Numerik der partiellen Differentialgleichungen

Skript zur Vorlesung
im
Herbstsemester 2015
und
Frühjahrsemester 2016

Helmut Harbrecht

Stand: 29. Mai 2016

Vorwort

Diese Mitschrift kann und soll nicht ganz den Wortlaut der Vorlesung wiedergeben. Sie soll das Nacharbeiten des Inhalts der Vorlesung erleichtern. Kapitel, die mit einem Stern markiert sind, beinhalten ergänzendes Material. Die Vorlesung orientiert sich hauptsächlich an dem unten genannten Buch von Dietrich Braess.

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*

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 $g : \Gamma \to \mathbb{R}$ beschreibe eine Drahtschlinge, die eine Seifenhaut aufspannt. Diese Seifenhaut lässt sich als Funktion $u : \overline{\Omega} \to \mathbb{R}$ beschreiben, deren Form minimale Oberfläche besitzt:

$$\int_\Omega \sqrt{1 + u_x^2 + u_y^2}\,\mathrm{d}x\,\mathrm{d}y \to \min.$$

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

$$F(u) := \frac{1}{2}\int_\Omega u_x^2 + u_y^2\,\mathrm{d}x\,\mathrm{d}y \to \min.$$

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

$$0 = \lim_{\varepsilon\to 0} \frac{F(u+\varepsilon v) - F(u)}{\varepsilon} = \int_\Omega u_x v_x + u_y v_y\,\mathrm{d}x\,\mathrm{d}y = \int_\Omega \langle \nabla u, \nabla v\rangle\,\mathrm{d}\mathbf{x}. \tag{1.1}$$

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

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

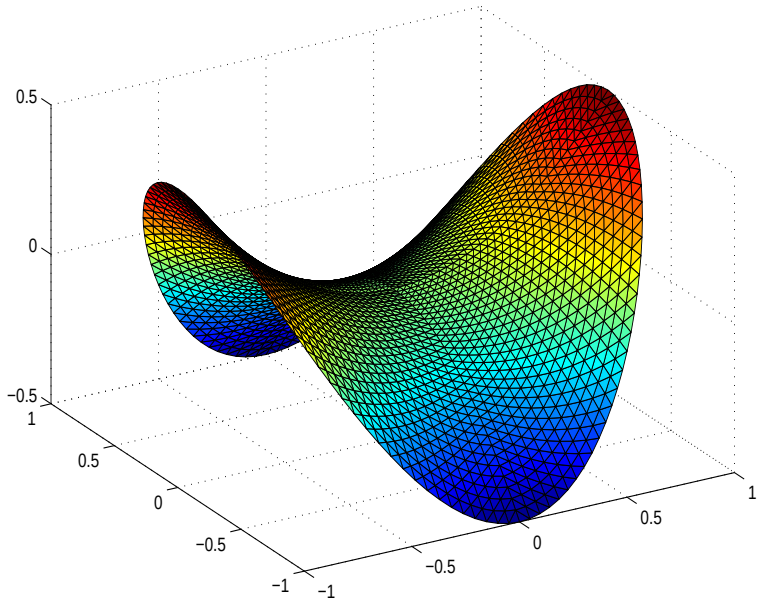
wobei $\Delta u = u_{xx} + u_{yy}$ den *Laplace-Operator* bezeichnet. Dies eingesetzt in (1.1) ergibt für u die Bedingung

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

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

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

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

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

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

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

so lässt sich die Lösung im Innern gemäß

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

darstellen.

Wärmeleitungsgleichung: In einem offenen, beschränkten Gebiet $\Omega \subset \mathbb{R}^d$ beschreibe die Funktion $u : \mathbb{R}_{\geq 0} \times \overline{\Omega} \to \mathbb{R}$ die Temperaturverteilung. Zum Zeitpunkt $t = 0$ liege die Anfangsverteilung $u(0, \mathbf{x}) = u_0(\mathbf{x}) \in C(\overline{\Omega})$ 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 $g \in C\big(\mathbb{R}_{>0} \times \Gamma\big)$ vorgegeben.
Aus dem Erhaltungssatz folgt nun für jedes Kontrollvolumen $V \subset \Omega$

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

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

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

Für die Temperaturverteilung folgt somit für alle Kontrollvolumen V die Gleichung

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

dies bedeutet

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

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

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

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

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

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

Diese Gleichung hat in der Elektrostatik ebenfalls eine immense Bedeutung: ist in Ω die Ladungsdichte $f : \Omega \to \mathbb{R}$ 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 $\int_{\partial V} \rho\langle \mathbf{v}, \mathbf{n}\rangle \,\mathrm{d}\sigma$. 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*

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

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 $\mathcal{L} : C^2(\Omega) \to C(\Omega)$ ein allgemeiner linearer Differentialoperator zweiter Ordnung

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

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

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

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 $\operatorname{rang}\big([\mathbf{A}(\mathbf{x}), \mathbf{b}(\mathbf{x})]\big) = d$ 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. △

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
$$\mathcal{L}u = f \text{ in } \Omega, \quad u = g \text{ auf } \Gamma. \tag{1.8}$$
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
$$\begin{aligned} u_t + \mathcal{L}u &= f \quad \text{in } \mathbb{R}_{>0} \times \Omega \\ u &= g \quad \text{auf } \mathbb{R}_{>0} \times \Gamma \quad \text{(Randbedingung)} \\ u(0,\cdot) &= u_0 \text{ auf } \overline{\Omega} \qquad \text{(Anfangsbedingung)} \end{aligned}$$
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
$$\begin{aligned} u_{tt} + \mathcal{L}u &= f \quad \text{in } \mathbb{R}_{>0} \times \Omega \\ u &= g \quad \text{auf } \mathbb{R}_{>0} \times \Gamma \quad \text{(Randbedingung)} \\ u(0,\cdot) = u_0,\ u_t(0,\cdot) &= u_1 \text{ auf } \overline{\Omega} \qquad \text{(Anfangsbedingungen)} \end{aligned}$$

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

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

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

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

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 $\mathbf{y}$ Extremalpunkt ist, gilt

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

Dies bedeutet

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

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. □

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. □

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 $\mathbf{x} \in \Omega$. □

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

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

   □

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

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

   □

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

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

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

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

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

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

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

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

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

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 $(\mathcal{L}u)(\mathbf{y}) - c(\mathbf{y})u(\mathbf{y}) \le (\mathcal{L}u)(\mathbf{y}) \le 0$. 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. □

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

$$\frac{\partial u}{\partial \mathbf{e}_j}(\mathbf{x}) = (\partial_j^{\pm h} u)(\mathbf{x}) + R_1^{\pm}, \quad |R_1^{\pm}| \leq \frac{h}{2} \|u\|_{C^2(\overline{\Omega})},$$

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

und

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

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

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

so folgt die zweite Aussage

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

Schließlich folgt aus Addition der drei Gleichungen

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

dass

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

□

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

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

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 $\overline{\Omega}$ die Vereinigung von Würfeln der Kantenlänge h, so sprechen wir von einem *Würfelgebiet*. In diesem Fall besitzen dann auch alle Rand- und randnahen 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*

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

Dieser ist für Würfelgebiete ausreichend.

In beliebigen Gebieten muss der Differenzenquotient für randnahe Punkte entsprechend modifiziert werden. Für $u \in C^3(\overline{\Omega})$ 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 $x_i = a + hi$, $i = 1, \ldots, n-1$, 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

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

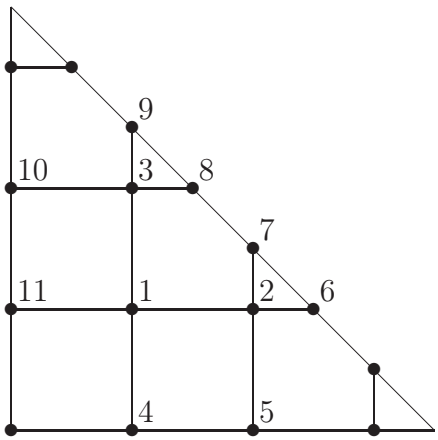
überziehen wir Ω mit einem Gitter der Maschenweite $h = 1/n$. Das entstehende Gleichungssystem wird übersichtlicher bei Benutzung von Doppelindizes $u_{i,j} = u(ih, jh)$, $1 \le i, j < n$. Es ergibt sich das Gleichungssystem

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

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

$$\left[\begin{array}{cc|cc} 4 & & & \\ & \ddots & & \mathbf{A} \\ & & 4 & \\ \hline & & 4 & \\ & \mathbf{A}^T & & \ddots \\ & & & 4 \end{array}\right] \mathbf{u} = \mathbf{f}.$$

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* $u_h : \overline{\Omega}_h \to \mathbb{R}$, so dass

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

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

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 $u \in C^4(\overline{\Omega})$ 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

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

△

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

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

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

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

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 $u \in C^4(\overline{\Omega})$. 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 Diskretiserung 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, $0 \le \ell \le k$, gelte $\alpha_\ell < 0$ für alle $\ell = 1, 2, \ldots, k$ und

$$\sum_{\ell=0}^{k} \alpha_\ell \ge 0, \qquad \sum_{\ell=0}^{k} \alpha_\ell p_\ell \le 0.$$

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

$$p_0 = p_1 = \cdots = p_k.$$

Beweis. Aus den Voraussetzungen folgt

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

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

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. △

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 herrührt. Der Differenzenstern zu jedem Gitterpunkt in Ω_h genüge folgenden drei Bedingungnen:

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

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

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 $p_1, p_2, \ldots, p_k$ 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 $p_0 = p_1 = \cdots = p_k$, 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. □

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

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

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

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. △

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. △

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 $\mathbf{x} \in \overline{\Omega}_h$. □

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. △

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

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

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

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

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

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

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

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

wobei $\mathbf{n} = (n_1, n_2, \ldots, n_d)$ 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 $\partial^{\boldsymbol{\alpha}} u \in L^2(\Omega)$ für alle $|\boldsymbol{\alpha}| \leq m$ besitzen.

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

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

und der zugehörigen Norm

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

ist ein Hilbert-Raum.

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

$$\|\partial^{\boldsymbol{\alpha}} v_n - v^{\boldsymbol{\alpha}}\|_{L^2(\Omega)} \overset{n\to\infty}{\longrightarrow} 0.$$

Wir müssen nun nur noch die Identität $\partial^{\boldsymbol{\alpha}} v = v^{\boldsymbol{\alpha}}$ 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 $(w_n, \phi)_{L^2(\Omega)} \to (w, \phi)_{L^2(\Omega)}$ für jede Testfunktion $\phi \in C_0^\infty(\Omega)$. Somit ergibt sich

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

dies bedeutet, es gilt tatsächlich $\partial^{\boldsymbol{\alpha}} v = v^{\boldsymbol{\alpha}}$. □

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

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

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. △

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 $\boldsymbol{H_0^m(\Omega)}$ 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:

$$\begin{array}{ccccccccc} L^2(\Omega) & = & H^0(\Omega) & \supset & H^1(\Omega) & \supset & H^2(\Omega) & \supset & \cdots \\ & & \| & & \cup & & \cup & & \\ & & H_0^0(\Omega) & \supset & H_0^1(\Omega) & \supset & H_0^2(\Omega) & \supset & \cdots \end{array}$$

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

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

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

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

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 $\Omega \subset \square := \{(x_1, x_2, \ldots, x_d) : 0 \le x_i \le s\}$ annehmen und $v(\mathbf{x}) = 0$ für $\mathbf{x} \in \square \setminus \Omega$ setzen. Es folgt

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

und mit der Cauchy-Schwarzschen Ungleichung weiter

$$|v(\mathbf{x})|^2 \le \left(\int_0^{x_1} 1^2 \,\mathrm{d}t \right) \left(\int_0^{x_1} |\partial_{x_1} v(t, x_2, \ldots, x_d)|^2 \,\mathrm{d}t \right) \le s \int_0^s |\partial_{x_1} v(t, x_2, \ldots, x_d)|^2 \,\mathrm{d}t.$$

Da die rechte Seite unabhängig von x_1 ist, ergibt sich

$$\int_0^s |v(\mathbf{x})|^2 \,\mathrm{d}x_1 \le s^2 \int_0^s |\partial_{x_1} v(\mathbf{x})|^2 \,\mathrm{d}x_1.$$

Schließlich wird über die anderen Koordinaten integriert:

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

□

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. △

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

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

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

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

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

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

□

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

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

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 $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \cdots \cup \Gamma_n$ 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

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

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

$$v\big(\phi(y), y\big) = v\big(\phi(y) + t, y\big) - \int_0^t \partial_x v\big(\phi(y) + s, y\big)\,\mathrm{d}s, \quad 0 \le t \le \delta.$$

Die Integration über t von 0 bis δ ergibt

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

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

$$\delta^2 v^2\big(\phi(y), y\big) \le 2\bigg(\int_0^\delta v\big(\phi(y) + t, y\big)\,\mathrm{d}t\bigg)^2 + 2\bigg(\int_0^\delta \partial_x v\big(\phi(y) + t, y\big)(\delta - t)\,\mathrm{d}t\bigg)^2.$$

Die Cauchy-Schwarzsche Ungleichung liefert nun

$$\begin{aligned}
\delta^2 v^2\big(\phi(y), y\big) &\le 2\bigg(\int_0^\delta 1\,\mathrm{d}t\bigg)\bigg(\int_0^\delta v^2\big(\phi(y) + t, y\big)\,\mathrm{d}t\bigg) \\
&\qquad + 2\bigg(\int_0^\delta (\delta - t)^2\,\mathrm{d}t\bigg)\bigg(\int_0^\delta \big|\partial_x v\big(\phi(y) + t, y\big)\big|^2\,\mathrm{d}t\bigg) \\
&= 2\delta\bigg(\int_0^\delta v^2\big(\phi(y) + t, y\big)\,\mathrm{d}t\bigg) + \frac{2}{3}\delta^3\bigg(\int_0^\delta \big|\partial_x v\big(\phi(y) + t, y\big)\big|^2\,\mathrm{d}t\bigg).
\end{aligned}$$

Wir dividieren durch δ^2 und integrieren bezüglich y

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

Das Kurvenelement auf Γ_m ist durch $\mathrm{d}o = \sqrt{1+(\phi'(y))^2}\,\mathrm{d}y$ gegeben. Deshalb haben wir mit $c_m := \max\left\{\sqrt{1+(\phi'(y))^2} : y_1 \le y \le y_2\right\}$:

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

Indem wir

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

setzen, erhalten wir schließlich

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

Die Restriktion $\gamma : H^1(\Omega) \to L^2(\Gamma)$ 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 $\|v - v_k\|_{H^1(\Omega)} \to 0$ für $k \to \infty$. Aus

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

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

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

□

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

3.2 Variationsformulierung von Dirichlet-Problemen

Gegeben sei die partielle Differentialgleichung

$$\begin{aligned} -\sum_{i,j=1}^{d}\partial_{x_i}\big(a_{i,j}(\mathbf{x})\partial_{x_j}u(\mathbf{x})\big) + c(\mathbf{x})u(\mathbf{x}) &= f(\mathbf{x}) \quad \text{für } \mathbf{x}\in\Omega,\\ u(\mathbf{x}) &= 0 \qquad \text{für } \mathbf{x}\in\Gamma \end{aligned}$$

mit $c(\mathbf{x}) \ge 0$ 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

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

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

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

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

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

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

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

Außerdem gibt es höchstens eine Minimallösung.

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

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

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. □

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

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

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

Beweis. Der Beweis erfolgt mit Hilfe des Gaußschen Integralsatzes:

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

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

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

Wir setzen nun

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

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

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

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. □

Mit der selben 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 $a : H\times H \to \mathbb{R}$ heißt **stetig**, wenn es ein $c_S > 0$ gibt, so dass

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

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

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

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 $\underline{c}, \overline{c} > 0$ derart, dass gilt

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

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 $a : H \times H \to \mathbb{R}$ eine stetige und V-elliptische Bilinearform. Für jedes $\ell \in V'$ hat das Variationsproblem

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

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

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

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

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

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

Bemerkung Im Spezialfall $V = H$ ergibt sich gerade der Rieszsche Darstellungsatz: Zu jedem $\ell \in H'$ gibt es ein Element $u \in H$ mit

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

△

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

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

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

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

in $H_0^1(\Omega)$.

Beweis. Aufgrund der Abschätzung

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

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

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

Wegen der Poincaré-Friedrichsschen Ungleichung sind $|\cdot|_{H^1(\Omega)}$ und $\|\cdot\|_{H^1_0(\Omega)}$ äquivalente Normen und damit ist $a(\cdot,\cdot)$ eine $H^1_0(\Omega)$-elliptische Bilinearform. Weiterhin ist auch die Linearform $\ell : H^1(\Omega) \to \mathbb{R}$ stetig:

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

Gemäß dem Satz von Lax-Milgram 3.11 existiert folglich eine eindeutige schwache Lösung, die zugleich das Variationsproblem löst. □

Bemerkungen

1. Das Funktional $\ell : H^1_0(\Omega) \to \mathbb{R}$ ist sogar stetig für alle $f \in H^{-1}(\Omega) := \big(H^1_0(\Omega)\big)' \supset L^2(\Omega)$. 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 $u_g\big|_\Gamma = g$ ist im Sinne des Spursatzes 3.7. Der Ansatz $u = u_0 + u_g$ führt dann auf die folgende Variationsformulierung:

   $$\text{suche } u_0 \in H^1_0(\Omega), \text{ so dass} \quad a(u_0,v) = \ell(v) - a(u_g,v) \quad \text{für alle } v \in H^1_0(\Omega).$$

   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 $v \in H^1_0(\Omega)$.
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

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

$$\sum_{i,j=1}^{d} a_{i,j}n_i(\mathbf{x})\partial_{x_j}u(\mathbf{x}) = g(\mathbf{x}) \quad \text{für } \mathbf{x}\in\Gamma$$

mit

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

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

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

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

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

Wir erhalten demnach

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

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

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

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:

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

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

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

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

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

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

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

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

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

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

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 $\int_\Gamma w^2 \,\mathrm{d}\sigma > 0$. 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 $\int_\Gamma vw \,\mathrm{d}\sigma > 0$. Dies ist ein Widerspruch, und die Randbedingung ist erfüllt. □

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

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*

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

erfüllt sein muss, damit überhaupt eine Lösung existiert.

Es bezeichne

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

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

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

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

$$\|v\|_{L^2(\Omega)} \le c\big(|\overline{v}| + |v|_{H^1(\Omega)}\big)$$

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

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

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

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

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 $\widetilde{f} = f - \eta$ und erhält

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

das heißt,

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

△

4. Galerkin-Verfahren

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

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

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

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

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

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

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

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

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). △

Satz 4.1 (Céa-Lemma) Die Bilinearform $a : V \times V \to \mathbb{R}$ 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

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

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

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

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

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

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 $\{\varphi_1, \varphi_2, \ldots, \varphi_N\}$ von V_h. Dann ist (4.2) äquivalent zu

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

Der Ansatz

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

führt zu dem linearen Gleichungssystem

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

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,

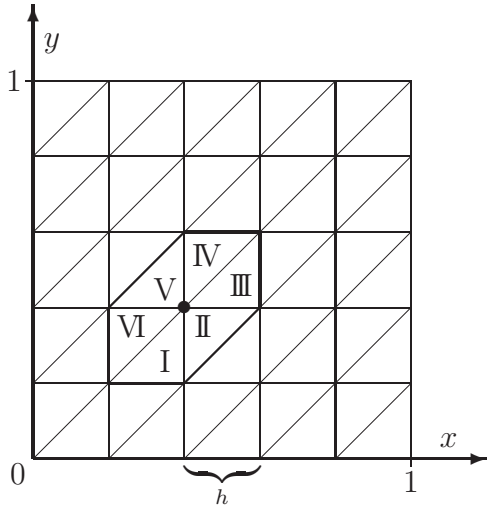
$$\mathbf{z}_h^T\mathbf{A}_h\mathbf{z}_h = \sum_{i,j=1}^{N} z_i a(\varphi_j,\varphi_i)z_j = a\Big(\sum_{j=1}^{N} z_j\varphi_j, \sum_{i=1}^{N} z_i\varphi_i\Big) \ge c_E\Big\|\sum_{i=1}^{N} z_i\varphi_i\Big\|_V^2,$$

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 $\overline{\Omega}$ wie folgt mit einem gleichmäßigen Dreiecksnetz der Maschenweite h überzogen:



Wir wählen

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

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

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

Weiter ergibt sich

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

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

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

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:

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

△

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.

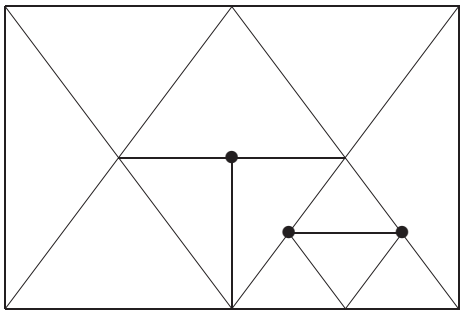
Ein *Finite-Element-Raum* ist charakterisiert durch

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.

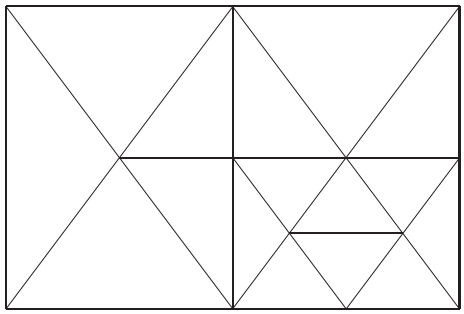
Definition 5.1 Eine Zerlegung $\mathcal{T} = \{T_1, T_2, \ldots, T_M\}$ von Ω in Dreiecks- oder Viereckselemente heißt **zulässig**, wenn folgende Eigenschaften erfüllt sind:

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

unzulässige Zerlegung



zulässige Zerlegung



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

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

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

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

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

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

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

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

folgt dann aus der Cauchy-Schwarzschen Ungleichung

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

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. □

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 $\ell = 1+2+\cdots+(m+1)$ Punkte $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_\ell$ angeordnet. Dann gibt es zu jedem $f \in C(T)$ genau ein Polynom p vom Grad m, das die Interpolationsaufgabe

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

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 $i = 1, 2, \ldots, m+1$ auf der x-Achse liegen. Die eindimensionale Theorie liefert ein Polynom $p_0 = p_0(x)$ vom Grad m mit

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

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

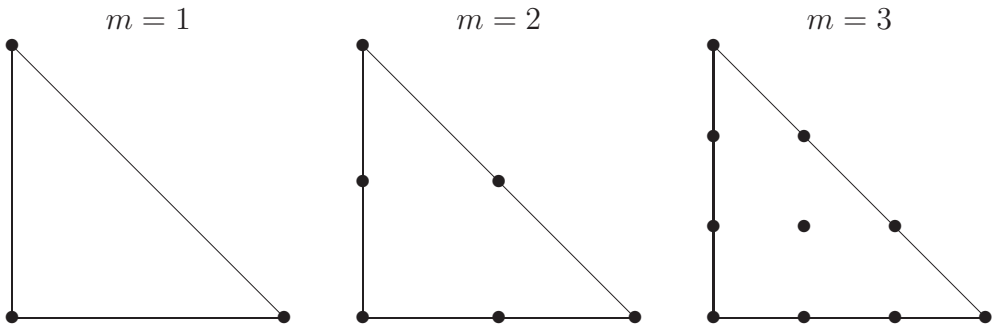
$$q(x_i, y_i) = \frac{1}{y_i}\big\{f(x_i, y_i) - p_0(x_i)\big\}, \quad i = m+2, m+3, \ldots, \ell.$$

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

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

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

verwendet.

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

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

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

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

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

löst.

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

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

Ferner mögen

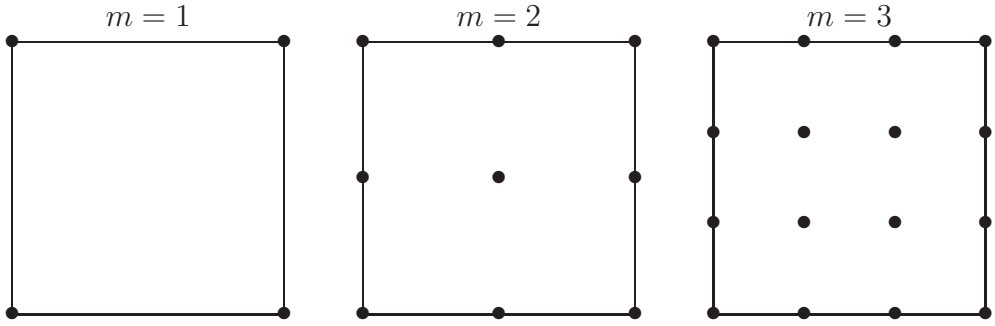
$$L_k(y) := \prod_{\substack{j=0 \\ j \ne k}}^{m} \frac{y - bt_j}{bt_k - bt_j}$$

die Lagrange-Polynome zu den Knoten $\{bt_0, bt_1, \ldots, bt_m\}$ bezeichnen, das heißt, für jedes $k = 0,1,\ldots,m$ 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 Paralellogramme 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

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

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. △

5.4 Dreidimensionaler Fall

Sei $\Omega \subset \mathbb{R}^3$ ein Polyeder. Eine Zerlegung $\mathcal{T} = \{T_1, T_2, \ldots, , T_M\}$ dieses Polyeders ist zulässig, wenn $\overline{\Omega} = \bigcup_{i=1}^{M} T_i$ 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

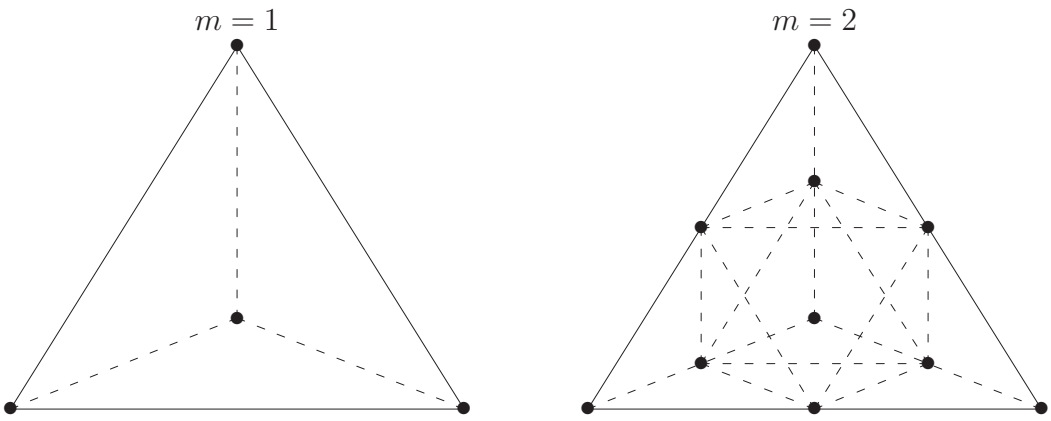
$$v|_T \in \mathcal{P}_m := \left\{ v(x,y,z) = \sum_{0 \le i+j+k \le m} \alpha_{i,j,k} x^i y^j z^k \right\}$$

oder

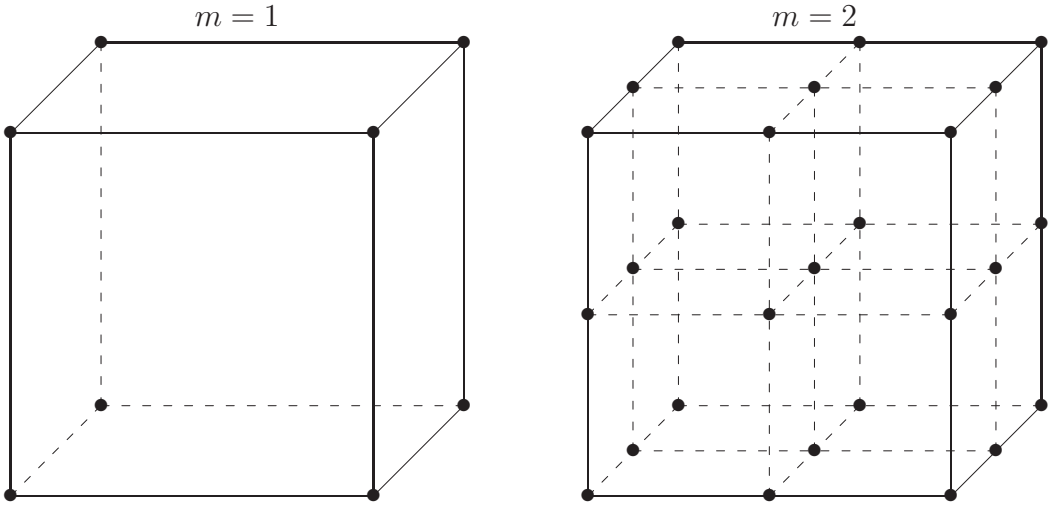
$$v|_T \in \mathcal{Q}_m := \left\{ v(x,y,z) = \sum_{0 \le i,j,k \le m} \alpha_{i,j,k} x^i y^j z^k \right\}$$

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 Interpolationspunkten vorgegeben:



Auf Quadern wählt man die Interpolationspunkte wie folgt:



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

5.5 Approximationseigenschaften

Das Céa-Lemma besagt, dass die Güte der Galerkin-Approximation davon abhängt, wie gut sich die Lösung im Finite-Element-Raum approximieren lässt. Daher müssen wir die Approximationseigenschaften der Finite-Element-Räume untersuchen.

Definition 5.6 Eine Familie von Finite-Element-Räumen V_h mit Zerlegungen $\mathcal{T}_h$ von $\Omega \subset \mathbb{R}^d$ heißt **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. $\mathcal{P}_{\text{ref}}$ ist ein Unterraum von $C(T_{\text{ref}})$ mit endlicher Dimension ℓ.
3. Σ ist eine Menge von ℓ linear unabhängigen Funktionalen auf $\mathcal{P}_{\text{ref}}$. Jedes $p \in \mathcal{P}_{\text{ref}}$ ist durch die Werte der ℓ Funktionale aus Σ eindeutig bestimmt. Da sich die Funktionale 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 $F_T : T_{\text{ref}} \to T$, so dass für jedes $v \in V_h$ gilt
$$v(\mathbf{x}) = p(F_T^{-1}\mathbf{x}) \text{ mit } p \in \mathcal{P}_{\text{ref}} \text{ für alle } \mathbf{x} \in T.$$
5. Es sei $V_h \subset C^k(\overline{\Omega})$. Die Restriktion von $p \in \mathcal{P}_{\text{ref}}$ auf jede Kante bzw. Fläche von T_{ref} sowie die Ableitungen bis zur Ordnung k sind durch solche Interpolationsbedingungen aus Σ eindeutig bestimmt, die sich nur auf Größen an Punkten auf dieser Kante bzw. Fläche beziehen.

