P_{j-1})S_j^K v_j\|_{H^1(\Omega)}^2 \\ &\leq \frac{c_S}{\underline{c}} \|\mathbf{S}_j^K \mathbf{v}_j\|_{2,j}^2 \\ &= \frac{c_S}{\underline{c}} \mathbf{v}_j^T \mathbf{S}_j^K \mathbf{A}_j^2 \mathbf{S}_j^K \mathbf{v}_j. \end{aligned}$$

Aus der Beziehung $\mathbf{A}_j = (\mathbf{I} - \mathbf{S}_j)/\alpha_j$ folgt weiter

$$\begin{aligned} a((I - P_{j-1})S_j^K v_j, (I - P_{j-1})S_j^K v_j) &\leq \frac{c_S}{\alpha_j \underline{c}} \mathbf{v}_j^T \mathbf{S}_j^K (\mathbf{I} - \mathbf{S}_j) \mathbf{A}_j \mathbf{S}_j^K \mathbf{v}_j \\ &= \frac{c_S}{\alpha_j \underline{c}} a((I - S_j)S_j^K v_j, S_j^K v_j) \\ &= \frac{c_S}{\alpha_j \underline{c}} a((I - S_j)S_j^{2K} v_j, v_j). \end{aligned}$$

Mit (8.3) erhalten wir daher die Abschätzung

$$a((I - P_{j-1})S_j^K v_j, (I - P_{j-1})S_j^K v_j) \leq \frac{1}{K} \underbrace{\frac{c_S}{2\alpha_j \underline{c}}}_{=: c^*} a((I - S_j^{2K})v_j, v_j).$$

(iii.) Der restliche Beweis geschieht mittels vollständiger Induktion. Für $j = 0$ ist die Behauptung wegen $a(E_0 v_0, v_0) = 0$ trivialerweise erfüllt. Wir zeigen nun den Induktionsschritt $j - 1 \mapsto j$. Wegen der Rekursionsformel aus Lemma 8.12 gilt

$$a(E_j v_j, v_j) = \underbrace{a((I - P_{j-1})S_j^K v_j, S_j^K v_j)}_{=a((I - P_{j-1})S_j^K v_j, (I - P_{j-1})S_j^K v_j)} + \underbrace{a(E_{j-1}P_{j-1}S_j^K v_j, S_j^K v_j)}_{=a(E_{j-1}P_{j-1}S_j^K v_j, P_{j-1}S_j^K v_j)}.$$

Daher folgt aus der Induktionsannahme

$$\begin{aligned} a(E_j v_j, v_j) &\leq a((I - P_{j-1})S_j^K v_j, (I - P_{j-1})S_j^K v_j) \\ &\quad + \frac{c^*}{K + c^*} \underbrace{a(P_{j-1}S_j^K v_j, P_{j-1}S_j^K v_j)}_{=a(S_j^K v_j, S_j^K v_j) - a((I - P_{j-1})S_j^K v_j, (I - P_{j-1})S_j^K v_j)} \\ &= \left(1 - \frac{c^*}{K + c^*}\right) a((I - P_{j-1})S_j^K v_j, (I - P_{j-1})S_j^K v_j) \\ &\quad + \frac{c^*}{K + c^*} a(S_j^K v_j, S_j^K v_j), \end{aligned}$$

und schließlich mit (ii.) die Behauptung:

$$\begin{aligned} a(E_j v_j, v_j) &\leq \underbrace{\left(1 - \frac{c^*}{K + c^*}\right) \frac{c^*}{K}}_{=c^*/(K+c^*)} a((I - S_j^{2K})v_j, v_j) + \frac{c^*}{K + c^*} \underbrace{a(S_j^K v_j, S_j^K v_j)}_{=a(S_j^{2K} v_j, v_j)} \\ &= \frac{c^*}{K + c^*} a(v_j, v_j). \end{aligned}$$

□

Satz 8.15 (Konvergenz des V-Zyklus) Der symmetrische V-Zyklus mit jeweils K A-priori- und K A-posteriori-Glättungsschritten erfüllt

$$\|\mathbf{u}_j - \mathbf{u}_j^{\text{neu}}\|_{1,j} \leq \frac{c^*}{K + c^*} \|\mathbf{u}_j - \mathbf{u}_j^{\text{alt}}\|_{1,j}$$

mit einer von j unabhängigen Konstanten c^* .

Beweis. Zu zeigen ist

$$a(u_j - u_j^{\text{neu}}, u_j - u_j^{\text{neu}}) \leq \left(\frac{c^*}{K + c^*}\right)^2 a(u_j - u_j^{\text{alt}}, u_j - u_j^{\text{alt}}).$$

Nach Lemma 8.13 existieren N_j Eigenpaare $\{(\mu_k, z_k)\}$ von E_j bezüglich des Innenprodukts $a(\cdot, \cdot)$ mit

$$E_j z_k = \mu_k z_k \quad \text{und} \quad a(z_k, z_\ell) = \delta_{k,\ell} \quad \text{für alle } k, \ell = 1, 2, \dots, N_j.$$

Lemma 8.14 impliziert

$$\mu_k = a(E_j z_k, z_k) \leq \frac{c^*}{K + c^*} a(z_k, z_k) = \frac{c^*}{K + c^*} \quad \text{für alle } k = 1, 2, \dots, N_j$$

und folglich für alle $v_j = \sum_{k=1}^{N_j} \sigma_k z_k \in V_j$

$$\begin{aligned} a(E_j v_j, E_j v_j) &= a\left(\sum_{k=1}^{N_j} \sigma_k E_j z_k, \sum_{k=1}^{N_j} \sigma_k E_j z_k\right) \\ &= \sum_{k=1}^{N_j} \sigma_k^2 \mu_k^2 \\ &\leq \left(\frac{c^*}{K + c^*}\right)^2 \sum_{k=1}^{N_j} \sigma_k^2 \\ &= \left(\frac{c^*}{K + c^*}\right)^2 a(v_j, v_j). \end{aligned}$$

□

Satz 8.15 besagt, dass der symmetrische V-Zyklus bereits mit einem A-priori- beziehungsweise A-posteriori-Glättungsschritt mit einer von j unabhängigen Konvergenzrate konvergiert. Konvergenz liegt in der Praxis auch bei Verwendung anderer Glättungsverfahren, wie beispielsweise dem Gauß-Seidel- oder dem Jacobi-Verfahren, vor.

8.6 Geschachtelte Iteration

Optimale, das heißt, unabhängig von j , konvergente Iterationsverfahren führen zu folgender Abschätzung

$$\rho^{R_j} \leq \varepsilon_j.$$

Hierin bezeichnet R_j die Anzahl der durchzuführenden Iterationschritte, ε_j die gewünschte Genauigkeit und ρ den Kontraktionsfaktor. Da die Genauigkeit ε_j von der Schrittweite h_j abhängt, im allgemeinen ist $\varepsilon_j \sim h_j^\sigma$ mit einem $\sigma > 0$, ist die Anzahl der Iterationsschritte nicht von j unabhängig beschränkt:

$$R_j \geq c |\log(h_j)|.$$

Dabei ist beispielsweise $\sigma = 1$ bei linearen Finiten Elementen. Wie wir sehen werden, kann mit Hilfe der geschachtelten Iteration lineare Komplexität erzielt werden.

Ausgehend von der größten Zerlegung führt man sukzessive auf jeder Gitterebene R V-Zyklen durch. Dadurch lässt sich der beim Lösen des entsprechenden linearen Gleichungssystems auftretende Fehler auf die Größe des jeweiligen Diskretisierungsfehlers reduzieren.

Algorithmus 8.16 (geschachtelte Iteration)

input: Diskretisierungslevel J

output: Näherungslösung $\hat{\mathbf{u}}_J \approx \mathbf{A}_J^{-1} \mathbf{f}_J$

① setze $\hat{\mathbf{u}}_0 := \mathbf{A}_0^{-1} \mathbf{f}_0$

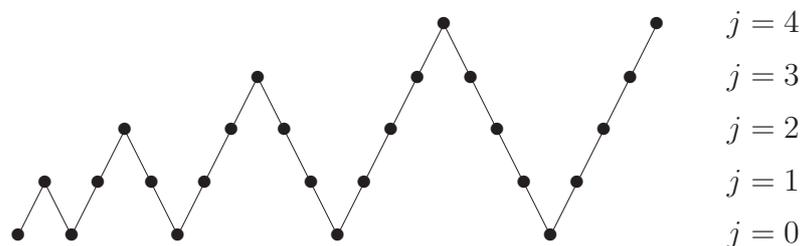
② für alle $j = 1, 2, \dots, J$

setze $\hat{\mathbf{u}}_j^{(0)} := \mathbf{I}_{j-1}^j \hat{\mathbf{u}}_{j-1}$

für alle $r = 1, 2, \dots, R$ berechne $\hat{\mathbf{u}}_j^{(r)} := \text{V-Zyklus}(\hat{\mathbf{u}}_j^{(r-1)}, \mathbf{f}_j)$

setze $\hat{\mathbf{u}}_j := \hat{\mathbf{u}}_j^{(R)}$

Im Fall $R = 1$ ergibt sich etwa folgendes Schema:



Bemerkung Analog kann man statt V-Zyklen auch W-Zyklen oder sogar ein vorkonditioniertes CG-Verfahren verwenden. \triangle

Wir haben in Satz 6.4 bewiesen, dass, unter der Voraussetzung der $H^2(\Omega)$ -Regularität, für den Diskretisierungsfehler gilt

$$\|u - u_j\|_{H^1(\Omega)} \leq ch_j \|f\|_{L^2(\Omega)}.$$

Der nachfolgende Satz besagt, dass die geschachtelte Iteration den algebraischen Fehler auf die Größe des Diskretisierungsfehlers reduziert, was hinreichend für die Konvergenz der Finite-Element-Methode ist.

Satz 8.17 (Optimalität der geschachtelten Iteration) Das Gebiet Ω sei ein konvexes Polygonebiet und $f \in L^2(\Omega)$. Ferner gelten die Voraussetzungen von Satz 8.15. Dann genügt die Lösung $\widehat{\mathbf{u}}_j$ der geschachtelten Iteration mit R V-Zyklen auf jeder Gitterebene der Fehlerabschätzung

$$\|\mathbf{u}_j - \widehat{\mathbf{u}}_j\|_{1,j} \leq ch_j \|f\|_{L^2(\Omega)}$$

mit einer von j unabhängigen Konstanten, vorausgesetzt R ist groß genug.

Beweis. Mit $\gamma = c^*/(K + c^*)$ folgt aus Satz 8.15

$$\begin{aligned} \|\mathbf{u}_j - \widehat{\mathbf{u}}_j\|_{1,j} &\leq \gamma^R \|\mathbf{u}_j - \mathbf{I}_{j-1}^j \widehat{\mathbf{u}}_{j-1}\|_{1,j} \\ &\leq \gamma^R \left\{ \|\mathbf{u}_j - \mathbf{I}_{j-1}^j \mathbf{u}_{j-1}\|_{1,j} + \|\mathbf{I}_{j-1}^j (\mathbf{u}_{j-1} - \widehat{\mathbf{u}}_{j-1})\|_{1,j} \right\} \\ &= \gamma^R \left\{ \sqrt{a(u_j - u_{j-1}, u_j - u_{j-1})} + \|\mathbf{u}_{j-1} - \widehat{\mathbf{u}}_{j-1}\|_{1,j-1} \right\} \\ &\leq \gamma^R \left\{ \sqrt{c_S} (\|u - u_j\|_{H^1(\Omega)} + \|u - u_{j-1}\|_{H^1(\Omega)}) + \|\mathbf{u}_{j-1} - \widehat{\mathbf{u}}_{j-1}\|_{1,j-1} \right\}. \end{aligned}$$

Satz 6.4 liefert $\|u - u_j\|_{H^1(\Omega)} \leq ch_j \|f\|_{L^2(\Omega)}$ und $\|u - u_{j-1}\|_{H^1(\Omega)} \leq 2ch_j \|f\|_{L^2(\Omega)}$, das heißt

$$\|\mathbf{u}_j - \widehat{\mathbf{u}}_j\|_{1,j} \leq c\gamma^R \{h_j \|f\|_{L^2(\Omega)} + \|\mathbf{u}_{j-1} - \widehat{\mathbf{u}}_{j-1}\|_{1,j-1}\}.$$

Aufsummieren und Verwenden von $\mathbf{u}_0 - \widehat{\mathbf{u}}_0 = \mathbf{0}$ ergibt

$$\begin{aligned} \|\mathbf{u}_j - \widehat{\mathbf{u}}_j\|_{1,j} &\leq \{c\gamma^R h_j + c^2 \gamma^{2R} h_{j-1} + \dots + c^{j+1} \gamma^{(j+1)R} h_0\} \|f\|_{L^2(\Omega)} \\ &= c\gamma^R h_j \{1 + 2c\gamma^R + \dots + 2^j c^j \gamma^{jR}\} \|f\|_{L^2(\Omega)} \\ &\leq \frac{c\gamma^R h_j}{1 - 2c\gamma^R} \|f\|_{L^2(\Omega)}, \end{aligned}$$

vorausgesetzt es ist $c\gamma^R < 1/2$. □

Satz 8.18 (Komplexität der geschachtelten Iteration) Der Aufwand der geschachtelten Iteration skaliert linear in der Anzahl der Unbekannten.

Beweis. Bezeichnet W_j die Anzahl der Rechenoperationen für den V-Zyklus auf der Gitterebene j , dann ergibt sich die Rekursionsformel

$$W_0 = cN_0, \quad W_j \leq c(K + L)N_j + W_{j-1}, \quad j \in \mathbb{N}.$$

Da $\dim V_j \sim 2^{jd}$ gilt, ist

$$W_j \leq c(K + L) \{N_j + N_{j-1} + \dots + N_0\} \leq c(K + L) \frac{N_j}{2^d - 1} \leq cN_j,$$

das heißt, der Aufwand eines V-Zyklus wächst linear mit der Anzahl der Unbekannten N_j .