Beispiel 5.7 Stückweise lineare Finite Elemente auf Dreiecken erhält man durch die Wahl

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

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

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

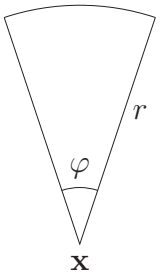
△

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

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

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

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

Im letzten Schritt haben wir benutzt, dass gilt

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

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

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

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. □

Satz 5.9 (Rellichscher Auswahlsatz) Es sei $m \ge 0$ 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. □

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

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

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

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

die Behauptung. Dabei haben wir ausgenutzt, dass Iv mit v in den Interpolationspunkten übereinstimmt und dass $\partial^{\boldsymbol{\alpha}}_{\mathbf{x}}(Iv) = 0$ ist für alle $|\boldsymbol{\alpha}| = m$.
Beim Nachweis der Äquivalenz ist eine Richtung schnell erbracht. Nach Lemma 5.8 ist die Einbettung $H^m(\Omega) \hookrightarrow C(\overline{\Omega})$ stetig. Das bewirkt

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

und weiter $|\!|\!|v|\!|\!| \le (1 + c\ell)\|v\|_{H^m(\Omega)}$.
Angenommen, die Umkehrung

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

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

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

Nach dem Rellichschen Auswahlsatz (Satz 5.9) konvergiert eine Teilfolge in $H^{m-1}(\Omega)$. Ohne Beschränkung der Allgmeinheit 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 $|v_i|_{H^m(\Omega)} \le |\!|\!|v_i|\!|\!| \to 0$ und

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

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

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

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

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

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

dann gilt

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

Beweis. Es sei $I : H^m(\Omega) \to \mathcal{P}_{m-1}$ 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

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

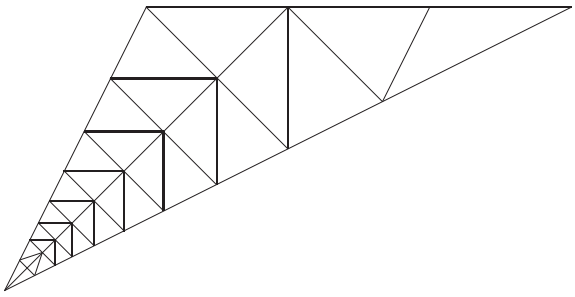
Hierbei bezeichnet $h_T \leq h := \max_{T \in \mathcal{T}_h} h_T$ den halben Durchmesser des Elements T. Ist die untere Schranke für den Innkreisradius ρ_T sogar unabhängig von T, gilt also

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

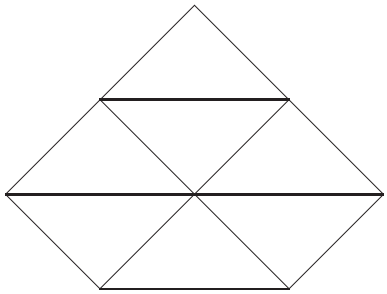
dann heißt die Familie $\{\mathcal{T}_h\}$ **quasi-uniform**.

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

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

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

$$v(\mathbf{x}) = v\big(F_T(\widehat{\mathbf{x}})\big) = \widehat{v}(\widehat{\mathbf{x}})$$

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

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

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

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

Hieraus folgt

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

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

$$\left|\partial_{\widehat{\mathbf{x}}}^{\boldsymbol{\alpha}} \widehat{v}(\widehat{\mathbf{x}})\right| \le \|\mathbf{B}\|_1^{|\boldsymbol{\alpha}|} \max_{\boldsymbol{\beta} \text{ mit } |\boldsymbol{\beta}|=|\boldsymbol{\alpha}|} \left|\partial_{\mathbf{x}}^{\boldsymbol{\beta}} v(\mathbf{x})\right|.$$

Deshalb ist

$$\sum_{|\boldsymbol{\alpha}|=m} \left|\partial_{\widehat{\mathbf{x}}}^{\boldsymbol{\alpha}} \widehat{v}(\widehat{\mathbf{x}})\right|^2 \le c\|\mathbf{B}\|_2^{2m} \sum_{|\boldsymbol{\beta}|=m} \left|\partial_{\mathbf{x}}^{\boldsymbol{\beta}} v(\mathbf{x})\right|^2$$

und somit

$$\begin{aligned}
|\widehat{v}|_{H^m(T_{\text{ref}})}^2 &= \sum_{|\boldsymbol{\alpha}|=m} \int_{T_{\text{ref}}} \left|\partial_{\widehat{\mathbf{x}}}^{\boldsymbol{\alpha}} \widehat{v}(\widehat{\mathbf{x}})\right|^2 \mathrm{d}\widehat{\mathbf{x}} \\
&\le c\|\mathbf{B}\|_2^{2m} \sum_{|\boldsymbol{\beta}|=m} \int_{T_{\text{ref}}} \left|\partial_{\mathbf{x}}^{\boldsymbol{\beta}} v\big(F_T(\widehat{\mathbf{x}})\big)\right|^2 \mathrm{d}\widehat{\mathbf{x}} \\
&= c\|\mathbf{B}\|_2^{2m} \sum_{|\boldsymbol{\beta}|=m} \int_{T} \left|\partial_{\mathbf{x}}^{\boldsymbol{\beta}} v\big(\mathbf{x}\big)\right|^2 \left|\det(\mathbf{B}^{-1})\right| \mathrm{d}\mathbf{x} \\
&= c\|\mathbf{B}\|_2^{2m} |\det \mathbf{B}|^{-1} |v|_{H^m(T)}^2.
\end{aligned}$$

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

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

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

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

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

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 $\widehat{\mathbf{x}}_0 \in T_{\text{ref}}$, so dass $\widehat{\mathbf{x}} \in T_{\text{ref}}$ für alle $\|\widehat{\mathbf{x}} - \widehat{\mathbf{x}}_0\|_2 \le \rho_{\text{ref}}$.

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

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

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

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

□

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

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

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

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

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

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

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

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, $k \ge 2$, $0 \le m \le k$, 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

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

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

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

$$|v - I_h v|_{H^\ell(T)} \le c \underbrace{\left\|\mathbf{B}^{-1}\right\|_2^\ell}_{\le (c\kappa/h)^\ell} |\det \mathbf{B}|^{1/2} |\widehat{v} - I\widehat{v}|_{H^\ell(T_{\mathrm{ref}})} \le ch^{-\ell} |\det \mathbf{B}|^{1/2} |\widehat{v} - I\widehat{v}|_{H^\ell(T_{\mathrm{ref}})}.$$

Aus Lemma 5.10 folgt

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

und damit die Abschätzung

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

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

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

Durch Aufsummation ergibt schließlich

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

und daher die Behauptung. □

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

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

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

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 k, 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,

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

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 $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\ell \in T_{\text{ref}}$ 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 ebenfalls $|\!|\!| \cdot |\!|\!|$ und $\|\cdot\|_{H^t(T_{\text{ref}})}$ 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 Ansatzsatzfunktionen 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. Fehleranalysis

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

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

heißt $\boldsymbol{H^s(\Omega)}$**-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

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

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

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

Demnach erhalten wir

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

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

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

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. △

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

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

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

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

$H^2(\Omega)$-regulär.

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

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. △

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

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

wobei die Bilinearform $a : V \times V \to \mathbb{R}$ 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

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

zugeordnet wird.

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

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

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*. △

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

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

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. □

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 $\widehat{u}$ die durch die affine Transformation resultierende Funktion auf T_{ref} und $I\widehat{u}$ die Interpolierende in $\mathcal{P}_{\text{ref}}$. Wegen der Einbettung $H^2(T_{\text{ref}}) \hookrightarrow C(T_{\text{ref}})$ folgt

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

Lemmata 5.10 und 5.14 liefern

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

und somit die gewünschte Abschätzung

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

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

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

Dies bedeutet gemäß Lemma 5.14, dass

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

und folglich ist

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

Wir kombinieren nun (6.4) und (6.5):

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

Hieraus folgt wegen

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

die Behauptung. □

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

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

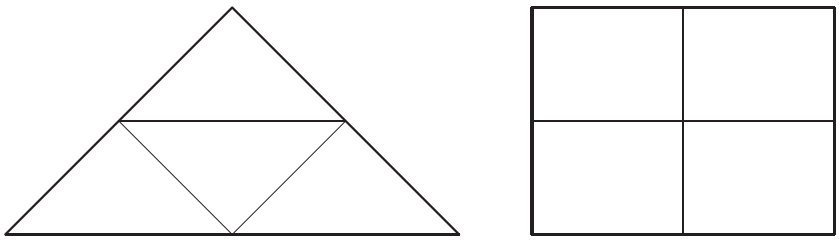
beweisen. Der logarithmische Term verschwindet im Fall $d = 3$ sogar. △

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 Fehleranalysis: 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 Dreicke 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

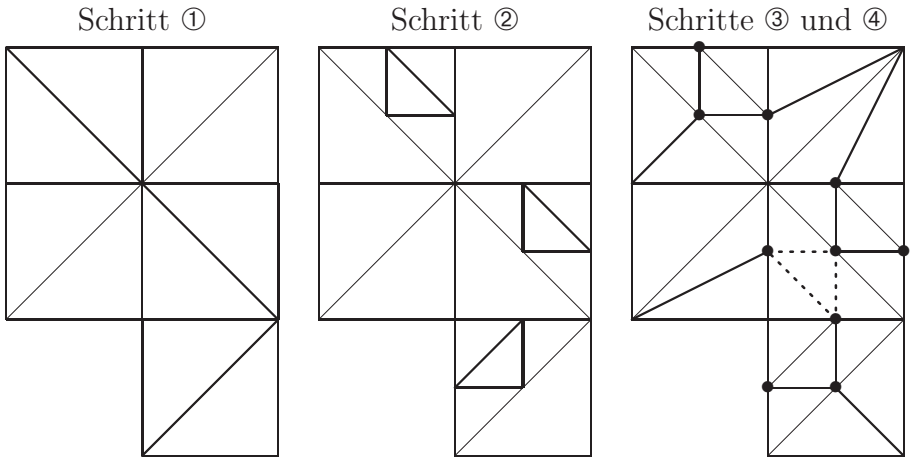
① verfeinere alle markierten Dreiecke in vier Dreiecke

② liegen auf den Kanten eines Dreicks 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

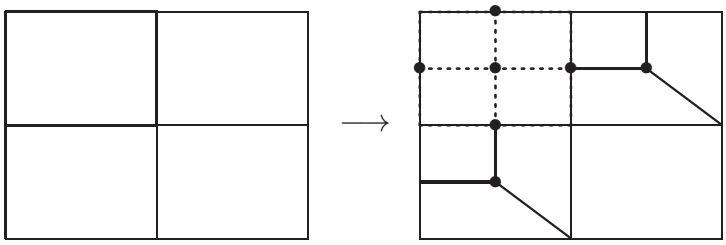
Beispiel 7.2 (Netzverfeinerung)



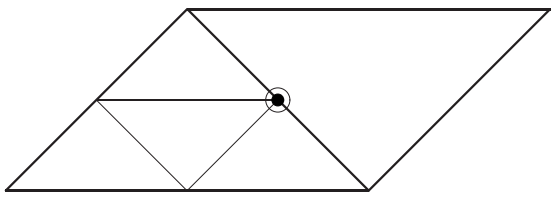
△

Bemerkungen

1. Grüne Kanten heißen *Transitionskanten* und die Dreiecke entsprechend *Transitionselemente*. 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

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

dann folgt

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

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

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

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

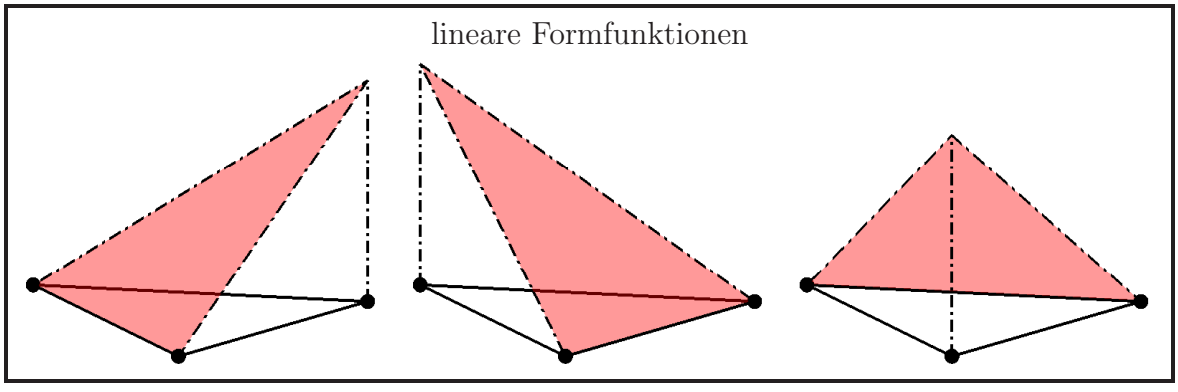
$$a_T(\varphi_{i_k},\varphi_{i_\ell}) = \frac{|T|}{|T_{\mathrm{ref}}|} \int_{T_{\mathrm{ref}}} \langle \mathbf{A}\mathbf{B}^{-T}\nabla_{\widehat{\mathbf{x}}}\psi_k, \mathbf{B}^{-T}\nabla_{\widehat{\mathbf{x}}}\psi_\ell\rangle \,\mathrm{d}\widehat{\mathbf{x}}. \tag{7.2}$$

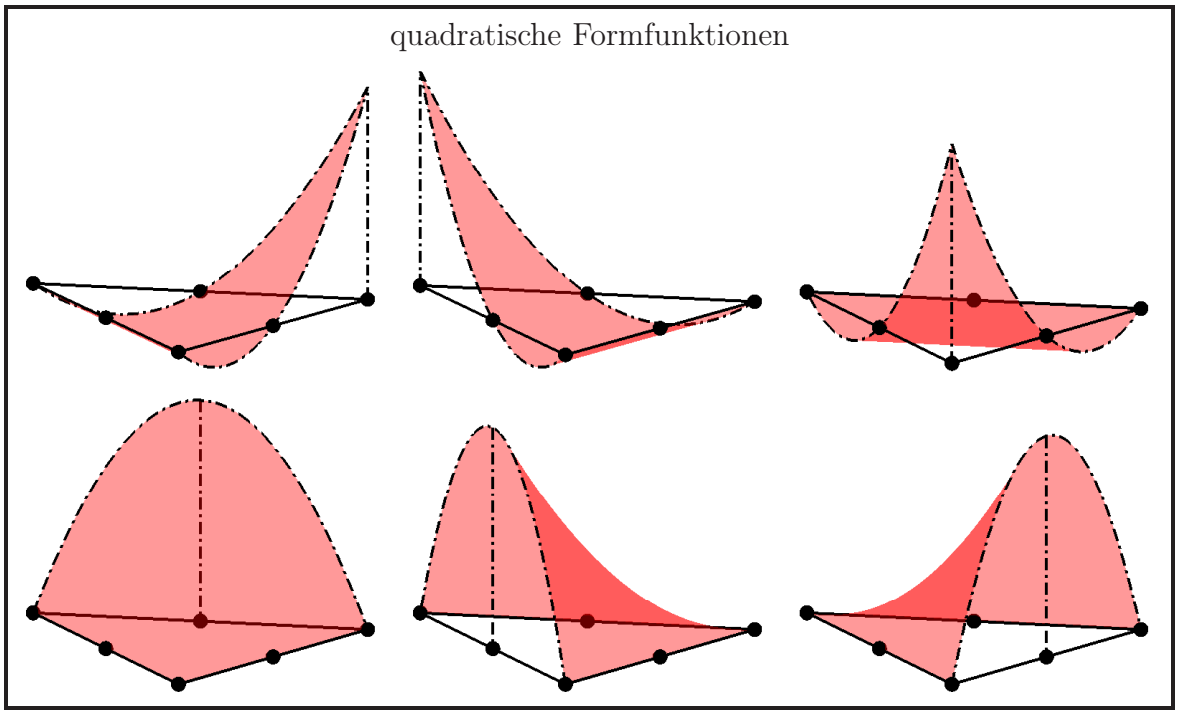
Jede Funktion aus der nodalen Basis fällt nach der Transformation auf das Referenzdreieck mit einer der normierten *Formfunktionen* $\psi_1,\psi_2,\ldots,\psi_n$ zusammen. Wählt man als Referenzelement das Dreieick $\triangle\big((0,0),(1,0),(0,1)\big)$, so ist bei linearen Elementen

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

und bei quadratischen

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





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

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

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

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

ausrechen. 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 inneren Knoten gehören. So hat zum Beispiel das quadratische Viereckselement oder das kubische Dreielement jeweils einen inneren Knoten. Die Elimination einer solchen 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

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

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 Differentialgleichgungen 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

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

- *Gauß-Seidel-Verfahren* oder *Einzelschrittverfahren* mit der Iterationsvorschrift

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

- *Richardson-Verfahren* mit der Iterationsvorschrift

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

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

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

und das Gauß-Seidel-Verfahren in der Form

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

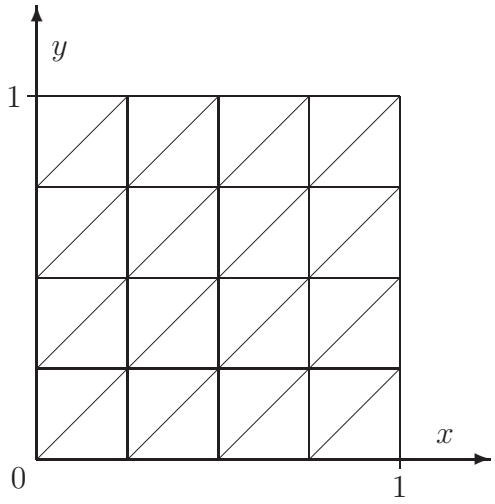
schreiben. Besitzt $\mathbf{A}_j$ den konstanten Eintrag d_j auf der Diagonalen, so ist das Richardson-Verfahren mit $\alpha_j := 1/d_j$ äquivalent zum Jacobi-Verfahren.

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 $\overline{\Omega}$ wie folgt mit einem gleichmäßigen Dreiecksnetz der Maschenweite $h_j := 2^{-(j+1)}$ überzogen:



Die Diskretisierung der Bilinearform

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

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

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

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

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

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

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

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 $h_j \to 0$.
Die Eigenwerte mit einem hochfrequenten Anteil in mindestens einer Richtung sind gerade $\{\mu_{k,\ell} : k \ge (N_j+1)/2 \text{ oder } \ell \ge (N_j+1)/2\}$. Für diese Eigenwerte gilt

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

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

Mehrgitterprinzip. Es sei

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

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 bezeichne $\mathbf{X}_j = [\mathbf{x}_{j,1}, \mathbf{x}_{j,2}, \ldots, \mathbf{x}_{j,N_j}]$ die aus den normierten Eigenvektoren $\{\mathbf{x}_{j,k}\}_{k=1}^{N_j}$ von $\mathbf{A}_j$ zusammengesetzte orthogonale Matrix.

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

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

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

Lemma 8.3 Es gilt die verallgemeinerte Cauchy-Schwarzsche Ungleichung

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

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

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

Dann folgt

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

das heißt, die Behauptung. □

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 orthonormierten 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

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

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

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

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

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

□

8.2 Prolongation und Restriktion

Wir betrachten die geschachtelte Folge von Finite-Element-Räumen

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

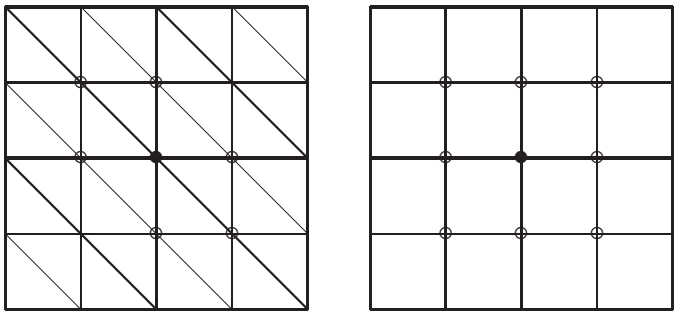
die durch uniformes Verfeinern erzeugt wird. Das Variationsproblem

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

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

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



Beispiel 8.5 Im eindimensionalen Fall gilt

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

△

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

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

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 Zweitgitterverfahrens setzt sich dann wie folgt zusammen:

1. *A-priori-Glättung.* Setze
$$\mathbf{u}_j^{(\mathrm{pre},0)} = \mathbf{u}_j^{\mathrm{alt}}$$
und führe K Schritte eines Iterationsverfahrens (zum Beispiel des Richardson-Verfahrens) durch
$$\mathbf{u}_j^{(\mathrm{pre},k)} = \mathbf{u}_j^{(\mathrm{pre},k-1)} + \alpha_j(\mathbf{f}_j - \mathbf{A}_j\mathbf{u}_j^{(\mathrm{pre},k-1)}), \quad k = 1, 2, \ldots, K.$$
2. *Grobgitterkorrektur.* Löse die Defektgleichung auf der nächstgröberen Triangulierung
$$\mathbf{A}_{j-1}\mathbf{e}_{j-1} = \mathbf{I}_j^{j-1}(\mathbf{f}_j - \mathbf{A}_j\mathbf{u}_j^{(\mathrm{pre},K)})$$
und addiere die erhaltene Näherung
$$\mathbf{u}_j^{(\mathrm{post},0)} = \mathbf{u}_j^{(\mathrm{pre},K)} + \mathbf{I}_{j-1}^{j}\mathbf{e}_{j-1}.$$
3. *A-posteriori-Glättung.* Führe L Schritte eines Iterationsverfahrens durch
$$\mathbf{u}_j^{(\mathrm{post},\ell)} = \mathbf{u}_j^{(\mathrm{post},\ell-1)} + \alpha_j(\mathbf{f}_j - \mathbf{A}_j\mathbf{u}_j^{(\mathrm{post},\ell-1)}), \quad \ell = 1, 2, \ldots, L$$
und setze
$$\mathbf{u}_j^{\mathrm{neu}} = \mathbf{u}_j^{(\mathrm{post},L)}.$$

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
$$\mathbf{u}_j - \mathbf{u}_j^{\mathrm{neu}} = \mathbf{S}_j^L(\mathbf{I} - \mathbf{I}_{j-1}^{j}\mathbf{A}_{j-1}^{-1}\mathbf{I}_j^{j-1}\mathbf{A}_j)\mathbf{S}_j^K(\mathbf{u}_j - \mathbf{u}_j^{\mathrm{alt}}). \tag{8.2}$$

Neben der Glättungseigenschaft des Iterationsverfahrens brauchen wir noch eine Aussage über die Approximationsgüte der Grobgitterkorrektur.

Lemma 8.6 Es seien $u_j, u_j^{(\mathrm{pre},K)} \in V_j$ und $e_{j-1} \in V_{j-1}$ die zu $\mathbf{u}_j, \mathbf{u}_j^{(\mathrm{pre},K)} \in \mathbb{R}^{N_j}$ und $\mathbf{e}_{j-1} \in \mathbb{R}^{N_{j-1}}$ gehörigen Funktionen. Dann ist
$$e_{j-1} = P_{j-1}(u_j - u_j^{(\mathrm{pre},K)})$$
die Galerkin-Projektion des nach der A-priori-Glättung erhaltenen Fehlers, das heißt, es gilt
$$a(e_{j-1}, v_{j-1}) = a(u_j - u_j^{(\mathrm{pre},K)}, v_{j-1}) \quad \text{für alle } v_{j-1} \in V_{j-1}.$$

Beweis. Für alle $v_{j-1} \in V_{j-1}$ gilt
$$\begin{aligned} a(e_{j-1}, v_{j-1}) &= \mathbf{v}_{j-1}^T\mathbf{A}_{j-1}\mathbf{e}_{j-1} \\ &= \mathbf{v}_{j-1}^T\mathbf{I}_j^{j-1}(\mathbf{f}_j - \mathbf{A}_j\mathbf{u}_j^{(\mathrm{pre},K)}) \\ &= (\mathbf{I}_{j-1}^{j}\mathbf{v}_{j-1})^T(\mathbf{f}_j - \mathbf{A}_j\mathbf{u}_j^{(\mathrm{pre},K)}) \\ &= (\mathbf{I}_{j-1}^{j}\mathbf{v}_{j-1})^T\mathbf{A}_j(\mathbf{u}_j - \mathbf{u}_j^{(\mathrm{pre},K)}) \\ &= a(u_j - u_j^{(\mathrm{pre},K)}, v_{j-1}). \end{aligned}$$

□

Bezeichnet $v_{j-1} = P_{j-1}v_j \in V_{j-1}$ die Galerkin-Projektion von $v_j \in V_j$, dann gilt

$$a(v_j, w_{j-1}) = a\big(P_{j-1}v_j, w_{j-1}\big) = a\big(v_{j-1}, w_{j-1}\big)$$

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 Polygongebiet. Dann ist

$$|||\mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1}|||_{0,j} \le c|||\mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1}|||_{1,j}$$

mit einer von j unabhängigen Konstanten c.

Beweis. Das Aubin-Nitsche-Lemma (Satz 6.5) liefert

$$\|v_j - P_{j-1}v_j\|_{L^2(\Omega)} \le c_S \|v_j - P_{j-1}v_j\|_{H^1(\Omega)} \sup_{g \in L^2(\Omega)\setminus\{0\}} \left\{ \frac{1}{\|g\|_{L^2(\Omega)}} \inf_{w_{j-1} \in V_{j-1}} \|\varphi_g - w_{j-1}\|_{H^1(\Omega)} \right\},$$

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

$$\|v_j - P_{j-1}v_j\|_{L^2(\Omega)} \le c_A h_{j-1} \|v_j - P_{j-1}v_j\|_{H^1(\Omega)}.$$

Wegen $h_{j-1} = 2h_j$ und (8.1) ergibt sich schließlich

$$\begin{aligned} |||\mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1}|||_{0,j} &\le \frac{\overline{c}}{h_j} \|v_j - P_{j-1}v_j\|_{L^2(\Omega)} \\ &\le 2c_A\overline{c} \|v_j - P_{j-1}v_j\|_{H^1(\Omega)} \\ &\le 2\underline{c}c_A\overline{c} |||\mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1}|||_{1,j}. \end{aligned}$$

□

Satz 8.8 (Approximationseigenschaft) Unter den Voraussetzungen von Lemma 8.7 gilt

$$|||\mathbf{v}_j - \mathbf{I}_{j-1}^j \mathbf{v}_{j-1}|||_{1,j} \le c|||\mathbf{v}_j|||_{2,j}.$$

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. □

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

$$\|\mathbf{u}_j - \mathbf{u}_j^{\text{neu}}\|_{1,j} \leq \frac{c}{\sqrt{K}} \|\mathbf{u}_j - \mathbf{u}_j^{\text{alt}}\|_{1,j}.$$

□

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 lässt sich auch auf den Fall von A-posteriori-Glättung übertragen.
3. In der Praxis ist das Gauß-Seidel-Verfahren der beste Glätter.

△

8.4 Mehrgitterverfahren

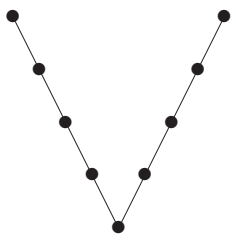
Es sei

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

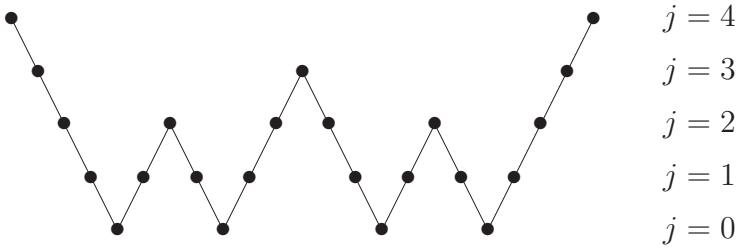
eine geschachtelte Folge von Finite-Element-Räumen, die durch uniformes Verfeinern des Grobgitterraums V_0 entstehen. Das allgemeine Mehrgitterverfahren lässt sich dann wie folgt beschreiben:

1. *A-priori-Glättung.* Setze
$$\mathbf{u}_j^{(\mathrm{pre},0)} = \mathbf{u}_j^{\mathrm{alt}}$$
und führe K Glättungsschritte durch
$$\mathbf{u}_j^{(\mathrm{pre},k)} = \mathbf{u}_j^{(\mathrm{pre},k-1)} + \alpha_j(\mathbf{f}_j - \mathbf{A}_j\mathbf{u}_j^{(\mathrm{pre},k-1)}), \quad k = 1, 2, \ldots, K.$$
2. *Restriktion.* Restringiere das Residuum auf das nächstgröbere Gitter
$$\mathbf{r}_{j-1} = \mathbf{I}_j^{j-1}(\mathbf{f}_j - \mathbf{A}_j\mathbf{u}_j^{(\mathrm{pre},K)}).$$
3. *Grobgitterkorrektur.* Falls $j = 1$ ist, dann löse das Gleichungssytem $\mathbf{A}_0\mathbf{e}_0 = \mathbf{r}_0$ 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
$$\mathbf{u}_j^{(\mathrm{post},0)} = \mathbf{u}^{(\mathrm{pre},K)} + \mathbf{I}_{j-1}^{j}\mathbf{e}_{j-1}.$$
5. *A-posteriori-Glättung.* Führe L Glättungsschritte durch
$$\mathbf{u}_j^{(\mathrm{post},\ell)} = \mathbf{u}_j^{(\mathrm{post},\ell-1)} + \alpha_j(\mathbf{f}_j - \mathbf{A}_j\mathbf{u}_j^{(\mathrm{post},\ell-1)}), \quad \ell = 1, 2, \ldots, L$$
und setze
$$\mathbf{u}_j^{\mathrm{neu}} = \mathbf{u}_j^{(\mathrm{post},L)}.$$

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
$$\|\mathbf{S}_j^\ell \mathbf{v}_j\|_{1,j} \leq \|\mathbf{v}_j\|_{1,j}, \quad \ell = 1, 2, \ldots$$
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

$$\|\mathbf{S}_j^\ell \mathbf{v}_j\|_{1,j}^2 = (\mathbf{S}_j^\ell \mathbf{v}_j)^T \mathbf{A}_j \mathbf{S}_j^\ell \mathbf{v}_j = \sum_{k=1}^{N_j} \sigma_k^2 (1-\alpha_j\lambda_k)^{2\ell}\lambda_k \le \sum_{k=1}^{N_j} \sigma_k^2 \lambda_k = \mathbf{v}_j^T \mathbf{A}_j \mathbf{v}_j = \|\mathbf{v}_j\|_{1,j}^2.$$

□

Satz 8.11 (Konvergenz des W-Zyklus) Zu jedem $\rho \in (0,1)$ gibt es ein K_0 derart, dass bei der Verwendung von $K \ge K_0$ A-priori-Glättungsschritten für die mit dem W-Zyklus erhaltenen Iterierten gilt

$$\|\mathbf{u}_j - \mathbf{u}_j^{\text{neu}}\|_{1,j} \le \rho \|\mathbf{u}_j - \mathbf{u}_j^{\text{alt}}\|_{1,j}$$

für alle $j = 1, 2, \ldots$.

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 $\widehat{\mathbf{e}}_{j-1}$ die exakte und $\mathbf{e}_{j-1}$ die durch den W-Zyklus berechnete Grobgitterkorrektur. Aus

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

folgt

$$\|\mathbf{u}_j - \mathbf{u}_j^{\text{neu}}\|_{1,j}^2 = \|\mathbf{u}_j - \mathbf{u}_j^{(\text{pre},K)} - \mathbf{I}_{j-1}^j \widehat{\mathbf{e}}_{j-1}\|_{1,j}^2 + \|\mathbf{I}_{j-1}^j (\widehat{\mathbf{e}}_{j-1} - \mathbf{e}_{j-1})\|_{1,j}^2.$$

Dabei fällt der gemischte Term wegen Lemma 8.6 weg:

$$\begin{aligned}
&(\mathbf{u}_j - \mathbf{u}_j^{(\text{pre},K)} - \mathbf{I}_{j-1}^j \widehat{\mathbf{e}}_{j-1})^T \mathbf{A}_j \mathbf{I}_{j-1}^j (\widehat{\mathbf{e}}_{j-1} - \mathbf{e}_{j-1}) \\
&\quad = a(u_j - u_j^{(\text{pre},K)} - \widehat{e}_{j-1}, \widehat{e}_{j-1} - e_{j-1}) \\
&\quad = a\big(u_j - u_j^{(\text{pre},K)} - P_{j-1}(u_j - u_j^{(\text{pre},K)}), \underbrace{\widehat{e}_{j-1} - e_{j-1}}_{\in V_{j-1}}\big) \\
&\quad = 0.
\end{aligned}$$

Mit

$$\|\mathbf{I}_{j-1}^j (\widehat{\mathbf{e}}_{j-1} - \mathbf{e}_{j-1})\|_{1,j}^2 = (\widehat{\mathbf{e}}_{j-1} - \mathbf{e}_{j-1})^T \underbrace{\mathbf{I}_j^{j-1} \mathbf{A}_j \mathbf{I}_{j-1}^j}_{=\mathbf{A}_{j-1}} (\widehat{\mathbf{e}}_{j-1} - \mathbf{e}_{j-1}) = \|\widehat{\mathbf{e}}_{j-1} - \mathbf{e}_{j-1}\|_{1,j-1}^2,$$

Satz 8.9 und der Induktionsvoraussetzung ergibt sich

$$\begin{aligned}
\|\mathbf{u}_j - \mathbf{u}_j^{\text{neu}}\|_{1,j}^2 &= \|\mathbf{u}_j - \mathbf{u}_j^{(\text{pre},K)} - \mathbf{I}_{j-1}^j \widehat{\mathbf{e}}_{j-1}\|_{1,j}^2 + \|\widehat{\mathbf{e}}_{j-1} - \mathbf{e}_{j-1}\|_{1,j-1}^2 \\
&\le \frac{c^2}{K} \|\mathbf{u}_j - \mathbf{u}_j^{\text{alt}}\|_{1,j}^2 + \rho^4 \|\widehat{\mathbf{e}}_{j-1} - \underbrace{\mathbf{e}_{j-1}^{\text{alt}}}_{=\mathbf{0}}\|_{1,j-1}^2,
\end{aligned}$$

wobei wir hier verwendet haben, dass zwei W-Zyklen aufgerufen werden. Aus

$$\begin{aligned} 0 &\le a(v_j - P_{j-1}v_j, v_j - P_{j-1}v_j) \\ &= a(v_j, v_j) - 2\underbrace{a(v_j, P_{j-1}v_j)}_{=a(P_{j-1}v_j, P_{j-1}v_j)} + a(P_{j-1}v_j, P_{j-1}v_j) \\ &= a(v_j, v_j) - a(P_{j-1}v_j, P_{j-1}v_j) \end{aligned}$$

für alle $v_j \in V_j$ folgt, da $\widehat{e}_{j-1} = P_{j-1}(u_j - u_j^{(\mathrm{pre},K)})$ gemäß Lemma 8.6 gilt, dass

$$a(\widehat{e}_{j-1}, \widehat{e}_{j-1}) \le a(u_j - u_j^{(\mathrm{pre},K)}, u_j - u_j^{(\mathrm{pre},K)}).$$

Dies bedeutet aber

$$\|\widehat{\mathbf{e}}_{j-1}\|_{1,j-1} \le \|\mathbf{u}_j - \mathbf{u}_j^{(\mathrm{pre},K)}\|_{1,j} = \|\mathbf{S}_j^K(\mathbf{u}_j - \mathbf{u}_j^{\mathrm{alt}})\|_{1,j}.$$

Lemma 8.10 liefert dann

$$\|\widehat{\mathbf{e}}_{j-1}\|_{1,j-1} \le \|\mathbf{u}_j - \mathbf{u}_j^{\mathrm{alt}}\|_{1,j}$$

und wir erhalten schließlich

$$\|\mathbf{u}_j - \mathbf{u}_j^{\mathrm{neu}}\|_{1,j}^2 \le \left(\frac{c^2}{K} + \rho^4\right)\|\mathbf{u}_j - \mathbf{u}_j^{\mathrm{alt}}\|_{1,j}^2.$$

Wählen wir $K \ge c^2/\big(\rho^2(1-\rho^2)\big)$, dann folgt $c^2/K + \rho^4 \le \rho^2$ und damit die Behauptung. □

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 $A_j, S_j : V_j \to V_j$ die zu $\mathbf{A}_j, \mathbf{S}_j : \mathbb{R}^{N_j} \to \mathbb{R}^{N_j}$ 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 $E_j : V_j \to V_j$ beziehungsweise $\mathbf{E}_j : \mathbb{R}^{N_j} \to \mathbb{R}^{N_j}$, dies bedeutet, es gilt

$$u_j - u_j^{\mathrm{neu}} = E_j(u_j - u_j^{\mathrm{alt}}) \quad \text{bzw.} \quad \mathbf{u}_j - \mathbf{u}_j^{\mathrm{neu}} = \mathbf{E}_j(\mathbf{u}_j - \mathbf{u}_j^{\mathrm{alt}}).$$

Lemma 8.12 Der symmetrische V-Zyklus erfüllt die Rekursionsformel

$$E_0 = 0, \qquad E_j = S_j^K\big(I - (I - E_{j-1})P_{j-1}\big)S_j^K, \quad j = 1, 2, \ldots.$$

Beweis. Wir beweisen die Aussage mittels vollständiger Induktion. Ist $j = 1$, dann ist die Grobgitterkorrektur exakt und die Behauptung folgt aus Lemma 8.6:

$$E_1 = S_1^K(I - P_0)S_1^K = S_1^K\big(I - (I - E_0)P_0\big)S_1^K.$$

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 $\widehat{e}_{j-1} = P_{j-1}(u_j - u_j^{(\mathrm{pre},K)})$ und aus der Induktionsvoraussetzung $\widehat{e}_{j-1} - e_{j-1} = E_{j-1}(\widehat{e}_{j-1} - 0)$, also

$$e_{j-1} = (I - E_{j-1})\widehat{e}_{j-1} = (I - E_{j-1})P_{j-1}(u_j - u_j^{(\mathrm{pre},K)}).$$

Daraus ergibt sich schließlich die Behauptung:

$$\begin{aligned} u_j - u_j^{\text{neu}} &= S_j^K(u_j - u_j^{(\text{post},0)}) \\ &= S_j^K(u_j - u_j^{(\text{pre},K)} - e_{j-1}) \\ &= S_j^K\big(I - (I - E_{j-1})P_{j-1}\big)(u_j - u_j^{(\text{pre},K)}) \\ &= S_j^K\big(I - (I - E_{j-1})P_{j-1}\big)S_j^K(u_j - u_j^{\text{alt}}) \\ &= E_j(u_j - u_j^{\text{alt}}). \end{aligned}$$

□

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

$$a(P_{j-1}v_j, w_j) = a(v_j, P_{j-1}w_j), \quad a\big((I - P_{j-1})v_j, w_j\big) = a\big(v_j, (I - P_{j-1})w_j\big).$$

Zusammen folgt daher die Symmetrie:

$$\begin{aligned} a(E_j v_j, w_j) &= a\Big(S_j^K\big(I - (I - E_{j-1})P_{j-1}\big)S_j^K v_j, w_j\Big) \\ &= a\big((I - P_{j-1})S_j^K v_j, S_j^K w_j\big) + a(E_{j-1}P_{j-1}S_j^K v_j, S_j^K w_j) \\ &= a\big(S_j^K v_j, (I - P_{j-1})S_j^K w_j\big) + a(S_j^K v_j, E_{j-1}P_{j-1}S_j^K w_j) \\ &= a\Big(v_j, S_j^K\big(I - (I - E_{j-1})P_{j-1}\big)S_j^K w_j\Big) \\ &= a(v_j, E_j w_j). \end{aligned}$$

Schließlich ergibt sich die Nichtnegativität gemäß

$$\begin{aligned} a(E_j v_j, v_j) &= a\Big(S_j^K \big(I - (I - E_{j-1})P_{j-1}\big) S_j^K v_j, v_j\Big) \\ &= a\big((I - P_{j-1})S_j^K v_j, S_j^K v_j\big) + a(E_{j-1}P_{j-1}S_j^K v_j, S_j^K v_j) \\ &= \underbrace{a\big((I - P_{j-1})S_j^K v_j, (I - P_{j-1})S_j^K v_j\big)}_{\geq 0 \text{ da } a \text{ elliptisch}} + \underbrace{a(E_{j-1}P_{j-1}S_j^K v_j, P_{j-1}S_j^K v_j)}_{\geq 0 \text{ nach Induktionsannahme}} \\ &\geq 0. \end{aligned}$$

□

Lemma 8.14 Der Iterationsoperator E_j des symmetrischen V-Zyklus mit jeweils K A-priori- und K A-posteriori-Glättungsschritten genügt der Abschätzung

$$a(E_j v_j, v_j) \leq \frac{c^\star}{K + c^\star} a(v_j, v_j) \quad \text{für alle } v_j \in V_j$$

mit einer von j unabhängigen Konstanten $c^\star$.

Beweis. Wir führen den Beweis in drei Schritten.
(i.) Wir zeigen zunächst, dass

$$a\big((I - S_j)S_j^{2K} v_j, v_j\big) \leq \frac{1}{2K} a\big((I - S_j^{2K})v_j, v_j\big) \tag{8.3}$$

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

$$\begin{aligned} a\big((I - S_j)S_j^\ell v_j, v_j\big) &= \mathbf{v}_j^T \mathbf{A}_j (\underbrace{\mathbf{I} - \mathbf{S}_j}_{=\alpha_j \mathbf{A}_j}) \mathbf{S}_j^\ell \mathbf{v}_j \\ &= \alpha_j \mathbf{v}_j^T \mathbf{A}_j^2 \mathbf{S}_j^\ell \mathbf{v}_j \\ &= \alpha_j \sum_{k=1}^{N_j} \sigma_k^2 \lambda_k^2 (\underbrace{1 - \alpha_j \lambda_k}_{\leq 1})^\ell \\ &\leq \alpha_j \sum_{k=1}^{N_j} \sigma_k^2 \lambda_k^2 (1 - \alpha_j \lambda_k)^{\ell - 1} \\ &= a\big((I - S_j)S_j^{\ell-1} v_j, v_j\big). \end{aligned}$$

Daraus ergibt sich dann mit Hilfe einer Teleskopsumme die Zwischenbehauptung (8.3)

$$\begin{aligned} a\big((I - S_j)S_j^{2K} v_j, v_j\big) &= \frac{1}{2K} \Big\{ \underbrace{a\big((I - S_j)S_j^{2K} v_j, v_j\big) + \cdots + a\big((I - S_j)S_j^{2K} v_j, v_j\big)}_{2K\text{-mal}} \Big\} \\ &\leq \frac{1}{2K} \sum_{\ell=0}^{2K-1} a\big((I - S_j)S_j^\ell v_j, v_j\big) \\ &= \frac{1}{2K} a\big((I - S_j^{2K})v_j, v_j\big). \end{aligned}$$

(ii.) Die Stetigkeit der Bilinearform $a(\cdot,\cdot)$ liefert zusammen mit der Approximationseigenschaft (Satz 8.8) und (8.1), dass

$$\begin{aligned} a\big((I-P_{j-1})S_j^K v_j,(I-P_{j-1})S_j^K v_j\big) &\le c_S\|(I-P_{j-1})S_j^K v_j\|^2_{H^1(\Omega)} \\ &\le \frac{c_S}{\underline{c}}\|\!|\mathbf{S}_j^K\mathbf{v}_j|\!\|^2_{2,j} \\ &= \frac{c_S}{\underline{c}}\mathbf{v}_j^T\mathbf{S}_j^K\mathbf{A}_j^2\mathbf{S}_j^K\mathbf{v}_j. \end{aligned}$$

Aus der Beziehung $\mathbf{A}_j=(\mathbf{I}-\mathbf{S}_j)/\alpha_j$ folgt weiter

$$\begin{aligned} a\big((I-P_{j-1})S_j^K v_j,(I-P_{j-1})S_j^K v_j\big) &\le \frac{c_S}{\alpha_j\underline{c}}\mathbf{v}_j^T\mathbf{S}_j^K(\mathbf{I}-\mathbf{S}_j)\mathbf{A}_j\mathbf{S}_j^K\mathbf{v}_j \\ &= \frac{c_S}{\alpha_j\underline{c}}a\big((I-S_j)S_j^K v_j,S_j^K v_j\big) \\ &= \frac{c_S}{\alpha_j\underline{c}}a\big((I-S_j)S_j^{2K} v_j,v_j\big). \end{aligned}$$

Mit (8.3) erhalten wir daher die Abschätzung

$$a\big((I-P_{j-1})S_j^K v_j,(I-P_{j-1})S_j^K v_j\big) \le \frac{1}{K}\underbrace{\frac{c_S}{2\alpha_j\underline{c}}}_{=:c^\star}a\big((I-S_j^{2K})v_j,v_j\big).$$

(iii.) Der restliche Beweis geschieht mittels vollständiger Induktion. Für $j=0$ ist die Behauptung wegen $a(E_0v_0,v_0)=0$ trivialerweise erfüllt. Wir zeigen nun den Induktionsschritt $j-1\mapsto j$. Wegen der Rekursionsformel aus Lemma 8.12 gilt

$$a(E_jv_j,v_j) = \underbrace{a\big((I-P_{j-1})S_j^Kv_j,S_j^Kv_j\big)}_{=a((I-P_{j-1})S_j^Kv_j,(I-P_{j-1})S_j^Kv_j)} + \underbrace{a(E_{j-1}P_{j-1}S_j^Kv_j,S_j^Kv_j)}_{=a(E_{j-1}P_{j-1}S_j^Kv_j,P_{j-1}S_j^Kv_j)}.$$

Daher folgt aus der Induktionsannahme

$$\begin{aligned} a(E_jv_j,v_j) &\le a\big((I-P_{j-1})S_j^Kv_j,(I-P_{j-1})S_j^Kv_j\big) \\ &\qquad + \frac{c^\star}{K+c^\star}\underbrace{a(P_{j-1}S_j^Kv_j,P_{j-1}S_j^Kv_j)}_{=a(S_j^Kv_j,S_j^Kv_j)-a((I-P_{j-1})S_j^Kv_j,(I-P_{j-1})S_j^Kv_j)} \\ &= \left(1-\frac{c^\star}{K+c^\star}\right)a\big((I-P_{j-1})S_j^Kv_j,(I-P_{j-1})S_j^Kv_j\big) \\ &\qquad + \frac{c^\star}{K+c^\star}a(S_j^Kv_j,S_j^Kv_j), \end{aligned}$$

und schließlich mit *(ii.)* die Behauptung:

$$\begin{aligned} a(E_jv_j,v_j) &\le \underbrace{\left(1-\frac{c^\star}{K+c^\star}\right)\frac{c^\star}{K}}_{=c^\star/(K+c^\star)}a\big((I-S_j^{2K})v_j,v_j\big) + \frac{c^\star}{K+c^\star}\underbrace{a(S_j^Kv_j,S_j^Kv_j)}_{=a(S_j^{2K}v_j,v_j)} \\ &= \frac{c^\star}{K+c^\star}a(v_j,v_j). \end{aligned}$$

□

Satz 8.15 (Konvergenz des V-Zyklus) Der symmetrische V-Zyklus mit jeweils K A-priori- und K A-posteriori-Glättungsschritten erfüllt

$$\|\mathbf{u}_j - \mathbf{u}_j^{\text{neu}}\|_{1,j} \leq \frac{c^\star}{K + c^\star} \|\mathbf{u}_j - \mathbf{u}_j^{\text{alt}}\|_{1,j}$$

mit einer von j unabhängigen Konstanten $c^\star$.

Beweis. Zu zeigen ist

$$a(u_j - u_j^{\text{neu}}, u_j - u_j^{\text{neu}}) \leq \left(\frac{c^\star}{K + c^\star}\right)^2 a(u_j - u_j^{\text{alt}}, u_j - u_j^{\text{alt}}).$$

Nach Lemma 8.13 existieren N_j Eigenpaare $\{(\mu_k, z_k)\}$ von E_j bezüglich des Innenprodukts $a(\cdot, \cdot)$ mit

$$E_j z_k = \mu_k z_k \quad \text{und} \quad a(z_k, z_\ell) = \delta_{k,\ell} \quad \text{für alle } k, \ell = 1, 2, \ldots, N_j.$$

Lemma 8.14 impliziert

$$\mu_k = a(E_j z_k, z_k) \leq \frac{c^\star}{K + c^\star} a(z_k, z_k) = \frac{c^\star}{K + c^\star} \quad \text{für alle } k = 1, 2, \ldots, N_j$$

und folglich für alle $v_j = \sum_{k=1}^{N_j} \sigma_k z_k \in V_j$

$$\begin{aligned} a(E_j v_j, E_j v_j) &= a\left(\sum_{k=1}^{N_j} \sigma_k E_j z_k, \sum_{k=1}^{N_j} \sigma_k E_j z_k\right) \\ &= \sum_{k=1}^{N_j} \sigma_k^2 \mu_k^2 \\ &\leq \left(\frac{c^\star}{K + c^\star}\right)^2 \sum_{k=1}^{N_j} \sigma_k^2 \\ &= \left(\frac{c^\star}{K + c^\star}\right)^2 a(v_j, v_j). \end{aligned}$$

□

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:

$$R_j \geq c|\log(h_j)|.$$

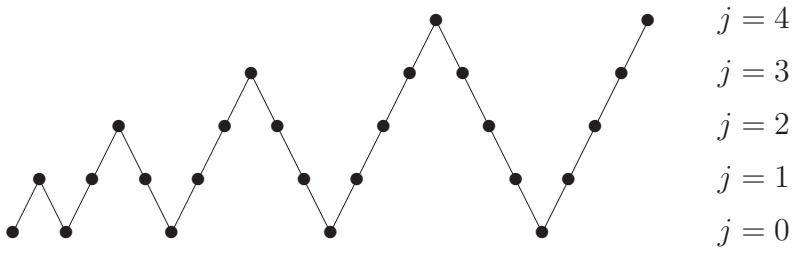
Dabei ist beispielsweise $\sigma = 1$ bei linearen Finiten Elementen. Wie wir sehen werden, kann mit Hilfe der geschachtelten Iteration lineare Komplexität erzielt werden.

Ausgehend von der gröbsten 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 $\widehat{\mathbf{u}}_J \approx \mathbf{A}_J^{-1}\mathbf{f}_J$

① setze $\widehat{\mathbf{u}}_0 := \mathbf{A}_0^{-1}\mathbf{f}_0$
② für alle $j = 1, 2, \ldots, J$
setze $\widehat{\mathbf{u}}_j^{(0)} := \mathbf{I}_{j-1}^{j}\widehat{\mathbf{u}}_{j-1}$
für alle $r = 1, 2, \ldots, R$ berechne $\widehat{\mathbf{u}}_j^{(r)} := \texttt{V-Zyklus}(\widehat{\mathbf{u}}_j^{(r-1)}, \mathbf{f}_j)$
setze $\widehat{\mathbf{u}}_j := \widehat{\mathbf{u}}_j^{(R)}$

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. △

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 Polygongebiet 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

$$\|\!|\mathbf{u}_j - \widehat{\mathbf{u}}_j|\!\|_{1,j} \le ch_j \|f\|_{L^2(\Omega)}$$

mit einer von j unabhängigen Konstanten, vorausgesetzt R ist groß genug.

Beweis. Mit $\gamma = c^\star/(K + c^\star)$ folgt aus Satz 8.15

$$\begin{aligned}
\|\!|\mathbf{u}_j - \widehat{\mathbf{u}}_j|\!\|_{1,j} &\le \gamma^R \|\!|\mathbf{u}_j - \mathbf{I}_{j-1}^j \widehat{\mathbf{u}}_{j-1}|\!\|_{1,j} \\
&\le \gamma^R \big\{ \|\!|\mathbf{u}_j - \mathbf{I}_{j-1}^j \mathbf{u}_{j-1}|\!\|_{1,j} + \|\!|\mathbf{I}_{j-1}^j (\mathbf{u}_{j-1} - \widehat{\mathbf{u}}_{j-1})|\!\|_{1,j} \big\} \\
&= \gamma^R \Big\{ \sqrt{a(u_j - u_{j-1}, u_j - u_{j-1})} + \|\!|\mathbf{u}_{j-1} - \widehat{\mathbf{u}}_{j-1}|\!\|_{1,j-1} \Big\} \\
&\le \gamma^R \big\{ \sqrt{c_S} \big(\|u - u_j\|_{H^1(\Omega)} + \|u - u_{j-1}\|_{H^1(\Omega)} \big) + \|\!|\mathbf{u}_{j-1} - \widehat{\mathbf{u}}_{j-1}|\!\|_{1,j-1} \big\}.
\end{aligned}$$

Satz 6.4 liefert $\|u - u_j\|_{H^1(\Omega)} \le ch_j \|f\|_{L^2(\Omega)}$ und $\|u - u_{j-1}\|_{H^1(\Omega)} \le 2ch_j \|f\|_{L^2(\Omega)}$, das heißt

$$\|\!|\mathbf{u}_j - \widehat{\mathbf{u}}_j|\!\|_{1,j} \le c\gamma^R \big\{ h_j \|f\|_{L^2(\Omega)} + \|\!|\mathbf{u}_{j-1} - \widehat{\mathbf{u}}_{j-1}|\!\|_{1,j-1} \big\}.$$

Aufsummieren und Verwenden von $\mathbf{u}_0 - \widehat{\mathbf{u}}_0 = \mathbf{0}$ ergibt

$$\begin{aligned}
\|\!|\mathbf{u}_j - \widehat{\mathbf{u}}_j|\!\|_{1,j} &\le \big\{ c\gamma^R h_j + c^2 \gamma^{2R} h_{j-1} + \cdots + c^{j+1} \gamma^{(j+1)R} h_0 \big\} \|f\|_{L^2(\Omega)} \\
&= c\gamma^R h_j \big\{ 1 + 2c\gamma^R + \cdots + 2^j c^j \gamma^{jR} \big\} \|f\|_{L^2(\Omega)} \\
&\le \frac{c\gamma^R h_j}{1 - 2c\gamma^R} \|f\|_{L^2(\Omega)},
\end{aligned}$$

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

$$W_0 = cN_0, \qquad W_j \le c(K + L)N_j + W_{j-1}, \quad j \in \mathbb{N}.$$

Da $\dim V_j \sim 2^{jd}$ gilt, ist

$$W_j \le c(K + L)\big\{ N_j + N_{j-1} + \cdots + N_0 \big\} \le c(K + L) \frac{N_j}{2^d - 1} \le cN_j,$$

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

$$\widehat{W}_0 = W_0, \qquad \widehat{W}_j = \widehat{W}_{j-1} + RW_j, \quad j \in \mathbb{N}.$$

Wegen $W_j \le cN_j$ bedeutet dies

$$\widehat{W}_j \le cR\big\{ N_j + N_{j-1} + \cdots + N_0 \big\} \le cN_j.$$

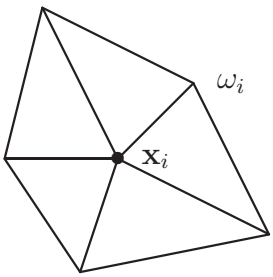
□

9. Residuale Fehlerschätzer

9.1 Cléments-Operator

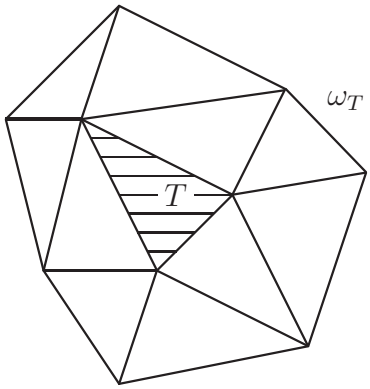
Der Einsatz des Interpolationsoperators I_h in Satz 5.16 setzt H^2-Funktionen voraus. Von P. Cléments stammt ein Approximationsprozess, mit dem auch H^1-Funktionen erfasst werden. Dazu sei $\mathcal{T}$ eine Triangulierung von Ω. Zu jedem Gitterknoten $\mathbf{x}_i$ definieren wir die Vereinigung der 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

$$C_h : H^1(\Omega) \to V_h, \quad C_h v(\mathbf{x}) := \sum_{i=1}^{N} (P_i v)\varphi_i(\mathbf{x}).$$

Hierbei bezeichnet $P_i : L^2(\omega_i) \to \mathcal{P}_0$ 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 $e \subset \partial T$

$$\|v - C_h v\|_{H^m(T)} \le c h_T^{1-m} \|v\|_{H^1(\omega_T)}, \quad m = 0, 1$$
$$\|v - C_h v\|_{L^2(e)} \le c h_T^{1/2} \|v\|_{H^1(\omega_T)}.$$

Beweis. Das Bramble-Hilbert-Lemma impliziert

$$\|v - P_i v\|_{H^1(\omega_i)} \le c \|v\|_{H^1(\omega_i)} \tag{9.1}$$

und mit einem Skalierungsargument

$$\|v - P_i v\|_{L^2(\omega_i)} \le c \operatorname{diam}(\omega_i) \|v\|_{H^1(\omega_i)} \le c h_T \|v\|_{H^1(\omega_i)}. \tag{9.2}$$

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:

$$\|v - P_i v\|_{H^m(T)} \le \|v - P_i v\|_{H^m(\omega_i)} \le c h_T^{1-m} \|v\|_{H^1(\omega_i)} \le c h_T^{1-m} \|v\|_{H^1(\omega_T)}, \quad m = 0, 1.$$

Da $C_h v$ punktweise eine Konvexkombination darstellt, folgt die erste Behauptung. Die zweite Behauptung ergibt sich aus einer verschärften Form des Spursatzes. □

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

$$R_e := R_e(u_h) := \left[\frac{\partial u_h}{\partial \mathbf{n}}\right] = \left.\frac{\partial u_h}{\partial \mathbf{n}}\right|_{T_1} - \left.\frac{\partial u_h}{\partial \mathbf{n}}\right|_{T_2} \quad \text{für } e = T_1 \cap T_2 \in \mathcal{E}.$$

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

$$\eta_{T,R}^2 := h_T^2 \|R_T\|_{L^2(T)}^2 + \frac{1}{2} \sum_{e \in \mathcal{E}_T} h_e \|R_e\|_{L^2(e)}^2,$$

die wir zu einem globalen Fehlerschätzer zusammenbauen

$$\eta_R^2 := \sum_{T \in \mathcal{T}} \eta_{T,R}^2 = \sum_{T \in \mathcal{T}} h_T^2 \|R_T\|_{L^2(T)} + \sum_{e \in \mathcal{E}} h_e \|R_e\|_{L^2(e)}^2.$$

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

$$|u - u_h|_{H^1(\Omega)} = \sup_{w \in H_0^1(\Omega) : |w|_{H^1(\Omega)} = 1} \big(\nabla(u - u_h), \nabla w\big)_{L^2(\Omega)}. \tag{9.3}$$

Für ein spezielles $w \in H_0^1(\Omega)$ folgt

$$\begin{aligned}
\ell(w) &:= \big(\nabla(u - u_h), \nabla w\big)_{L^2(\Omega)} \\
&= (f, w)_{L^2(\Omega)} - \sum_{T \in \mathcal{T}} (\nabla u_h, \nabla w)_{L^2(T)} \\
&= (f, w)_{L^2(\Omega)} - \sum_{T \in \mathcal{T}} \left\{ (-\Delta u_h, w)_{L^2(T)} + \sum_{e \in \mathcal{E}_T} \Big(\frac{\partial u_h}{\partial \mathbf{n}}, w \Big)_{L^2(e)} \right\} \\
&= \sum_{T \in \mathcal{T}} (\Delta u_h + f, w)_{L^2(T)} + \sum_{e \in \mathcal{E}} \Big(\Big[\frac{\partial u_h}{\partial \mathbf{n}} \Big], w \Big)_{L^2(e)} \\
&= \sum_{T \in \mathcal{T}} (R_T, w)_{L^2(T)} + \sum_{e \in \mathcal{E}} (R_e, w)_{L^2(e)} \\
&= \sum_{T \in \mathcal{T}} \left\{ (R_T, w)_{L^2(T)} + \frac{1}{2} \sum_{e \in \mathcal{E}_T} (R_e, w)_{L^2(e)} \right\}.
\end{aligned}$$

Wir setzen nun $w_h := C_h w$ und benutzen die Galerkin-Orthogonalität

$$\big(\nabla(u - u_h), \nabla v_h\big)_{L^2(\Omega)} = 0 \quad \text{für alle } v_h \in V_h.$$

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

$$\begin{aligned}\ell(w) &\le c\sum_{T\in\mathcal{T}}\underbrace{\left\{h_T\|R_T\|_{L^2(T)}+\frac{1}{2}\sum_{e\in\mathcal{E}_T}h_e^{1/2}\|R_e\|_{L^2(e)}\right\}}_{|\cdot|\le c\eta_{T,R}}\|w\|_{H^1(\omega_T)}\\ &\le c\left(\sum_{T\in\mathcal{T}}\eta_{T,R}^2\right)^{1/2}\left(\sum_{T\in\mathcal{T}}\|w\|_{H^1(\omega_T)}^2\right)^{1/2}\\ &\le c\eta_R\|w\|_{H^1(\Omega)}.\end{aligned}$$

Diese Abschätzung eingesetzt in das Dualitätsargument (9.3) liefert mit der Friedrichsschen Ungleichung die Behauptung. □

Bemerkung Das flächenbezogene Residuum ist in dieser Form numerisch nicht berechnenbar, 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

$$\|\Delta u_h + f\|_{L^2(T)} \le \|\Delta u_h + f_h\|_{L^2(T)} + \|f - f_h\|_{L^2(T)}$$

erhält man die Schranke

$$\|u-u_h\|_{H^1(\Omega)} \le c\left\{\left(\sum_{T\in\mathcal{T}}\eta_{T,R}^2\right)^{1/2}+\left(\sum_{T\in\mathcal{T}}h_T^2\|f-f_h\|_{L^2(T)}^2\right)^{1/2}\right\}.$$

Der letzte Summand wird dabei *Datenoszillation* genannt. △

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*

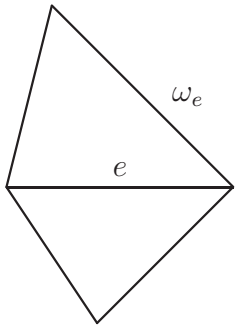
$$\operatorname{supp}\psi_T = T, \quad \psi_T \in \mathcal{P}_3, \quad 0 \le \psi_T \le 1 = \max\psi_T.$$

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

$$\operatorname{supp}\psi_e = T_1 \cup T_2, \quad \psi_e\big|_{T_i} \in \mathcal{P}_2, \quad 0 \le \psi_e \le 1 = \max\psi_e.$$

Ferner setzen wir

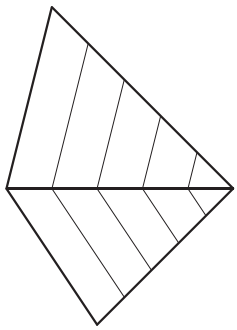
$$\omega_e := \bigcup_{T\in\mathcal{T}:e\subset\mathcal{E}_T} T$$



und definieren eine Abbildung $E : L^2(e) \to L^2(\omega_e)$, die jede auf einer Kante e definierte Funktion auf die benachbarten Dreiecke ω_e fortsetzt. Wir setzen

$$\big(E(\sigma)\big)(\mathbf{x}) = \sigma(\mathbf{z}) \text{ in } T, \text{ wenn } \mathbf{z} \in e \text{ der Punkt aus } e \text{ mit } \lambda_j(\mathbf{x}) = \lambda_j(\mathbf{z}) \text{ ist.}$$