Der Aufwand \widehat{W}_j der geschachtelten Iteration ergibt sich nun gemäß

$$\widehat{W}_0 = W_0, \quad \widehat{W}_j = \widehat{W}_{j-1} + RW_j, \quad j \in \mathbb{N}.$$

Wegen $W_j \leq cN_j$ bedeutet dies

$$\widehat{W}_j \leq cR \{N_j + N_{j-1} + \dots + N_0\} \leq cN_j.$$

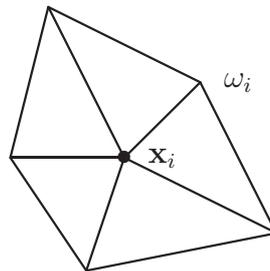
□

9. Residuale Fehlerschätzer

9.1 Clément-Operator

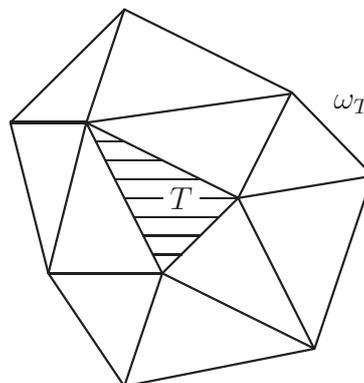
Der Einsatz des Interpolationsoperators I_h in Satz 5.16 setzt H^2 -Funktionen voraus. Von P. Clément stammt ein Approximationsprozess, mit dem auch H^1 -Funktionen erfasst werden. Dazu sei \mathcal{T} eine Triangulierung von Ω . Zu jedem Gitterknoten \mathbf{x}_i definieren wir die Vereinigung der angrenzenden Elemente

$$\omega_i := \bigcup_{T \in \mathcal{T}: \mathbf{x}_i \in T} T.$$



Ähnlich definiert sich zu jedem Element $T \in \mathcal{T}$ der Patch ω_T als die Vereinigung aller Elemente, die mindestens einen Eckpunkt mit T gemeinsam haben:

$$\omega_T := \bigcup_{T' \in \mathcal{T}: T \cap T' \neq \emptyset} T' = \bigcup_{\mathbf{x}_i \in T} \omega_i.$$



Auf nicht entarteten Triangulierungen gilt offensichtlich

$$|\omega_T| \leq c|T| \leq ch_T^2.$$

Die *Clément-Approximation* im linearen Finite-Element-Raum V_h ist nun in der nodalen Basis $\{\varphi_i\}$ gegeben durch

$$C_h : H^1(\Omega) \rightarrow V_h, \quad C_h v(\mathbf{x}) := \sum_{i=1}^N (P_i v) \varphi_i(\mathbf{x}).$$

Hierbei bezeichnet $P_i : L^2(\omega_i) \rightarrow \mathcal{P}_0$ die L^2 -Projektion von v auf die Konstanten.

Satz 9.1 (Clément) Sei \mathcal{T} eine nicht entartete Triangulierung. Dann gilt für jedes Element $T \in \mathcal{T}$ und dessen Kanten $e \subset \partial T$

$$\begin{aligned} \|v - C_h v\|_{H^m(T)} &\leq ch_T^{1-m} \|v\|_{H^1(\omega_T)}, \quad m = 0, 1 \\ \|v - C_h v\|_{L^2(e)} &\leq ch_T^{1/2} \|v\|_{H^1(\omega_T)}. \end{aligned}$$

Beweis. Das Bramble-Hilbert-Lemma impliziert

$$\|v - P_i v\|_{H^1(\omega_i)} \leq c \|v\|_{H^1(\omega_i)} \quad (9.1)$$

und mit einem Skalierungsargument

$$\|v - P_i v\|_{L^2(\omega_i)} \leq c \operatorname{diam}(\omega_i) \|v\|_{H^1(\omega_i)} \leq ch_T \|v\|_{H^1(\omega_i)}. \quad (9.2)$$

Für jedes Element $T \in \mathcal{T}$ erscheinen drei Summanden in $C_h v$, die jeweils den Ecken zugeordnet sind. Für jeden dieser Summanden gilt nach (9.1), (9.2) eine Fehlerabschätzung in der gewünschten Form:

$$\|v - P_i v\|_{H^m(T)} \leq \|v - P_i v\|_{H^m(\omega_i)} \leq ch_T^{1-m} \|v\|_{H^1(\omega_i)} \leq ch_T^{1-m} \|v\|_{H^1(\omega_T)}, \quad m = 0, 1.$$

Da $C_h v$ punktweise eine Konvexkombination darstellt, folgt die erste Behauptung. Die zweite Behauptung ergibt sich aus einer verschärften Form des Spursatzes. \square

9.2 A-posteriori-Fehlerschätzung

Bei der Darstellung von A-posteriori-Fehlerschätzern beschränken wir uns auf die Poisson-Gleichung $-\Delta u = f$ mit homogenen Dirichlet-Randbedingungen. Wenn man die Galerkin-Lösung u_h in die Poisson-Gleichung in klassischer Form einsetzt, erhält man ein Residuum. Außerdem unterscheidet sich u_h von der klassischen Lösung durch Sprünge in den Ableitungen an den Elementgrenzen. Wir haben folglich flächenbezogene Residuen

$$R_T := R_T(u_h) := \Delta u_h + f \quad \text{für } T \in \mathcal{T}$$

und kantenbezogene Sprünge

$$R_e := R_e(u_h) := \left[\frac{\partial u_h}{\partial \mathbf{n}} \right] = \frac{\partial u_h}{\partial \mathbf{n}} \Big|_{T_1} - \frac{\partial u_h}{\partial \mathbf{n}} \Big|_{T_2} \quad \text{für } e = T_1 \cap T_2 \in \mathcal{E}.$$

Hier bezeichnet \mathcal{E} die Menge der inneren Kanten (also Kanten, die nicht auf dem Rand Γ liegen). Im weiteren sei ferner \mathcal{E}_T die Menge der inneren Kanten eines Elementes T .

Wir definieren nun die lokalen Fehlerschätzer

$$\eta_{T,R}^2 := h_T^2 \|R_T\|_{L^2(T)}^2 + \frac{1}{2} \sum_{e \in \mathcal{E}_T} h_e \|R_e\|_{L^2(e)}^2,$$

welche wir zu einem globalen Fehlerschätzer zusammenbauen

$$\eta_R^2 := \sum_{T \in \mathcal{T}} \eta_{T,R}^2 = \sum_{T \in \mathcal{T}} h_T^2 \|R_T\|_{L^2(T)}^2 + \sum_{e \in \mathcal{E}} h_e \|R_e\|_{L^2(e)}^2.$$

Dabei bezeichne h_e die Länge der Kante e .

Satz 9.2 Sei \mathcal{T} eine nicht entartete Triangulierung mit Regularitätsparameter κ . Dann gibt es eine Konstante $c = c(\Omega, \kappa)$, so dass

$$\|u - u_h\|_{H^1(\Omega)} \leq c \sqrt{\sum_{T \in \mathcal{T}} \eta_{T,R}^2}.$$

Beweis. Der Ausgangspunkt ist das Dualitätsargument

$$|u - u_h|_{H^1(\Omega)} = \sup_{w \in H_0^1(\Omega): |w|_{H^1(\Omega)}=1} (\nabla(u - u_h), \nabla w)_{L^2(\Omega)}. \quad (9.3)$$

Für ein spezielles $w \in H_0^1(\Omega)$ folgt

$$\begin{aligned} \ell(w) &:= (\nabla(u - u_h), \nabla w)_{L^2(\Omega)} \\ &= (f, w)_{L^2(\Omega)} - \sum_{T \in \mathcal{T}} (\nabla u_h, \nabla w)_{L^2(T)} \\ &= (f, w)_{L^2(\Omega)} - \sum_{T \in \mathcal{T}} \left\{ (-\Delta u_h, w)_{L^2(T)} + \sum_{e \in \mathcal{E}_T} \left(\frac{\partial u_h}{\partial \mathbf{n}}, w \right)_{L^2(e)} \right\} \\ &= \sum_{T \in \mathcal{T}} (\Delta u_h + f, w)_{L^2(T)} + \sum_{e \in \mathcal{E}} \left(\left[\frac{\partial u_h}{\partial \mathbf{n}} \right], w \right)_{L^2(e)} \\ &= \sum_{T \in \mathcal{T}} (R_T, w)_{L^2(T)} + \sum_{e \in \mathcal{E}} (R_e, w)_{L^2(e)} \\ &= \sum_{T \in \mathcal{T}} \left\{ (R_T, w)_{L^2(T)} + \frac{1}{2} \sum_{e \in \mathcal{E}_T} (R_e, w)_{L^2(e)} \right\}. \end{aligned}$$

Wir setzen nun $w_h := C_h w$ und benutzen die Galerkin-Orthogonalität

$$(\nabla(u - u_h), \nabla v_h)_{L^2(\Omega)} = 0 \quad \text{für alle } v_h \in V_h.$$

Dann folgt

$$\ell(w) = \ell(w - w_h) \leq \sum_{T \in \mathcal{T}} \left\{ \|R_T\|_{L^2(T)} \|w - w_h\|_{L^2(T)} + \frac{1}{2} \sum_{e \in \mathcal{E}_T} \|R_e\|_{L^2(e)} \|w - w_h\|_{L^2(e)} \right\}.$$

Da $\bigcup_{T \in \mathcal{T}} \omega_T$ das Gebiet Ω nur endlich oft überdeckt, ergibt sich mit Hilfe von Satz 9.1

$$\begin{aligned} \ell(w) &\leq c \sum_{T \in \mathcal{T}} \left\{ \underbrace{h_T \|R_T\|_{L^2(T)} + \frac{1}{2} \sum_{e \in \mathcal{E}_T} h_e^{1/2} \|R_e\|_{L^2(e)}}_{|\cdot| \leq c\eta_{T,R}} \right\} \|w\|_{H^1(\omega_T)} \\ &\leq c \left(\sum_{T \in \mathcal{T}} \eta_{T,R}^2 \right)^{1/2} \left(\sum_{T \in \mathcal{T}} \|w\|_{H^1(\omega_T)}^2 \right)^{1/2} \\ &\leq c\eta_R \|w\|_{H^1(\Omega)}. \end{aligned}$$

Diese Abschätzung eingesetzt in das Dualitätsargument (9.3) liefert mit der Friedrichschen Ungleichung die Behauptung. \square

Bemerkung Das flächenbezogene Residuum ist in dieser Form numerisch nicht berechenbar, da man im allgemeinen f nicht exakt integrieren kann. Daher spaltet man f auf in $f = f_h + f - f_h$, wobei nun das flächenbezogene Residuum $\Delta u_h + f_h$ exakt berechnet werde. Wegen

$$\|\Delta u_h + f\|_{L^2(T)} \leq \|\Delta u_h + f_h\|_{L^2(T)} + \|f - f_h\|_{L^2(T)}$$

erhält man die Schranke

$$\|u - u_h\|_{H^1(\Omega)} \leq c \left\{ \left(\sum_{T \in \mathcal{T}} \eta_{T,R}^2 \right)^{1/2} + \left(\sum_{T \in \mathcal{T}} h_T^2 \|f - f_h\|_{L^2(T)}^2 \right)^{1/2} \right\}.$$

Der letzte Summand wird dabei *Datenoszillation* genannt. \triangle

9.3 Untere Abschätzung

Wir wollen auch eine untere lokale Schranke für den Fehler beweisen. Ein wesentliches Hilfsmittel bilden dabei Abschneidefunktionen ψ_T und ψ_e .

Es ist ψ_T die kubische *Blasenfunktion*

$$\text{supp } \psi_T = T, \quad \psi_T \in \mathcal{P}_3, \quad 0 \leq \psi_T \leq 1 = \max \psi_T.$$

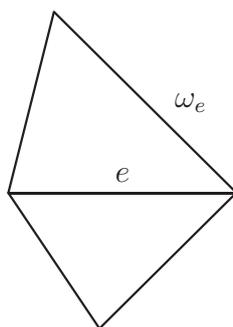
In baryzentrischen Koordinaten ist sie gegeben durch $\psi_T(\mathbf{x}) = 27\lambda_1(\mathbf{x})\lambda_2(\mathbf{x})\lambda_3(\mathbf{x})$.

Dagegen ist $\psi_e \in C(\Omega)$ aus quadratischen Polynomen zusammengesetzt, die auf je zwei Seiten der Dreiecke aus dem Träger verschwinden

$$\text{supp } \psi_e = T_1 \cup T_2, \quad \psi_e|_{T_i} \in \mathcal{P}_2, \quad 0 \leq \psi_e \leq 1 = \max \psi_e.$$

Ferner setzen wir

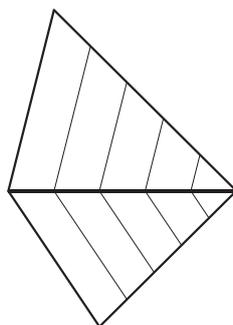
$$\omega_e := \bigcup_{T \in \mathcal{T}: e \in \mathcal{E}_T} T$$



und definieren eine Abbildung $E : L^2(e) \rightarrow L^2(\omega_e)$, die jede auf einer Kante e definierte Funktion auf die benachbarten Dreiecke ω_e fortsetzt. Wir setzen

$$(E(\sigma))(\mathbf{x}) = \sigma(\mathbf{z}) \text{ in } T, \text{ wenn } \mathbf{z} \in e \text{ der Punkt aus } e \text{ mit } \lambda_j(\mathbf{x}) = \lambda_j(\mathbf{z}) \text{ ist.}$$

Dabei ist λ_j eine der zwei baryzentrischen Koordinaten in T , die auf e nicht konstant ist. Die Niveaulinien dieser Fortsetzung sind demnach wie folgt gegeben:



Lemma 9.3 Sei \mathcal{T} eine nicht entartete Triangulierung. Dann gilt für $T \in \mathcal{T}$ und $e \in \mathcal{E}_T$ mit einer nur vom Parameter κ abhängigen Zahl c :

$$\|\psi_T v\|_{L^2(T)} \leq \|v\|_{L^2(T)} \quad \text{für alle } v \in L^2(T) \quad (9.4)$$

$$\|\psi_T^{1/2} p\|_{L^2(T)} \geq c \|p\|_{L^2(T)} \quad \text{für alle } p \in \mathcal{P}_2 \quad (9.5)$$

$$|\psi_T p|_{H^1(T)} \leq c h_T^{-1} \|\psi_T p\|_{L^2(T)} \quad \text{für alle } p \in \mathcal{P}_2 \quad (9.6)$$

$$\|\psi_e^{1/2} \sigma\|_{L^2(e)} \geq c \|\sigma\|_{L^2(e)} \quad \text{für alle } \sigma \in \mathcal{P}_2 \quad (9.7)$$

$$c h_e^{1/2} \|\sigma\|_{L^2(e)} \leq \|\psi_e E(\sigma)\|_{L^2(T)} \leq C h_e^{1/2} \|\sigma\|_{L^2(e)} \quad \text{für alle } \sigma \in \mathcal{P}_2 \quad (9.8)$$

$$|\psi_e E(\sigma)|_{H^1(T)} \leq c h_T^{-1} \|\psi_e E(\sigma)\|_{L^2(T)} \quad \text{für alle } \sigma \in \mathcal{P}_2 \quad (9.9)$$

Beweis. Die Abschätzung (9.4) folgt sofort aus $0 \leq \psi_T \leq 1$. Die anderen Behauptungen sind wegen der endlichen Dimension von \mathcal{P}_2 für ein festes Referenzelement klar. Die Übertragung auf beliebige Dreiecke ergibt sich mit den üblichen Skalierungsargumenten. \square

Satz 9.4 Sei \mathcal{T} eine nicht entartete Triangulierung mit Regularitätsparameter κ und sei

$$\tilde{\omega}_T := \bigcup_{e \in \mathcal{E}_T} \omega_e.$$

Dann gibt es eine Konstante $c = c(\Omega, \kappa)$, so dass

$$\eta_{T,R} \leq c \left\{ \|u - u_h\|_{H^1(\tilde{\omega}_T)}^2 + \sum_{T' \subset \tilde{\omega}_T} h_{T'}^2 \|f - f_h\|_{L^2(T')}^2 \right\}^{1/2}, \quad (9.10)$$

wobei f_h die L^2 -Projektion von f in V_h ist.