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:

$$\|\psi_T v\|_{L^2(T)} \le \|v\|_{L^2(T)} \qquad \text{für alle } v \in L^2(T) \tag{9.4}$$

$$\|\psi_T^{1/2} p\|_{L^2(T)} \ge c\|p\|_{L^2(T)} \qquad \text{für alle } p \in \mathcal{P}_2 \tag{9.5}$$

$$|\psi_T p|_{H^1(T)} \le c h_T^{-1} \|\psi_T p\|_{L^2(T)} \qquad \text{für alle } p \in \mathcal{P}_2 \tag{9.6}$$

$$\|\psi_e^{1/2} \sigma\|_{L^2(e)} \ge c\|\sigma\|_{L^2(e)} \qquad \text{für alle } \sigma \in \mathcal{P}_2 \tag{9.7}$$

$$c h_e^{1/2} \|\sigma\|_{L^2(e)} \le \|\psi_e E(\sigma)\|_{L^2(T)} \le C h_e^{1/2} \|\sigma\|_{L^2(e)} \qquad \text{für alle } \sigma \in \mathcal{P}_2 \tag{9.8}$$

$$|\psi_e E(\sigma)|_{H^1(T)} \le c h_T^{-1} \|\psi_e E(\sigma)\|_{L^2(T)} \qquad \text{für alle } \sigma \in \mathcal{P}_2 \tag{9.9}$$

Beweis. Die Abschätzung (9.4) folgt sofort aus $0 \le \psi_T \le 1$. 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. □

Satz 9.4 Sei $\mathcal{T}$ eine nicht entartete Triangulierung mit Regularitätsparameter κ und sei

$$\widetilde{\omega}_T := \bigcup_{e \in \mathcal{E}_T} \omega_e.$$

Dann gibt es eine Konstante $c = c(\Omega, \kappa)$, so dass

$$\eta_{T,R} \le c\Big\{ \|u - u_h\|^2_{H^1(\widetilde{\omega}_T)} + \sum_{T' \in \omega_T} h^2_{T'} \|f - f_h\|^2_{L^2(T')} \Big\}^{1/2}, \tag{9.10}$$

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

$$R_{T,\mathrm{red}} := R_{T,\mathrm{red}}(u_h) := \Delta u_h + f_h \in \mathcal{P}_2$$

und setzen

$$w := w_T := \psi_T \cdot R_{T,\mathrm{red}}.$$

Mit (9.5) folgt dann

$$\begin{aligned} c\|R_{T,\mathrm{red}}\|^2_{L^2(T)} &\le \|\psi_T^{1/2} R_{T,\mathrm{red}}\|^2_{L^2(T)} \\ &= (R_{T,\mathrm{red}}, w)_{L^2(T)} \\ &= \big(\Delta(u_h - u), w\big)_{L^2(T)} + (f_h - f, w)_{L^2(T)} \\ &= \big(\nabla(u - u_h), \nabla w\big)_{L^2(T)} + (f_h - f, w)_{L^2(T)} \\ &\le |u - u_h|_{H^1(T)} |w|_{H^1(T)} + \|f - f_h\|_{L^2(T)} \|w\|_{L^2(T)}. \end{aligned}$$

Man beachte, dass wegen (9.6) gilt $|w|_{H^1(T)} \le c h_T^{-1} \|w\|_{L^2(T)}$ und wegen (9.4) gilt $\|w\|_{L^2(T)} \le \|R_{T,\mathrm{red}}\|_{L^2(T)}$. Daher schließen wir, dass

$$\|R_{T,\mathrm{red}}\|^2_{L^2(T)} \le c\big\{ h_T^{-1} |u - u_h|_{H^1(T)} \|R_{T,\mathrm{red}}\|_{L^2(T)} + \|f - f_h\|_{L^2(T)} \|R_{T,\mathrm{red}}\|_{L^2(T)} \big\},$$

das heißt

$$\|R_{T,\mathrm{red}}\|_{L^2(T)} \le c\big\{ h_T^{-1} |u - u_h|_{H^1(T)} + \|f - f_h\|_{L^2(T)} \big\}.$$

Wegen

$$\begin{aligned} \|R_T\|_{L^2(T)} &= \|\Delta u_h + f\|_{L^2(T)} \\ &\le \|\Delta u_h + f_h\|_{L^2(T)} + \|f - f_h\|_{L^2(T)} \\ &= \|R_{T,\mathrm{red}}\|_{L^2(T)} + \|f - f_h\|_{L^2(T)} \end{aligned}$$

ergibt sich hieraus die Abschätzung

$$h_T \|R_T\|_{L^2(T)} \le c\big\{ |u - u_h|_{H^1(T)} + h_T \|f - f_h\|_{L^2(T)} \big\}. \tag{9.11}$$

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 $\operatorname{supp} w = \omega_e$. Außerdem ist $R_e \in \mathcal{P}_2$. Mit (9.7) erhalten wir

$$\begin{aligned} c\|R_e\|^2_{L^2(e)} &\le \|\psi_e^{1/2} R_e\|^2_{L^2(e)} \\ &= (R_e, w)_{L^2(e)} \\ &= \big(\nabla(u_h - u), \nabla w\big)_{L^2(\omega_e)} + \sum_{T\in\omega_e} (\Delta u_h + f, w)_{L^2(T)} \\ &= \big(\nabla(u_h - u), \nabla w\big)_{L^2(\omega_e)} + \sum_{T\in\omega_e} (R_T, w)_{L^2(T)} \\ &\le |u - u_h|_{H^1(\omega_e)} |w|_{H^1(\omega_e)} + \sum_{T\in\omega_e} \|R_T\|_{L^2(T)} \|w\|_{L^2(T)}. \end{aligned}$$

Mit (9.9) schließen wir $\|w\|_{H^1(T)} \le c h_T^{-1} \|w\|_{L^2(T)} \le c h_e^{-1} \|w\|_{L^2(T)}$ für $T \in \omega_e$ und mit (9.8) folgt $\|w\|_{L^2(T)} \le C h_e^{1/2} \|R_e\|_{L^2(e)}$. Also ergibt sich

$$\|R_e\|^2_{L^2(e)} \le c h_e^{-1/2} |u - u_h|_{H^1(\omega_e)} \|R_e\|_{L^2(e)} + c h_e^{1/2} \sum_{T\in\omega_e} \|R_T\|_{L^2(T)} \|R_e\|_{L^2(e)},$$

das heißt

$$h_e^{1/2} \|R_e\|_{L^2(e)} \le c \bigg\{ |u - u_h|_{H^1(\omega_e)} + h_e \sum_{T\in\omega_e} \|R_T\|_{L^2(T)} \bigg\}.$$

Wegen $h_e \le h_T$ für $T \in \omega_e$ erhalten wir im Hinblick auf (9.11)

$$h_e^{1/2} \|R_e\|_{L^2(e)} \le c \bigg\{ |u - u_h|_{H^1(\omega_e)} + \sum_{T\in\omega_e} h_T \|f - f_h\|_{L^2(T)} \bigg\}. \tag{9.12}$$

Wenn wir noch $\widetilde{\omega}_T = \bigcup_{e\in\mathcal{E}_T} \omega_e$ beachten, folgt aus (9.11) und (9.12) schließlich die Behauptung. □

Bemerkung Aufsummation der lokalen unteren Fehlerschranke (9.10) liefert die Abschätzung

$$\eta_R^2 \le c \bigg\{ \|u - u_h\|^2_{H^1(\Omega)} + \sum_{T\in\mathcal{T}} h_T^2 \|f - f_h\|^2_{L^2(T)} \bigg\},$$

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. △

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 $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ zur Differentialgeichung

$$-\Delta u + \langle \mathbf{b}, \nabla u \rangle + u = f \text{ in } \Omega, \quad u = 0 \text{ auf } \Gamma$$

auf dem beschränkten Gebiet $\Omega \subset \mathbb{R}^d$ lautet

$$a(u,v) = \int_\Omega \left\{ \langle \nabla u, \nabla v \rangle + \langle \mathbf{b}, \nabla u \rangle v + uv \right\} \mathrm{d}\mathbf{x}.$$

Diese unsymmetrische Bilinearform ist offensichtlich stetig. Wegen $|\langle \mathbf{b}, \nabla u \rangle| \leq \|\nabla u\|$ ist sie auch elliptisch:

$$\begin{aligned} a(u,u) &\geq \int_\Omega \left\{ \|\nabla u\|^2 - \|\nabla u\||u| + u^2 \right\} \mathrm{d}\mathbf{x} \\ &= \frac{1}{2} \int_\Omega \left\{ \|\nabla u\|^2 + u^2 \right\} \mathrm{d}\mathbf{x} + \frac{1}{2} \int_\Omega \underbrace{\left\{ \|\nabla u\| - |u| \right\}^2}_{\geq 0} \mathrm{d}\mathbf{x} \\ &\geq \frac{1}{2} \|u\|_{H^1(\Omega)}^2. \end{aligned}$$

△

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 $a : V \times V \to \mathbb{R}$, sowie ein beschränktes lineares Funktional $\ell : V \to \mathbb{R}$. 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

$$A : V \to V', \quad v \mapsto A_v$$

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 Rieszschen 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 Abbildung $\tau : V' \to V$ 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

$$T : V \to V, \quad v \mapsto Tv = v - \alpha(\tau A_v - \tau\ell).$$

Ist $\alpha > 0$ so gewählt, dass T eine Kontraktion darstellt, das heißt $\|T\|_{V \to V} = \gamma < 1$, 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. □

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. △

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

$$\frac{\partial}{\partial t}\big(u(t),v\big)_{L^2(\Omega)} + \big(\nabla u(t),\nabla v\big)_{L^2(\Omega)} = \big(f(t),v\big)_{L^2(\Omega)} \quad \text{für alle } v \in H_0^1(\Omega)$$

liefert die Einschränkung auf den Finite-Elemente-Raum $V_h \subset H_0^1(\Omega)$

suche $u_h(t) \in V_h$, so dass

$$\frac{\partial}{\partial t}\big(u_h(t),v_h\big)_{L^2(\Omega)} + \big(\nabla u_h(t),\nabla v_h\big)_{L^2(\Omega)} = \big(f(t),v_h\big)_{L^2(\Omega)} \quad \text{für alle } v_h \in V_h$$

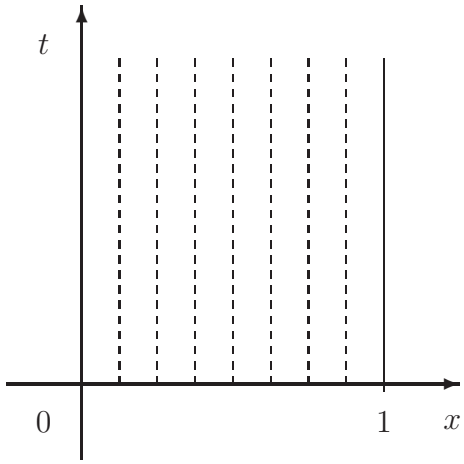
ein lineares Gleichungssystem

$$\mathbf{M}\frac{\partial}{\partial t}\mathbf{u}(t) + \mathbf{A}\mathbf{u}(t) = \mathbf{f}(t), \quad \mathbf{u}(0) = \mathbf{g}. \tag{11.1}$$

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 $\mathbf{f}(t) = \big[\big(f(t),\varphi_i\big)_{L^2(\Omega)}\big]_{i,j}$ 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 Trinagulierung $\mathcal{T}_h$. $\triangle$



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

$$\mathbf{M}\frac{\mathbf{u}_{i+1} - \mathbf{u}_i}{k_i} + \mathbf{A}\mathbf{u}_i = \mathbf{f}(t_i) \quad \text{bzw.} \quad \mathbf{M}\mathbf{u}_{i+1} = (\mathbf{M} - k_i\mathbf{A})\mathbf{u}_i + k_i\mathbf{f}(t_i). \tag{11.2}$$

Hingegen liefert das *implizite Euler-Verfahren*

$$\mathbf{M}\frac{\mathbf{u}_{i+1} - \mathbf{u}_i}{k_i} + \mathbf{A}\mathbf{u}_{i+1} = \mathbf{f}(t_{i+1}) \quad \text{bzw.} \quad (\mathbf{M} + k_i\mathbf{A})\mathbf{u}_{i+1} = \mathbf{M}\mathbf{u}_i + k_i\mathbf{f}(t_{i+1}). \tag{11.3}$$

Der Startwert lautet in beiden Fällen $\mathbf{u}_0 = \mathbf{g}$. Kombinieren wir (11.2) und (11.3), so erhalten wir das *θ-Schema*

$$\mathbf{M}\frac{\mathbf{u}_{i+1} - \mathbf{u}_i}{k_i} + (1-\theta)\mathbf{A}\mathbf{u}_i + \theta\mathbf{A}\mathbf{u}_{i+1} = (1-\theta)\mathbf{f}(t_i) + \theta\mathbf{f}(t_{i+1}), \tag{11.4}$$

beziehungsweise

$$(\mathbf{M} + k_i\theta\mathbf{A})\mathbf{u}_{i+1} = \big(\mathbf{M} - k_i(1-\theta)\mathbf{A}\big)\mathbf{u}_i + k_i(1-\theta)\mathbf{f}(t_i) + k_i\theta\mathbf{f}(t_{i+1}).$$

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

$$\|u_{h,M}\|^2_{L^2(\Omega)} + \sum_{i=0}^{M-1} \left\{ k_i |u_{h,i+\theta}|^2_{H^1(\Omega)} + (2\theta - 1)\|u_{h,i+1} - u_{h,i}\|^2_{L^2(\Omega)} \right\} \leq \|u_{h,0}\|^2_{L^2(\Omega)} + c \sum_{i=0}^{M-1} k_i \|f_{i+\theta}\|^2_{L^2(\Omega)},$$

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

$$(u_{h,i+1} - u_{h,i}, u_{h,i+\theta})_{L^2(\Omega)} + k_i \underbrace{(\nabla u_{h,i+\theta}, \nabla u_{h,i+\theta})_{L^2(\Omega)}}_{=|u_{h,i+\theta}|^2_{H^1(\Omega)}} = k_i (f_{i+\theta}, u_{h,i+\theta})_{L^2(\Omega)}, \tag{11.5}$$

vergleiche (11.4). Andererseits gilt

$$\begin{aligned}
&(u_{h,i+1} - u_{h,i}, u_{h,i+\theta})_{L^2(\Omega)} \\
&\quad = \big(u_{h,i+1} - u_{h,i}, (1-\theta)u_{h,i} + \theta u_{h,i+1}\big)_{L^2(\Omega)} \\
&\quad = \left(u_{h,i+1} - u_{h,i}, \frac{1}{2}u_{h,i+1} + \frac{1}{2}u_{h,i} + \left(\theta - \frac{1}{2}\right)(u_{h,i+1} - u_{h,i})\right)_{L^2(\Omega)} \\
&\quad = \frac{1}{2}\|u_{h,i+1}\|^2_{L^2(\Omega)} - \frac{1}{2}\|u_{h,i}\|^2_{L^2(\Omega)} + \left(\theta - \frac{1}{2}\right)\|u_{h,i+1} - u_{h,i}\|^2_{L^2(\Omega)}.
\end{aligned}$$

Hieraus folgt zusammen mit (11.5), der Cauchy-Schwarzschen Ungleichung und der Poincaré-Friedrichsschen Ungleichung

$$\begin{aligned}
&\|u_{h,i+1}\|^2_{L^2(\Omega)} - \|u_{h,i}\|^2_{L^2(\Omega)} + (2\theta - 1)\|u_{h,i+1} - u_{h,i}\|^2_{L^2(\Omega)} + 2k_i |u_{h,i+\theta}|^2_{H^1(\Omega)} \\
&\quad = 2k_i (f_{i+\theta}, u_{h,i+\theta})_{L^2(\Omega)} \\
&\quad \leq 2k_i c_\Omega \|f_{i+\theta}\|_{L^2(\Omega)} |u_{h,i+\theta}|_{H^1(\Omega)} \\
&\quad \leq k_i c_\Omega^2 \|f_{i+\theta}\|^2_{L^2(\Omega)} + k_i |u_{h,i+\theta}|^2_{H^1(\Omega)},
\end{aligned}$$

dies bedeutet

$$\|u_{h,i+1}\|^2_{L^2(\Omega)} - \|u_{h,i}\|^2_{L^2(\Omega)} + (2\theta - 1)\|u_{h,i+1} - u_{h,i}\|^2_{L^2(\Omega)} + k_i |u_{h,i+\theta}|^2_{H^1(\Omega)} \leq k_i c_\Omega^2 \|f_{i+\theta}\|^2_{L^2(\Omega)}.$$

Summation über i liefert die Behauptung. □

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, produziert es aber unphysikalische Oszillationen und sollte daher gemieden werden. Am besten wählt man daher $\theta = 1/2 + \xi$ mit einem kleinen ξ. △

11.3 Fehleranalysis

Wir wollen speziell das implizite Euler-Verfahren (also $\theta = 1$) betrachten, das wir leicht modifizieren, indem wir die rechte Seite entsprechend

$$\overline{f}(t_{i+1}) := \frac{1}{k_i} \int_{t_i}^{t_{i+1}} f(t)\,\mathrm{d}t = f(t_{i+1}) + \mathcal{O}(k_i)$$

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 Polygongebiet 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

$$\|u - u_h\|_{h,\infty} \le c\left\{ \sqrt{T} h^2 k^{-1/2} \|\Delta u\|_{h,\infty} + \left(\sum_{i=0}^{M-1} k_i^2 \int_{t_i}^{t_{i+1}} \left| \frac{\partial u}{\partial t} \right|_{H^1(\Omega)}^2 \mathrm{d}t \right)^{1/2} \right\},$$

vorausgesetzt es ist $\min_{i=1}^{M}\{k_i\} \ge k$.

Beweis. Mit $R_h u_i$ bezeichnen wir die Ritz-Projektion von $u_i := u(t_i)$ auf den Raum V_h, die per Galerkin-Verfahren berechnet wird:

$$\text{suche } R_h u_i \in V_h, \text{ so dass } \quad (\nabla R_h u_i, \nabla v_h)_{L^2(\Omega)} = (\nabla u_i, \nabla v_h)_{L^2(\Omega)} \quad \text{für alle } v_h \in V_h.$$

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

$$\|\xi_i\|_{L^2(\Omega)} = \|(R_h - I)u_i\|_{L^2(\Omega)} \le c h^2 |u_i|_{H_2(\Omega)}. \tag{11.6}$$

Hieraus folgt unmittelbar

$$\|(R_h - I)u\|_{h,\infty} \le c h^2 \|\Delta u\|_{h,\infty}.$$

Für den Fehleranteil η_{i+1} verwenden wir die folgende Identität:

$$\|\eta_{i+1}\|_{L^2(\Omega)}^2 - \|\eta_i\|_{L^2(\Omega)}^2 = 2(\eta_{i+1} - \eta_i, \eta_{i+1})_{L^2(\Omega)} - \|\eta_{i+1} - \eta_i\|_{L^2(\Omega)}^2. \tag{11.7}$$

Für den rechten Term gilt für beliebiges $v_h \in V_h$ unter Ausnutzung der Gleichung für die Ritz-Projektion

$$\begin{aligned}
&(\eta_{i+1} - \eta_i, \eta_{i+1})_{L^2(\Omega)} \\
&\quad = k_i(f_{i+1}, \eta_{i+1})_{L^2(\Omega)} - k_i(\nabla u_{h,i+1}, \nabla \eta_{i+1})_{L^2(\Omega)} - (R_h u_{i+1} - R_h u_i, \eta_{i+1})_{L^2(\Omega)} \\
&\quad = k_i(f_{i+1}, \eta_{i+1})_{L^2(\Omega)} - k_i\big(\nabla(\eta_{i+1} + u_{i+1}), \nabla \eta_{i+1}\big)_{L^2(\Omega)} - (u_{i+1} - u_i, \eta_{i+1})_{L^2(\Omega)} \\
&\qquad - \underbrace{(R_h u_{i+1} - u_{i+1} - R_h u_i + u_i, \eta_{i+1})_{L^2(\Omega)}}_{=(\xi_{i+1} - \xi_i, \eta_{i+1})_{L^2(\Omega)}}.
\end{aligned}$$

Die ersten drei Terme lassen sich unter Verwendung der Wärmeleitungsgleichung beschränken gemäß

$$\begin{aligned}
E_1 &:= k_i(f_{i+1}, \eta_{i+1})_{L^2(\Omega)} - k_i\big(\nabla(\eta_{i+1} + u_{i+1}), \nabla \eta_{i+1}\big)_{L^2(\Omega)} - (u_{i+1} - u_i, \eta_{i+1})_{L^2(\Omega)} \\
&= \int_{t_i}^{t_{i+1}} \left(f - \frac{\partial u}{\partial t}, \eta_{i+1} \right)_{L^2(\Omega)} \mathrm{d}t - k_i(\nabla u_{i+1}, \nabla \eta_{i+1})_{L^2(\Omega)} - k_i \|\nabla \eta_{i+1}\|^2_{L^2(\Omega)} \\
&= \int_{t_i}^{t_{i+1}} \big(\nabla(u - u_{i+1}), \nabla \eta_{i+1}\big)_{L^2(\Omega)} \mathrm{d}t - k_i \|\nabla \eta_{i+1}\|^2_{L^2(\Omega)}.
\end{aligned}$$

Mit dem Hauptsatz der Integralrechnung folgt

$$\int_x^y \{g(z) - g(x)\} \,\mathrm{d}z = \int_x^y \int_x^z g'(t) \,\mathrm{d}t \,\mathrm{d}z.$$

Wegen

$$\{(t,z) : x \le z \le y,\ x \le t \le z\} = \{(t,z) : x \le t \le y,\ t \le z \le y\}$$

folgt daher die Identität

$$\int_x^y \{g(z) - g(x)\} \,\mathrm{d}z = \int_x^y \int_t^y g'(t) \,\mathrm{d}z \,\mathrm{d}t = \int_x^y g'(t)(y - t) \,\mathrm{d}t.$$

Dies eingesetzt liefert

$$\begin{aligned}
E_1 &= \int_{t_i}^{t_{i+1}} (t_i - t) \left(\nabla \frac{\partial u}{\partial t}, \nabla \eta_{i+1} \right)_{L^2(\Omega)} \mathrm{d}t - k_i \|\nabla \eta_{i+1}\|^2_{L^2(\Omega)} \\
&\le \int_{t_i}^{t_{i+1}} (t - t_i) \left| \frac{\partial u}{\partial t} \right|_{H^1(\Omega)} \|\nabla \eta_{i+1}\|_{L^2(\Omega)} \,\mathrm{d}t - k_i \|\nabla \eta_{i+1}\|^2_{L^2(\Omega)} \\
&\le \left(\frac{k_i^2}{2} \int_{t_i}^{t_{i+1}} \left| \frac{\partial u}{\partial t} \right|^2_{H^1(\Omega)} \mathrm{d}t \right)^{1/2} \left(2 \int_{t_i}^{t_{i+1}} \|\nabla \eta_{i+1}\|^2_{L^2(\Omega)} \,\mathrm{d}t \right)^{1/2} - k_i \|\nabla \eta_{i+1}\|^2_{L^2(\Omega)}.
\end{aligned}$$

Mit der Ungleichung vom arithmetischen und geometrischen Mittel, $\sqrt{ab} \le (a+b)/2$ folgt schließlich

$$E_1 \le \frac{k_i^2}{4} \int_{t_i}^{t_{i+1}} \left| \frac{\partial u}{\partial t} \right|^2_{H^1(\Omega)} \mathrm{d}t.$$

Weiter gilt

$$\begin{aligned}
E_2 &:= (\xi_{i+1} - \xi_i, \eta_{i+1})_{L^2(\Omega)} \\
&= (\xi_{i+1}, \eta_{i+1})_{L^2(\Omega)} - (\xi_i, \eta_i)_{L^2(\Omega)} + \underbrace{(\xi_i, \eta_i - \eta_{i+1})_{L^2(\Omega)}}_{\le \|\eta_i - \eta_{i+1}\|_{L^2(\Omega)} \|\xi_i\|_{L^2(\Omega)}} \\
&\le (\xi_{i+1}, \eta_{i+1})_{L^2(\Omega)} - (\xi_i, \eta_i)_{L^2(\Omega)} + \frac{1}{2} \|\eta_i - \eta_{i+1}\|^2_{L^2(\Omega)} + c h^4 |u_i|^2_{H^2(\Omega)},
\end{aligned}$$

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

$$\begin{aligned}\|\eta_{i+1}\|^2_{L^2(\Omega)} - \|\eta_i\|^2_{L^2(\Omega)} &= 2E_1 + 2E_2 - \|\eta_{i+1} - \eta_i\|^2_{L^2(\Omega)} \\ &\leq 2(\xi_{i+1}, \eta_{i+1})_{L^2(\Omega)} - 2(\xi_i, \eta_i)_{L^2(\Omega)} + \frac{k_i^2}{2}\int_{t_i}^{t_{i+1}} \left|\frac{\partial u}{\partial t}\right|^2_{H^1(\Omega)} \mathrm{d}t + ch^4|u_i|^2_{H^2(\Omega)}.\end{aligned}$$

Aufsummation über alle i liefert dann

$$\begin{aligned}\|\eta_M\|^2_{L^2(\Omega)} &\leq \|\eta_0\|^2_{L^2(\Omega)} + 2 \underbrace{(\xi_M, \eta_M)_{L^2(\Omega)}}_{\leq(\sqrt{2}\|\xi_M\|_{L^2(\Omega)})(\|\eta_M\|_{L^2(\Omega)}/\sqrt{2})} - 2 \underbrace{(\xi_0, \eta_0)_{L^2(\Omega)}}_{\leq(\sqrt{2}\|\xi_0\|_{L^2(\Omega)})(\|\eta_0\|_{L^2(\Omega)}/\sqrt{2})} \\ &\quad + \sum_{i=0}^{M-1} \frac{k_i^2}{2}\int_{t_i}^{t_{i+1}} \left|\frac{\partial u}{\partial t}\right|^2_{H^1(\Omega)} \mathrm{d}t + ch^4 \sum_{i=0}^{M-1} |u_i|^2_{H^2(\Omega)} \\ &\leq \|\eta_0\|^2_{L^2(\Omega)} + 2\|\xi_M\|^2_{L^2(\Omega)} + \frac{1}{2}\|\eta_M\|^2_{L^2(\Omega)} + 2\|\xi_0\|^2_{L^2(\Omega)} + \frac{1}{2}\|\eta_0\|^2_{L^2(\Omega)} \\ &\quad + \sum_{i=0}^{M-1} \frac{k_i^2}{2}\int_{t_i}^{t_{i+1}} \left|\frac{\partial u}{\partial t}\right|^2_{H^1(\Omega)} \mathrm{d}t + ch^4 \sum_{i=0}^{M-1} k_i \underbrace{|k_i^{-1/2}u_i|^2_{H^2(\Omega)}}_{\leq\|k^{-1/2}\Delta u\|^2_{h,\infty}} \\ &\leq \frac{3}{2}\|\eta_0\|^2_{L^2(\Omega)} + 2\|\xi_M\|^2_{L^2(\Omega)} + \frac{1}{2}\|\eta_M\|^2_{L^2(\Omega)} + 2\|\xi_0\|^2_{L^2(\Omega)} \\ &\quad + \sum_{i=0}^{M-1} \frac{k_i^2}{2}\int_{t_i}^{t_{i+1}} \left|\frac{\partial u}{\partial t}\right|^2_{H^1(\Omega)} \mathrm{d}t + ch^4 T\|k^{-1/2}\Delta u\|^2_{h,\infty}.\end{aligned}$$

Projizieren wir den Startwert $u_0 = g$ durch eine geeignete Projektion $P_h : H_0^1(\Omega) \to V_h$, 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

$$\begin{aligned}\frac{1}{2}\|\eta_M\|^2_{L^2(\Omega)} &\leq 2\|\xi_M\|^2_{L^2(\Omega)} + 2\|\xi_0\|^2_{L^2(\Omega)} + ch^4(T+1)\|k^{-1/2}\Delta u\|^2_{h,\infty} \\ &\quad + \sum_{i=0}^{M-1} \frac{k_i^2}{2}\int_{t_i}^{t_{i+1}} \left|\frac{\partial u}{\partial t}\right|^2_{H^1(\Omega)} \mathrm{d}t.\end{aligned}$$

Wegen $\|u - u_h\|_{h,\infty} \leq \|\xi\|_{h,\infty} + \|\eta\|_{h,\infty}$ ergibt sich endlich die Behauptung. □

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 $h \leq ck^{4/3}$ 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. △

12. Nichtkonforme Finite Elemente

12.1 Lemmata von Strang

Bisher haben wir stets den Fall betrachtet, dass $V_h \subset V$ gilt. Diese Eigenschaft des Ansatzraums V_h heißt *Konformität*. Ist diese Bedingung verletzt, dann spricht man von *nichtkonformen Elementen*. Neben dem Approximationsfehler entsteht ein weiterer Fehler, der Konsistenzfehler genannt wird. Die Konvergenz ist deshalb keineswegs selbstverständlich. Die Lemmata von Strang schätzen diesen Konsistenzfehler ab.

Das Variationsproblem

$$a(u,v) = \ell(v) \quad \text{für } v \in V \tag{12.1}$$

wird durch eine Folge finiter Probleme ersetzt: Gesucht wird $u_h \in V_h$ mit

$$a_h(u_h,v) = \ell_h(v) \quad \text{für } v \in V_h. \tag{12.2}$$

Wir betrachten zunächst den Fall, dass $V_h \subset V$ ist. Die Bilinearformen a_h seien gleichgradig elliptisch, das heißt, es mit einer von h unabhängigen Zahl α gelte

$$a_h(v,v) \geq \alpha \|v\|_V^2 \quad \text{für } v \in V_h.$$

Man beachte, dass sowohl die Bilinearform a_h, als auch das Funktional ℓ_h, nicht für alle $v \in V$ definiert sein müssen. Beispielsweise können a_h und ℓ_h durch Quadraturformeln aus a und ℓ hervorgehen.

Satz 12.1 (erstes Lemma von Strang) Unter den obigen Voraussetzungen ist mit einer von h unabhängigen Zahl c

$$\|u-u_h\|_V \leq c \inf_{v_h \in V_h} \left\{ \|u-v_h\|_V + \sup_{w_h \in V_h} \frac{|a(v_h,w_h)-a_h(v_h,w_h)|}{\|w_h\|_V} + \sup_{w_h \in V_h} \frac{|\ell(w_h)-\ell_h(w_h)|}{\|w_h\|_V} \right\}.$$

Beweis. Sei $v_h \in V_h$. Wir setzen zur Abkürzung $w_h = u_h - v_h$ und schließen aus der gleichgradigen Elliptizität zusammen mit (12.1) und (12.2)

$$\begin{aligned}
\alpha\|u_h - v_h\|_V^2 &\leq a_h(u_h - v_h, u_h - v_h) \\
&= a_h(u_h - v_h, w_h) \\
&= a(u - v_h, w_h) + \{a(v_h,w_h) - a_h(v_h,w_h)\} + \{a_h(u_h,w_h) - a(u,w_h)\} \\
&= a(u - v_h, w_h) + \{a(v_h,w_h) - a_h(v_h,w_h)\} + \{\ell_h(w_h) - \ell(w_h)\}.
\end{aligned}$$

Wir dividieren durch $\|w_h\|_V = \|u_h - v_h\|_V$ und nutzen die Stetigkeit von a aus:

$$\|u_h - v_h\|_V \le \frac{c_S}{\alpha}\left(\|u - v_h\|_V + \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|_V} + \frac{|\ell(w_h) - \ell_h(w_h)|}{\|w_h\|_V}\right).$$

Es ist v_h ein beliebiges Element aus V_h, so dass zusammen mit der Dreiecksungleichung

$$\|u - u_h\|_V \le \|u - v_h\|_V + \|u_h - v_h\|_V$$

die Behauptung folgt mit $c = 1 + c_S/\alpha$. □

Beim Verzicht auf die Konformitätsbedingung $V_h \subset V$ muss im allgemeinen eine gitterabhängige Norm $\|\cdot\|_{V_h}$ betrachtet werden. Ferner wird vorausgesetzt, dass die Bilinearform a_h für Funktionen in $V_h \oplus V$ erklärt ist. Außerdem wird die Elliptizität

$$a_h(v, v) \ge \alpha \|v\|_{V_h}^2 \quad \text{für } v \in V_h \tag{12.3}$$

und Stetigkeit

$$|a_h(u, v)| \le \beta \|u\|_{V_h} \|v\|_{V_h} \quad \text{für } u \in V \oplus V_h,\ v \in V_h \tag{12.4}$$

mit von h unabhängigen Konstanten α und β benötigt.