Beweis. Sei $T \in \mathcal{T}$. Wir bilden das reduzierte flächenbezogene Residuum

$$R_{T,\text{red}} := R_{T,\text{red}}(u_h) := \Delta u_h + f_h \in \mathcal{P}_2$$

und setzen

$$w := w_T := \psi_T \cdot R_{T,\text{red}}.$$

Mit (9.5) folgt dann

$$\begin{aligned} c \|R_{T,\text{red}}\|_{L^2(T)}^2 &\leq \|\psi_T^{1/2} R_{T,\text{red}}\|_{L^2(T)}^2 \\ &= (R_{T,\text{red}}, w)_{L^2(T)} \\ &= (\Delta(u_h - u), w)_{L^2(T)} + (f_h - f, w)_{L^2(T)} \\ &= (\nabla(u - u_h), \nabla w)_{L^2(T)} + (f_h - f, w)_{L^2(T)} \\ &\leq |u - u_h|_{H^1(T)} |w|_{H^1(T)} + \|f - f_h\|_{L^2(T)} \|w\|_{L^2(T)}. \end{aligned}$$

Man beachte, dass wegen (9.6) gilt $|w|_{H^1(T)} \leq ch_T^{-1} \|w\|_{L^2(T)}$ und wegen (9.4) gilt $\|w\|_{L^2(T)} \leq \|R_{T,\text{red}}\|_{L^2(T)}$. Daher schließen wir, dass

$$\|R_{T,\text{red}}\|_{L^2(T)}^2 \leq c \{ h_T^{-1} |u - u_h|_{H^1(T)} \|R_{T,\text{red}}\|_{L^2(T)} + \|f - f_h\|_{L^2(T)} \|R_{T,\text{red}}\|_{L^2(T)} \},$$

das heißt

$$\|R_{T,\text{red}}\|_{L^2(T)} \leq c \{ h_T^{-1} |u - u_h|_{H^1(T)} + \|f - f_h\|_{L^2(T)} \}.$$

Wegen

$$\begin{aligned} \|R_T\|_{L^2(T)} &= \|\Delta u_h + f\|_{L^2(T)} \\ &\leq \|\Delta u_h + f_h\|_{L^2(T)} + \|f - f_h\|_{L^2(T)} \\ &= \|R_{T,\text{red}}\|_{L^2(T)} + \|f - f_h\|_{L^2(T)} \end{aligned}$$

ergibt sich hieraus die Abschätzung

$$h_T \|R_T\|_{L^2(T)} \leq c \{ |u - u_h|_{H^1(T)} + h_T \|f - f_h\|_{L^2(T)} \}. \quad (9.11)$$

In ähnlicher Weise werden die kantenbezogenen Terme des Fehlerschätzers behandelt. Sei $e \in \mathcal{E}$, dann folgt für

$$w := w_e := \psi_e \cdot E(R_e)$$

dass $\text{supp } w = \omega_e$. Außerdem ist $R_e \in \mathcal{P}_2$. Mit (9.7) erhalten wir

$$\begin{aligned}
c \|R_e\|_{L^2(e)}^2 &\leq \|\psi_e^{1/2} R_e\|_{L^2(e)}^2 \\
&= (R_e, w)_{L^2(e)} \\
&= (\nabla(u_h - u), \nabla w)_{L^2(\omega_e)} + \sum_{T \subset \omega_e} (\Delta u_h + f, w)_{L^2(T)} \\
&= (\nabla(u_h - u), \nabla w)_{L^2(\omega_e)} + \sum_{T \subset \omega_e} (R_T, w)_{L^2(T)} \\
&\leq |u - u_h|_{H^1(\omega_e)} |w|_{H^1(\omega_e)} + \sum_{T \subset \omega_e} \|R_T\|_{L^2(T)} \|w\|_{L^2(T)}.
\end{aligned}$$

Mit (9.9) schließen wir $|w|_{H^1(T)} \leq ch_T^{-1} \|w\|_{L^2(T)} \leq ch_e^{-1} \|w\|_{L^2(T)}$ für $T \subset \omega_e$ und mit (9.8) folgt $\|w\|_{L^2(T)} \leq Ch_e^{1/2} \|R_e\|_{L^2(e)}$. Also ergibt sich

$$\|R_e\|_{L^2(e)}^2 \leq ch_e^{-1/2} |u - u_h|_{H^1(\omega_e)} \|R_e\|_{L^2(e)} + ch_e^{1/2} \sum_{T \subset \omega_e} \|R_T\|_{L^2(T)} \|R_e\|_{L^2(e)},$$

das heißt

$$h_e^{1/2} \|R_e\|_{L^2(e)} \leq c \left\{ |u - u_h|_{H^1(\omega_e)} + h_e \sum_{T \subset \omega_e} \|R_T\|_{L^2(T)} \right\}.$$

Wegen $h_e \leq h_T$ für $T \subset \omega_e$ erhalten wir im Hinblick auf (9.11)

$$h_e^{1/2} \|R_e\|_{L^2(e)} \leq c \left\{ |u - u_h|_{H^1(\omega_e)} + \sum_{T \subset \omega_e} h_T \|f - f_h\|_{L^2(T)} \right\}. \quad (9.12)$$

Wenn wir noch $\tilde{\omega}_T = \bigcup_{e \in \mathcal{E}_T} \omega_e$ beachten, folgt aus (9.11) und (9.12) schließlich die Behauptung. \square

Bemerkung Aufsummation der lokalen unteren Fehlerschranke (9.10) liefert die Abschätzung

$$\eta_R^2 \leq c \left\{ \|u - u_h\|_{H^1(\Omega)}^2 + \sum_{T \in \mathcal{T}} h_T^2 \|f - f_h\|_{L^2(T)}^2 \right\},$$

das heißt, der globale Fehlerschätzer η_R ist bis auf die Datenoszillation äquivalent zum Fehler der Lösung. Die lokale untere Fehlerschranke (9.10) zeigt jedoch, dass der Fehler wirklich in den lokalen Schätzern $\eta_{T,R}$ lokalisiert ist. In der Praxis berechnet man alle lokalen Fehlergrößen $\eta_{T,R}$ und wählt einen festen Prozentsatz der größten Beiträge davon aus. Die zugehörigen Elemente werden dann zur Verfeinerung markiert. \triangle

10. Nichtsymmetrische Bilinearformen

Bei den bisher behandelten Variationsproblemen war die zugrundeliegende Bilinearform stets symmetrisch. Bei nichtsymmetrischen Bilinearformen ergeben sich folgende Schwierigkeiten:

- Die Bilinearform $a(\cdot, \cdot)$ definiert kein Skalarprodukt, daher ist ein neuer Zugang zum Beweis von Existenz und Eindeutigkeit einer Lösung notwendig.
- Die Variationsaufgabe ist nicht äquivalent zu einem Minimierungsproblem.
- Die Steifigkeitsmatrix ist nicht mehr symmetrisch. Daher ist beispielsweise ein CG-Verfahren nicht mehr anwendbar. Insbesondere ist im allgemeinen eine Konvergenzanalyse durch Eigenwerte nicht mehr gültig.

Beispiel 10.1 Der Vektor $\mathbf{b} \in \mathbb{R}^d$ besitze konstante Länge 1. Die Bilinearform $a : H_0^1(\Omega) \times H_0^1(\Omega) \rightarrow \mathbb{R}$ zur Differentialgleichung

$$-\Delta u + \langle \mathbf{b}, \nabla u \rangle + u = f \text{ in } \Omega, \quad u = 0 \text{ auf } \Gamma$$

auf dem beschränkten Gebiet $\Omega \subset \mathbb{R}^d$ lautet

$$a(u, v) = \int_{\Omega} \{ \langle \nabla u, \nabla v \rangle + \langle \mathbf{b}, \nabla u \rangle v + uv \} \, d\mathbf{x}.$$

Diese unsymmetrische Bilinearform ist offensichtlich stetig. Wegen $|\langle \mathbf{b}, \nabla u \rangle| \leq \|\nabla u\|$ ist sie auch elliptisch:

$$\begin{aligned} a(u, u) &\geq \int_{\Omega} \{ \|\nabla u\|^2 - \|\nabla u\||u| + u^2 \} \, d\mathbf{x} \\ &= \frac{1}{2} \int_{\Omega} \{ \|\nabla u\|^2 + u^2 \} \, d\mathbf{x} + \frac{1}{2} \int_{\Omega} \underbrace{\{ \|\nabla u\| - |u| \}^2}_{\geq 0} \, d\mathbf{x} \\ &\geq \frac{1}{2} \|u\|_{H^1(\Omega)}^2. \end{aligned}$$

△

Satz 10.2 (Lax-Milgram) Gegeben seien ein Hilbert-Raum V mit Innenprodukt (\cdot, \cdot) , eine in V stetige und elliptische, nicht notwendigerweise symmetrische Bilinearform $a : V \times V \rightarrow \mathbb{R}$, sowie ein beschränktes lineares Funktional $\ell : V \rightarrow \mathbb{R}$. Dann gibt es genau ein $u \in V$ mit

$$a(u, v) = \ell(v) \quad \text{für alle } v \in V.$$

Beweis. Wir vollziehen den Beweis in vier Schritten.

(i.) Für festes $v \in V$ definieren wir das lineare Funktional $A_v \in V'$ durch $A_v(w) = a(v, w)$. Die Abbildung

$$A : V \rightarrow V', \quad v \mapsto A_v$$

ist linear

$$A_{\alpha v_1 + \beta v_2}(w) = a(\alpha v_1 + \beta v_2, w) = \alpha a(v_1, w) + \beta a(v_2, w) = \alpha A_{v_1}(w) + \beta A_{v_2}(w)$$

und stetig

$$\frac{\|A_v\|_{V'}}{\|v\|_V} = \sup_{w \in V} \frac{A_v(w)}{\|v\|_V \|w\|_V} = \sup_{w \in V} \frac{a(v, w)}{\|v\|_V \|w\|_V} \leq c_S.$$

(ii.) Nach dem Riesz'schen Darstellungssatz gibt es zu jedem $\ell \in V'$ genau ein $v \in V$ derart, dass $(v, w) = \ell(w)$ für alle $w \in V$. Dies definiert eine bijektive lineare Abbildung $\tau : V' \rightarrow V$ mit $(\tau\ell, w) = \ell(w)$ für alle $w \in V$, die isometrisch ist:

$$\|\ell\|_{V'} = \sup_{w \in V} \frac{\ell(w)}{\|w\|_V} = \sup_{w \in V} \frac{(\tau\ell, w)}{\|w\|_V} = \|\tau\ell\|_V.$$

(iii.) Wir zeigen nun, dass genau ein $u \in V$ existiert, so dass $A_u(v) = \ell(v)$ ist für alle $v \in V$. Dies ist äquivalent zur Aussage, dass es genau ein $u \in V$ gibt mit $A_u = \ell$ in V' , beziehungsweise nach (ii.) mit $\tau A_u = \tau\ell$ in V . Betrachte dazu im folgenden die Abbildung

$$T : V \rightarrow V, \quad v \mapsto Tv = v - \alpha(\tau A_v - \tau\ell).$$

Ist $\alpha > 0$ so gewählt, dass T eine Kontraktion darstellt, das heißt $\|T\|_{V \rightarrow V} = \gamma < 1$, dann liefert der Banachsche Fixpunktsatz die Existenz und Eindeutigkeit der Lösung $Tu = u$, beziehungsweise $\tau A_u = \tau\ell$.

(iv.) Um die Kontraktionseigenschaft zu zeigen, seien $u_1, u_2 \in V$. Mit $v := u_1 - u_2$ folgt dann

$$\begin{aligned} \|Tu_1 - Tu_2\|_V^2 &= \|u_1 - u_2 - \alpha(\tau A_{u_1} - \tau A_{u_2})\|_V^2 \\ &= \|v - \alpha\tau A_v\|_V^2 \\ &= \|v\|_V^2 - 2\alpha(v, \tau A_v) + \alpha^2 \|\tau A_v\|_V^2. \end{aligned}$$

Nach Definition von τ gilt $(v, \tau A_v) = A_v(v) = a(v, v)$, also

$$\|Tu_1 - Tu_2\|_V^2 = \|v\|_V^2 - 2\alpha a(v, v) + \alpha^2 \|\tau A_v\|_V^2.$$

Weiter ist nach (i.) $\|\tau A_v\|_V = \|A_v\|_{V'} \leq c_S \|v\|_V$, so dass zusammen mit der Elliptizität der Bilinearform folgt

$$\|Tu_1 - Tu_2\|_V^2 \leq \|v\|_V^2 - 2\alpha c_E \|v\|_V^2 + \alpha^2 c_S^2 \|v\|_V^2 = \underbrace{(1 - 2\alpha c_E + \alpha^2 c_S^2)}_{=:\gamma^2} \|u_1 - u_2\|_V^2.$$

Setzen wir etwa $\alpha = c_E/c_S^2$, so ist $\gamma < 1$ und folglich T kontrahierend. Gemäß (iii.) ergibt sich daher das Behauptete. \square

Bemerkung Aus der Elliptizität ergibt sich die Stabilitätsabschätzung

$$\|u\|_V^2 \leq \frac{1}{c_E} a(u, u) = \frac{1}{c_E} \ell(u) \leq \frac{1}{c_E} \|\ell\|_{V'} \|u\|_V,$$

das heißt, $\|u\|_V \leq \|\ell\|_{V'}/c_E$. Die Variationsaufgabe bezüglich einer elliptischen und stetigen Bilinearform ist demnach sachgemäß gestellt. \triangle

Der Beweis des Céa-Lemmas (Satz 4.1) behält in dieser Form auch im Fall einer unsymmetrischen elliptischen Bilinearform seine Gültigkeit. Das gleiche gilt für die komplette Konvergenztheorie: alle Aussagen bleiben gültig.