Satz 12.2 (zweites Lemma von Strang) Unter den obigen Voraussetzungen ist mit einer von h unabhängigen Zahl c

$$\|u - u_h\|_{V_h} \le c\left(\inf_{v_h \in V_h} \|u - v_h\|_{V_h} + \sup_{w_h \in V_h} \frac{|a_h(u, w_h) - \ell_h(w_h)|}{\|w_h\|_{V_h}}\right).$$

Beweis. Sei $v_h \in V_h$. Aus (12.3) schließen wir

$$\begin{aligned}\alpha \|u_h - v_h\|_{V_h}^2 &\le a_h(u_h - v_h, u_h - v_h) \\ &= a_h(u - v_h, u_h - v_h) + \{\ell_h(u_h - v_h) - a_h(u, u_h - v_h)\}.\end{aligned}$$

Wir dividieren wieder durch $\|u_h - v_h\|_{V_h}$ und kürzen $w_h = u_h - v_h$ ab. Aus (12.4) folgt dann

$$\|u_h - v_h\|_{V_h} \le \frac{\beta}{\alpha}\left(\|u - v_h\|_{V_h} + \sup_{w_h \in V_h} \frac{|a_h(u, w_h) - \ell_h(w_h)|}{\|w_h\|_{V_h}}\right).$$

Wie im Beweis des ersten Lemma folgt die Behauptung mittels der Dreiecksungleichung. □

Im Fall nichtkonformer Elemente ergeben sich im Aubin-Nitsche-Lemma zwei zusätzliche Terme.

Satz 12.3 (Aubin-Nitsche-Lemma) Die Hilbert-Räume V und H mögen den Voraussetzungen des Aubin-Nitsche-Lemmas 6.5 genügen. Ferner seien $V_h \subset H$ und die Bilinearform a_h auf $V \oplus V_h$ derart definiert, dass a_h auf V mit a übereinstimmt. Dann gilt für die Finite-Elemente-Lösung u_h von (12.2)

$$\begin{aligned}\|u-u_h\|_H \le \sup_{g\in H\setminus\{0\}} \frac{1}{\|g\|_H}\Big\{&\beta\|u-u_h\|_{V_h}\|\varphi_g-\varphi_{g,h}\|_{V_h}\\ &+|a_h(u-u_h,\varphi_g)-(u-u_h,g)_H|\\ &+|a_h(u,\varphi_g-\varphi_{g,h})-\ell(\varphi_g)+\ell_h(\varphi_{g,h})|\Big\}.\end{aligned}$$

Dabei werden jedem $g \in H$ die schwachen Lösungen $\varphi_g \in V$ und $\varphi_{g,h} \in V_h$ von $a(w,\varphi) = (w,g)_H$ und $a_h(w,\varphi_h) = (w,g)_H$ zugeordnet.

Beweis. Nach Definition von u_h, φ_g und $\varphi_{g,h}$ ist für jedes $g \in H$

$$\begin{aligned}(u-u_h,g)_H &= \underbrace{a(u,\varphi_g)}_{=a_h(u,\varphi_g)} - a_h(u_h,\varphi_{g,h})\\ &= a_h(u-u_h,\varphi_g-\varphi_{g,h}) + a_h(u_h,\varphi_g-\varphi_{g,h}) + a_h(u-u_h,\varphi_{g,h}).\end{aligned}$$

Mit

$$\begin{aligned}a_h(u_h,\varphi_g-\varphi_{g,h}) &= a_h(u_h-u,\varphi_g) + \underbrace{a_h(u,\varphi_g)}_{=(u,g)_H} - \underbrace{a_h(u_h,\varphi_{g,h})}_{=(u_h,g)_H}\\ &= a_h(u_h-u,\varphi_g) + (u-u_h,g)_H\end{aligned}$$

und

$$\begin{aligned}a_h(u-u_h,\varphi_{g,h}) &= a_h(u,\varphi_{g,h}-\varphi_g) + \underbrace{a_h(u,\varphi_g)}_{=\ell(\varphi_g)} - \underbrace{a_h(u_h,\varphi_{g,h})}_{=\ell_h(\varphi_{g,h})}\\ &= a_h(u,\varphi_{g,h}-\varphi_g) + \ell(\varphi_g) - \ell_h(\varphi_{g,h})\end{aligned}$$

folgt

$$\begin{aligned}(u-u_h,g)_H = a_h(u-u_h,\varphi_g-\varphi_{g,h}) &- \{a_h(u-u_h,\varphi_g)-(u-u_h,g)_H\}\\ &- \{a_h(u,\varphi_g-\varphi_{g,h})-\ell(\varphi_g)+\ell_h(\varphi_{g,h})\}.\end{aligned}$$

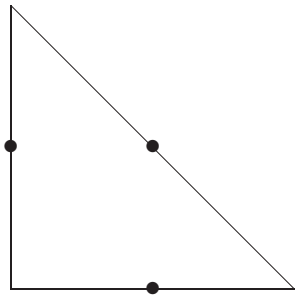
Aus der Stetigkeit von a_h und der Identität

$$\|u-u_h\|_H = \sup_{g\in H\setminus\{0\}} \frac{(u-u_h,g)_H}{\|g\|_H}$$

folgt hieraus die Behauptung. □

12.2 Crouzeix-Raviart-Element

Das einfachste nichtkonforme Element stellt das *Crouzeix-Raviart-Element* dar, das auch als *nichtkonformes $\mathcal{P}_1$-Element* bezeichnet wird:



Der Ansatzraum ist dann

$$V_h = \{v \in L^2(\Omega) : v|_T \text{ ist linear für jedes } T \in \mathcal{T}_h, \\ v \text{ ist stetig in den Mittelpunkten der Dreiecksseiten}\}.$$

Bei Nullrandbedingungen wird zusätzlich gefordert, dass $v = 0$ gilt in den Mittelpunkten der Dreiecksseiten auf $\partial\Omega$.
Zur Lösung des homogenen Dirichlet-Problems zur Poisson-Gleichung

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

in einem konvexen Polygongebiet $\Omega \subset \mathbb{R}^2$ sei

$$a_h(u,v) = \sum_{T\in\mathcal{T}} \int_T \langle \nabla u, \nabla v\rangle \,\mathrm{d}\mathbf{x} \quad \text{und} \quad \ell_h(v) := \int_\Omega fv \,\mathrm{d}\mathbf{x} \quad \text{für } u, v \in H^1(\Omega) \oplus V_h$$

und

$$\|v\|_{V_h} := \sqrt{a_h(v,v)}$$

gesetzt.

Satz 12.4 Sei $\{\mathcal{T}_h\}$ eine Familie quasi-uniformer Triangulierungen eines konvexen Polygongebiets $\Omega \subset \mathbb{R}^2$ und $f \in L^2(\Omega)$. Dann gilt für den Diskretisierungsfehler des Poisson-Problems mit Crouzeix-Raviart-Elementen

$$\|u - u_h\|_{V_h} \le ch\|f\|_{L^2(\Omega)}.$$

Beweis. Im Hinblick auf das zweite Lemma von Strang (Satz 12.2) berechnen wir für $v_h \in V_h$

$$\begin{aligned} L_u(v_h) &:= a_h(u, v_h) - \ell_h(v_h) \\ &= \sum_{T\in\mathcal{T}_h} \int_T \langle \nabla u, \nabla v\rangle \,\mathrm{d}\mathbf{x} - \int_\Omega f v_h \,\mathrm{d}\mathbf{x} \\ &= \sum_{T\in\mathcal{T}_h} \left\{ \int_{\partial T} \frac{\partial u}{\partial \mathbf{n}} v_h \,\mathrm{d}\sigma - \int_T \Delta u v_h \,\mathrm{d}\mathbf{x} \right\} - \int_\Omega f v_h \,\mathrm{d}\mathbf{x} \\ &= \sum_{T\in\mathcal{T}_h} \int_{\partial T} \frac{\partial u}{\partial \mathbf{n}} v_h \,\mathrm{d}\sigma - \int_\Omega \underbrace{(\Delta u + f)}_{=0} v_h \,\mathrm{d}\mathbf{x} \\ &= \sum_{T\in\mathcal{T}_h} \int_{\partial T} \frac{\partial u}{\partial \mathbf{n}} v_h \,\mathrm{d}\sigma. \end{aligned}$$

Da $\partial u/\partial \mathbf{n}$ lediglich das Vorzeichen wechselt, je nachdem von welchem Element man eine innere Kante betrachtet, ändern sich die Integrale nicht, wenn man auf jeder Seite einer Kante e den Integralmittelwert $\overline{v_h}$ abzieht:

$$L_u(v_h) = \sum_{T\in\mathcal{T}_h}\sum_{e\in\partial T}\int_e \frac{\partial u}{\partial \mathbf{n}}\big(v_h\big|_T - \overline{v_h}(e)\big)\,\mathrm{d}\sigma.$$

Sei $I_h u \in V_h \cap C(\Omega)$ die stetige, stückweise lineare Funktion, die $u \in H^2(\Omega)$ in den Eckpunkten interpoliert. Da $\partial I_h u/\partial \mathbf{n}$ auf jeder Kante konstant ist, folgt nach Definition von $\overline{v_h}$, dass

$$\int_e \frac{\partial I_h u}{\partial \mathbf{n}}\big(v_h\big|_{T_i} - \overline{v_h}(e)\big)\,\mathrm{d}\sigma = 0, \quad i = 1, 2.$$

Da dies ebenso für die Kanten auf dem Rand gilt, schließen wir

$$L_u(v_h) = \sum_{T\in\mathcal{T}_h}\sum_{e\in\partial T}\int_e \frac{\partial (u - I_h u)}{\partial \mathbf{n}}\big(v_h\big|_T - \overline{v_h}(e)\big)\,\mathrm{d}\sigma.$$

Die Anwendung der Cauchy-Schwarzschen Ungleichung liefert nun

$$|L_u(v_h)| \le \sum_{T\in\mathcal{T}_h}\sum_{e\in\partial T}\left\|\frac{\partial (u - I_h u)}{\partial \mathbf{n}}\right\|_{L^2(e)} \|v_h - \overline{v_h}(e)\|_{L^2(e)}. \tag{12.5}$$

Die auftretenden Normausdrücke beschränken wir nun separat. Für $v \in H^2(T_{\mathrm{ref}})$ ist gemäß Spursatz 3.7 und Lemma 5.10

$$\int_{\partial T_{\mathrm{ref}}} \|\nabla(v - I_h v)\|_2^2\,\mathrm{d}\sigma \le c\|\nabla(v - I_h v)\|^2_{H^1(T_{\mathrm{ref}})} \le c\|v - I_h v\|^2_{H^2(T_{\mathrm{ref}})} \le c|v|^2_{H^2(T_{\mathrm{ref}})}.$$

Mit dem Transformationsformel aus Lemma 5.14 folgt deshalb für $T \in \mathcal{T}_h$

$$\int_{\partial T} \|\nabla(u - I_h u)\|^2\,\mathrm{d}\sigma \le ch|u|^2_{H^2(T)}. \tag{12.6}$$

Weil $\int_{e_{\mathrm{ref}}} |v_h - \overline{v_h}(e)|^2\,\mathrm{d}\sigma = 0$ für all konstanten Funktionen v_h gilt, können wir das Bramble-Hilbert-Lemma (Proposition 5.11) für jede Kante e_{ref} von T_{ref} anwenden und erhalten

$$\int_{e_{\mathrm{ref}}} |v_h - \overline{v_h}(e)|^2\,\mathrm{d}\sigma \le c|v_h|^2_{H^1(T_{\mathrm{ref}})} \quad \text{für } v_h \in \mathcal{P}_1.$$

Die Transformationsformel liefert daher für eine Kante e aus $T \in \mathcal{T}_h$

$$\int_e |v_h - \overline{v_h}(e)|^2\,\mathrm{d}\sigma \le ch|v_h|^2_{H^1(T)} \quad \text{für } v_h \in V_h. \tag{12.7}$$

Die Abschätzungen (12.6) und (12.7) setzen wir nun in (12.5) ein. Außerdem benutzen wir die Cauchy-Schwarzsche Ungleichung für Skalarprodukte

$$\begin{aligned}
|L_u(v_h)| &\le \sum_{T\in\mathcal{T}_h} ch|u|_{H^2(T)}|v_h|_{H^1(T)} \\
&\le ch\sqrt{\sum_{T\in\mathcal{T}_h}|u|^2_{H^2(T)}}\sqrt{\sum_{T\in\mathcal{T}_h}|v_h|^2_{H^1(T)}} \\
&\le ch|u|_{H^2(\Omega)}\|v_h\|_{V_h}.
\end{aligned} \tag{12.8}$$

Beachten wir schließlich noch, dass die stückweise linearen, konformen Elemente in V_h enthalten sind, liefern die bereits bekannten Approximationsätze zusammen mit dem zweiten Lemma von Strang die Fehlerabschätzung

$$\|u - u_h\|_{V_h} \leq ch|u|_{H^2(\Omega)} \leq ch\|f\|_{L^2(\Omega)}.$$

□

Wir wenden nun das Aubin-Nitsche-Lemma (Satz 12.3) an.

Satz 12.5 Sei $\{\mathcal{T}_h\}$ eine Familie quasi-uniformer Triangulierungen eines konvexen Polygongebiets $\Omega \subset \mathbb{R}^2$ und $f \in L^2(\Omega)$. Dann gilt für den Diskretisierungsfehler des Poisson-Problems mit Crouzeix-Raviart-Elementen

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

Beweis. Um das Aubin-Nitsche-Lemma anzuwenden, setzen wir $H := L^2(\Omega)$ und $V := H_0^1(\Omega)$. Zunächst müssen wir die Differenz $\varphi_g - \varphi_{g,h}$ der Lösungen des jeweils dualen Problems abschätzen. Dazu greifen wir auf obigen Satz 12.4 zurück und erhalten

$$\|\varphi_g - \varphi_{g,h}\|_{V_h} \leq ch\|g\|_{L^2(\Omega)}.$$

Eine wesentliche Beobachtung ist, dass (12.8) sogar für alle $w \in V_h \oplus H_0^1(\Omega)$ gilt. Dies erkennt man sofort, wenn man die Herleitung der Formel überprüft, wobei zu beachten ist, dass der Mittelwert ϕ für beliebige Funktionen aus $H_0^1(\Omega)$ wohldefiniert ist. Daher folgt für die beiden Zusatzterme des Aubin-Nitsche-Lemma

$$\begin{aligned}\left|a_h(u-u_h,\varphi_g) - (u-u_h,g)_{L^2(\Omega)}\right| &= |L_{\varphi_g}(u-u_h)| \\ &\leq ch|\varphi_g|_{H^2(\Omega)}\|u-u_h\|_{V_h} \\ &\leq ch\|g\|_{L^2(\Omega)}\|u-u_h\|_{V_h}\end{aligned}$$

und

$$\begin{aligned}\left|a_h(u,\varphi_g-\varphi_{g,h}) - (f,\varphi_g-\varphi_{g,h})_{L^2(\Omega)}\right| &= |L_u(\varphi_g-\varphi_{g,h})| \\ &\leq ch|u|_{H^2(\Omega)}\|\varphi_g-\varphi_{g,h}\|_{V_h} \\ &\leq ch\|g\|_{L^2(\Omega)}\|\varphi_g-\varphi_{g,h}\|_{V_h}.\end{aligned}$$

Die Kombination dieser drei Abschätzungen mit Satz 12.4 ergibt schließlich die Behauptung. □

12.3 Polygonale Approximation krummliniger Ränder

Zur Triangulierung eines nichtpolygonalen Gebiets Ω benötigt man krummlinige Dreiecke zur Beschreibung des Randes $\Gamma = \partial\Omega$. Die Triangulierung $\mathcal{T}_h$ von Ω soll dabei zulässig sein. Ersetzt man die krummlinigen Kanten durch eine Strecke, so erhält man eine polygonale Approximation Ω_h an das ursprüngliche Gebiet Ω.

Als Finite Elemente wählen wir lineare Dreieckselemente, wobei Nullrandbedingungen nur für die Knoten auf dem Rand verlangt werden

$$V_h = \{v \in C(\Omega) : v|_T \text{ ist linear für jedes } T \in \mathcal{T}_h, \\ v(\mathbf{x}) = 0 \text{ für jeden Knoten } \mathbf{x} \in \Gamma\}.$$

Es ist also $V_h \not\subset H_0^1(\Omega)$. Trotzdem braucht man wegen $V_h \subset H^1(\Omega)$ keine neuen gitterabhängigen Normen einzuführen und kann

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

und

$$\ell_h(v) = \ell(v) = \int_\Omega fv \,\mathrm{d}\mathbf{x}$$

setzen.

Lemma 12.6 Sei Ω ein Gebiet mit C^2-Rand und durch $\mathcal{T}_h$ sei eine Folge quasi-uniformer Triangulierungen gegeben. Dann gilt

$$\|v_h\|_{L^2(\Gamma)} \le ch^{3/2}|v_h|_{H^1(\Omega)} \quad \text{für alle } v_h \in V_h.$$

Beweis. Sei $T \in \mathcal{T}_h$ ein Element mit der krummlinigen Kante $\Gamma_T = T \cap \Gamma$. Wir werden

$$\int_{\Gamma_T} v_h^2 \,\mathrm{d}\sigma \le ch_T^3 \int_T \left\{|\partial_x v_h|^2 + |\partial_y v_h|^2\right\} \mathrm{d}\mathbf{x}$$

zeigen. Dann folgt die Behauptung durch Summation über alle Dreiecke.

Das Koordinatensystem wählen wir so, dass die ξ-Achse durch die beiden auf Γ liegenden Ecken von T läuft. Die Koordinaten der Eckpunkte seien $(\xi_1, 0)$, $(\xi_2, 0)$, und der Rand werde durch $\eta = g(\xi)$ beschrieben. Wegen $g(\xi_1) = g(\xi_2) = 0$, $|\xi_1 - \xi_2| \le h_T$ und $g''(\xi) \le c$ ist

$$|g(\xi)| \le ch_T^2 \quad \text{für } \xi_1 \le \xi \le \xi_2. \tag{12.9}$$

Da $v_h \in V_h$ in T linear ist und entlang der ξ-Achse verschwindet, hat v_h in T die Gestalt

$$v_h(\xi, \eta) = b\eta.$$

Der Gradient ist in T konstant: $\|\nabla v_h\|_2 = b$. Die Fläche von T kann nach unten durch die des Inkreises abgeschätzt werden. Sein Radius beträgt mindestens h_T/κ. Deshalb ist

$$\int_T \|\nabla v_h\|_2^2 \,\mathrm{d}\mathbf{x} \ge \pi \frac{h_T^2}{\kappa^2} b^2.$$

Andererseits gilt

$$\int_{\Gamma_T} v_h^2 \,\mathrm{d}\sigma = \int_{\xi_1}^{\xi_2} \big(bg(\xi)\big)^2 \sqrt{1 + \underbrace{|g'(\xi)|^2}_{\le c}} \,\mathrm{d}\xi \le cb^2 h_T^4 \int_{\xi_1}^{\xi_2} 1 \,\mathrm{d}\xi = c^2 b^2 h_T^5.$$

Durch den Vergleich der beiden letzten Resultate ergibt sich die Behauptung. □

Satz 12.7 Sei Ω ein Gebiet mit C^2-Rand und $f \in L^2(\Omega)$. Für die Finite-Element-Approximation durch lineare Dreieckselemente gilt bei quasi-uniformen Triangulierungen

$$\|u - u_h\|_{H^1(\Omega)} \le ch\|u\|_{H^2(\Omega)} \le ch\|f\|_{L^2(\Omega)}.$$

Beweis. Das durch die Sehne abgeschnitte Flächenstück $F := T \cap (\Omega \setminus \Omega_h)$ besitzt wegen (12.9) nur den Inhalt

$$|F| \le h_T \max_{\xi_1 \le \xi \le \xi_2} |g(\xi)| \le ch_T^3 \le ch_T|T|.$$

Sei $u \in H^2(\Omega) \cap H_0^1(\Omega)$ die Lösung von $\mathcal{L}u = f$ und u_h die zugehörige schwache Lösung in V_h, das heißt

$$a(u_h, v_h) = (f, v_h)_{L^2(\Omega)} \quad \text{für alle } v_h \in V_h.$$

Dann ergibt sich mit partieller Integration

$$(f, v_h)_{L^2(\Omega)} = (\mathcal{L}u, v_h)_{L^2(\Omega)} = a(u, v_h) - \int_\Gamma \langle \mathbf{A}\nabla u, \mathbf{n}\rangle v_h \, \mathrm{d}\sigma.$$

Mit der Cauchy-Schwarzschen Ungleichung, dem Spursatz und Lemma 12.6 folgt

$$|a(u, v_h) - (f, v_h)_{L^2(\Omega)}| \le c\|\nabla u\|_{L^2(\Gamma)}\|v_h\|_{L^2(\Gamma)} \le ch^{3/2}\|u\|_{H^2(\Omega)}\|v_h\|_{H^1(\Omega)}.$$

Das zweite Lemma von Strang impliziert nun die Behauptung. □

Im vorliegenden Fall führt das Aubin-Nitsche-Lemma in der Form von Satz 12.3 leider nicht zum Ziel. Daher müssen wir nachfolgendes Ergebnis zu Fuß beweisen.

Satz 12.8 Unter den Voraussetzungen des Satzes 12.7 ist

$$\|u - u_h\|_{L^2(\Omega)} \le ch^{3/2}\|u\|_{H^2(\Omega)}.$$

Beweis. Um keine Doppelsummen in den Randintegralen mitführen zu müssen, beschränken wir uns im folgenden auf die Poisson-Gleichung. Zu $w := u - u_h$ sei φ die Lösung von

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

Weil Ω glatt ist, ist das Problem $H^2(\Omega)$-regulär, das heißt, es ist $\varphi = H^2(\Omega) \cap H_0^1(\Omega)$ und

$$\|\varphi\|_{H^2(\Omega)} \le c\|w\|_{L^2(\Omega)}.$$

Im Gegensatz zu Umformungen bei konformen Finiten Elementen erhalten wir wegen $w \notin H_0^1(\Omega)$ Randterme aus der Greenschen Formel

$$\|w\|_{L^2(\Omega)}^2 = (w, -\Delta\varphi)_{L^2(\Omega)} = a(w, \varphi) - \left(w, \frac{\partial\varphi}{\partial\mathbf{n}}\right)_{L^2(\Gamma)}.$$

Da für beliebiges $v_h \in V_h$ gilt

$$a(u-u_h,-v_h) = \left(\frac{\partial u}{\partial \mathbf{n}}, -v_h\right)_{L^2(\Gamma)} = \left(\frac{\partial u}{\partial \mathbf{n}}, \underbrace{\varphi}_{=0} - v_h\right)_{L^2(\Gamma)},$$

folgt

$$\|w\|^2_{L^2(\Omega)} = a(w,\varphi - v_h) - \left(\frac{\partial u}{\partial \mathbf{n}}, \varphi - v_h\right)_{L^2(\Gamma)} - \left(w, \frac{\partial \varphi}{\partial \mathbf{n}}\right)_{L^2(\Gamma)}. \tag{12.10}$$

Wir wählen $v_h := I_h\varphi$ und schätzen ab

$$\begin{aligned} a(w,\varphi - v_h) &\le c_S \|w\|_{H^1(\Omega)} \|\varphi - v_h\|_{H^1(\Omega)} \\ &\le ch\|w\|_{H^1(\Omega)}\|\varphi\|_{H^2(\Omega)} \\ &\le ch\|w\|_{H^1(\Omega)}\|w\|_{L^2(\Omega)}. \end{aligned}$$

Für die Behandlung des zweiten Terms aus (12.10) benötigen wir die Approximationsaussage $\|\varphi - I_h\varphi\|_{L^2(\Gamma)} \le ch^{3/2}\|\varphi\|_{H^2(\Omega)}$, die mit dem Spursatz und dem Transformationssatz gezeigt werden kann. Außerdem wird der Spursatz auf ∇u angewandt, weshalb folgt

$$\begin{aligned} \left|\left(\frac{\partial u}{\partial \mathbf{n}}, \varphi - v_h\right)_{L^2(\Gamma)}\right| &\le \|\nabla u\|_{L^2(\Gamma)}\|\varphi - v_h\|_{L^2(\Gamma)} \\ &\le ch^{3/2}\|u\|_{H^2(\Omega)}\|\varphi\|_{H^2(\Omega)} \\ &\le ch^{3/2}\|u\|_{H^2(\Omega)}\|w\|_{L^2(\Omega)}. \end{aligned}$$

Für den letzten Term greifen wir auf Lemma 12.6 und den Spursatz zurück:

$$\begin{aligned} \left|\left(w, \frac{\partial \varphi}{\partial \mathbf{n}}\right)_{L^2(\Gamma)}\right| &\le \|w\|_{L^2(\Gamma)}\|\nabla\varphi\|_{L^2(\Gamma)} \\ &= \|\underbrace{u}_{=0} - u_h\|_{L^2(\Gamma)}\|\nabla\varphi\|_{L^2(\Gamma)} \\ &\le ch^{3/2}\|u_h\|_{H^1(\Omega)}\|\varphi\|_{H^2(\Omega)} \\ &\le ch^{3/2}\big\{\|u\|_{H^1(\Omega)} + \|u-u_h\|_{H^1(\Omega)}\big\}\|w\|_{L^2(\Omega)} \\ &\le ch^{3/2}\|u\|_{H^2(\Omega)}\|w\|_{L^2(\Omega)}. \end{aligned}$$

Zusammen haben wir demnach

$$\|u-u_h\|^2_{L^2(\Omega)} = \|w\|^2_{L^2(\Omega)} \le c\|w\|_{L^2(\Omega)}\big\{h\|w\|_{H^1(\Omega)} + h^{3/2}\|u\|_{H^2(\Omega)}\big\}$$

gezeigt, woraus die Behauptung folgt. □

13. Gemischte Finite Elemente

13.1 Gemischte Formulierung des Poisson-Problems

Wir wollen das Poisson-Problem als System erster Ordnung schreiben und dann das so erhaltene System mit Finiten Elementen diskretisieren. Es stellt sich allerdings heraus, dass die bislang von uns verwendete Theorie nicht ausreicht. Das resultierende System hat *Sattelpunktstruktur*. Was das heißt, werden wir nun entwickeln.

Beispiel 13.1 (Strömung in einem porösen Medium) In einem völlig gesättigtem, inkompressiblen Medium (etwa der Boden nach starkem Niederschlag oder eine grundwasserführende Bodenschicht) gelten für den Fluss $\mathbf{v}$ und das Druckpotential p im stationären Zustand

$$\begin{aligned} \operatorname{div}(\rho \mathbf{v}) &= 0 && \text{(Massenerhaltung)} \\ \mathbf{v} &= -K\nabla p && \text{(Darcey-Gesetz)} \end{aligned}$$

Dabei bezeichnet ρ die Dichte und K die Durchlässigkeit des porösen Mediums.
Eine Möglichkeit zur Lösung dieses Systems ist die Lösung der Gleichung

$$\operatorname{div}(\rho K \nabla p) = 0$$

unter Verwendung von Finiten Elementen für p und danach die Bildung von $\mathbf{v} = -K\nabla p$. Dabei ist jedoch die Approximationsgüte an den Fluss $\mathbf{v}$, an dem man vornehmlich interessiert ist, schlechter als für das Druckpotential p. Als Abhilfe verwendet man einen gemischten Finite-Element-Ansatz zur simultanen Approximation von $\mathbf{v}$ und p. △

Zur Übersetzung der Poisson-Gleichung in ein System erster Ordnung verwenden wir $-\Delta p = \operatorname{div}(\nabla p)$ und setzen $\mathbf{v} := \nabla p$. Wir erhalten dann das System

$$-\operatorname{div}(\mathbf{v}) = f,\ \nabla p - \mathbf{v} = \mathbf{0} \text{ in } \Omega, \quad p = 0 \text{ auf } \Gamma.$$

Um hierzu die Variationsformulierung aufzustellen, werden wir die erste Gleichung partiell integrieren

$$\begin{aligned} (\mathbf{v}, \nabla q)_{L^2(\Omega)} &= (f, q)_{L^2(\Omega)} && \text{für alle } q \in H_0^1(\Omega), \\ -(\nabla p, \mathbf{w})_{L^2(\Omega)} + (\mathbf{v}, \mathbf{w})_{L^2(\Omega)} &= 0 && \text{für alle } \mathbf{w} \in [L^2(\Omega)]^d. \end{aligned}$$

Die Lösung $(\mathbf{v}, p)$ lebt demnach im *Produktraum* $X = [L^2(\Omega)]^d \times H_0^1(\Omega)$. In X ist eine geeignete Norm gegeben durch

$$\|(\mathbf{v}, p)\|_X := \|\mathbf{v}\|_{L^2(\Omega)} + |p|_{H^1(\Omega)}.$$

Führen wir die Bilinearform

$$A : X \times X \to \mathbb{R}, \quad A\big((\mathbf{v},p),(\mathbf{w},q)\big) := (\mathbf{v},\nabla q)_{L^2(\Omega)} - (\nabla p,\mathbf{w})_{L^2(\Omega)} + (\mathbf{v},\mathbf{w})_{L^2(\Omega)}$$

und das Funktional

$$\ell : X \to \mathbb{R}, \quad \ell(\mathbf{w},q) := (f,q)_{L^2(\Omega)}$$

ein, dann suchen wir $(\mathbf{v},p) \in X$, so dass gilt

$$A\big((\mathbf{v},p),(\mathbf{w},q)\big) = \ell(\mathbf{w},q) \quad \text{für alle } (\mathbf{w},q) \in X. \tag{13.1}$$

Die Bilinearform A ist stetig, denn es gilt

$$\begin{aligned} |A\big((\mathbf{v},p),(\mathbf{w},q)\big)| &\le \|\mathbf{v}\|_{L^2(\Omega)}|q|_{H^1(\Omega)} + |p|_{H^1(\Omega)}\|\mathbf{w}\|_{L^2(\Omega)} + \|\mathbf{v}\|_{L^2(\Omega)}\|\mathbf{w}\|_{L^2(\Omega)} \\ &\le \big\{\|\mathbf{v}\|_{L^2(\Omega)} + |p|_{H^1(\Omega)}\big\}\big\{\|\mathbf{w}\|_{L^2(\Omega)} + |q|_{H^1(\Omega)}\big\} \\ &= \|(\mathbf{v},p)\|_X \|(\mathbf{w},q)\|_X. \end{aligned}$$

Sie ist jedoch nicht elliptisch, denn es ist

$$A\big((\mathbf{v},p),(\mathbf{v},p)\big) = (\mathbf{v},\nabla p)_{L^2(\Omega)} - (\nabla p,\mathbf{v})_{L^2(\Omega)} + (\mathbf{v},\mathbf{v})_{L^2(\Omega)} = \|\mathbf{v}\|^2_{L^2(\Omega)}.$$

Es fehlt offensichtlich die Kontrolle über den Anteil $|p|_{H^1(\Omega)}$. Folglich lassen sich die bisher verwendeten Techniken nicht verwenden.
Unsere Bilinearform ist offensichtlich von der Form

$$A\big((\mathbf{v},p),(\mathbf{w},q)\big) = a(\mathbf{v},\mathbf{w}) + b(\mathbf{v},q) - b(p,\mathbf{w})$$

mit einer symmetrischen $L^2(\Omega)$-elliptischen Bilinearform

$$a(\cdot,\cdot) : [L^2(\Omega)]^d \times [L^2(\Omega)]^d \to \mathbb{R}, \quad a(\mathbf{v},\mathbf{w}) = (\mathbf{v},\mathbf{w})_{L^2(\Omega)}$$

und einer gemischten Bilinearform

$$b : [L^2(\Omega)]^d \times H^1_0(\Omega) \to \mathbb{R}, \quad b(\mathbf{v},q) = (\mathbf{v},\nabla q)_{L^2(\Omega)}.$$

Wir werden dies im folgenden abstrakt formulieren, da viele partielle Differentialgleichungen von dieser Form sind.