Der Vollständigkeit halber zeigen wir hingegen nun, dass sich das Céa-Lemma im *symmetrischen* Fall verschärfen lässt. Dabei wird explizit die Äquivalenz der Variationsformulierung zu einem Minimierungsproblem ausgenutzt.

Satz 10.3 (Céa-Lemma) Die symmetrische Bilinearform $a : V \times V \rightarrow \mathbb{R}$ sei stetig und elliptisch, und $u \in V$ und $u_h \in V_h \subset V$ seien die Lösungen der Variationsprobleme (4.1) und (4.2). Dann gilt

$$\|u - u_h\|_V \leq \sqrt{\frac{c_S}{c_E}} \inf_{v_h \in V_h} \|u - v_h\|_V.$$

Beweis. Im Fall einer symmetrischen Bilinearform ergibt der Charakterisierungssatz 3.8 in Verbindung mit dem Satz von Lax-Milgram 3.11, dass

$$a(u_h, u_h) - 2\ell(u_h) \leq a(v_h, v_h) - 2\ell(v_h) \quad \text{für alle } v_h \in V_h.$$

Daher folgt

$$\begin{aligned} a(u - u_h, u - u_h) &= a(u, u) - 2a(u, u_h) + a(u_h, u_h) \\ &= a(u, u) - 2\ell(u_h) + a(u_h, u_h) \\ &\leq a(u, u) - 2\ell(v_h) + a(v_h, v_h) \\ &= a(u, u) - 2a(u, v_h) + a(v_h, v_h) \\ &= a(u - v_h, u - v_h). \end{aligned}$$

Für beliebiges $v_h \in V_h$ ergibt sich daraus

$$\|u - u_h\|_V^2 \leq \frac{1}{c_E} a(u - u_h, u - u_h) \leq \frac{1}{c_E} a(u - v_h, u - v_h) \leq \frac{c_S}{c_E} \|u - v_h\|_V^2,$$

das ist die Behauptung. □

Bemerkung Als Konsequenz dieser verschärften Form des Céa-Lemmas lassen sich alle Konstanten in der zuvor vorgestellten Konvergenztheorie verbessern. △

11. Parabolische Differentialgleichungen

11.1 Linienmethode

Wir betrachten die Wärmeleitungsgleichung

$$\frac{\partial}{\partial t} u(t, \mathbf{x}) - \Delta u(t, \mathbf{x}) = f(t, \mathbf{x}), \quad (t, \mathbf{x}) \in [0, T] \times \Omega$$

für ein Gebiet $\Omega \subset \mathbb{R}^d$ und einen Endzeitpunkt T . Der Einfachheit halber geben wir uns homogene Dirichlet-Randwerte vor

$$u(t, \mathbf{x}) = 0 \quad \text{für alle } (t, \mathbf{x}) \in [0, T] \times \Gamma.$$

Zum Zeitpunkt $t = 0$ verlangen wir die Anfangsbedingung

$$u(0, \mathbf{x}) = g(\mathbf{x}) \quad \text{für alle } \mathbf{x} \in \Omega.$$

Im stationären Fall $\partial u / \partial t \equiv 0$ erhalten wir die übliche Poisson-Gleichung.

Um die Wärmeleitungsgleichung numerisch zu lösen, führen wir zunächst eine Semidiskretisierung im Ort durch, indem wir wie bei einer elliptischen Differentialgleichung vorgehen. Ausgehend von der Variationsformulierung

suche $u(t) \in H_0^1(\Omega)$, so dass

$$\frac{\partial}{\partial t} (u(t), v)_{L^2(\Omega)} + (\nabla u(t), \nabla v)_{L^2(\Omega)} = (f(t), v)_{L^2(\Omega)} \quad \text{für alle } v \in H_0^1(\Omega)$$

liefert die Einschränkung auf den Finite-Elemente-Raum $V_h \subset H_0^1(\Omega)$

suche $u_h(t) \in V_h$, so dass

$$\frac{\partial}{\partial t} (u_h(t), v_h)_{L^2(\Omega)} + (\nabla u_h(t), \nabla v_h)_{L^2(\Omega)} = (f(t), v_h)_{L^2(\Omega)} \quad \text{für alle } v_h \in V_h$$

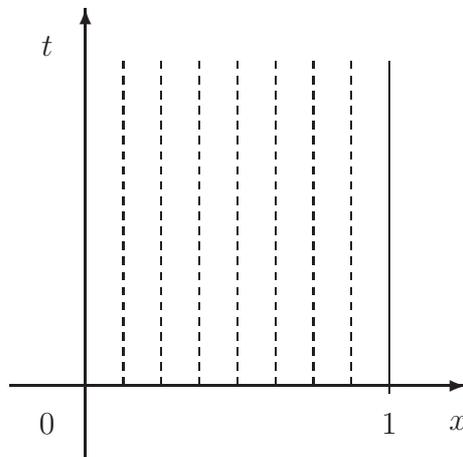
ein lineares Gleichungssystem

$$\mathbf{M} \frac{\partial}{\partial t} \mathbf{u}(t) + \mathbf{A} \mathbf{u}(t) = \mathbf{f}(t), \quad \mathbf{u}(0) = \mathbf{g}. \quad (11.1)$$

Hierin sind die Massenmatrix $\mathbf{M} = [(\varphi_i, \varphi_j)_{L^2(\Omega)}]_{i,j}$ und die Steifigkeitsmatrix $\mathbf{A} = [(\nabla \varphi_i, \nabla \varphi_j)_{L^2(\Omega)}]_{i,j}$ unabhängig von der Zeit, während die Koeffizienten des Lösungsvektors $\mathbf{u}(t) = [u_i(t)]_i$ und die rechte Seite $\mathbf{f}(t) = [(f(t), \varphi_i)_{L^2(\Omega)}]_i$ zeitabhängig sind. Die approximative Lösung besitzt demnach die Form

$$u_h(t, \mathbf{x}) = \sum_i u_i(t) \varphi_i(\mathbf{x}).$$

Bemerkung Aus anschaulichen Gründen wird die Semidiskretisierung (11.1) *Linienmethode* genannt. Für alle $t \geq 0$ enthält nämlich die vektorwertige Funktion $\mathbf{u}(t)$ die Funktionswerte der Approximation $u_h(t, \mathbf{x})$ bezüglich einer räumlichen Triangulierung \mathcal{T}_h . \triangle

Linienmethode für $\Omega = (0, 1)$

11.2 θ -Schema

Wir benötigen nun noch eine geeignete Zeitdiskretisierung. Dazu unterteilen wir das Zeitintervall $[0, T]$ in M Intervalle $[t_i, t_{i+1}]$ und setzen $k_i := t_{i+1} - t_i$. Das *explizite Euler-Verfahren* führt auf

$$\mathbf{M} \frac{\mathbf{u}_{i+1} - \mathbf{u}_i}{k_i} + \mathbf{A} \mathbf{u}_i = \mathbf{f}(t_i) \quad \text{bzw.} \quad \mathbf{M} \mathbf{u}_{i+1} = (\mathbf{M} - k_i \mathbf{A}) \mathbf{u}_i + k_i \mathbf{f}(t_i). \quad (11.2)$$

Hingegen liefert das *implizite Euler-Verfahren*

$$\mathbf{M} \frac{\mathbf{u}_{i+1} - \mathbf{u}_i}{k_i} + \mathbf{A} \mathbf{u}_{i+1} = \mathbf{f}(t_{i+1}) \quad \text{bzw.} \quad (\mathbf{M} + k_i \mathbf{A}) \mathbf{u}_{i+1} = \mathbf{M} \mathbf{u}_i + k_i \mathbf{f}(t_{i+1}). \quad (11.3)$$

Der Startwert lautet in beiden Fällen $\mathbf{u}_0 = \mathbf{g}$. Kombinieren wir (11.2) und (11.3), so erhalten wir das θ -Schema

$$\mathbf{M} \frac{\mathbf{u}_{i+1} - \mathbf{u}_i}{k_i} + (1 - \theta) \mathbf{A} \mathbf{u}_i + \theta \mathbf{A} \mathbf{u}_{i+1} = (1 - \theta) \mathbf{f}(t_i) + \theta \mathbf{f}(t_{i+1}), \quad (11.4)$$

beziehungsweise

$$(\mathbf{M} + k_i \theta \mathbf{A}) \mathbf{u}_{i+1} = (\mathbf{M} - k_i (1 - \theta) \mathbf{A}) \mathbf{u}_i + k_i (1 - \theta) \mathbf{f}(t_i) + k_i \theta \mathbf{f}(t_{i+1}).$$

Dabei gilt

$$\theta = \begin{cases} 0, & \text{explizites Euler-Verfahren} \\ 1/2, & \text{Trapez-Methode} \\ 1, & \text{implizites Euler-Verfahren} \end{cases}$$

wobei die Trapez-Methode im Zusammenhang mit parabolischen Differentialgleichungen auch *Crank-Nicolson-Verfahren* genannt wird. Das θ -Schema ist konsistent von erster Ordnung, im Falle $\theta = 0.5$ sogar von zweiter Ordnung.

Satz 11.1 Für die Wärmeleitungsgleichung und $1/2 \leq \theta \leq 1$ ist das θ -Schema stabil, das heißt, es gilt

$$\begin{aligned} \|u_{h,M}\|_{L^2(\Omega)}^2 + \sum_{i=0}^{M-1} \{k_i |u_{h,i+\theta}|_{H^1(\Omega)}^2 + (2\theta - 1) \|u_{h,i+1} - u_{h,i}\|_{L^2(\Omega)}^2\} \\ \leq \|u_{h,0}\|_{L^2(\Omega)}^2 + c \sum_{i=0}^{M-1} k_i \|f_{i+\theta}\|_{L^2(\Omega)}^2, \end{aligned}$$

wobei wir

$$u_{h,i+\theta} := (1 - \theta)u_{h,i} + \theta u_{h,i+1}, \quad f_{i+\theta} := (1 - \theta)f(t_i) + \theta f(t_{i+1})$$

gesetzt haben.

Beweis. Einsetzen der Testfunktion $u_{h,i+\theta}$ in die Variationsformulierung liefert

$$(u_{h,i+1} - u_{h,i}, u_{h,i+\theta})_{L^2(\Omega)} + k_i \underbrace{(\nabla u_{h,i+\theta}, \nabla u_{h,i+\theta})_{L^2(\Omega)}}_{=|u_{h,i+\theta}|_{H^1(\Omega)}^2} = k_i (f_{i+\theta}, u_{h,i+\theta})_{L^2(\Omega)}, \quad (11.5)$$

vergleiche (11.4). Andererseits gilt

$$\begin{aligned} (u_{h,i+1} - u_{h,i}, u_{h,i+\theta})_{L^2(\Omega)} &= (u_{h,i+1} - u_{h,i}, (1 - \theta)u_{h,i} + \theta u_{h,i+1})_{L^2(\Omega)} \\ &= \left(u_{h,i+1} - u_{h,i}, \frac{1}{2}u_{h,i+1} + \frac{1}{2}u_{h,i} + \left(\theta - \frac{1}{2}\right)(u_{h,i+1} - u_{h,i}) \right)_{L^2(\Omega)} \\ &= \frac{1}{2} \|u_{h,i+1}\|_{L^2(\Omega)}^2 - \frac{1}{2} \|u_{h,i}\|_{L^2(\Omega)}^2 + \left(\theta - \frac{1}{2}\right) \|u_{h,i+1} - u_{h,i}\|_{L^2(\Omega)}^2. \end{aligned}$$

Hieraus folgt zusammen mit (11.5), der Cauchy-Schwarzschen Ungleichung und der Poincaré-Friedrichsschen Ungleichung

$$\begin{aligned} \|u_{h,i+1}\|_{L^2(\Omega)}^2 - \|u_{h,i}\|_{L^2(\Omega)}^2 + (2\theta - 1) \|u_{h,i+1} - u_{h,i}\|_{L^2(\Omega)}^2 + 2k_i |u_{h,i+\theta}|_{H^1(\Omega)}^2 \\ = 2k_i (f_{i+\theta}, u_{h,i+\theta})_{L^2(\Omega)} \\ \leq 2k_i c_\Omega \|f_{i+\theta}\|_{L^2(\Omega)} |u_{h,i+\theta}|_{H^1(\Omega)} \\ \leq k_i c_\Omega^2 \|f_{i+\theta}\|_{L^2(\Omega)}^2 + k_i |u_{h,i+\theta}|_{H^1(\Omega)}^2, \end{aligned}$$

dies bedeutet

$$\begin{aligned} \|u_{h,i+1}\|_{L^2(\Omega)}^2 - \|u_{h,i}\|_{L^2(\Omega)}^2 + (2\theta - 1) \|u_{h,i+1} - u_{h,i}\|_{L^2(\Omega)}^2 \\ + k_i |u_{h,i+\theta}|_{H^1(\Omega)}^2 \leq k_i c_\Omega^2 \|f_{i+\theta}\|_{L^2(\Omega)}^2. \end{aligned}$$

Summation über i liefert die Behauptung. \square

Das θ -Schema ist demnach stabil für alle $\theta \in [0.5, 1]$. Fehler werden daher nicht exponentiell verstärkt. Für $\theta < 0.5$ ist das Verfahren nur stabil, wenn gilt $k_i \sim h^2$. Dies ist die sogenannte CFL-Bedingung (CFL steht für Courant, Friedrichs und Lewy) für parabolische Probleme. Im Falle $\theta > 0.5$ werden sogar die hochfrequenten Anteile in der diskreten Lösung exponentiell in der Zeit gedämpft. Ebenso werden lokale Störungen in den Daten $\mathbf{f}(t_i)$ und \mathbf{g} deutlich gedämpft.

Bemerkung Da das Crank-Nicolson-Verfahren das einfachste Verfahren von zweiter Ordnung ist, ist es unheimlich populär. Da Fehler nicht exponentiell gedämpft werden, kann es allerdings unphysikalische Oszillationen produzieren und sollte daher gemieden werden. Am besten wählt man daher $\theta = 1/2 + \xi$ mit einem $\xi > 0$, das die Größe der Dämpfung steuert. \triangle

11.3 Fehleranalysis

Wir wollen speziell das implizite Euler-Verfahren (also $\theta = 1$) betrachten, das wir leicht modifizieren, indem wir die rechte Seite entsprechend

$$\bar{f}(t_{i+1}) := \frac{1}{k_i} \int_{t_i}^{t_{i+1}} f(t) dt = f(t_{i+1}) + \mathcal{O}(k_i)$$

mitteln. Um den Fehler abzuschätzen, benutzen wir folgende, diskrete Seminorm

$$\|u\|_{h,\infty} := \max_{i=1}^M \|u(t_i)\|_{L^2(\Omega)}.$$

Für zeitlich diskrete Funktionen $u_h \in V_h$ ist dies sogar eine Norm.

Satz 11.2 Sei Ω ein konvexes Polygonebiet und $\{\mathcal{T}_h\}$ eine Familie quasi-uniformer Gitter, welche zeitlich konstant sind. Für die kontinuierliche Lösung der Wärmeleitungsgleichung gelte $u \in H^1(0, T) \otimes H^2(\Omega)$. Dann genügt das implizite Euler-Verfahren mit einer Ortsdiskretisierung durch \mathcal{P}_1 - beziehungsweise \mathcal{Q}_1 -Elemente der Fehlerabschätzung

$$\|u - u_h\|_{h,\infty} \leq c \left\{ \sqrt{T} h^2 k^{-1/2} \|\Delta u\|_{h,\infty} + \left(\sum_{i=0}^{M-1} k_i^2 \int_{t_i}^{t_{i+1}} \left| \frac{\partial u}{\partial t} \right|_{H^1(\Omega)}^2 dt \right)^{1/2} \right\},$$

vorausgesetzt es ist $\min_{i=1}^M \{k_i\} \geq k$.

Beweis. Mit $R_h u_i$ bezeichnen wir die Galerkin-Projektion von $u_i := u(t_i)$ auf den Raum V_h , die per Galerkin-Verfahren berechnet wird:

$$\text{suche } R_h u_i \in V_h, \text{ so dass } (\nabla R_h u_i, \nabla v_h)_{L^2(\Omega)} = (\nabla u_i, \nabla v_h)_{L^2(\Omega)} \quad \text{für alle } v_h \in V_h.$$

Wir spalten den Fehler $u_i - u_{h,i}$ auf in $\xi_i = (R_h - I)u_i$ und $\eta_i = u_{h,i} - R_h u_i \in V_h$. Nach Proposition 6.6 haben wir für ξ_i die Abschätzung

$$\|\xi_i\|_{L^2(\Omega)} = \|(R_h - I)u_i\|_{L^2(\Omega)} \leq ch^2 |u_i|_{H^2(\Omega)}. \quad (11.6)$$

Hieraus folgt unmittelbar

$$\|(R_h - I)u\|_{h,\infty} \leq ch^2 \|\Delta u\|_{h,\infty}.$$

Für den Fehleranteil η_{i+1} verwenden wir die folgende Identität:

$$\|\eta_{i+1}\|_{L^2(\Omega)}^2 - \|\eta_i\|_{L^2(\Omega)}^2 = 2(\eta_{i+1} - \eta_i, \eta_{i+1})_{L^2(\Omega)} - \|\eta_{i+1} - \eta_i\|_{L^2(\Omega)}^2. \quad (11.7)$$

Für den rechten Term gilt für beliebiges $v_h \in V_h$ unter Ausnutzung der Gleichung für die Galerkin-Projektion

$$\begin{aligned} & (\eta_{i+1} - \eta_i, \eta_{i+1})_{L^2(\Omega)} \\ &= k_i (f_{i+1}, \eta_{i+1})_{L^2(\Omega)} - k_i (\nabla u_{h,i+1}, \nabla \eta_{i+1})_{L^2(\Omega)} - (R_h u_{i+1} - R_h u_i, \eta_{i+1})_{L^2(\Omega)} \\ &= k_i (f_{i+1}, \eta_{i+1})_{L^2(\Omega)} - k_i (\nabla(\eta_{i+1} + u_{i+1}), \nabla \eta_{i+1})_{L^2(\Omega)} - (u_{i+1} - u_i, \eta_{i+1})_{L^2(\Omega)} \\ &\quad - \underbrace{(R_h u_{i+1} - u_{i+1} - R_h u_i + u_i, \eta_{i+1})_{L^2(\Omega)}}_{=(\xi_{i+1} - \xi_i, \eta_{i+1})_{L^2(\Omega)}}. \end{aligned}$$

Die ersten drei Terme lassen sich unter Verwendung der Wärmeleitungsgleichung beschränken gemäß

$$\begin{aligned} E_1 &:= k_i (f_{i+1}, \eta_{i+1})_{L^2(\Omega)} - k_i (\nabla(\eta_{i+1} + u_{i+1}), \nabla \eta_{i+1})_{L^2(\Omega)} - (u_{i+1} - u_i, \eta_{i+1})_{L^2(\Omega)} \\ &= \int_{t_i}^{t_{i+1}} \left(f - \frac{\partial u}{\partial t}, \eta_{i+1} \right)_{L^2(\Omega)} dt - k_i (\nabla u_{i+1}, \nabla \eta_{i+1})_{L^2(\Omega)} - k_i \|\nabla \eta_{i+1}\|_{L^2(\Omega)}^2 \\ &= \int_{t_i}^{t_{i+1}} (\nabla(u - u_{i+1}), \nabla \eta_{i+1})_{L^2(\Omega)} dt - k_i \|\nabla \eta_{i+1}\|_{L^2(\Omega)}^2. \end{aligned}$$

Mit dem Hauptsatz der Integralrechnung folgt

$$\int_x^y \{g(z) - g(x)\} dz = \int_x^y \int_x^z g'(t) dt dz.$$

Wegen

$$\{(t, z) : x \leq z \leq y, x \leq t \leq z\} = \{(t, z) : x \leq t \leq y, t \leq z \leq y\}$$

folgt daher die Identität

$$\int_x^y \{g(z) - g(x)\} dz = \int_x^y \int_t^y g'(t) dz dt = \int_x^y g'(t)(y - t) dt.$$

Dies eingesetzt liefert

$$\begin{aligned} E_1 &= \int_{t_i}^{t_{i+1}} (t_i - t) \left(\nabla \frac{\partial u}{\partial t}, \nabla \eta_{i+1} \right)_{L^2(\Omega)} dt - k_i \|\nabla \eta_{i+1}\|_{L^2(\Omega)}^2 \\ &\leq \int_{t_i}^{t_{i+1}} (t - t_i) \left| \frac{\partial u}{\partial t} \right|_{H^1(\Omega)} \|\nabla \eta_{i+1}\|_{L^2(\Omega)} dt - k_i \|\nabla \eta_{i+1}\|_{L^2(\Omega)}^2 \\ &\leq \left(\frac{k_i^2}{2} \int_{t_i}^{t_{i+1}} \left| \frac{\partial u}{\partial t} \right|_{H^1(\Omega)}^2 dt \right)^{1/2} \left(2 \int_{t_i}^{t_{i+1}} \|\nabla \eta_{i+1}\|_{L^2(\Omega)}^2 dt \right)^{1/2} - k_i \|\nabla \eta_{i+1}\|_{L^2(\Omega)}^2. \end{aligned}$$

Mit der Ungleichung vom arithmetischen und geometrischen Mittel, $\sqrt{ab} \leq (a+b)/2$ folgt schließlich

$$E_1 \leq \frac{k_i^2}{4} \int_{t_i}^{t_{i+1}} \left| \frac{\partial u}{\partial t} \right|_{H^1(\Omega)}^2 dt.$$

Weiter gilt

$$\begin{aligned} E_2 &:= (\xi_{i+1} - \xi_i, \eta_{i+1})_{L^2(\Omega)} \\ &= (\xi_{i+1}, \eta_{i+1})_{L^2(\Omega)} - (\xi_i, \eta_i)_{L^2(\Omega)} + \underbrace{(\xi_i, \eta_i - \eta_{i+1})_{L^2(\Omega)}}_{\leq \|\eta_i - \eta_{i+1}\|_{L^2(\Omega)} \|\xi_i\|_{L^2(\Omega)}} \\ &\leq (\xi_{i+1}, \eta_{i+1})_{L^2(\Omega)} - (\xi_i, \eta_i)_{L^2(\Omega)} + \frac{1}{2} \|\eta_i - \eta_{i+1}\|_{L^2(\Omega)}^2 + ch^4 |u_i|_{H^2(\Omega)}^2, \end{aligned}$$

wobei wir im letzten Schritt erst (11.6) und danach die Ungleichung vom arithmetischen und geometrischen Mittel verwendet haben.

Die Gleichung $(\eta_{i+1} - \eta_i, \eta_{i+1})_{L^2(\Omega)} = E_1 + E_2$ eingesetzt in (11.7) ergibt

$$\begin{aligned} \|\eta_{i+1}\|_{L^2(\Omega)}^2 - \|\eta_i\|_{L^2(\Omega)}^2 &= 2E_1 + 2E_2 - \|\eta_{i+1} - \eta_i\|_{L^2(\Omega)}^2 \\ &\leq 2(\xi_{i+1}, \eta_{i+1})_{L^2(\Omega)} - 2(\xi_i, \eta_i)_{L^2(\Omega)} + \frac{k_i^2}{2} \int_{t_i}^{t_{i+1}} \left| \frac{\partial u}{\partial t} \right|_{H^1(\Omega)}^2 dt + ch^4 |u_i|_{H^2(\Omega)}^2. \end{aligned}$$

Aufsummation über alle i liefert dann

$$\begin{aligned} \|\eta_M\|_{L^2(\Omega)}^2 &\leq \|\eta_0\|_{L^2(\Omega)}^2 + 2 \underbrace{(\xi_M, \eta_M)_{L^2(\Omega)}}_{\leq (\sqrt{2}\|\xi_M\|_{L^2(\Omega)})(\|\eta_M\|_{L^2(\Omega)}/\sqrt{2})} - 2 \underbrace{(\xi_0, \eta_0)_{L^2(\Omega)}}_{\leq (\sqrt{2}\|\xi_0\|_{L^2(\Omega)})(\|\eta_0\|_{L^2(\Omega)}/\sqrt{2})} \\ &\quad + \sum_{i=0}^{M-1} \frac{k_i^2}{2} \int_{t_i}^{t_{i+1}} \left| \frac{\partial u}{\partial t} \right|_{H^1(\Omega)}^2 dt + ch^4 \sum_{i=0}^{M-1} |u_i|_{H^2(\Omega)}^2 \\ &\leq \|\eta_0\|_{L^2(\Omega)}^2 + 2\|\xi_M\|_{L^2(\Omega)}^2 + \frac{1}{2}\|\eta_M\|_{L^2(\Omega)}^2 + 2\|\xi_0\|_{L^2(\Omega)}^2 + \frac{1}{2}\|\eta_0\|_{L^2(\Omega)}^2 \\ &\quad + \sum_{i=0}^{M-1} \frac{k_i^2}{2} \int_{t_i}^{t_{i+1}} \left| \frac{\partial u}{\partial t} \right|_{H^1(\Omega)}^2 dt + ch^4 \sum_{i=0}^{M-1} k_i \underbrace{|k_i^{-1/2} u_i|_{H^2(\Omega)}^2}_{\leq \|k^{-1/2} \Delta u\|_{h,\infty}^2} \\ &\leq \frac{3}{2}\|\eta_0\|_{L^2(\Omega)}^2 + 2\|\xi_M\|_{L^2(\Omega)}^2 + \frac{1}{2}\|\eta_M\|_{L^2(\Omega)}^2 + 2\|\xi_0\|_{L^2(\Omega)}^2 \\ &\quad + \sum_{i=0}^{M-1} \frac{k_i^2}{2} \int_{t_i}^{t_{i+1}} \left| \frac{\partial u}{\partial t} \right|_{H^1(\Omega)}^2 dt + ch^4 T \|k^{-1/2} \Delta u\|_{h,\infty}^2. \end{aligned}$$

Projizieren wir den Startwert $u_0 = g$ durch eine geeignete Projektion $P_h : H_0^1(\Omega) \rightarrow V_h$, dann gilt

$$\|\eta_0\|_{L^2(\Omega)} = \|P_h u_0 - R_h u_0\|_{L^2(\Omega)} \leq \|(I - P_h)u_0\|_{L^2(\Omega)} + \|(I - R_h)u_0\|_{L^2(\Omega)} \leq ch^2 |u_0|_{H^2(\Omega)}.$$

Wir erhalten damit

$$\begin{aligned} \frac{1}{2}\|\eta_M\|_{L^2(\Omega)}^2 &\leq 2\|\xi_M\|_{L^2(\Omega)}^2 + 2\|\xi_0\|_{L^2(\Omega)}^2 + ch^4(T+1) \|k^{-1/2} \Delta u\|_{h,\infty}^2 \\ &\quad + \sum_{i=0}^{M-1} \frac{k_i^2}{2} \int_{t_i}^{t_{i+1}} \left| \frac{\partial u}{\partial t} \right|_{H^1(\Omega)}^2 dt. \end{aligned}$$

Wegen $\|u - u_h\|_{h,\infty} \leq \|\xi\|_{h,\infty} + \|\eta\|_{h,\infty}$ ergibt sich endlich die Behauptung. \square

Bemerkung Im Fall von uniformen Zeitschritten $k \leq ck_i$ besagt die Fehlerabschätzung aus Satz 11.2, dass gilt

$$\|u - u_h\|_{h,\infty} = \mathcal{O}(h^2 k^{-1/2} + k).$$

Aufgrund des Faktors $k^{-1/2}$ ist dies nicht optimal. Unter der Bedingung $h \leq ck^{3/4}$ ergibt sich aber die zeitoptimale Konvergenzordnung $\mathcal{O}(k)$. Beim Crank-Nicolson-Verfahren ist der Diskretisierungsfehler hingegen $\mathcal{O}(k^2 + h^2)$, das entspricht quadratischer Konvergenz $\mathcal{O}(k^2)$, wenn $h \sim k$ gewählt wird. \triangle

12. Lineare Elastizität*

12.1 Herleitung

In der Elastizitätstheorie werden Verformungen und Spannungen von Körpern betrachtet. Ausgegangen wird von einem elastischen Körper, der mit einem beschränkten und zusammenhängenden Gebiet $\Omega \subset \mathbb{R}^d$ identifiziert werde. Wirken auf diesen Körper *Oberflächenkräfte* $\mathbf{g} : \Gamma_N \rightarrow \mathbb{R}^d$ und *Volumenkräfte* $\mathbf{f} : \Omega \rightarrow \mathbb{R}^d$, so erzeugen diese einen *Spannungszustand* im Innern des Körpers. Dieser wird beschrieben durch den *Spannungstensor*

$$\boldsymbol{\sigma} = [\sigma_{i,j}] : \Omega \rightarrow \mathbb{R}_{\text{sym}}^{d \times d}.$$