13.2 Satz vom abgeschlossenen Bild

Seien X und Y zwei Banach-Räume, deren duale Paarungen mit X' beziehungsweise Y' mit $\langle\cdot,\cdot\rangle$ bezeichnet werden. Sei $T : X \to Y$ ein stetiger linearer Operator. Dann ist der zu T adjungierte Operator $T^\star : Y' \to X'$ gegeben durch

$$\langle x, T^\star y\rangle = \langle Tx, y\rangle \quad \text{für alle } y \in Y' \text{ und } x \in X.$$

Der dazugehörige Kern wird mit

$$\operatorname{kern}(T^\star) := \{y \in Y' : T^\star y = 0\} \subset Y'$$

bezeichnet. Das *orthogonale Komplement* ist definert als

$$\operatorname{kern}(T^\star)^\perp := \{y \in Y : \langle x,y\rangle = 0 \text{ für alle } x \in \operatorname{kern}(T^\star)\} \subset Y.$$

Im folgenden wird mehrfach folgender Satz herangezogen, der in der englischsprachigen Literatur als *closed range theorem* bekannt ist.

Satz 13.2 (Satz vom abgeschlossenen Bild) Seien X und Y zwei Banach-Räume und $T : X \to Y$ stetig. Dann ist das Bild $T(X)$ genau dann abgeschlossen in Y, wenn $T(X) = \ker(T^\star)^\perp$ ist.

Beweis. Einen Beweis des Satzes findet der geneigte Leser in K. Yosida "Functional Analysis". □

Korollar 13.3 Seien X und Y zwei Banach-Räume und $T : X \to Y$ stetig. Ist T injektiv derart, dass T^{-1} auf dem Bild $T(X)$ stetig und $T^\star$ injektiv ist, dann ist T ein Isomorphismus von X nach Y.

Beweis. Aufgrund der Stetigkeit von T und der von T^{-1} auf dem Bild $T(X)$ ist $T(X)$ abgeschlossen. Nach dem Satz vom abgeschlossenen Bild folgt nun mit der Injektivität von $T^\star$, dass

$$T(X) = \ker(T^\star)^\perp = \{0\}^\perp = Y.$$

Also ist T eine Bijektion und zusammen mit der Stetigkeit von T und T^{-1} ein Isomorphismus. □

13.3 inf-sup-Bedingung

Seien U und V zwei Hilbert-Räume und

$$A : U \times V \to \mathbb{R}$$

eine Bilinearform. Wir fragen bei gegebenen $f \in V'$ nach der Lösbarkeit des Variationsproblems

$$\text{suche } u \in U, \text{ so dass} \quad A(u, v) = \langle f, v \rangle \quad \text{für alle } v \in V. \tag{13.2}$$

Satz 13.4 Das Problem (13.2) besitzt zu jedem $f \in V'$ eine eindeutige Lösung $u \in U$, wenn folgende Bedingungen erfüllt sind:

(*i.*) A ist stetig, das heißt, es ist

$$|A(u, v)| \leq c_S \|u\|_U \|v\|_V \quad \text{für alle } u \in U,\ v \in V.$$

(*ii.*) Es gilt die *inf-sup-Bedingung*

$$\inf_{u \in U \setminus \{0\}} \sup_{v \in V \setminus \{0\}} \frac{A(u, v)}{\|u\|_U \|v\|_V} \geq c_E > 0.$$

(*iii.*) Zu jedem $v \in V \setminus \{0\}$ existiert ein $u \in U$ mit $A(u, v) \neq 0$.

Ferner erfüllt die Lösung u in diesem Fall die Stabilitätsaussage

$$\|u\|_U \leq \frac{1}{c_E} \|f\|_{V'}.$$

Beweis. Die Eindeutigkeit einer etwaigen Lösung $u \in U$ folgt unmittelbar aus der inf-sup-Bedingung. Denn ist $\widetilde{u} \in U$ eine weitere Lösung, so gilt aufgrund der Linearität $A(u - \widetilde{u}, v) = 0$ für alle $v \in V$. Daraus folgt aber

$$\sup_{v \in V \setminus \{0\}} \frac{A(u - \widetilde{u}, v)}{\|v\|_V} = 0,$$

was wegen der inf-sup-Bedingung $u - \widetilde{u} = 0$ bedeutet.
Die Stabilitätsaussage folgt ebenfalls aus der inf-sup-Bedingung, denn

$$c_E \|u\|_U \leq \sup_{v \in V \setminus \{0\}} \frac{A(u, v)}{\|v\|_V} = \sup_{v \in V \setminus \{0\}} \frac{\langle f, v \rangle}{\|v\|_V} \leq \sup_{v \in V \setminus \{0\}} \frac{\|f\|_{V'} \|v\|_V}{\|v\|_V} = \|f\|_{V'}.$$

Es verbleibt also noch, die Existenz einer Lösung nachzuweisen. Hierzu betrachten wir die Menge

$$W := \{w \in V : A(u, w) = 0 \text{ für alle } u \in U\} \subset V.$$

Wir führen ferner den linearen Operator $L : U \to V'$ ein, der definiert ist durch

$$\langle Lu, v \rangle := A(u, v).$$

Der Operator L ist stetig, da A es ist. Ebenso ist L^{-1} auf dem Bild $L(U)$ stetig, denn aus $c_E \|u\|_U \leq \|Lu\|_{V'}$ folgt nämlich

$$\|L^{-1} f\|_U \leq \frac{1}{c_E} \|f\|_{V'} \quad \text{für alle } f \in L(U).$$

Folglich ist $L(U)$ abgeschlossen in V'. Um den Satz vom abschlossenen Bild anzuwenden, betrachten wir den zu L adjungierten Operator $L^\star : V \to U'$, gegeben durch $\langle u, L^\star v \rangle = \langle Lu, v \rangle$. Hier verwenden wir die Reflexivität von Hilbert-Räumen, also $V'' = V$. Offensichtlich ist $\operatorname{kern}(L^\star) = W$. Daher ist gemäß Satz 13.2

$$L(U) = \operatorname{kern}(L^\star)^\perp = W^\perp = \{f \in V' : \langle f, w \rangle = 0 \text{ für alle } w \in W\}.$$

Aufgrund der Bedingung (*iii.*) gilt aber $W = \{0\}$. Daher ist $L(U) = \{0\}^\perp = V'$, das heißt, L ist surjektiv. Die Lösung erhalten wir daher durch $u := L^{-1} f$ für beliebig vorgegebenes $f \in V'$. □

Bemerkung Dieser Satz ist eine Verallgemeinerung des Satzes von Lax-Milgram, denn im Fall $U = V$ und einer V-elliptischen Bilinearform mit Elliptizitätskonstante $c_E > 0$ folgt die inf-sup Bedingung gemäß

$$\inf_{u \in V \setminus \{0\}} \sup_{v \in V \setminus \{0\}} \frac{A(u, v)}{\|u\|_V \|v\|_V} \geq \inf_{u \in V \setminus \{0\}} \frac{A(u, u)}{\|u\|_V^2} \geq c_E.$$

Die Bedingung (*iii.)* im vorherigen Satz ergibt sich ebenso unmittelbar, indem man $u = v$ wählt und die Elliptizität verwendet. △

Satz 13.5 Die Bedingungen aus Satz 13.4 seien gleichermaßen für die Hilbert-Räume U und V sowie für die konformen Finite-Element-Räume $U_h \subset U$ und $V_h \subset V$ mit Konstanten $c_E, c_S > 0$ erfüllt. Dann gilt für die eindeutig definierte Lösung $u_h \in U_h$ die Fehlerabschätzung

$$\|u - u_h\|_U \leq \left(1 + \frac{c_S}{c_E}\right) \inf_{w_h \in U_h} \|u - w_h\|_U.$$

Beweis. Zu beliebigem $w_h \in U_h$ wenden wir die Dreiecksungleichung an:

$$\|u - u_h\|_U \leq \|u - w_h\|_U + \|u_h - w_h\|_U.$$

Wir müssen offensichtlich den letzten Term beschränken. Dazu definieren wir das Funktional $g \in V'$ durch

$$\langle g, v\rangle := A(u - w_h, v) \quad \text{für alle } v \in V$$

und den stetigen linearen Operator $L : U_h \to V_h'$ durch

$$\langle Lu_h, v_h\rangle := A(u_h, v_h) \quad \text{für alle } v_h \in V.$$

Aufgrund der Galerkin-Orthogonalität gilt für $v_h \in V_h$

$$\langle g, v_h\rangle = A(u - w_h, v_h) = A(u_h - w_h, v_h) = \langle L(u_h - w_h), v_h\rangle.$$

Mit anderen Worten, es ist

$$g|_{V_h} = L(u_h - w_h),$$

beziehungsweise

$$u_h - w_h = L^{-1}g.$$

Aus der Stetigkeit von A folgt $\|g\|_{V'} \leq c_S\|u - w_h\|_U$. Da, wie im Beweis von Satz 13.4 gezeigt, gilt $\|L^{-1}\|_{V_h' \to U_h} \leq 1/c_E$, ergibt sich

$$\|u_h - w_h\|_U = \|L^{-1}g\|_U \leq \frac{1}{c_E}\|g\|_{V'} \leq \frac{c_S}{c_E}\|u - w_h\|_U.$$

Zusammen erhalten wir die Behauptung. □

13.4 LBB-Bedingung für Sattelpunktprobleme

Wenn wir nun von einem Sattelpunktproblem ausgehen, wie wir es unter anderem im Fall des Poisson-Problems als System erster Ordnung vorliegen haben, so muss die inf-sup-Bedingung für den gemischten Anteil der Bilinearform erfüllt sein. Man spricht in diesem Zusammenhang auch von der LBB-Bedingung, benannt nach den Mathematikern Ladyshenskaja, Babuška und Brezzi.

Seien U und V zwei Hilbert-Räume. Den Produktraum bezeichnen wir wie zuvor mit $X := U \times V$. Wir betrachten allgemein das Sattelpunktproblem

$$\begin{aligned} a(u, v) - b(v, p) &= \langle f, v\rangle \quad \text{für alle } v \in U \\ b(u, q) &= \langle g, q\rangle \quad \text{für alle } q \in V \end{aligned} \tag{13.3}$$

mit stetigen Bilinearformen

$$a : U \times U \to \mathbb{R}, \quad b : U \times V \to \mathbb{R}.$$

Satz 13.6 Das Sattelpunktproblem (13.3) mit einer symmetrischen und stetigen Bilinearform a und einer stetigen Bilinearform b hat genau dann für beliebige $f \in U'$ und $g \in V'$ eine eindeutige Lösung $(u, p) \in X$, falls gilt:

(*i.*) Die Bilinearform a ist U_0-elliptisch, wobei

$$U_0 := \{v \in U : b(v, q) = 0 \text{ für alle } q \in V\}$$

der Kern von b ist.

(*ii.*) Die LBB-Bedingung ist erfüllt

$$\inf_{q \in V \setminus \{0\}} \sup_{v \in U \setminus \{0\}} \frac{b(v, q)}{\|v\|_U \|q\|_V} \geq \gamma > 0.$$

Beweis. Wir nehmen zunächst an, die Bedingungen (*i.*) und (*ii.*) seien erfüllt, und zeigen die Existenz einer Lösung des Sattelpunktproblems. Die LBB-Bedingung ist eine inf-sup Bedingung wie in Satz 13.4 für die Bilinearform b. Wie im dortigen Beweis gezeigt wurde, ist der Operator $B^\star : V \to U_0^\perp = \{v \in U' : \langle v, w\rangle = 0 \text{ für alle } w \in U_0\} \subset U'$, definiert durch

$$\langle v, B^\star q\rangle := b(v, q),$$

ein Isomorphismus. Dann ist aber auch der adjungierte Operator $B : W \to V'$ ein Isomorphismus, falls wir setzen

$$W = \{v \in U : (v, w)_U = 0 \text{ für alle } w \in U_0\} \subset U.$$

Er ist definiert durch

$$\langle Bv, q\rangle := b(v, q).$$

Folglich existiert zu $g \in V'$ ein $v_g \in W$ mit $Bv_g = g$ beziehungsweise

$$b(v_g, q) = \langle g, q\rangle \text{ für alle } q \in V.$$

Daher ist das Sattelpunktproblem äquivalent zur Suche nach einem $w := v - v_g \in U$ und $p \in V$ mit

$$\begin{aligned} a(w, v) - b(v, p) &= \langle f, v\rangle - a(v_g, v) && \text{für alle } v \in U, \\ b(w, q) &= 0 && \text{für alle } q \in V. \end{aligned}$$

Nun nutzen wir die U_0-Elliptizität der Bilinearform a. Sie impliziert die Existenz einer eindeutigen Lösung $w \in U_0$ mit

$$a(w, v) = \langle f, v\rangle - a(v_g, v) \quad \text{für alle } v \in U_0. \tag{13.4}$$

Nun ist noch ein $p \in V$ zu finden mit

$$b(v, p) = -\langle f, v\rangle + a(v_g + w, v) \quad \text{für alle } v \in U.$$

Hier ist die rechte Seite ein Funktional in U':

$$\langle \ell, v\rangle := -\langle f, v\rangle + a(v_g + w, v) \quad \text{für alle } v \in U.$$

Für $v \in U_0$ gilt nun aber wegen (13.4) $\langle \ell, v\rangle = 0$, das heißt, es ist $\ell \in U_0^\perp$. Nun benutzen wir nocheinmal die Isometrie-Eigenschaft des Operators $B^\star$, der uns jetzt eine eindeutige Lösung $p \in V$ liefert mit $b(v,p) = \langle \ell, v\rangle$ für alle $v \in U$.

Um die Eindeutigkeit der Lösung $(u,p) \in X$ zu zeigen, weisen wir die Stabilität der Lösung nach. Da $B : W \to V'$ ein Isomorphismus ist, folgt

$$\|v_g\|_U \le \frac{1}{\gamma}\|g\|_{V'}.$$

Die U_0-Elliptizität der Bilinearform a liefert

$$c_E\|w\|_U \le \|f\|_{U'} + \|a(v_g,\cdot)\|_{U'} \le \|f\|_{U'} + c_S\|v_g\|_U \le \|f\|_{U'} + \frac{c_S}{\gamma}\|g\|_{V'}.$$

Hierbei sind $c_S > 0$ die Beschränktheitskonstante und $c_E > 0$ die Elliptizitätskonstante von a. Schließlich ergibt sich die Eindeutigkeit aus der Eindeutigkeit der homogenen Gleichung, bei der die rechten Seiten verschwinden.

Um die Rückrichtung zu zeigen, nehmen wir an, dass das Sattelpunktproblem ein Isomorphismus $L : U \times V \to U' \times V'$ ist. Insbesondere ist $\|L^{-1}\| \le c$. Zu einem gegebenen Funktional $f \in U_0'$ gibt es nach dem Satz von Hahn-Banach eine Erweiterung $\widetilde{f} : U \to \mathbb{R}$ mit $\|\widetilde{f}\|_{U'} = \|f\|_{U_0'}$. Wir setzen $(u,p) = L^{-1}(\widetilde{f}, 0)$ und beachten, dass $u \in U_0$ gelten muss. Da die Abbildung $f \mapsto u \in U_0$ beschränkt ist, ist a auch U_0-elliptisch.

Schließlich wird jedem $g \in V'$ durch $(u,p) := L^{-1}(0,g)$ ein $u \in U$ zugeordnet mit $\|u\|_U \le c\|g\|_{V'}$. Wir bestimmen zu diesem u die Orthoprojektion $\Pi u \in W$. Wegen $\|\Pi u\|_U \le \|u\|_U$, ist die Abbildung $g \mapsto u \mapsto \Pi u$ beschränkt. Es ist $B\Pi u = g$. Also ist $B : W \to V'$ ein Isomorphismus, und b erfüllt nach Satz 13.4 die LBB-Bedingung. □

Korollar 13.7 Es sei $\{\mathcal{T}_h\}$ eine Familie quasi-uniformer Triangulierungen, so dass die Bedingungen aus Satz 13.6 für konforme Finite-Element-Räume $U_h \subset U$ und $V_h \subset V$ erfüllt sind mit von h unabhängig Konstanten c_E, c_S und γ. Dann existiert eine eindeutige Lösung $(u_h, p_h) \in U_h \times V_h$ und es gilt die Fehlerabschätzung

$$\|u-u_h\|_U + \|p-p_h\|_V \le c\left\{\inf_{w_h\in U_h}\|u-w_h\|_U + \inf_{q_h\in V_h}\|p-q_h\|_V\right\}.$$

Beweis. Das Behauptete folgt unmittelbar aus den Sätzen 13.5 und 13.6. □

Das folgende Hilfsmittel ist oftmals nützlich zum Nachweis der diskreten inf-sup-Bedingung.

Satz 13.8 (Fortins Kriterium) Sei $b : U \times V \to \mathbb{R}$ eine Bilinearform, die der LBB-Bedingung genügt. Seien $U_h \times V_h \subset U \times V$ eine Familie von Teilräumen derart, dass lineare Projektoren $\Pi_h : U \to U_h$ existieren mit

$$b(v - \Pi_h v, p_h) = 0 \quad \text{für alle } v \in U \text{ und } p_h \in V_h,$$

und

$$\|\Pi_h v\|_U \le c_\Pi \|v\|_U \quad \text{für alle } v \in U$$

mit einer von h unabhängigen Konstante c_Π. Dann erfüllt b auch die LBB-Bedingung für die Räume $U_h \times V_h$ mit einer von h unabhängigen Konstanten $\widetilde{\gamma}$.

Beweis. Nach Voraussetzung gilt wegen $\Pi_h v \in U_h$ für beliebiges $p_h \in V_h$

$$\begin{aligned}\gamma\|p_h\|_V &\le \sup_{v\in U}\frac{b(v,p_h)}{\|v\|_U}\\ &= \sup_{v\in U}\frac{\overbrace{b(v-\Pi_h v,p_h)}^{=0}+b(\Pi_h v,p_h)}{\|v\|_U}\\ &\le c_\Pi \sup_{v\in U}\frac{b(\Pi_h v,p_h)}{\|\Pi_h v\|_U}\\ &\le c_\Pi \sup_{v\in U_h}\frac{b(v_h,p_h)}{\|v_h\|_U}.\end{aligned}$$

Hieraus folgt die Behauptung mit $\widetilde{\gamma} := \gamma/c_\Pi$. □

13.5 Lösbarkeit der gemischten Formulierung des Poisson-Problems

Primal-gemischte Formulierung: Wir untersuchen jetzt die sogenannte *primal-gemischte Formulierung* des Poisson-Problems

$$\begin{aligned}(\mathbf{v},\mathbf{w})_{L^2(\Omega)} - (\nabla p,\mathbf{w})_{L^2(\Omega)} &= 0 && \text{für alle } \mathbf{w}\in[L^2(\Omega)]^d,\\ (\mathbf{v},\nabla q)_{L^2(\Omega)} &= (f,q)_{L^2(\Omega)} && \text{für alle } q\in H_0^1(\Omega).\end{aligned} \tag{13.5}$$

Die Existenz und Eindeutigkeit von Lösungen ergibt sich aus dem Satz 13.6.

Satz 13.9 Zum Poisson-Problem in der primal-gemischten Formulierung (13.5) existiert für jedes $f \in \left(H_0^1(\Omega)\right)'$ eine eindeutige Lösung $(\mathbf{v},p)\in[L^2(\Omega)]^d\times H_0^1(\Omega)$.

Beweis. Wir setzen $U := [L^2(\Omega)]^d$, $V := H_0^1(\Omega)$, ausgestattet mit der Norm $|\cdot|_{H^1(\Omega)}$, und

$$a(\mathbf{v},\mathbf{w}) := (\mathbf{v},\mathbf{w})_{L^2(\Omega)},\quad b(\mathbf{v},p) := (\mathbf{v},\nabla p)_{L^2(\Omega)}.$$

Die Bilinearform a ist U-elliptisch, denn für $\mathbf{v}\in V$ gilt $a(\mathbf{v},\mathbf{v}) = \|\mathbf{v}\|_{L^2(\Omega)}$. Somit ist sie auch auf dem Kern von b elliptisch.

Um die inf-sup-Bedingung für b zu zeigen, wählen wir speziell $\mathbf{v} := \nabla p$. Wäre dabei $\mathbf{v} = \nabla p = \mathbf{0}$, so wäre p konstant, was wegen $p\in H_0^1(\Omega)$ wiederum $p = 0$ impliziert. Somit ist für $p \neq 0$ auch $\mathbf{v} = \nabla p \in U\setminus\{\mathbf{0}\}$ und es ergibt sich

$$\inf_{p\in V\setminus\{0\}}\sup_{\mathbf{v}\in U\setminus\{\mathbf{0}\}}\frac{b(\mathbf{v},p)}{\|\mathbf{v}\|_U\|p\|_V} \ge \inf_{p\in V\setminus\{0\}}\frac{b(\nabla p,p)}{\|\nabla p\|_U\|p\|_V} = \inf_{p\in V\setminus\{0\}}\frac{(\nabla p,\nabla p)_{L^2(\Omega)}}{|p|^2_{H^1(\Omega)}} = 1.$$

Folglich sind die Voraussetzungen von Satz 13.6 erfüllt, woraus sich die Behauptung ergibt. □

Dual-gemischte Formulierung: Wir betrachten den Hilbert-Raum

$$H_{\mathrm{div}}(\Omega) := \{\mathbf{v} \in [L^2(\Omega)]^d : \operatorname{div} \mathbf{v} \in L^2(\Omega)\}$$

mit dem Skalarprodukt und der Norm

$$\begin{aligned}(\mathbf{v}, \mathbf{w})_{H_{\mathrm{div}}(\Omega)} &:= (\mathbf{v}, \mathbf{w})_{L^2(\Omega)} + (\operatorname{div} \mathbf{v}, \operatorname{div} \mathbf{w})_{L^2(\Omega)},\\ \|\mathbf{v}\|^2_{H_{\mathrm{div}}(\Omega)} &:= \|\mathbf{v}\|^2_{L^2(\Omega)} + \|\operatorname{div} \mathbf{v}\|^2_{L^2(\Omega)}.\end{aligned}$$

Eine weitere gemischte Formulierung, die sogenannte *dual-gemische Formulierung*, lautet: suche $(\mathbf{v}, p) \in H_{\mathrm{div}}(\Omega) \times L^2(\Omega)$, so dass

$$\begin{aligned}(\mathbf{v}, \mathbf{w})_{L^2(\Omega)} + (p, \operatorname{div} \mathbf{w})_{L^2(\Omega)} &= 0 && \text{für alle } \mathbf{w} \in H_{\mathrm{div}}(\Omega),\\ -(\operatorname{div} \mathbf{v}, q)_{L^2(\Omega)} &= (f, q)_{L^2(\Omega)} && \text{für alle } q \in L^2(\Omega).\end{aligned} \tag{13.6}$$

Hierbei fällt auf, dass die homogene Dirichlet-Bedingung $p|_\Gamma = 0$ gar nicht explizit auftritt. Wir können sie auch (zunächst) nicht in den Raum V einbauen, da er ja nur aus $L^2(\Omega)$-Funktionen besteht, welche auf dem Rand nicht stetig zu sein brauchen. Es stellt sich allerdings heraus, dass die Randbedingungen bereits eingebaut sind:

Satz 13.10 Beim Poisson-Problem in der dual-gemischten Formulierung (13.6) existiert für jedes $f \in \big(H^1_0(\Omega)\big)'$ eine eindeutige Lösung $(\mathbf{v}, p) \in H_{\mathrm{div}}(\Omega) \times L^2(\Omega)$. Für p gilt sogar $p \in H^1_0(\Omega)$.

Beweis. Das Problem (13.6) besitzt Sattelpunktform (13.3) in den Hilbert-Räumen $U = H_{\mathrm{div}}(\Omega)$ und $V := L^2(\Omega)$ und den Bilinearformen

$$a(\mathbf{v}, \mathbf{w}) := (\mathbf{v}, \mathbf{w})_{L^2(\Omega)}, \quad b(\mathbf{v}, p) := -(\operatorname{div} \mathbf{v}, p)_{L^2(\Omega)}.$$

Dabei ist der Kern von b gerade

$$U_0 := \{\mathbf{v} \in U : (\operatorname{div} \mathbf{v}, p)_{L^2(\Omega)} = 0 \text{ für alle } p \in L^2(\Omega)\},$$

das heißt, er enthält alle Funktionen mit $\operatorname{div} \mathbf{v} = 0$.
Die Bilinearform a ist in U beschränkt. Außerdem ist a auch U_0-elliptisch, denn für $\mathbf{v} \in U_0$ gilt $\|\operatorname{div} \mathbf{v}\|_{L^2(\Omega)} = 0$ und somit

$$a(\mathbf{v}, \mathbf{v}) = \|\mathbf{v}\|^2_{L^2(\Omega)} = \|\mathbf{v}\|^2_{L^2(\Omega)} + \|\operatorname{div} \mathbf{v}\|^2_{L^2(\Omega)} = \|\mathbf{v}\|^2_{H_{\mathrm{div}}(\Omega)}.$$

Nun ist noch die inf-sup-Bedingung für b zu zeigen. Sei hierzu $p \in L^2(\Omega) \setminus \{0\}$ beliebig. Wir konstruieren uns ein geeignetes $\mathbf{v} \in U \setminus \{\mathbf{0}\}$. Es existiert ein $\widetilde{p} \in C^\infty_0(\Omega)$ mit

$$\|p - \widetilde{p}\|^2_{L^2(\Omega)} \le \frac{1}{2}\|p\|^2_{L^2(\Omega)}.$$

Nun setzen wir die erste Komponente von $\mathbf{v}$ wie folgt

$$v_1(x_1, x_2, \ldots, x_d) := -\int_{-\infty}^{x_1} \widetilde{p}(t, x_2, \ldots, \ldots, x_d)\,\mathrm{d}t.$$

Das Integral ist wohldefiniert, da Ω beschränkt ist. Alle übrigen Komponenten setzen wir zu Null: $v_2 = \cdots = v_d := 0$. Per Konstruktion gilt punktweise

$$\operatorname{div} \mathbf{v} = \frac{\partial v_1}{\partial x_1} = -\widetilde{p}.$$

Daher folgt

$$\begin{aligned} b(\mathbf{v}, p) &= -(\operatorname{div} \mathbf{v}, p)_{L^2(\Omega)} \\ &= (\widetilde{p}, p)_{L^2(\Omega)} \\ &= \frac{1}{2}\Big\{ \|\widetilde{p}\|^2_{L^2(\Omega)} + \|p\|^2_{L^2(\Omega)} - \underbrace{\|\widetilde{p} - p\|^2_{L^2(\Omega)}}_{\leq \frac{1}{2}\|p\|^2_{L^2(\Omega)}} \Big\} \\ &\geq \frac{1}{4}\Big\{ \|\widetilde{p}\|^2_{L^2(\Omega)} + \|p\|^2_{L^2(\Omega)} \Big\} \\ &\geq \frac{1}{2}\|\widetilde{p}\|_{L^2(\Omega)}\|p\|_{L^2(\Omega)}. \end{aligned}$$

Eine Argumentation wie im Beweis der Poincaré-Friedrichsschen Ungleichung (Satz 3.5) liefert

$$\|\mathbf{v}\|_{L^2(\Omega)} \leq c\|\operatorname{div} \mathbf{v}\|_{L^2(\Omega)},$$

woraus sich

$$\begin{aligned} \|\mathbf{v}\|^2_{H_{\operatorname{div}}(\Omega)} &= \|\mathbf{v}\|^2_{L^2(\Omega)} + \|\operatorname{div} \mathbf{v}\|^2_{L^2(\Omega)} \\ &\leq (1 + c^2)\|\operatorname{div} \mathbf{v}\|^2_{L^2(\Omega)} \\ &= (1 + c^2)\|\widetilde{p}\|^2_{L^2(\Omega)} \end{aligned}$$

ergibt. Folglich gilt

$$\frac{b(\mathbf{v}, p)}{\|\mathbf{v}\|_U \|p\|_V} = \frac{(\widetilde{p}, p)_{L^2(\Omega)}}{\|\mathbf{v}\|_{H_{\operatorname{div}}(\Omega)}\|p\|_{L^2(\Omega)}} \geq \frac{1}{\sqrt{1 + c^2}} \frac{(\widetilde{p}, p)_{L^2(\Omega)}}{\|\widetilde{p}\|_{L^2(\Omega)}\|p\|_{L^2(\Omega)}} \geq \frac{1}{2\sqrt{1 + c^2}}.$$

Also ist die LBB-Bedingung erfüllt. Satz 13.6 impliziert deshalb die Existenz und Eindeutigkeit einer Lösung $(\mathbf{v}, p) \in H_{\operatorname{div}}(\Omega) \times L^2(\Omega)$.
Es fehlt nun noch der Nachweis $p \in H^1_0(\Omega)$. Da $[C^\infty_0(\Omega)]^d \subset U$ ist, gilt für die Lösung p gemäß (13.6)

$$(\mathbf{v}, \boldsymbol{\phi})_{L^2(\Omega)} = -(p, \operatorname{div} \boldsymbol{\phi})_{L^2(\Omega)} \quad \text{für alle } \boldsymbol{\phi} \in [C^\infty_0(\Omega)]^d.$$

Dies ist aber gerade das Kriterium für $p \in H^1(\Omega)$ mit schwacher Ableitung $\nabla p = \mathbf{v}$. Dass die Spur von p auf dem Gebietsrand Γ verschwindet, sieht man wieder anhand von (13.6) durch Testen mit $\boldsymbol{\phi} \in [C^\infty(\Omega)]^d \subset U$:

$$\begin{aligned} 0 &= (\mathbf{v}, \boldsymbol{\phi})_{L^2(\Omega)} + (p, \operatorname{div} \boldsymbol{\phi})_{L^2(\Omega)} \\ &= (\nabla p, \boldsymbol{\phi})_{L^2(\Omega)} + (p, \operatorname{div} \boldsymbol{\phi})_{L^2(\Omega)} \\ &= \int_\Gamma \langle \boldsymbol{\phi}, \mathbf{n} \rangle \, p \, \mathrm{d}\sigma. \end{aligned}$$

Damit haben wir $p \in H^1_0(\Omega)$ gezeigt. □

13.6 Raviart-Thomas-Element

Wir stellen jetzt das *Raviart-Thomas-Element* zur Diskretisierung der dual-gemischten Formulierung (13.6) in zwei Dimensionen vor. Dazu sei

$$U_h := \big\{\mathbf{v}_h \in [L^2(\Omega)]^2 : \mathbf{v}_h(x,y)|_T = (a_T, b_T) + c_T(x,y) \text{ mit } a_T, b_T, c_T \in \mathbb{R} \text{ für alle } T \in \mathcal{T}_h, \\ \langle \mathbf{v}_h, \mathbf{n}\rangle \text{ ist stetig an den Elementgrenzen}\big\}.$$

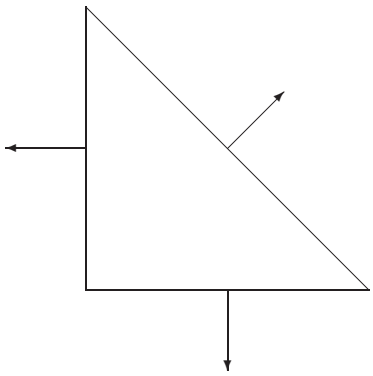
Als endlich-dimensionalen Teilraum von $V = L^2(\Omega)$ wählen wir hierbei elementweise konstante Funktionen

$$V_h := \big\{p \in L^2(\Omega) : p|_T \in \mathcal{P}_0 \text{ für alle } T \in \mathcal{T}_h\big\}.$$

Eine Raviart-Thomas-Funktion $\mathbf{v}_h \in U_h$ erfüllt offenbar $\mathbf{v}_h|_T \in [\mathcal{P}_1]^2$ und besitzt drei Freiheitsgrade pro Element. Die Normalkomponente einer Raviart-Thomas-Funktion ist auf jeder Dreieckskante konstant, was man wie folgt einsieht. Die Kante e sei gegeben durch $(\alpha, \beta) + t(\gamma, \delta)$; die dazugehörige Normale ist also $\mathbf{n}_e = (-\delta, \gamma)$. Es folgt

$$\begin{aligned}\langle \mathbf{v}_h(\alpha + t\gamma, \beta + t\delta), \mathbf{n}_e\rangle &= \langle \big(a_T + c_T(\alpha + t\gamma), b_T + c_T(\beta + t\delta)\big), \mathbf{n}_e\rangle \\ &= \langle (a_T, b_T) + c_T(\alpha, \beta), \mathbf{n}_e\rangle,\end{aligned}$$

das heißt, die Normalkomponenten sind jeweils konstant. Die drei Freiheitgrade auf jeder Zelle können daher repräsentiert werden durch die drei Normalenkomponenten $\langle \mathbf{v}_h, \mathbf{n}_e\rangle$ zu den drei Kanten $e \in \mathcal{E}_T$:



Wir wollen nun einmal erläutern, wie man aus der Vorgabe der drei Normalenkomponenten den Ansatz $\mathbf{v}_h|_T$ ermittelt: Zu jedem Eckpunkt $\mathbf{x}_i$, $i = 1, 2, 3$, des Dreiecks T bestimmt man die beiden Komponenten von $\mathbf{v}_h(\mathbf{x}_i)$ aus den Normalenkomponenten der jeweiligen beiden angrenzenden Kanten. Zu den sechs Werten $v_{h,1}(\mathbf{x}_i)$, $v_{h,2}(\mathbf{x}_i)$, $i = 1, 2, 3$, gibt es nun ein eindeutiges lineares Polynom $\mathbf{v}_h|_T \in \mathcal{P}_1^2$.