Die Spannung erzeugt eine *Deformation* $\boldsymbol{\varphi} : \Omega \rightarrow \mathbb{R}^d$ des Körpers. Durch diese Deformation wird jeder Massenpunkt $\mathbf{x} \in \Omega$ auf $\boldsymbol{\varphi}(\mathbf{x}) \in \mathbb{R}^d$ abgebildet. Die Abbildung $\boldsymbol{\varphi}$ wird im folgenden stets als genügend glatt und lokal injektiv vorausgesetzt. Daher muss für den Deformationsgradienten $\nabla \boldsymbol{\varphi}$ gelten:

$$\det(\nabla \boldsymbol{\varphi}) = \det \begin{bmatrix} \frac{\partial \varphi_1}{\partial x_1} & \frac{\partial \varphi_1}{\partial x_2} & \dots & \frac{\partial \varphi_1}{\partial x_d} \\ \frac{\partial \varphi_2}{\partial x_1} & \frac{\partial \varphi_2}{\partial x_2} & \dots & \frac{\partial \varphi_2}{\partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \varphi_d}{\partial x_1} & \frac{\partial \varphi_d}{\partial x_2} & \dots & \frac{\partial \varphi_d}{\partial x_d} \end{bmatrix} > 0.$$

Wichtig sind die durch die Deformation $\boldsymbol{\varphi}$ erzeugten Änderungen des Linienelements. Es ist

$$\boldsymbol{\varphi}(\mathbf{x} + \mathbf{y}) = \boldsymbol{\varphi}(\mathbf{x}) + \nabla \boldsymbol{\varphi}(\mathbf{x})\mathbf{y} + \mathcal{O}(\|\mathbf{y}\|^2).$$

Also gilt für den Euklidischen Abstand

$$\|\boldsymbol{\varphi}(\mathbf{x} + \mathbf{y}) - \boldsymbol{\varphi}(\mathbf{x})\|^2 = \|\nabla \boldsymbol{\varphi}(\mathbf{x})\mathbf{y}\|^2 + \mathcal{O}(\|\mathbf{y}\|^3) = \mathbf{y}^T (\nabla \boldsymbol{\varphi}(\mathbf{x}))^T \nabla \boldsymbol{\varphi}(\mathbf{x}) \mathbf{y} + \mathcal{O}(\|\mathbf{y}\|^3).$$

Für die lokale Änderung von Längen ist also der *Cauchy-Greensche Verzerrungstensor* ausschlaggebend:

$$\mathbf{C} = (\nabla \boldsymbol{\varphi}(\mathbf{x}))^T \nabla \boldsymbol{\varphi}(\mathbf{x}).$$

Die durch

$$\mathbf{E} = \frac{1}{2}(\mathbf{C} - \mathbf{I}) \tag{12.1}$$

definierte Abweichung von der Identität bezeichnet man als *Verzerrung*.

In der Praxis ist die *Verschiebung* $\mathbf{u} : \Omega \rightarrow \mathbb{R}^d$ wichtiger als die Deformation. Sie ist gegeben durch

$$\mathbf{u}(\mathbf{x}) = \boldsymbol{\varphi}(\mathbf{x}) - \mathbf{x}, \quad \mathbf{x} \in \Omega, \tag{12.2}$$

Durch Einsetzen von (12.2) in (12.1) erhält man

$$\mathbf{E} = \frac{1}{2}[\nabla \mathbf{u} + (\nabla \mathbf{u})^T] + \frac{1}{2}(\nabla \mathbf{u})^T \nabla \mathbf{u}.$$

Bei kleinen Verzerrungen kann der letzte Term vernachlässigt werden und so entsteht der linearisierte *Verzerrungstensor*

$$\boldsymbol{\varepsilon} = [\varepsilon_{i,j}] : \Omega \rightarrow \mathbb{R}_{\text{sym}}^{d \times d},$$

welcher gegeben ist durch

$$\boldsymbol{\varepsilon}(\mathbf{u}) := \frac{1}{2}[\nabla \mathbf{u} + (\nabla \mathbf{u})^T]. \quad (12.3)$$

Gemäß des Hookeschen Gesetzes wird der Zusammenhang zwischen den Verzerrungen $\boldsymbol{\varepsilon}$ und den Spannungen $\boldsymbol{\sigma}$ durch den (konstanten) *Elastizitätstensor* $\mathbf{A} : \mathbb{R}_{\text{sym}}^{d \times d} \rightarrow \mathbb{R}_{\text{sym}}^{d \times d}$ beschrieben:

$$\boldsymbol{\sigma} = \mathbf{A}\boldsymbol{\varepsilon} := \frac{E\nu}{(1+\nu)(1-2\nu)} \text{tr}(\boldsymbol{\varepsilon})\mathbf{I} + \frac{E}{1+\nu}\boldsymbol{\varepsilon}.$$

Hier ist $E > 0$ das *Elastizitätsmodul* und $0 < \nu < 1/2$ die *Poisson-Zahl*.

Beispiel 12.1 Aufgrund der Symmetrie des Verzerrungs- und des Spannungstensors lautet in Matrix-Schreibweise der Zusammenhang zwischen beiden in zwei Raumdimensionen

$$\begin{bmatrix} \sigma_{1,1} \\ \sigma_{2,2} \\ \sigma_{1,2} \end{bmatrix} = \frac{E}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1-\nu & \nu & & \\ \nu & 1-\nu & & \\ & & & 1-2\nu \end{bmatrix} \begin{bmatrix} \varepsilon_{1,1} \\ \varepsilon_{2,2} \\ \varepsilon_{1,2} \end{bmatrix}$$

und in drei Raumdimensionen

$$\begin{bmatrix} \sigma_{1,1} \\ \sigma_{2,2} \\ \sigma_{3,3} \\ \sigma_{1,2} \\ \sigma_{1,3} \\ \sigma_{2,3} \end{bmatrix} = \frac{E}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1-\nu & \nu & \nu & & & \\ \nu & 1-\nu & \nu & & & \\ \nu & \nu & 1-\nu & & & \\ & & & 1-2\nu & & \\ & & & & 1-2\nu & \\ & & & & & 1-2\nu \end{bmatrix} \begin{bmatrix} \varepsilon_{1,1} \\ \varepsilon_{2,2} \\ \varepsilon_{3,3} \\ \varepsilon_{1,2} \\ \varepsilon_{1,3} \\ \varepsilon_{2,3} \end{bmatrix}.$$

In zwei Raumdimensionen sieht man sofort mit Hilfe von Gerschgorin-Kreisen

$$|\lambda - (1 - \nu)| \leq \nu \quad \text{oder} \quad \lambda = 1 - 2\nu$$

ein, wobei wir den Faktor vor der Kopplungsmatrix weggelassen haben, dass diese für $0 < \nu < 1/2$ positiv definit ist. In drei Raumdimension zeigt man dies, indem man die Gerschgorin-Kreise für deren Inverse betrachtet:

$$\frac{1}{E} \begin{bmatrix} 1 & -\nu & -\nu & & & \\ -\nu & 1 & -\nu & & & \\ -\nu & -\nu & 1 & & & \\ & & & 1+\nu & & \\ & & & & 1+\nu & \\ & & & & & 1+\nu \end{bmatrix}.$$

Denn für die Inverse gilt modulo Skalierung

$$|\lambda - 1| \leq 2\nu \quad \text{oder} \quad \lambda = 1 + \nu.$$

Im Gleichgewichtszustand gilt im Körper

$$-\operatorname{div}(\boldsymbol{\sigma}(\mathbf{u})) = \mathbf{f} \quad \text{in } \Omega, \quad (12.4)$$

wobei die Divergenz einer matrixwertigen Funktion $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_d]^T$ gegeben ist durch

$$\operatorname{div} \mathbf{S} = [\operatorname{div} \mathbf{s}_1, \operatorname{div} \mathbf{s}_2, \dots, \operatorname{div} \mathbf{s}_d]^T.$$

Zusätzlich zu (12.4) gelten Dirichlet- bzw. Verschiebungsrandbedingungen

$$\mathbf{u} = \mathbf{0} \quad \text{auf } \Gamma_D \subset \partial\Omega \quad (12.5)$$

oder Neumann- bzw. Spannungsrandbedingungen

$$\boldsymbol{\sigma}(\mathbf{u})\mathbf{n} = \mathbf{g} \quad \text{auf } \Gamma_N = \partial\Omega \setminus \bar{\Gamma}_D. \quad (12.6)$$

Wir wollen abschließend noch eine äquivalente Schreibweise der linken Seite von (12.4) herleiten. Mit Hilfe der *Lamé-Konstanten*

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}, \quad \mu = \frac{E}{2(1+\nu)} \quad (12.7)$$

und (12.3) folgt der direkte Zusammenhang zwischen den Verschiebungen \mathbf{u} und dem Spannungstensor $\boldsymbol{\sigma}$:

$$\boldsymbol{\sigma}(\mathbf{u}) = \lambda \operatorname{div}(\mathbf{u})\mathbf{I} + \mu [\nabla\mathbf{u} + (\nabla\mathbf{u})^T]. \quad (12.8)$$

Demnach haben wir die Beziehung

$$\operatorname{div}(\boldsymbol{\sigma}(\mathbf{u})) = \mu\Delta\mathbf{u} + (\lambda + \mu)\nabla(\operatorname{div} \mathbf{u})$$

mit dem Laplace-Operator für vektorwertige Funktionen:

$$\Delta\mathbf{u} = \frac{\partial}{\partial x_1^2}\mathbf{u} + \frac{\partial}{\partial x_2^2}\mathbf{u} + \dots + \frac{\partial}{\partial x_d^2}\mathbf{u}.$$

12.2 Variationsformulierung

Um die Variationsformulierung zu erhalten, müssen wir (12.4) mit einer Testfunktion $\mathbf{v} : \Omega \rightarrow \mathbb{R}^d$ multiplizieren und über das Gebiet Ω integrieren:

$$-\int_{\Omega} \langle \operatorname{div}(\boldsymbol{\sigma}(\mathbf{u})), \mathbf{v} \rangle \, d\mathbf{x} = \int_{\Omega} \langle \mathbf{f}, \mathbf{v} \rangle \, d\mathbf{x}. \quad (12.9)$$

Die partielle Integration der linken Seite via komponentenweisem Anwenden des Gaußschen Integralsatzes liefert

$$-\int_{\Omega} \langle \operatorname{div}(\boldsymbol{\sigma}(\mathbf{u})), \mathbf{v} \rangle \, d\mathbf{x} = \int_{\Omega} \boldsymbol{\sigma}(\mathbf{u}) : \nabla\mathbf{v} \, d\mathbf{x} - \int_{\partial\Omega} \langle \boldsymbol{\sigma}(\mathbf{u})\mathbf{n}, \mathbf{v} \rangle \, d\sigma. \quad (12.10)$$

Hierbei ist für zwei Matrizen $\mathbf{B} = [b_{i,j}]$, $\mathbf{D} = [d_{i,j}] \in \mathbb{R}^{d \times d}$ das *Frobenius-Skalarprodukt* definiert durch

$$\mathbf{B} : \mathbf{D} = \sum_{i,j=1}^d b_{i,j}d_{i,j}.$$

In Anbetracht von (12.5) erfüllen alle kinematisch zulässigen Verschiebungen die Bedingung $\mathbf{v} = \mathbf{0}$ auf Γ_D . Demnach lautet der Energierraum

$$V := \left\{ \mathbf{v} \in [H^1(\Omega)]^d : \mathbf{v}|_{\Gamma_D} = \mathbf{0} \right\}.$$

Definieren wir also die Bilinearform $a : V \times V \rightarrow \mathbb{R}$ durch

$$a(\mathbf{u}, \mathbf{v}) := \int_{\Omega} \boldsymbol{\sigma}(\mathbf{u}) : \nabla \mathbf{v} \, d\mathbf{x}$$

und die Linearform $\ell : V \rightarrow \mathbb{R}$ durch

$$\ell(\mathbf{v}) := \int_{\Omega} \langle \mathbf{f}, \mathbf{v} \rangle \, d\mathbf{x} + \int_{\Gamma_N} \langle \mathbf{g}, \mathbf{v} \rangle \, d\sigma,$$

so erhalten wir wegen (12.6) offensichtlich die Variationsformulierung:

$$\text{suche } \mathbf{u} \in V, \text{ so dass } a(\mathbf{u}, \mathbf{v}) = \ell(\mathbf{v}) \quad \text{für alle } \mathbf{v} \in V. \quad (12.11)$$

Hierbei können wir wegen der Symmetrie von $\boldsymbol{\sigma}(\mathbf{u})$ und (12.8) die Bilinearform $a(\cdot, \cdot)$ auch umschreiben gemäß

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) &= \int_{\Omega} \boldsymbol{\sigma}(\mathbf{u}) : \frac{1}{2} [\nabla \mathbf{v} + (\nabla \mathbf{v})^T] \, d\mathbf{x} = \int_{\Omega} \mathbf{A} \boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}) \, d\mathbf{x} \\ &= 2\mu \int_{\Omega} \boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}) \, d\mathbf{x} + \lambda \int_{\Omega} \operatorname{div}(\mathbf{u}) \operatorname{div}(\mathbf{v}) \, d\mathbf{x}. \end{aligned} \quad (12.12)$$

Hieraus ist sofort ersichtlich, dass die Bilinearform $a(\cdot, \cdot)$ symmetrisch ist. Mit Hilfe der Cauchy-Schwarzschen Ungleichung weist man ausserdem leicht

$$\begin{aligned} |a(\mathbf{u}, \mathbf{v})| &\leq \frac{1}{2} \mu \underbrace{\|\nabla \mathbf{u} + (\nabla \mathbf{u})^T\|_{L^2(\Omega)}}_{\leq 2\|\nabla \mathbf{u}\|_{L^2(\Omega)}} \underbrace{\|\nabla \mathbf{v} + (\nabla \mathbf{v})^T\|_{L^2(\Omega)}}_{\leq 2\|\nabla \mathbf{v}\|_{L^2(\Omega)}} + d\lambda \|\nabla \mathbf{u}\|_{L^2(\Omega)} \|\nabla \mathbf{v}\|_{L^2(\Omega)} \\ &\leq (2\mu + d\lambda) |\mathbf{u}|_{H^1(\Omega)} |\mathbf{v}|_{H^1(\Omega)} \end{aligned}$$

nach, das heißt die Bilinearform $a(\cdot, \cdot)$ ist stetig. Auch das Funktional $\ell(\cdot)$ ist stetig:

$$|\ell(\mathbf{v})| \leq \|\mathbf{f}\|_{L^2(\Omega)} \|\mathbf{v}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\Gamma_N)} \|\mathbf{v}\|_{L^2(\Gamma_N)} \leq c \|\mathbf{v}\|_{H^1(\Omega)}.$$

Die Existenz und die Eindeutigkeit der Lösung der Variationsformulierung (12.11) folgt deshalb aus dem Satz von Lax-Milgram (Satz 3.11), vorausgesetzt $a(\cdot, \cdot)$ ist elliptisch.

12.3 Elliptizitätsabschätzung

Um die Elliptizität der Bilinearform $a(\cdot, \cdot)$ im Fall eines Dirichlet-Randwertproblems nachzuweisen, benötigen wir folgendes Lemma.

Lemma 12.2 (erste Kornsche Ungleichung) Für alle $\mathbf{v} \in [H_0^1(\Omega)]^d$ gilt

$$\int_{\Omega} \boldsymbol{\varepsilon}(\mathbf{v}) : \boldsymbol{\varepsilon}(\mathbf{v}) \, d\mathbf{x} \geq \frac{1}{2} |\mathbf{v}|_{H^1(\Omega)}^2.$$

Beweis. Für $\varphi \in [C_0^\infty(\Omega)]^d$ folgt

$$\begin{aligned} \int_{\Omega} \varepsilon(\varphi) : \varepsilon(\varphi) \, dx &= \frac{1}{4} \int_{\Omega} [\nabla \varphi + (\nabla \varphi)^T] : [\nabla \varphi + (\nabla \varphi)^T] \, dx \\ &= \frac{1}{2} \int_{\Omega} \nabla \varphi : \nabla \varphi \, dx + \frac{1}{2} \int_{\Omega} \nabla \varphi : (\nabla \varphi)^T \, dx. \end{aligned}$$

Zweimaliges partielles integrieren ergibt

$$\begin{aligned} \int_{\Omega} \nabla \varphi : (\nabla \varphi)^T \, dx &= \sum_{i,j=1}^d \int_{\Omega} \frac{\partial \varphi_i}{\partial x_j} \frac{\partial \varphi_j}{\partial x_i} \, dx = \sum_{i,j=1}^d \int_{\Omega} \frac{\partial \varphi_i}{\partial x_i} \frac{\partial \varphi_j}{\partial x_j} \, dx = \int_{\Omega} \left[\sum_{i=1}^d \frac{\partial \varphi_i}{\partial x_i} \right]^2 \, dx \\ &= \int_{\Omega} (\operatorname{div}(\varphi))^2 \, dx \geq 0. \end{aligned}$$

Daher erhalten wir

$$\int_{\Omega} \varepsilon(\varphi) : \varepsilon(\varphi) \, dx \geq \frac{1}{2} \int_{\Omega} \nabla \varphi : \nabla \varphi \, dx = \frac{1}{2} |\varphi|_{H^1(\Omega)}^2.$$

Diese Abschätzung impliziert die Behauptung aufgrund der Dichtheit von $[C_0^\infty(\Omega)]^d$ in $[H_0^1(\Omega)]^d$. \square

Aus (12.12) folgern wir

$$a(\mathbf{v}, \mathbf{v}) \geq 2\mu \int_{\Omega} \varepsilon(\mathbf{v}) : \varepsilon(\mathbf{v}) \, dx \geq \mu |\mathbf{v}|_{H^1(\Omega)}^2.$$

Aufgrund der Poincaré-Friedrichsschen Ungleichung (Satz 3.5), welche auf jede Komponente von $\mathbf{v} \in [H_0^1(\Omega)]^d$ angewendet werden kann, folgt nun offensichtlich die Elliptizität der Bilinearform

$$a(\mathbf{v}, \mathbf{v}) \geq c_E \|\mathbf{v}\|_{H^1(\Omega)}^2 \quad \text{für alle } \mathbf{v} \in [H_0^1(\Omega)]^d. \quad (12.13)$$

Bemerkung Die Elliptizitätsabschätzung (12.13) behält auch dann ihre Gültigkeit, wenn für \mathbf{v} nur auf einem Teil des Rands Nullrandbedingungen vorgeschrieben sind, also wenn $\mathbf{v} \in V$ und $|\Gamma_D| \neq 0$ ist. \triangle

12.4 Starrkörperbewegungen

Im Fall eines reinen Neumann-Randwertproblems ist der Körper nirgendwo fixiert. Es ist anschaulich klar, dass dann *Starrkörperbewegungen*, also Translationen und orthogonale Transformationen, den Spannungszustand eines Körpers nicht ändern. Diese Starrkörperbewegungen werden gerade durch den Raum

$$\mathcal{R} := \{\mathbf{v}(\mathbf{x}) = \mathbf{B}\mathbf{x} + \mathbf{d} : \mathbf{B} = -\mathbf{B}^T, \mathbf{B} \in \mathbb{R}^{d \times d}, \mathbf{d} \in \mathbb{R}^d\} \subset [L^2(\Omega)]^d \quad (12.14)$$

charakterisiert, denn es gilt:

Lemma 12.3 Es gilt $\boldsymbol{\varepsilon}(\mathbf{v}) = \mathbf{0}$ genau dann, wenn $\mathbf{v} \in \mathcal{R}$.

Beweis. Man rechnet leicht nach, dass $\boldsymbol{\varepsilon}(\mathbf{v}) = \mathbf{0}$ für $\mathbf{v} \in \mathcal{R}$. Wir müssen also nur die Rückrichtung zeigen. Dazu beachte man die Identität

$$\frac{\partial v_k}{\partial x_i \partial x_j} = \frac{\partial \varepsilon_{j,k}(\mathbf{v})}{\partial x_i} + \frac{\partial \varepsilon_{i,k}(\mathbf{v})}{\partial x_j} - \frac{\partial \varepsilon_{i,j}(\mathbf{v})}{\partial x_k}.$$

Im Fall $\boldsymbol{\varepsilon}(\mathbf{v}) = \mathbf{0}$ folgt hieraus $\partial v_k / (\partial x_i \partial x_j) = 0$. Deshalb ist $\mathbf{v} = \mathbf{B}\mathbf{x} + \mathbf{d}$ eine lineare Funktion. Wegen $\mathbf{0} = 2\boldsymbol{\varepsilon}(\mathbf{v}) = \nabla \mathbf{v} + (\nabla \mathbf{v})^T = \mathbf{B} + \mathbf{B}^T$ ergibt sich $\mathbf{B} = -\mathbf{B}^T$ und $\mathbf{d} \in \mathbb{R}^d$. Dies impliziert $\mathbf{v} \in \mathcal{R}$. \square

Bemerkung Die durch den Vektor \mathbf{d} charakterisierten Verschiebungen heißen *Starrkörpertranslationen*. Die durch die Matrix \mathbf{B} charakterisierten orthogonalen Drehungen heißen *Starrkörperrotationen*. Wie man sich leicht überlegt, ist die Dimension von \mathcal{R} gerade $d(d+1)/2$. Dies bedeutet, es gilt in $d = 2$ Raumdimensionen $\dim \mathcal{R} = 3$ und

$$\mathcal{R} = \text{span} \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} x_2 \\ -x_1 \end{bmatrix} \right\},$$

während in $d = 3$ Raumdimensionen $\dim \mathcal{R} = 6$ ist und

$$\mathcal{R} = \text{span} \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} x_2 \\ -x_1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ x_3 \\ -x_2 \end{bmatrix}, \begin{bmatrix} -x_3 \\ 0 \\ x_1 \end{bmatrix} \right\}.$$

\triangle

Beschränkt man sich auf Funktionen $\mathbf{v} \in [H^1(\Omega)]^d$, die orthogonal zu den Starrkörperbewegungen sind, das heißt, für die $\mathbf{v} \perp \mathcal{R}$ gilt, so erhält man die folgende Elliptizitätsabschätzung.

Lemma 12.4 (zweite Kornsche Ungleichung) Für alle $\mathbf{v} \in [H^1(\Omega)]^d \cap \mathcal{R}^\perp$ gilt

$$\int_{\Omega} \boldsymbol{\varepsilon}(\mathbf{v}) : \boldsymbol{\varepsilon}(\mathbf{v}) \, d\mathbf{x} \geq c \|\mathbf{v}\|_{H^1(\Omega)}^2.$$

Beweis. Den Beweis dieser Abschätzung findet der geneigte Leser beispielsweise in O.A. Oleinik, A.S. Shamaev and G.A. Yosifian “Mathematical Problems in Elasticity and Homogenization”. \square

Weil insbesondere $\text{div}(\mathbf{v}) = 0$ gilt für alle $\mathbf{v} \in \mathcal{R}$, schließen wir aus Lemma 12.3, dass

$$a(\mathbf{v}, \mathbf{v}) = 0 \quad \text{für alle } \mathbf{v} \in \mathcal{R}. \quad (12.15)$$

Deshalb benötigen wir im Fall eines reinen Neumann-Randwertproblems eine Kompatibilitätsbedingung an die rechte Seite:

$$\ell(\mathbf{v}) = \int_{\Omega} \langle \mathbf{f}, \mathbf{v} \rangle \, d\mathbf{x} + \int_{\Gamma_N} \langle \mathbf{g}, \mathbf{v} \rangle \, d\sigma = 0 \quad \text{für alle } \mathbf{v} \in \mathcal{R}. \quad (12.16)$$

Definieren wir nun

$$V := \{\mathbf{v} \in [H^1(\Omega)]^d : \mathbf{v} \perp \mathcal{R}\},$$

so folgt aus der zweiten Kornschen Ungleichung die V -Elliptizität der Bilinearform $a(\cdot, \cdot)$ auch im Fall eines reinen Neumann-Randwertproblems:

$$a(\mathbf{v}, \mathbf{v}) \geq 2\mu \int_{\Omega} \boldsymbol{\varepsilon}(\mathbf{v}) : \boldsymbol{\varepsilon}(\mathbf{v}) \, d\mathbf{x} \geq 2\mu c \|\mathbf{v}\|_{H^1(\Omega)}^2 \quad \text{für alle } \mathbf{v} \in V.$$

Deshalb besitzt das Variationsproblem (12.11) unter der Voraussetzung (12.16) eine eindeutige Lösung in V .

12.5 Lagrange-Multiplikatoren

Im Fall eines reinen Neumann-Randwertproblems ist die Variationsformulierung nur dann eindeutig lösbar, falls die Lösung orthogonal zu den Starrkörperbewegungen gesucht wird. Diese Nebenbedingung kann man mittels Lagrange-Parametern erzwingen. Dazu sei $\{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_R\}$ mit $R = \dim \mathcal{R}$ eine Basis von \mathcal{R} . Wir suchen nun $\mathbf{u} \in [H^1(\Omega)]^d$ und $\lambda_1, \lambda_2, \dots, \lambda_R \in \mathbb{R}$, so dass

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) + \sum_{k=1}^R \lambda_k \langle \mathbf{r}_k, \mathbf{v} \rangle &= \ell(\mathbf{v}) \quad \text{für alle } \mathbf{v} \in [H^1(\Omega)]^d, \\ \langle \mathbf{r}_k, \mathbf{u} \rangle &= 0 \quad \text{für alle } k = 1, 2, \dots, R. \end{aligned} \quad (12.17)$$

Wählen wir als Testfunktionen gerade die \mathbf{r}_k , $k = 1, 2, \dots, R$, so folgt wegen (12.15), (12.16) und der linearen Unabhängigkeit der Funktionen $\{\mathbf{r}_k\}$ sofort $\lambda_k = 0$ für alle $k = 1, 2, \dots, R$. Folglich können wir $\lambda_k = 0$ in der ersten Gleichung aus (12.17) durch die entsprechende zweite Gleichung aus (12.17) ersetzen und erhalten

$$a(\mathbf{u}, \mathbf{v}) + \sum_{k=1}^R \langle \mathbf{r}_k, \mathbf{u} \rangle \langle \mathbf{r}_k, \mathbf{v} \rangle = \ell(\mathbf{v}) \quad \text{für alle } \mathbf{v} \in [H^1(\Omega)]^d. \quad (12.18)$$

Die modifizierte Bilinearform

$$a_{\text{mod}}(\mathbf{u}, \mathbf{v}) := a(\mathbf{u}, \mathbf{v}) + \sum_{k=1}^R \langle \mathbf{r}_k, \mathbf{u} \rangle \langle \mathbf{r}_k, \mathbf{v} \rangle$$

ist nun elliptisch auf ganz $[H^1(\Omega)]^d$, da in Lemma 12.5 gezeigt wird, dass durch

$$\|\mathbf{v}\|^2 := |\mathbf{v}|_{H^1(\Omega)}^2 + \sum_{k=1}^R |\langle \mathbf{r}_k, \mathbf{v} \rangle|^2$$

eine zur $[H^1(\Omega)]^d$ -Norm äquivalente Norm definiert wird. Folglich ist das Variationsproblem (12.18) eindeutig lösbar in $[H^1(\Omega)]^d$.

Lemma 12.5 Es gilt

$$\underline{c} \|\mathbf{v}\|_{H^1(\Omega)} \leq \|\mathbf{v}\| \leq \bar{c} \|\mathbf{v}\|_{H^1(\Omega)}$$

für alle $\mathbf{v} \in [H^1(\Omega)]^d$.

Beweis. Wegen

$$\sum_{k=1}^R |\langle \mathbf{r}_k, \mathbf{v} \rangle|^2 \leq \sum_{k=1}^R \|\mathbf{r}_k\|_{L^2(\Omega)}^2 \|\mathbf{v}\|_{L^2(\Omega)}^2 \leq c \|\mathbf{v}\|_{H^1(\Omega)}^2,$$

ist die Abschätzung nach oben trivial. Die Abschätzung nach unten wird hingegen indirekt nachgewiesen.