Lemma 13.11 Die Abbildung $\mathrm{div} : U_h \to V_h$ ist surjektiv und für jedes $p_h \in V_h$ genügt das Urbild $\mathbf{v}_h \in U_h$ der Abschätzung

$$\|\mathbf{v}_h\|_{H_{\mathrm{div}}(\Omega)} \le c\|p_h\|_{L^2(\Omega)}. \tag{13.7}$$

Beweis. Sei $p_h \in V_h$ gegeben. Wir wählen ein konvexes Polygongebiet $\widetilde{\Omega} \supset \Omega$ und setzen p_h trivial auf $\widetilde{\Omega}$ fort. Dann existiert ein $u \in H^2(\widetilde{\Omega})$ mit $\Delta u = p_h$ auf $\widetilde{\Omega}$ und $u|_{\partial\widetilde{\Omega}} = 0$. Für $\mathbf{v} := \nabla u \in H^1(\Omega)$ gilt dann $\operatorname{div} \mathbf{v} = p_h$ auf Ω. Nun müssen wir dieses $\mathbf{v}$ noch geeignet auf den Raum U_h projezieren. Hierzu betrachten wir auf jeder Kante $e \in \mathcal{E}_h$ die Mittelwerte der Normalkomponenten von $\mathbf{v}$. Als $\mathbf{v}_h \in U_h$ wählen wir nun diejenige Raviart-Thomas-Funktion, deren Normalkomponenten an den Kanten genau diesen Mittelwerten entspricht. Nach Konstruktion gilt für jedes Element $T \in \mathcal{T}_h$

$$\int_T p_h \,\mathrm{d}\mathbf{x} = \int_T \operatorname{div} \mathbf{v} \,\mathrm{d}\mathbf{x} = \int_{\partial T} \langle \mathbf{v}, \mathbf{n} \rangle \,\mathrm{d}\sigma = \int_{\partial T} \langle \mathbf{v}_h, \mathbf{n} \rangle \,\mathrm{d}\sigma = \int_T \operatorname{div} \mathbf{v}_h \,\mathrm{d}\mathbf{x}.$$

Da $p_h \in V_h$ und $\operatorname{div} \mathbf{v}_h$ auf jedem Element T konstant sind, folgt $\operatorname{div} \mathbf{v}_h(\mathbf{x}) = p_h(\mathbf{x})$ auf Ω. Mittels Transformation auf das Referenzelement zeigt man schnell, dass gilt

$$\|\mathbf{v}_h\|_{L^2(T)} \le c\|\mathbf{v}\|_{H^1(T)} \quad \text{für alle } T \in \mathcal{T}_h$$

mit einer Konstanten $c = c(\kappa)$. Hieraus folgt dann

$$\begin{aligned} \|\mathbf{v}_h\|^2_{H_{\operatorname{div}}(\Omega)} &= \|\mathbf{v}_h\|^2_{L^2(\Omega)} + \|\operatorname{div} \mathbf{v}_h\|^2_{L^2(\Omega)} \\ &= \sum_{T \in \mathcal{T}_h} \|\mathbf{v}_h\|^2_{L^2(T)} + \|p_h\|^2_{L^2(\Omega)} \\ &\le c\|\mathbf{v}\|^2_{H^1(\Omega)} + \|p_h\|^2_{L^2(\Omega)}. \end{aligned}$$

Aufgrund der Konvexität von $\widetilde{\Omega}$ folgt aus Satz 6.3

$$\|\mathbf{v}\|^2_{H^1(\Omega)} = \|u\|^2_{H^2(\Omega)} \le \|u\|^2_{H^2(\widetilde{\Omega})} \le c\|p_h\|^2_{L^2(\widetilde{\Omega})}.$$

Wegen der trivialen Fortsetzung von p_h auf $\widetilde{\Omega}$ folgt die gewünschte Abschätzung. □

Satz 13.12 Die Diskretisierung mit Raviart-Thomas-Elementen liefert für das Poisson-Problem zu jeder rechten Seite $f \in \big(H^1_0(\Omega)\big)'$ eine eindeutige Lösung $(v_h, p_h) \in U_h \times V_h$.

Beweis. Wir hatten bereits gesehen, dass die kontinuierliche Sattelpunktformulierung im Raum $U \times V = H_{\operatorname{div}}(\Omega) \times L^2(\Omega)$ für das Poisson-Problem die Bedingungen von Satz 13.6 erfüllt und daher stets eine eindeutige Lösung liefert. Wir wollen jetzt nachprüfen, ob die Bedingungen auch für den diskreten Ansatzraum $U_h \times V_h$ gelten.

Hierzu werden wir zunächst überprüfen, ob $U_h \times V_h \subset H_{\operatorname{div}}(\Omega) \times L^2(\Omega)$ gilt. Hierbei ist $V_h \subset L^2(\Omega)$ offensichtlich. Es bleibt nachzuweisen, dass für jedes $\mathbf{v}_h \in U_h$ gilt $\operatorname{div} \mathbf{v}_h \in L^2(\Omega)$. Da $\mathbf{v}_h$ elementweise polynomial ist, gilt $(\operatorname{div} \mathbf{v}_h)|_T \in L^2(T)$ für alle $T \in \mathcal{T}_h$. Hieraus folgt sofort $\operatorname{div} \mathbf{v}_h \in L^2(\Omega)$.

Da die Bilinearform $a(\cdot,\cdot)$ auf $H_{\operatorname{div}}(\Omega)$ beschränkt und auf U_0 elliptisch ist, ist sie auch auf U_h beschränkt und auf $U_{0,h} := U_h \cap U_0$ elliptisch. Es verbleibt demnach, die inf-sup-Bedingung für $b(\cdot,\cdot)$ zu zeigen. Dies geschieht durch das Kriterium von Fortin (Satz 13.8). Wir müssen dazu eine lineare Projektion $\Pi_h : H_{\operatorname{div}}(\Omega) \to U_h$ konstruieren mit h-unabhängiger Operatornorm und

$$b(\mathbf{v} - \Pi_h \mathbf{v}, p_h) = \big(\operatorname{div}(\mathbf{v} - \Pi_h \mathbf{v}), p_h\big)_{L^2(\Omega)} = 0 \quad \text{für alle } \mathbf{v} \in H_{\operatorname{div}}(\Omega) \text{ und } p_h \in V_h. \tag{13.8}$$

Der Raum V_h besteht aus elementweise konstanten Funktionen. Daher lässt sich Bedingung (13.8) auch lokal formulieren:

$$\int_T \operatorname{div}(\Pi_h \mathbf{v})\,\mathrm{d}\mathbf{x} = \int_T \operatorname{div}\mathbf{v}\,\mathrm{d}\mathbf{x} \quad \text{für alle } \mathbf{v} \in H_{\operatorname{div}}(\Omega) \text{ und } T \in \mathcal{T}_h.$$

Dass dies möglich ist, folgt nun aus dem vorhergehenden Lemma, indem wir $p_h \in V_h$ auf jedem Element T konstant wählen als

$$p_T = p_h|_T := \frac{1}{|T|}\int_T \operatorname{div}\mathbf{v}\,\mathrm{d}\mathbf{x}$$

und $\Pi_h\mathbf{v}$ als das zugehörige Urbild. Es folgt dann nämlich

$$\int_T \operatorname{div}(\Pi_h\mathbf{v})\,\mathrm{d}\mathbf{x} = \int_T p_h\,\mathrm{d}\mathbf{x} = |T|\,p_T = \int_T \operatorname{div}\mathbf{v}\,\mathrm{d}\mathbf{x}.$$

Außerdem folgt die h-Unabhängigkeit von der Norm der Projektion aus (13.7). Es ist

$$\|\Pi_h\mathbf{v}\|_{H_{\operatorname{div}}(\Omega)} \le c\|p_h\|_{L^2(\Omega)}$$

und weiter

$$\begin{aligned}
\|p_h\|^2_{L^2(\Omega)} &= \sum_{T\in\mathcal{T}_h} \|p_h\|^2_{L^2(T)} \\
&= \sum_{T\in\mathcal{T}_h} |p_T|^2|T| \\
&\le \sum_{T\in\mathcal{T}_h} \|\operatorname{div}\mathbf{v}\|^2_{L^1(T)}|T|^{-1} \\
&\le \sum_{T\in\mathcal{T}_h} \|\operatorname{div}\mathbf{v}\|^2_{L^2(T)}\|1\|^2_{L^2(T)}|T|^{-1} \\
&= \|\operatorname{div}\mathbf{v}\|^2_{L^2(\Omega)},
\end{aligned}$$

also insgesamt

$$\|\Pi_h\mathbf{v}\|_{H_{\operatorname{div}}(\Omega)} \le c\|\operatorname{div}\mathbf{v}\|_{L^2(\Omega)} \le c\|\mathbf{v}\|_{H_{\operatorname{div}}(\Omega)}.$$

Damit ist der Beweis vollständig erbracht. □

Satz 13.13 Wenn für die Lösung gilt $\mathbf{v} \in [H^1(\Omega)]^d$ und $p \in H^1(\Omega)$, so gilt auf quasiuniformen Triangulierungen für den Diskretisierungsfehler des Raviart-Thomas-Elements

$$\|\mathbf{v}-\mathbf{v}_h\|_{H_{\operatorname{div}}(\Omega)} + \|p-p_h\|_{L^2(\Omega)} \le ch\big\{|\mathbf{v}|_{H^1(\Omega)} + \|p\|_{H^1(\Omega)}\big\} + \inf_{f_h\in V_h}\|f-f_h\|_{L^2(\Omega)}.$$

Beweis. Aufgrund von Korollar 13.7 wissen wir

$$\|\mathbf{v}-\mathbf{v}_h\|_{H_{\operatorname{div}}(\Omega)} + \|p-p_h\|_{L^2(\Omega)} \le c\Big\{\inf_{\mathbf{w}_h\in U_h}\|\mathbf{v}-\mathbf{w}_h\|_{H^1(\Omega)} + \inf_{q_h\in V_h}\|p-q_h\|_{L^2(\Omega)}\Big\}.$$

Für elementweise konstante Ansatzfunktionen aus V_h erhalten wir

$$\inf_{q_h\in V_h}\|p-q_h\|_{L^2(\Omega)} \le ch\|p\|_{H^1(\Omega)}.$$

Es fehlt also nur noch eine geeignete Schranke für die Interpolation von $\mathbf{v} \in [H^1(\Omega)]^d$ zu finden.
Wie im Beweis von Lemma 13.11 existiert zu jedem $\mathbf{v} \in [H^1(\Omega)]^d$ eine Funktion $I_h\mathbf{v} \in U_h$ mit

$$\big(\operatorname{div}(\mathbf{v} - I_h\mathbf{v}), q_h\big)_{L^2(\Omega)} = 0 \quad \text{für alle } q_h \in V_h.$$

Falls $\mathbf{v}$ elementweise linear ist, gilt $\mathbf{v} = I_h\mathbf{v}$. Nach dem Bramble-Hilbert-Lemma folgt daher

$$\|\mathbf{v} - I_h\mathbf{v}\|_{L^2(\Omega)} \le ch|\mathbf{v}|_{H^1(\Omega)}.$$

Um den Fehler in der Divergenz abzuschätzen, zeigen wir folgende Minimaleigenschaft:

$$\|\operatorname{div}(\mathbf{v} - I_h\mathbf{v})\|_{L^2(\Omega)} = \inf_{q_h \in V_h} \|\operatorname{div}\mathbf{v} - q_h\|_{L^2(\Omega)}.$$

Wenn dies gezeigt ist, folgt wegen $\operatorname{div}\mathbf{v} = f$ die Behauptung.
Die gesuchte Minimaleigenschaft erfüllt offensichtlich gerade die L^2-Orthoprojektion von $\operatorname{div}\mathbf{v}$ auf den Raum U_h. Wir hatten aber bereits gesehen, dass

$$\big(\operatorname{div}(\mathbf{v} - I_h\mathbf{v}), q_h\big)_{L^2(\Omega)} = 0 \quad \text{für alle } q_h \in V_h.$$

Daher ist $\operatorname{div}(I_h\mathbf{v})$ gerade die L^2-Orthoprojektion von $\operatorname{div}\mathbf{v}$ auf V_h. Insgesamt folgt schließlich

$$\inf_{\mathbf{w}_h \in U_h} \|\mathbf{v} - \mathbf{w}_h\|_{H^1(\Omega)} \le c\|\mathbf{v} - I_h\mathbf{v}\|_{H^1(\Omega)} \le ch|\mathbf{v}|_{H^1(\Omega)} + \inf_{q_h \in V_h} \|f - q_h\|_{L^2(\Omega)}.$$

□

13.7 Bramble-Pasciak-CG

Bei der Diskretisierung eines Sattelpunktproblems erhält man ganz allgemein ein Gleichungssystem der Form

$$\underbrace{\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & -\mathbf{C} \end{bmatrix}}_{=:\mathbf{S}} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}. \tag{13.9}$$

Hierbei sei $\mathbf{A} \in \mathbb{R}^{m\times m}$ symmetrisch und positiv definit, während $\mathbf{C} \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit ist. Die inf-sup-Bedingung impliziert, dass auch $\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}$ positiv definit ist. Folglich ist die Systemmatrix symmetrisch, jedoch wegen

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{0} \end{bmatrix}^T \mathbf{S} \begin{bmatrix} \mathbf{x} \\ \mathbf{0} \end{bmatrix} = \mathbf{x}^T\mathbf{A}\mathbf{x} > \mathbf{0}, \quad \begin{bmatrix} \mathbf{0} \\ \mathbf{y} \end{bmatrix}^T \mathbf{S} \begin{bmatrix} \mathbf{0} \\ \mathbf{y} \end{bmatrix} = -\mathbf{y}^T\mathbf{C}\mathbf{y} \le \mathbf{0}, \quad \mathbf{x} \neq \mathbf{0}$$

indefinit.
Wir nehmen nun an, dass $\mathbf{A}_0 \in \mathbb{R}^{m\times m}$ eine symmetrische und positiv definite Matrix ist mit

$$0 < \mathbf{x}^T\mathbf{A}_0\mathbf{x} < \mathbf{x}^T\mathbf{A}\mathbf{x}, \quad \mathbf{x} \neq \mathbf{0}. \tag{13.10}$$

Wir multiplizieren (13.9) von links mit

$$\mathbf{T} := \begin{bmatrix} \mathbf{A}_0^{-1} & \mathbf{0} \\ \mathbf{B}^T\mathbf{A}_0^{-1} & -\mathbf{I} \end{bmatrix} \in \mathbb{R}^{(m+n)\times(m+n)}$$

und erhalten

$$\underbrace{\begin{bmatrix} \mathbf{A}_0^{-1}\mathbf{A} & \mathbf{A}_0^{-1}\mathbf{B} \\ \mathbf{B}^T\mathbf{A}_0^{-1}(\mathbf{A}-\mathbf{A}_0) & \mathbf{B}^T\mathbf{A}_0^{-1}\mathbf{B}+\mathbf{C} \end{bmatrix}}_{=:\widehat{\mathbf{S}}} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_0^{-1}\mathbf{f} \\ \mathbf{B}^T\mathbf{A}_0^{-1}\mathbf{f}-\mathbf{g} \end{bmatrix}.$$

Satz 13.14 Die Systemmatrix $\widehat{\mathbf{S}} = \mathbf{TS}$ ist symmetrisch und positiv definit bezüglich des Innenproduktes

$$\left\langle \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}, \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \right\rangle_{\mathbf{M}} := \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}^T \underbrace{\begin{bmatrix} \mathbf{A}-\mathbf{A}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}}_{=:\mathbf{M}} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}.$$

Beweis. Unter der Voraussetzung (13.10) ist die Matrix

$$\mathbf{M} = \begin{bmatrix} \mathbf{A}-\mathbf{A}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

symmetrisch und positiv definit, so dass $\langle \cdot, \cdot \rangle_{\mathbf{M}}$ in der Tat ein Innenprodukt definiert. Nun gilt

$$\begin{aligned} \left\langle \widehat{\mathbf{S}} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}, \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \right\rangle_{\mathbf{M}} &= \left\langle \begin{bmatrix} \mathbf{A}_0^{-1}(\mathbf{A}\mathbf{u}+\mathbf{B}\mathbf{v}) \\ \mathbf{B}^T\mathbf{A}_0^{-1}(\mathbf{A}-\mathbf{A}_0)\mathbf{u} + (\mathbf{B}^T\mathbf{A}_0^{-1}\mathbf{B}+\mathbf{C})\mathbf{v} \end{bmatrix}, \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \right\rangle_{\mathbf{M}} \\ &= (\mathbf{A}\mathbf{u}+\mathbf{B}\mathbf{v})^T\mathbf{A}_0^{-1}(\mathbf{A}-\mathbf{A}_0)\mathbf{x} \\ &\qquad + \big(\mathbf{B}^T\mathbf{A}_0^{-1}(\mathbf{A}-\mathbf{A}_0)\mathbf{u} + (\mathbf{B}^T\mathbf{A}_0^{-1}\mathbf{B}+\mathbf{C})\mathbf{v}\big)^T\mathbf{y} \\ &= \mathbf{u}^T(\mathbf{A}\mathbf{A}_0^{-1}\mathbf{A}-\mathbf{A})\mathbf{x} + \mathbf{v}^T\mathbf{B}^T\mathbf{A}_0^{-1}(\mathbf{A}-\mathbf{A}_0)\mathbf{x} \\ &\qquad + \mathbf{u}^T(\mathbf{A}-\mathbf{A}_0)\mathbf{A}_0^{-1}\mathbf{B}\mathbf{y} + \mathbf{v}^T(\mathbf{B}^T\mathbf{A}_0^{-1}\mathbf{B}+\mathbf{C})\mathbf{y}. \end{aligned}$$

Andererseits ergibt sich

$$\begin{aligned} \left\langle \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}, \widehat{\mathbf{S}} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \right\rangle_{\mathbf{M}} &= \left\langle \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}, \begin{bmatrix} \mathbf{A}_0^{-1}(\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{y}) \\ \mathbf{B}^T\mathbf{A}_0^{-1}(\mathbf{A}-\mathbf{A}_0)\mathbf{x} + (\mathbf{B}^T\mathbf{A}_0^{-1}\mathbf{B}+\mathbf{C})\mathbf{y} \end{bmatrix} \right\rangle_{\mathbf{M}} \\ &= \mathbf{u}^T(\mathbf{A}-\mathbf{A}_0)\mathbf{A}_0^{-1}(\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{y}) \\ &\qquad + \mathbf{v}^T\big(\mathbf{B}^T\mathbf{A}_0^{-1}(\mathbf{A}-\mathbf{A}_0)\mathbf{x} + (\mathbf{B}^T\mathbf{A}_0^{-1}\mathbf{B}+\mathbf{C})\mathbf{y}\big) \\ &= \mathbf{u}^T(\mathbf{A}\mathbf{A}_0^{-1}\mathbf{A}-\mathbf{A})\mathbf{x} + \mathbf{u}^T(\mathbf{A}-\mathbf{A}_0)\mathbf{A}_0^{-1}\mathbf{B}\mathbf{y} \\ &\qquad + \mathbf{v}^T\mathbf{B}^T\mathbf{A}_0^{-1}(\mathbf{A}-\mathbf{A}_0)\mathbf{x} + \mathbf{v}^T(\mathbf{B}^T\mathbf{A}_0^{-1}\mathbf{B}+\mathbf{C})\mathbf{y}. \end{aligned}$$

Damit haben wir die Symmetrie gezeigt.
Weiter ist

$$\begin{aligned} \left\langle \widehat{\mathbf{S}} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \right\rangle_{\mathbf{M}} &= \big(\mathbf{x}^T(\mathbf{A}-\mathbf{A}_0) + \mathbf{y}^T\mathbf{B}^T\big)\mathbf{A}_0^{-1}\big((\mathbf{A}-\mathbf{A}_0)\mathbf{x}+\mathbf{B}\mathbf{y}\big) \\ &\qquad + \mathbf{x}^T(\mathbf{A}-\mathbf{A}_0)\mathbf{x} + \mathbf{y}^T\mathbf{C}\mathbf{y}. \end{aligned}$$

Zerlegen wir nun

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{y} \end{bmatrix},$$

wobei $\mathbf{x}_1$ die eindeutige Lösung der Gleichung

$$\mathbf{A}\mathbf{x}_1 + \mathbf{B}\mathbf{y} = \mathbf{0}$$

sei, ergibt sich der Zusammenhang

$$\mathbf{x}_1^T \mathbf{A}\mathbf{x}_1 = \mathbf{y}^T \mathbf{B}^T \mathbf{A}^{-1} \mathbf{B}\mathbf{y}.$$

Damit erhalten wir

$$\left\langle \widehat{\mathbf{S}} \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{0} \end{bmatrix} \right\rangle_{\mathbf{M}} = \mathbf{x}_0^T (\mathbf{A}\mathbf{A}_0^{-1}\mathbf{A} - \mathbf{A})\mathbf{x}_0 \geq \underbrace{\lambda_{\min}(\mathbf{A}\mathbf{A}_0^{-1}\mathbf{A} - \mathbf{A})}_{=:c_1>0} \|\mathbf{x}_0\|_2^2$$

und

$$\begin{aligned} \left\langle \widehat{\mathbf{S}} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{y} \end{bmatrix}, \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{y} \end{bmatrix} \right\rangle_{\mathbf{M}} &= \mathbf{x}_1^T \mathbf{A}_0 \mathbf{x}_1 + \mathbf{x}_1^T (\mathbf{A} - \mathbf{A}_0)\mathbf{x}_1 + \mathbf{y}^T \mathbf{C}\mathbf{y} \\ &= \frac{1}{2}\mathbf{x}_1^T \mathbf{A}\mathbf{x}_1 + \mathbf{y}^T \left(\frac{1}{2}\mathbf{B}^T \mathbf{A}^{-1}\mathbf{B} + \mathbf{C} \right) \mathbf{y} \\ &\geq \underbrace{\frac{1}{2}\lambda_{\min}(\mathbf{A})}_{=:c_2>0} \|\mathbf{x}_1\|_2^2 + \underbrace{\frac{1}{2}\lambda_{\min}(\mathbf{B}^T \mathbf{A}^{-1}\mathbf{B})}_{=:c_3>0} \|\mathbf{y}\|_2^2. \end{aligned}$$

Wegen

$$\left\langle \widehat{\mathbf{S}} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{y} \end{bmatrix}, \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{0} \end{bmatrix} \right\rangle_{\mathbf{M}} = \left(\mathbf{x}_1^T \mathbf{A} + \mathbf{y}^T \mathbf{B}^T \right) \mathbf{A}_0^{-1} (\mathbf{A} - \mathbf{A}_0)\mathbf{x}_0 = 0,$$

schließen wir

$$\begin{aligned} \left\langle \widehat{\mathbf{S}} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \right\rangle_{\mathbf{M}} &= \left\langle \widehat{\mathbf{S}} \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{0} \end{bmatrix} \right\rangle_{\mathbf{M}} + \left\langle \widehat{\mathbf{S}} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{y} \end{bmatrix}, \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{y} \end{bmatrix} \right\rangle_{\mathbf{M}} \\ &\geq c_1 \|\mathbf{x}_0\|_2^2 + c_2 \|\mathbf{x}_1\|_2^2 + c_3 \|\mathbf{y}\|_2^2. \end{aligned}$$

Benutzen wir

$$\|\mathbf{x}\|_2^2 = \|\mathbf{x}_0 + \mathbf{x}_1\|_2^2 \leq 2\|\mathbf{x}_0\|_2^2 + 2\|\mathbf{x}_1\|_2^2,$$

so ergibt sich letztlich die positive Definitheit von $\widehat{\mathbf{S}}$:

$$\left\langle \widehat{\mathbf{S}} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \right\rangle_{\mathbf{M}} \geq \frac{1}{2} \min\{c_1, c_2\} \|\mathbf{x}\|_2^2 + c_3 \|\mathbf{y}\|_2^2.$$

□

Da $\widehat{\mathbf{S}}$ symmetrisch und positiv definit bezüglich des $\langle \cdot, \cdot \rangle_{\mathbf{M}}$-Innenprodukts ist, kann man das Verfahren der konjugierten Gradienten anwenden, um das diskrete Sattelpunktproblem (13.9) zu lösen. Es lässt sich so formulieren, dass nur einfache Matrix-Vektor-Produkte zur Verfügung gestellt werden müssen. Auch muss nur $\mathbf{A}_0^{-1}$ angewendet werden.

Algorithmus 13.15 (Bramble-Pasciak-CG)

input: Matrizen $\mathbf{A} \in \mathbb{R}^{m \times m}$, $\mathbf{B} \in \mathbb{R}^{m \times n}$, $\mathbf{C} \in \mathbb{R}^{n \times n}$, rechte Seiten $\mathbf{f} \in \mathbb{R}^m$, $\mathbf{g} \in \mathbb{R}^n$ und Startnäherungen $\mathbf{x}_0 \in \mathbb{R}^m$, $\mathbf{y}_0 \in \mathbb{R}^n$

output: Folge von Iterierten $\{(\mathbf{x}_k, \mathbf{y}_k)\}_{k>0}$

① Initialisierung: berechne

$$\begin{bmatrix}\mathbf{s}_0\\ \mathbf{t}_0\end{bmatrix} := \begin{bmatrix}\mathbf{f}\\ \mathbf{g}\end{bmatrix} - \begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & -\mathbf{C}\end{bmatrix}\begin{bmatrix}\mathbf{x}_0\\ \mathbf{y}_0\end{bmatrix}$$
$$\mathbf{c}_0 = \mathbf{p}_0 := \mathbf{A}_0^{-1}\mathbf{s}_0, \quad \mathbf{d}_0 = \mathbf{q}_0 := \mathbf{B}^T\mathbf{p}_0 - \mathbf{t}_0$$
$$r_0 := \mathbf{p}_0^T(\mathbf{A}\mathbf{p}_0 - \mathbf{s}_0) + \mathbf{q}_0^T\mathbf{q}_0$$

und setze $k := 0$

② berechne

$$\begin{bmatrix}\mathbf{u}_k\\ \mathbf{v}_k\end{bmatrix} := \begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & -\mathbf{C}\end{bmatrix}\begin{bmatrix}\mathbf{c}_k\\ \mathbf{d}_k\end{bmatrix}, \quad \mathbf{w}_k := \mathbf{A}_0^{-1}\mathbf{u}_k$$
$$\alpha_k := \frac{r_k}{\mathbf{c}_k^T(\mathbf{A}\mathbf{w}_k - \mathbf{u}_k) + \mathbf{d}_k^T(\mathbf{B}^T\mathbf{w}_k - \mathbf{v}_k)}$$
$$\begin{bmatrix}\mathbf{x}_{k+1}\\ \mathbf{y}_{k+1}\end{bmatrix} := \begin{bmatrix}\mathbf{x}_k\\ \mathbf{y}_k\end{bmatrix} + \alpha_k\begin{bmatrix}\mathbf{c}_k\\ \mathbf{d}_k\end{bmatrix}$$
$$\begin{bmatrix}\mathbf{s}_{k+1}\\ \mathbf{t}_{k+1}\end{bmatrix} := \begin{bmatrix}\mathbf{f}\\ \mathbf{g}\end{bmatrix} - \begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & -\mathbf{C}\end{bmatrix}\begin{bmatrix}\mathbf{x}_{k+1}\\ \mathbf{y}_{k+1}\end{bmatrix}$$
$$\mathbf{p}_{k+1} := \mathbf{A}_0^{-1}\mathbf{s}_{k+1}, \quad \mathbf{q}_{k+1} := \mathbf{B}^T\mathbf{p}_{k+1} - \mathbf{t}_{k+1}$$
$$r_{k+1} = \mathbf{p}_{k+1}^T(\mathbf{A}\mathbf{p}_{k+1} - \mathbf{s}_{k+1}) + \mathbf{q}_{k+1}^T\mathbf{q}_{k+1}$$
$$\begin{bmatrix}\mathbf{c}_{k+1}\\ \mathbf{d}_{k+1}\end{bmatrix} := \begin{bmatrix}\mathbf{p}_{k+1}\\ \mathbf{q}_{k+1}\end{bmatrix} + \frac{r_{k+1}}{r_k}\begin{bmatrix}\mathbf{c}_k\\ \mathbf{d}_k\end{bmatrix}$$

③ falls $\sqrt{r_{k+1}} > \varepsilon$ erhöhe $k := k + 1$ und gehe nach ②

Bemerkungen

1. Es ist
$$\begin{bmatrix}\mathbf{p}_k\\ \mathbf{q}_k\end{bmatrix} = \mathbf{T}\left\{\begin{bmatrix}\mathbf{f}\\ \mathbf{g}\end{bmatrix} - \begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & -\mathbf{C}\end{bmatrix}\begin{bmatrix}\mathbf{x}_k\\ \mathbf{y}_k\end{bmatrix}\right\}$$
das Residuum, $\sqrt{r_k}$ dessen Norm im $\langle\cdot,\cdot\rangle_{\mathbf{M}}$-Innenprodukt und $(\mathbf{c}_k, \mathbf{d}_k)$ eine bezüglich des Standardinnenprodukts **MTS**-konjugierte Suchrichtung.
2. Ist $\mathbf{A}_0$ zusätzlich zu (13.10) uniform spektral äquivalent zu $\mathbf{A}$, das heißt, gilt
$$\alpha\mathbf{x}^T\mathbf{A}_0\mathbf{x} \le \mathbf{x}^T\mathbf{A}\mathbf{x} \le \beta\mathbf{x}^T\mathbf{A}_0\mathbf{x} \quad \text{für alle } \mathbf{x} \in \mathbb{R}^m$$
mit von der Schrittweite h unabhängigen Konstanten $0 < \alpha \le \beta$, dann ist im Standardinnenprodukt die Systemmatrix **MTS** spektral äquivalent zu
$$\mathbf{N} = \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B} + \mathbf{C}\end{bmatrix}.$$
Man benötigt dann nur noch einen geeigneten Vorkonditionierer für das Schur-Komplement von $\mathbf{A}$.