Angenommen, die Abschätzung nach unten gilt nicht. Dann gibt es zu jedem $n \in \mathbb{N}$ eine Funktion $\mathbf{v}_n \in [H^1(\Omega)]^d$, so dass

$$\|\mathbf{v}_n\|_{H^1(\Omega)} = 1 \quad \text{und} \quad \|\mathbf{v}_n\| < \frac{1}{n}$$

gilt. Demnach muss

$$\lim_{n \rightarrow \infty} \sum_{k=1}^R |\langle \mathbf{r}_k, \mathbf{v}_n \rangle|^2 = 0 \quad \text{und} \quad \lim_{n \rightarrow \infty} |\mathbf{v}_n|_{H^1(\Omega)} = 0$$

gelten. Weil nach dem Rellichschen Auswahlssatz 5.9 die Einbettung $[H^1(\Omega)]^d \hookrightarrow [L^2(\Omega)]^d$ kompakt ist, existiert eine Teilfolge von $\{\mathbf{v}_n\}$, welche in $[L^2(\Omega)]^d$ konvergiert. Ohne Beschränkung der Allgemeinheit können wir annehmen, dass es sich dabei um die ganze Folge handelt. Aufgrund von

$$\|\mathbf{v}_n - \mathbf{v}_m\|_{H^1(\Omega)}^2 = \|\mathbf{v}_n - \mathbf{v}_m\|_{L^2(\Omega)}^2 + |\mathbf{v}_n - \mathbf{v}_m|_{H^1(\Omega)}^2$$

schließen wir, dass $\{\mathbf{v}_n\}$ eine Cauchy-Folge in $[H^1(\Omega)]^d$ ist. Wegen der Vollständigkeit des Raums existiert ein Grenzelement $\mathbf{v}^* \in [H^1(\Omega)]^d$ mit $\|\mathbf{v}_n - \mathbf{v}^*\|_{H^1(\Omega)} \rightarrow 0$ für $n \rightarrow \infty$. Aus Stetigkeitsgründen ergibt sich

$$\|\mathbf{v}^*\|_{H^1(\Omega)} = 1 \quad \text{und} \quad \|\mathbf{v}^*\| = 0.$$

Es folgt also $|\mathbf{v}^*|_{H^1(\Omega)} = 0$, weshalb \mathbf{v}^* konstant sein muss. Wegen $\sum_{k=1}^R |\langle \mathbf{r}_k, \mathbf{v}^* \rangle|^2 = 0$ muss aber $\mathbf{v}^* = \mathbf{0}$ sein, was ein Widerspruch zu $\|\mathbf{v}^*\|_{H^1(\Omega)} = 1$ darstellt. \square

Bemerkung Bei diesem Lemma handelt es sich um eine spezielle Version des Normierungssatzes von Sobolev, dessen Beweis auf derselben Beweistechnik beruht. Speziell haben wir zuvor diese Technik bereits zum Beweis von Lemma 5.10 verwendet. \triangle

12.6 Finite-Element-Approximation

Zunächst erörtern wir die Galerkin-Diskretisierung der Variationsformulierung (12.11) im Fall $|\Gamma_D| > 0$. Dazu sei $V_h \subset V$ ein endlichdimensionaler Teilraum von V . Speziell betrachten wir stückweise lineare Ansatzfunktionen auf einer quasi-uniformen Familie \mathcal{T}_h von Triangulierungen des Gebietes Ω :

$$V_h = \{ \mathbf{v} \in [C(\Omega)]^d : \mathbf{v}|_T \in [\mathcal{P}_1]^d \text{ für alle } T \in \mathcal{T}_h \text{ und } \mathbf{v}|_{\Gamma_D} = \mathbf{0} \}.$$

Man beachte, dass der Knoten am Übergang von den Dirichlet-Randbedingungen zu Neumann-Randbedingungen selbst ein Dirichlet-Randknoten ist, das heißt, Γ_D ist eine abgeschlossene Menge.

Die Galerkin-Diskretisierung der Variationsformulierung (12.11) lautet nun:

$$\text{suche } \mathbf{u}_h \in V_h, \text{ so dass } a(\mathbf{u}_h, \mathbf{v}_h) = \ell(\mathbf{v}_h) \quad \text{für alle } \mathbf{v}_h \in V_h. \quad (12.19)$$

Aufgrund von Stetigkeit und Elliptizität der Bilinearform $a(\cdot, \cdot)$ folgt sofort aus dem Céa-Lemma (Satz 4.1), dass

$$\|\mathbf{u} - \mathbf{u}_h\|_{H^k(\Omega)} \leq ch^{2-k} \|\mathbf{u}\|_{H^2(\Omega)}, \quad k = 0, 1, \quad (12.20)$$

vorausgesetzt die Lösung \mathbf{u} ist in $[H^2(\Omega)]^d$ enthalten. Dazu genügt im Fall eines reinen Dirichlet-Randwertproblems die Voraussetzung, dass Ω ein konvexes Polygonebiet und $\mathbf{f} \in [L^2(\Omega)]^d$ ist. Im Fall gemischter Randbedingungen ist die Regularitätstheorie wesentlich schwieriger, weil im allgemeinen Singularitäten am Übergang von Dirichlet- zu Neumann-Randbedingungen auftreten.

Wie im Abschnitt 12.5 gezeigt wurde, behält die Fehleransätzung (12.20) auch im Fall eines reinen Neumann-Randwertproblems ihre Gültigkeit, falls die modifizierte Bilinearform aus (12.18) zugrundegelegt wird.

Schließlich wollen wir die Bilinearform (12.12) auch noch explizit angeben. Dazu verwenden wir die unter Ingenieuren übliche *Voigtsche Notation*. Für

$$\widehat{\boldsymbol{\varepsilon}} := \begin{bmatrix} \varepsilon_{1,1} \\ \varepsilon_{2,2} \\ 2\varepsilon_{1,2} \end{bmatrix} = \begin{bmatrix} \frac{\partial u_1}{\partial x_1} \\ \frac{\partial u_2}{\partial x_2} \\ \frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1} \end{bmatrix}$$

gilt, ausgedrückt mit Hilfe der Lamé-Konstanten aus (12.7),

$$\widehat{\boldsymbol{\sigma}} := \begin{bmatrix} \sigma_{1,1} \\ \sigma_{2,2} \\ \sigma_{1,2} \end{bmatrix} = \begin{bmatrix} \lambda + 2\mu & \lambda & \\ \lambda & \lambda + 2\mu & \\ & & \mu \end{bmatrix} \widehat{\boldsymbol{\varepsilon}},$$

vergleiche Beispiel 12.1. Weil aufgrund der Symmetrie die Nebendiagonalelemente in der Summe $\boldsymbol{\sigma} : \boldsymbol{\varepsilon} = \sum_{i,j=1}^2 \sigma_{i,j} \varepsilon_{i,j}$ doppelt vorkommen, folgt in zwei Raumdimensionen

$$a(\mathbf{u}, \mathbf{v}) = \int_{\Omega} \begin{bmatrix} \frac{\partial u_1}{\partial x_1} \\ \frac{\partial u_2}{\partial x_2} \\ \frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1} \end{bmatrix}^T \begin{bmatrix} \lambda + 2\mu & \lambda & \\ \lambda & \lambda + 2\mu & \\ & & \mu \end{bmatrix} \begin{bmatrix} \frac{\partial v_1}{\partial x_1} \\ \frac{\partial v_2}{\partial x_2} \\ \frac{\partial v_1}{\partial x_2} + \frac{\partial v_2}{\partial x_1} \end{bmatrix} d\mathbf{x}.$$

Analog ergibt sich im dreidimensionalen Fall für

$$\widehat{\boldsymbol{\varepsilon}} := \begin{bmatrix} \varepsilon_{1,1} \\ \varepsilon_{2,2} \\ \varepsilon_{3,3} \\ 2\varepsilon_{1,2} \\ 2\varepsilon_{1,3} \\ 2\varepsilon_{2,3} \end{bmatrix} = \begin{bmatrix} \frac{\partial u_1}{\partial x_1} \\ \frac{\partial u_2}{\partial x_2} \\ \frac{\partial u_3}{\partial x_3} \\ \frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1} \\ \frac{\partial u_1}{\partial x_3} + \frac{\partial u_3}{\partial x_1} \\ \frac{\partial u_2}{\partial x_3} + \frac{\partial u_3}{\partial x_2} \end{bmatrix},$$

dass

$$\widehat{\boldsymbol{\sigma}} := \begin{bmatrix} \sigma_{1,1} \\ \sigma_{2,2} \\ \sigma_{3,3} \\ \sigma_{1,2} \\ \sigma_{1,3} \\ \sigma_{2,3} \end{bmatrix} = \begin{bmatrix} \lambda + 2\mu & \lambda & \lambda & & & \\ \lambda & \lambda + 2\mu & \lambda & & & \\ \lambda & \lambda & \lambda + 2\mu & & & \\ & & & \mu & & \\ & & & & \mu & \\ & & & & & \mu \end{bmatrix} \widehat{\boldsymbol{\varepsilon}}.$$

Daher ist

$$a(\mathbf{u}, \mathbf{v}) = \int_{\Omega} \begin{bmatrix} \frac{\partial u_1}{\partial x_1} \\ \frac{\partial u_2}{\partial x_2} \\ \frac{\partial u_3}{\partial x_3} \\ \frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1} \\ \frac{\partial u_1}{\partial x_3} + \frac{\partial u_3}{\partial x_1} \\ \frac{\partial u_2}{\partial x_3} + \frac{\partial u_3}{\partial x_2} \end{bmatrix}^T \begin{bmatrix} \lambda + 2\mu & \lambda & \lambda & & & \\ \lambda & \lambda + 2\mu & \lambda & & & \\ \lambda & \lambda & \lambda + 2\mu & & & \\ & & & \mu & & \\ & & & & \mu & \\ & & & & & \mu \end{bmatrix} \begin{bmatrix} \frac{\partial v_1}{\partial x_1} \\ \frac{\partial v_2}{\partial x_2} \\ \frac{\partial v_3}{\partial x_3} \\ \frac{\partial v_1}{\partial x_2} + \frac{\partial v_2}{\partial x_1} \\ \frac{\partial v_1}{\partial x_3} + \frac{\partial v_3}{\partial x_1} \\ \frac{\partial v_2}{\partial x_3} + \frac{\partial v_3}{\partial x_2} \end{bmatrix} dx.$$

Index

- $H^m(\Omega)$, 24
 - Regularität, 54
 - Seminorm, 25
- $H_0^m(\Omega)$, 25
- θ -Schema, 93
- affine Familie, 45
- Basis
 - Lagrange-, 42
 - nodale, 38, 42
- Bilinearform
 - H -elliptische, 30
 - stetige, 30
- Blasenfunktion, 85
- Céa-Lemma, 36, 91
- Cauchy-Greensche Verzerrungstensor, 98
- CFL-Bedingung, 94
- Clément-Approximation, 83
- Crank-Nicolson-Verfahren, 93
- Datenoszillation, 85
- Deformation, 98
- Delta-Distribution, 46
- Differentialgleichung
 - elliptische, 8
 - hyperbolische, 8
 - parabolische, 8
- Differentialoperator, 8
 - elliptischer, 8
 - hyperbolischer, 8
 - parabolischer, 8
- Differenz
 - linksseitige, 13
 - rechtsseitige, 13
 - zentrale, 13
- Differenzenstern, 15
 - 5-Punkte-Stern, 15
 - für beliebigen Differentialoperator, 17
- Differenzenverfahren, 16, 18
- duales Problem, 57
- Dualitätsargument, 56
- Einzelschrittverfahren, 64
- Elastizitätsmodul, 99
- Elastizitätstensor, 99
- Energienorm, 30
- Euler-Verfahren
 - explizites, 93
 - implizites, 93
- Familie von Zerlegungen, 40
 - affine, 45
 - nicht entartete, 49
 - quasi-uniforme, 49
- Finite-Element-Raum, 40
- Formfunktion, 61
- Frobenius-Innenprodukt, 100
- Funktion
 - harmonische, 13
- Galerkin
 - Orthogonalität, 37
 - Projektion, 70, 95
 - Verfahren, 36
- Gauß-Seidel-Verfahren, 64
- Gebiet, 5
 - diskret zusammenhängend, 18
- Gesamtschrittverfahren, 64
- Gitter, 15
- Gitterfunktion, 16
- Gitterpunkt, 15
 - randferner, 15
 - randnahe, 15
- Glättungseigenschaft, 67
- Gleichung
 - Laplace-, 5
 - Poisson-, 7
 - Potential-, 5
 - Wärmeleitungs-, 6, 92

- Wellen-, 8
- innere Kondensation, 62
- inverse Abschätzung, 52
- Iteration
 - geschachtelte, 80
- Jacobi-Verfahren, 64
- Knoten
 - hängende, 60
- Kompatibilitätsbedingung, 34
- Konistenz, 20
- Konvergenz, 20
 - quasi-optimale, 37
- Kornsche Ungleichung
 - erste, 101
 - zweite, 103
- Lösung
 - klassische, 13, 28
 - schwache, 31
- Lamé-Konstanten, 100
- Laplace-Gleichung, 5
- Laplace-Operator, 5, 7
- Maximumprinzip, 10
 - diskretes, 19
- Mehrgitterverfahren, 64, 73
 - geschachtelte Iteration, 80
 - V-Zyklus, 73
 - W-Zyklus, 73
- Minimumprinzip, 11
- Mittelwert der Funktion, 34
- Norm
 - Sobolev-, 24
- Operator
 - Laplace-, 5, 7
- Poincaré-Friedrichssche Ungleichung, 25
- Poisson-Gleichung, 7
- Poisson-Zahl, 99
- Polynome
 - \mathcal{P}_m , 41, 44
 - \mathcal{Q}_m , 43, 44
- Potentialgleichung, 5
- Problem
 - duales, 57
 - sachgemäß gestelltes, 9
 - schlecht gestelltes, 9
- Prolongation, 69
- Randbedingung
 - Dirichlet-, 9
 - natürliche, 34
 - Neumann-, 9
 - wesentliche, 33
- Randpunkt, 15
- Restriktion, 68
- Richardson-Verfahren, 64
- Satz
 - Aubin-Nitsche-Lemma, 56
 - Bramble-Hilbert-Lemma, 48
 - Céa-Lemma, 36, 91
 - Charakterisierungssatz, 29
 - Lemma von Sobolev, 46
 - Rellichscher Auswahlssatz, 47
 - Spursatz, 26
 - von Lax-Milgram, 30, 89
- Schema
 - θ -, 93
- schwache Ableitung, 23
- Shortley-Weller-Approximation, 15
- Skalarprodukt
 - Frobenius-, 100
- Sobolev-Raum
 - $H^m(\Omega)$, 24
 - $H_0^m(\Omega)$, 25
- Spannungstensor, 98
- Stabilität, 20
- Starrkörper
 - bewegungen, 102
 - rotationen, 103
 - translationen, 103
- Steifigkeitsmatrix, 37
- Transitions
 - element, 60
 - kante, 60
- Ungleichung
 - erste Kornsche, 101
 - zweite Kornsche, 103
- Verfahren
 - Crank-Nicolson-, 93
 - Differenzen-, 16, 18
 - Einzelsschritt-, 64

- explizites Euler-, 93
- Galerkin-, 36
- Gauß-Seidel-, 64
- Gesamtschritt-, 64
- implizites Euler-, 93
- Jacobi-, 64
- Mehrgitter-, 64, 73
- Richardson-, 64
- Vergleichsprinzip, 11
- Verschiebung, 98
- Verzerrung, 98
- Verzerrungstensor, 99
- Voigtsche Notation, 106

- Wärmeleitungsgleichung, 6, 92
- Wellengleichung, 8

- Zerlegung
 - nicht entartete, 49
 - quasi-uniforme, 49
 - zulässige, 40