△

14. Stokessche Gleichung

14.1 Herleitung

Die mitunter einfachste Modellgleichung zur Beschreibung einer Strömung ist die Stokessche Gleichung. Sie beschreibt ein extrem zähflüssiges Fluid, wie beispielsweise Honig. Im Gegensatz zu einem weniger viskosen Fluid treten hier keine kleinskaligen Wirbel auf.

Das Fluid sei inkompressibel, das heißt, die Dichte ρ ist unabhängig von $\mathbf{x}$ und t. Der Fluss $\mathbf{v}$ erfüllt aufgrund der Massenerhaltung

$$\operatorname{div} \mathbf{v} = 0. \tag{14.1}$$

Die Impulserhaltung des Fluids in einem kleinen Volumen $\Omega_t \subset \mathbb{R}^d$ führt auf

$$\rho \frac{\partial}{\partial t} \int_{\Omega_t} \mathbf{v}(\mathbf{x}, t)\, \mathrm{d}\mathbf{x} = \underbrace{\rho \int_{\Omega_t} f(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}}_{\text{äußere Kräfte}} + \underbrace{\int_{\partial\Omega_t} \mathbf{S}(\mathbf{x}, t) \cdot \mathbf{n}(\mathbf{x}, t)\, \mathrm{d}\sigma}_{\text{Oberflächenkräfte}}.$$

Der Spannungstensor $\mathbf{S} \in \mathbb{R}^{d\times d}$ hat im Fall zäher Fluide die Form

$$\mathbf{S} = -p\,\mathbf{I} + \mu\big[\nabla \mathbf{v} + (\nabla \mathbf{v})^T\big],$$

wobei der Gradient der vektorwertigen Funktion $\mathbf{v}$ gegeben ist als

$$\nabla \mathbf{v} = \big[\nabla v_1, \nabla v_2, \ldots, \nabla v_d\big]^T = \begin{bmatrix} \frac{\partial v_1}{\partial x_1} & \frac{\partial v_1}{\partial x_2} & \cdots & \frac{\partial v_1}{\partial x_d} \\ \frac{\partial v_2}{\partial x_1} & \frac{\partial v_2}{\partial x_2} & \cdots & \frac{\partial v_2}{\partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial v_d}{\partial x_1} & \frac{\partial v_d}{\partial x_2} & \cdots & \frac{\partial v_d}{\partial x_d} \end{bmatrix}.$$

Man beachte, dass sich das Gebiet Ω_t mit der Zeit ändert, das heißt, es ist

$$\frac{\partial}{\partial t} \int_{\Omega_t} v_i\, \mathrm{d}\mathbf{x} = \int_{\Omega_t} \frac{\partial}{\partial t} v_i\, \mathrm{d}\mathbf{x} + \int_{\Omega_t} \underbrace{\operatorname{div}(v_i \mathbf{v})}_{= v_i \operatorname{div} \mathbf{v} + \langle \nabla v_i, \mathbf{v} \rangle} \mathrm{d}\mathbf{x}, \quad i = 1, 2, \ldots, d.$$

Nun ist die Divergenz der matrixwertigen Funktion $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_d]^T$ gegeben durch

$$\operatorname{div} \mathbf{S} = \big[\operatorname{div} \mathbf{s}_1, \operatorname{div} \mathbf{s}_2, \ldots, \operatorname{div} \mathbf{s}_d\big]^T.$$

Setzt man die Definition des Spannungstensors ein, so folgt

$$\operatorname{div} \mathbf{S} = -\nabla p + \mu\{\Delta \mathbf{v} + \nabla(\underbrace{\operatorname{div} \mathbf{v}}_{=0})\}$$

mit

$$\Delta \mathbf{v} = \frac{\partial}{\partial x_1^2}\mathbf{v} + \frac{\partial}{\partial x_2^2}\mathbf{v} + \cdots + \frac{\partial}{\partial x_d^2}\mathbf{v}.$$

Daher liefert der Gaußsche Integralsatz

$$\int_{\partial\Omega_t} \mathbf{S}\cdot\mathbf{n}\,\mathrm{d}\sigma = \int_{\Omega_t} \operatorname{div}\mathbf{S}\,\mathrm{d}\mathbf{x} = \int_{\Omega_t} \{-\nabla p + \mu\Delta\mathbf{v}\}\,\mathrm{d}\mathbf{x},$$

schließlich

$$\rho\left[\frac{\partial}{\partial t}\mathbf{v} + \nabla\mathbf{v}\cdot\mathbf{v}\right] = \rho\,\mathbf{f} - \nabla p + \mu\Delta\mathbf{v}. \tag{14.2}$$

Die Gleichungen (14.1) und (14.2) bilden zusammen die *Navier-Stokes-Gleichung*. Bei zähen Fluiden kann der nichtlineare Term $\nabla\mathbf{v}\cdot\mathbf{v}$ vernachlässigt werden, außerdem beschränken wir uns auf stationäre Zustände. Dann erhalten wir die (stationäre) *Stokes-Gleichung*:

$$-\Delta\mathbf{v} + \nabla p = \mathbf{f}, \quad \operatorname{div}\mathbf{v} = 0 \quad \text{in } \Omega. \tag{14.3}$$

Am Rand schreiben wir uns den Fluss vor

$$\mathbf{v} = \mathbf{g} \quad \text{auf } \Gamma := \partial\Omega. \tag{14.4}$$

Damit zu den Randwerten $\mathbf{g}$ überhaupt eine divergenzfreie Strömung existieren kann, muss nach dem Gaußschen Integralsatz

$$\int_\Gamma \langle\mathbf{g},\mathbf{n}\rangle\,\mathrm{d}\sigma = \int_\Gamma \langle\mathbf{v},\mathbf{n}\rangle\,\mathrm{d}\sigma = \int_\Omega \operatorname{div}\mathbf{v}\,\mathrm{d}\mathbf{x} = \mathbf{0}$$

gelten. Diese *Kompatibilitätsbedingung* an $\mathbf{g}$ ist für homogene Randwerte selbstverständlich erfüllt.

Weiterhin fällt auf, dass nur der Gradient des Drucks p in (14.3) vorkommt und nicht der Druck selbst. Daher kann der Druck nicht eindeutig sein, weshalb man ihn normiert gemäß

$$\int_\Omega p\,\mathrm{d}\mathbf{x} = 0.$$

14.2 Variationsformulierung

Der Gaußsche Integralsatz liefert die beiden Gleichungen

$$-(\Delta\mathbf{v},\boldsymbol{\phi})_{L^2(\Omega)} = (\nabla\mathbf{v},\nabla\boldsymbol{\phi})_{L^2(\Omega)} - (\nabla\mathbf{v}\cdot\mathbf{n},\boldsymbol{\phi})_{L^2(\Gamma)},$$

$$-(\nabla p,\boldsymbol{\phi})_{L^2(\Omega)} = (p,\operatorname{div}\boldsymbol{\phi})_{L^2(\Omega)} + (p\mathbf{n},\boldsymbol{\phi})_{L^2(\Gamma)}.$$

Multiplizieren wir die erste Gleichung aus (14.3) mit einer Testfunktion $\boldsymbol{\phi} \in [C_0^\infty(\Omega)]^d$ und integrieren über Ω, so erhalten wir folglich

$$(\nabla\mathbf{v},\nabla\boldsymbol{\phi})_{L^2(\Omega)} - (p,\operatorname{div}\boldsymbol{\phi})_{L^2(\Omega)} = (\mathbf{f},\boldsymbol{\phi})_{L^2(\Omega)}. \tag{14.5}$$

Wir ziehen uns wie üblich auf homogene Randbedingungen (14.4) zurück und definieren die Bilinearformen

$$a : [H_0^1(\Omega)]^d \times [H_0^1(\Omega)]^d \to \mathbb{R},$$

$$a(\mathbf{v},\mathbf{w}) := (\nabla\mathbf{v},\nabla\mathbf{w})_{L^2(\Omega)} = \sum_{i,j=1}^{d} \int_\Omega \frac{\partial v_i}{\partial x_j}\frac{\partial w_i}{\partial x_j}\,\mathrm{d}\mathbf{x}$$

und

$$b : [H_0^1(\Omega)]^d \times L_0^2(\Omega) \to \mathbb{R}, \qquad b(\mathbf{v}, q) := -(\operatorname{div} \mathbf{v}, q)_{L^2(\Omega)},$$

wobei $L_0^2(\Omega)$ den Raum

$$L_0^2(\Omega) := \left\{ q \in L^2(\Omega) : \int_\Omega q \, \mathrm{d}\mathbf{x} = 0 \right\}$$

bezeichne. Aus (14.3) und (14.5) erhalten wir dann die Variationsformulierung der Stokesschen Gleichung: suche $\mathbf{v} \in [H_0^1(\Omega)]^d$ und $p \in L_0^2(\Omega)$, so dass

$$\begin{aligned} a(\mathbf{v}, \mathbf{w}) + b(\mathbf{w}, p) &= (\mathbf{f}, \mathbf{w})_{L^2(\Omega)} && \text{für alle } \mathbf{w} \in [H_0^1(\Omega)]^2, \\ b(\mathbf{v}, q) &= 0 && \text{für alle } q \in L_0^2(\Omega). \end{aligned} \tag{14.6}$$

Speziell für $d = 2$ erhalten wir ausgeschrieben die Gleichungen

$$\begin{aligned} \int_\Omega \left\{ \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} \right\} q \, \mathrm{d}\mathbf{x} &= 0 && \text{für alle } q \in L_0^2(\Omega), \\ \int_\Omega \left\{ \frac{\partial v_1}{\partial x}\frac{\partial w_1}{\partial x} + \frac{\partial v_1}{\partial y}\frac{\partial w_1}{\partial y} \right\} \mathrm{d}\mathbf{x} - \int_\Omega p \frac{\partial w_1}{\partial x} \, \mathrm{d}\mathbf{x} &= \int_\Omega f_1 w_1 \, \mathrm{d}\mathbf{x} && \text{für alle } w_1 \in H_0^1(\Omega), \\ \int_\Omega \left\{ \frac{\partial v_2}{\partial x}\frac{\partial w_2}{\partial x} + \frac{\partial v_2}{\partial y}\frac{\partial w_2}{\partial y} \right\} \mathrm{d}\mathbf{x} - \int_\Omega p \frac{\partial w_2}{\partial y} \, \mathrm{d}\mathbf{x} &= \int_\Omega f_2 w_2 \, \mathrm{d}\mathbf{x} && \text{für alle } w_2 \in H_0^1(\Omega). \end{aligned}$$

Satz 14.1 Für rechte Seiten $\mathbf{f} \in [L^2(\Omega)]^d$ sind schwache Lösungen $\mathbf{v} \in [H_0^1(\Omega)]^d$ und $p \in L_0^2(\Omega)$ von (14.6) mit $\mathbf{v} \in [C^2(\Omega)]^d$ und $p \in C^1(\Omega)$ klassische Lösungen von (14.3).

Beweis. Wir wählen als Testfunktion $q := \operatorname{div} \mathbf{v} \in L^2(\Omega)$. Dann gilt mit einer Konstanten c, dass $q - c \in L_0^2(\Omega)$. Aufgrund von (14.6) gilt nun

$$\| \operatorname{div} \mathbf{v} \|_{L^2(\Omega)}^2 = (\operatorname{div} \mathbf{v}, q)_{L^2(\Omega)} = (\operatorname{div} \mathbf{v}, c)_{L^2(\Omega)} = c \int_\Omega \operatorname{div} \mathbf{v} \, \mathrm{d}\mathbf{x}.$$

Aus dem Gaußschen Integralsatz schließen wir wegen $\mathbf{v}|_\Gamma = \mathbf{0}$, dass

$$\int_\Omega \operatorname{div} \mathbf{v} \, \mathrm{d}\mathbf{x} = \int_\Gamma \langle \mathbf{v}, \mathbf{n} \rangle \, \mathrm{d}\sigma = 0.$$

Also ist $\| \operatorname{div} \mathbf{v} \|_{L^2(\Omega)} = 0$ und aus $\mathbf{v} \in [C^1(\Omega)]^d$ folgt daher die Divergenzfreiheit in jedem Punkt $\mathbf{x} \in \Omega$.

Die punktweise Gültigkeit der ersten Gleichung aus (14.3) erhalten wir durch ein Zurückführen auf das Poisson-Problem. Für die schwache Lösung $\mathbf{v} \in [H_0^1(\Omega)]^d$ gilt

$$(\nabla \mathbf{v}, \nabla \mathbf{w})_{L^2(\Omega)} = (\mathbf{g}, \mathbf{w})_{L^2(\Omega)}$$

mit $(\mathbf{g}, \mathbf{w})_{L^2(\Omega)} := (\mathbf{f}, \mathbf{w})_{L^2(\Omega)} + (p, \operatorname{div} \mathbf{w})_{L^2(\Omega)}$. Somit ist $\mathbf{v}$ die klassische Lösung des Poisson-Problems

$$-\Delta \mathbf{v} = \mathbf{g} = \mathbf{f} - \nabla p \quad \text{in } \Omega$$

mit homogenen Dirichlet-Randbedingungen. Dies ist aber gerade die gesuchte Gleichung. □

Die Bilinearform $a(\cdot,\cdot)$ ist offensichtlich auf ganz $[H_0^1(\Omega)]^d$ elliptisch. Um die Existenz und Eindeutigkeit der Lösung der Stokesschen Gleichung zu garantieren, müssen wir nach Satz 13.4 also nur noch die LBB-Bedingung für die Bilinearform $b(\cdot,\cdot)$ nachweisen. Dies erfolgt über eine Abschätzung, deren Beweis allerdings den Rahmen sprengen würde. Dazu bezeichne $H^{-1}(\Omega) := \big(H_0^1(\Omega)\big)'$ den Dualraum von $H_0^1(\Omega)$, ausgestattet mit der Norm

$$\|p\|_{H^{-1}(\Omega)} := \sup_{q\in H_0^1(\Omega)} \frac{(p,q)_{L^2(\Omega)}}{\|q\|_{H^1(\Omega)}}.$$

Lemma 14.2 Sei Ω ein beschränktes, zusammenhängendes Gebiet mit Lipschitz-stetigem Rand.

(*i.*) Das Bild der linearen Abbildung

$$\nabla : L^2(\Omega) \to [H^{-1}(\Omega)]^d \tag{14.7}$$

ist abgeschlossen in $[H^{-1}(\Omega)]^d$.

(*ii.*) Es gilt mit einer Konstante $c = c(\Omega)$:

$$\|p\|_{L^2(\Omega)} \le c\big\{\|\nabla p\|_{H^{-1}(\Omega)} + \|p\|_{H^{-1}(\Omega)}\big\} \quad \text{für alle } p \in L^2(\Omega), \tag{14.8}$$

$$\|p\|_{L^2(\Omega)} \le c\|\nabla p\|_{H^{-1}(\Omega)} \quad \text{für alle } p \in L_0^2(\Omega). \tag{14.9}$$

Satz 14.3 In beschränkten Lipschitz-Gebieten Ω besitzt das Stokes-Problem (14.6) für beliebiges $\mathbf{f} \in \big([H_0^1(\Omega)]^d\big)'$ eine eindeutige Lösung $(\mathbf{v}, p) \in [H_0^1(\Omega)]^d \times L_0^2(\Omega)$.

Beweis. Wir müssen nur noch die LBB-Bedingung für die Bilinearform $b(\cdot,\cdot)$ nachweisen. Für $p \in L_0^2(\Omega)$ folgt aus (14.9)

$$\|\nabla p\|_{H^{-1}(\Omega)} \ge \frac{1}{c}\|p\|_{L^2(\Omega)}.$$

Nach Definition der $H^{-1}(\Omega)$-Norm gibt es ein $\mathbf{v} \in [H_0^1(\Omega)]^d$ mit $\|\mathbf{v}\|_{H^1(\Omega)} = 1$ und

$$(\nabla p, \mathbf{v})_{L^2(\Omega)} \ge \frac{1}{2}\|\mathbf{v}\|_{H^1(\Omega)}\|\nabla p\|_{H^{-1}(\Omega)} \ge \frac{1}{2c}\|p\|_{L^2(\Omega)}.$$

Wegen $b(\mathbf{v}, p) = -(\operatorname{div}\mathbf{v}, p)_{L^2(\Omega)} = (\mathbf{v}, \nabla p)_{L^2(\Omega)}$ folgt

$$\frac{b(\mathbf{v},p)}{\|\mathbf{v}\|_{H^1(\Omega)}} = (\mathbf{v}, \nabla p)_{L^2(\Omega)} \ge \frac{1}{2c}\|p\|_{L^2(\Omega)}$$

und damit die LBB-Bedingung. □

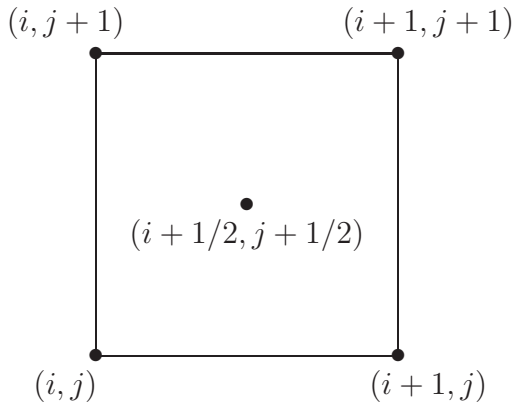
14.3 Instabile Elemente

Lange war wegen der Einfachheit das sogenannte $\mathcal{Q}_1$-$\mathcal{P}_0$-Element, ein Rechteckelement, sehr beliebt zur Diskretisierung der Stokes-Gleichung. Der Fluss wird hier mit bilinearen

Funktionen approximiert, der Druck mit elementweise konstanten Funktionen:

$$U_h := \big\{\mathbf{v} \in [C(\overline{\Omega})]^2 : \mathbf{v}|_\Gamma = \mathbf{0} \text{ und } \mathbf{v}|_T \in [\mathcal{Q}_1]^2, \text{ also bilinear für alle } T \in \mathcal{T}_h\big\},$$
$$V_h := \big\{p \in L_0^2(\Omega) : p|_T \in \mathcal{P}_0, \text{ also elementweise konstant für alle } T \in \mathcal{T}_h\big\}.$$

Dieses Element ist instabil, denn es erfüllt nicht die diskrete LBB-Bedingung. Ein Hinweis auf die Instabilität ist die Tatsache, dass der Kern von $B^\star : V_h \to U_h'$ nicht nur das Nullelement enthält.
Wir numerieren die Knoten im Element $T_{i,j}$ wie folgt:



Weil auf jedem Element $\operatorname{div}\mathbf{v}$ linear und p konstant ist, ergibt sich

$$\begin{aligned}\int_{T_{i,j}} q \operatorname{div}\mathbf{v}\,\mathrm{d}\mathbf{x} &= h^2 q_{(i+1/2,j+1/2)} \operatorname{div}\mathbf{v}_{(i+1/2,j+1/2)} \\ &= h^2 q_{(i+1/2,j+1/2)} \frac{1}{2h}\big\{v_{1,(i+1,j+1)} + v_{1,(i+1,j)} - v_{1,(i,j+1)} - v_{1,(i,j)} \\ &\qquad + v_{2,(i+1,j+1)} - v_{2,(i+1,j)} + v_{2,(i,j+1)} - v_{2,(i,j)}\big\}.\end{aligned}$$

Wir summieren über alle Elemente und sortieren die Terme nach den Gitterpunkten um

$$\int_{T_{i,j}} q \operatorname{div}\mathbf{v}\,\mathrm{d}\mathbf{x} = -h^2 \sum_{i,j} \big\{v_{1,(i,j)}(\nabla_1 q)_{i,j} + v_{2,(i,j)}(\nabla_2 q)_{i,j}\big\}$$

mit den Differenzenquotienten

$$(\nabla_1 q)_{i,j} := \frac{1}{2h}\big\{q_{(i+1/2,j+1/2)} + q_{(i+1/2,j-1/2)} - q_{(i-1/2,j+1/2)} - q_{(i-1/2,j-1/2)}\big\},$$
$$(\nabla_2 q)_{i,j} := \frac{1}{2h}\big\{q_{(i+1/2,j+1/2)} - q_{(i+1/2,j-1/2)} + q_{(i-1/2,j+1/2)} - q_{(i-1/2,j-1/2)}\big\}.$$

Wegen $\mathbf{v} \in [H_0^1(\Omega)]^2$ erstreckt sich die Summation über alle inneren Knoten. Es ist $q \in \operatorname{kern}(\mathbf{B}_h^T)$, wenn

$$\int_\Omega q \operatorname{div}\mathbf{v}\,\mathrm{d}\mathbf{x} = 0 \quad \text{für alle } \mathbf{v} \in U_h$$

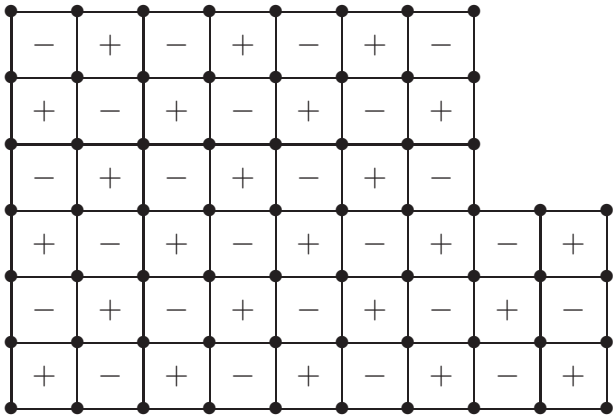
gilt, also $(\nabla_1 q)_{i,j}$ und $(\nabla_2 q)_{i,j}$ an allen inneren Knoten verschwinden. Dies tritt ein falls

$$q_{(i+1/2,j+1/2)} = q_{(i-1/2,j-1/2)}, \quad q_{(i+1/2,j-1/2)} = q_{(i-1/2,j+1/2)}.$$

Die beiden Gleichung bedeuten nicht, dass q konstant sein muss. Es braucht nur

$$q_{(i+1/2,j+1/2)} = \begin{cases} a, & \text{falls } i+j \text{ gerade,} \\ b, & \text{falls } i+j \text{ ungerade,} \end{cases}$$

zu sein. Die Zahlen a und b sind dabei so zu wählen, dass $\int_\Omega p \, \mathrm{d}\mathbf{x} = 0$ gilt. Insbesondere haben a und b entgegengesetzte Vorzeichen. Es bildet sich daher ein Schachbrettmuster, weshalb man auch von *Schachbrettinstabilität* spricht:



Bemerkung Es hilft nichts, den Raum V_h so einzuschränken, dass er orthogonal auf dem Kern von $\mathbf{B}_h^T$ ist. Man kann zeigen, dass dann die LBB-Bedingung zwar erfüllt ist, jedoch mit einer h-abhängigen Konstanten. Dies zeigt, dass die LBB-Bedingung eine analytische Eigenschaft ist und nicht rein algebraisch verstanden werden darf. △

14.4 MINI-Element

Das MINI-Element basiert auf einer $\mathcal{P}_1$-$\mathcal{P}_1$-Diskretisierung, wobei allerdings der Geschwindigkeitsraum angereichert ist. In zwei Dimensionen betrachten wir auf jedem Dreieckselement $T \in \mathcal{T}_h$ die kubische Blasenfunktion $b_T(\mathbf{x}) = 9\lambda_{T,1}\lambda_{T,2}\lambda_{T,3}$, wobei $(\lambda_{T,1}, \lambda_{T,2}, \lambda_{T,3})$ die baryzentrischen Koordinaten auf dem Dreieck T bezeichnen. Es ist also

$$\begin{aligned} U_h &:= \big\{ \mathbf{v} \in [C(\overline{\Omega})]^2 : \mathbf{v}|_\Gamma = \mathbf{0} \text{ und } \mathbf{v}|_T \in [\mathcal{P}_1 \oplus \operatorname{span}\{b_T\}]^2 \text{ für alle } T \in \mathcal{T}_h \big\}, \\ V_h &:= \big\{ p \in C(\overline{\Omega}) : p|_T \in \mathcal{P}_1 \text{ für alle } T \in \mathcal{T}_h \big\}. \end{aligned}$$

Satz 14.4 Sei Ω ein konvexes Polygongebiet. Dann erfüllt das MINI-Element die LBB-Bedingung mit einem $\beta > 0$.

Beweis. Wir weisen die Gültigkeit von Fortins Kriterium nach. Wir haben also die Existenz einer Projektion $\Pi_h : [H_0^1(\Omega)]^2 \to U_h$ nachzuweisen mit der Orthogonalitäts-Eigenschaft

$$\big(\operatorname{div}(\mathbf{v} - \Pi_h \mathbf{v}), q_h \big)_{L^2(\Omega)} = 0 \quad \text{für alle } \mathbf{v} \in [H_0^1(\Omega)]^2 \text{ und } q_h \in V_h$$

und deren Beschränktheit mit einer von h-unabhängigen Konstanten c_Π

$$\|\Pi_h \mathbf{v}\|_{H^1(\Omega)} \le c_\Pi \|\mathbf{v}\|_{H^1(\Omega)}.$$

Diese Projektion wird von der Form

$$\Pi_h = \pi_h^{(1)} + \pi_h^{(2)}(I - \pi_h^{(1)})$$

mit zwei weiteren Projektionen $\pi_h^{(1)}, \pi_h^{(2)} : [H_0^1(\Omega)]^2 \to U_h$ sein.
Die erste Projektion $\pi_h^{(1)}$ definieren wir über die Lösung der Helmholtz-Gleichung. Es ist

$$\big(\nabla(\pi_h^{(1)}\mathbf{v}), \nabla\mathbf{w}_h\big)_{L^2(\Omega)} + \big(\pi_h^{(1)}\mathbf{v}, \mathbf{w}_h\big)_{L^2(\Omega)} = (\nabla\mathbf{v}, \nabla\mathbf{w}_h)_{L^2(\Omega)} + (\mathbf{v}, \mathbf{w}_h)_{L^2(\Omega)}$$

für alle $\mathbf{w}_h \in U_h$. Offensichtlich ist dann

$$\big\|\pi_h^{(1)}\mathbf{v}\big\|_{H^1(\Omega)} \le \|\mathbf{v}\|_{H^1(\Omega)}.$$

Das übliche Dualitätsargument liefert sofort

$$\big\|\mathbf{v} - \pi_h^{(1)}\mathbf{v}\big\|_{L^2(\Omega)} \le ch\big\|\mathbf{v} - \pi_h^{(1)}\mathbf{v}\big\|_{H^1(\Omega)} \le ch\|\mathbf{v}\|_{H^1(\Omega)}. \tag{14.10}$$

Die zweite Projektion wird uns die Orthogonalitäts-Eigenschaft sichern:

$$\pi_h^{(2)}\mathbf{v} = \sum_{T\in\mathcal{T}_h} \boldsymbol{\beta}_T b_T \quad \text{mit} \quad \boldsymbol{\beta}_T := \left(\int_T b_T \,\mathrm{d}\mathbf{x}\right)^{-1} \int_T \mathbf{v} \,\mathrm{d}\mathbf{x}.$$

Man beachte, dass $(\pi_h^{(2)}\mathbf{v})|_T$ eine Blasenfunktion ist, die den Mittelwert von $\mathbf{v}|_T$ erhält. Aufgrund dieser Konstruktion besitzt diese Projektion die Eigenschaft

$$\int_T \big\{\mathbf{v} - \pi_h^{(2)}\mathbf{v}\big\} \,\mathrm{d}\mathbf{x} = \mathbf{0} \quad \text{für alle } T \in \mathcal{T}_h.$$

Hieraus folgt dann mit partieller Integration

$$\begin{aligned}\big(\operatorname{div}(\mathbf{v} - \pi_h^{(2)}\mathbf{v}), q_h\big)_{L^2(\Omega)} &= \sum_{T\in\mathcal{T}_h} \big(\operatorname{div}(\mathbf{v} - \pi_h^{(2)}\mathbf{v}), q_h\big)_{L^2(T)} \\ &= \sum_{T\in\mathcal{T}_h} \Big\{ -(\mathbf{v} - \pi_h^{(2)}\mathbf{v}, \nabla q_h)_{L^2(T)} + \int_{\partial T} \langle \mathbf{v} - \pi_h^{(2)}\mathbf{v}, \mathbf{n}\rangle q_h \,\mathrm{d}\sigma \Big\}.\end{aligned}$$

Nun ist zu bemerken, dass die inneren Kantenintegrale verschwinden aufgrund des wechselnden Vorzeichens der Normalen. Die Kantenintegrale entlang der äußeren Kanten verschwinden, da $\mathbf{v}|_\Gamma = \pi_h^{(2)}\mathbf{v}|_\Gamma = \mathbf{0}$. Da die Druckgradienten ∇p_h elementweise konstant sind, ergibt sich insgesamt:

$$\big(\operatorname{div}(\mathbf{v} - \pi_h^{(2)}\mathbf{v}), q_h\big)_{L^2(\Omega)} = -\sum_{T\in\mathcal{T}_h} \Big\langle (\nabla q_h)|_T, \underbrace{\int_T \big\{\mathbf{v} - \pi_h^{(2)}\mathbf{v}\big\} \,\mathrm{d}\mathbf{x}}_{=\mathbf{0}} \Big\rangle = 0.$$

Man kann sich die Projektion $\pi_h^{(2)}$ als zweistufigen Prozess vorstellen. Zunächst wird auf die elementweise konstante Funktionen projiziert. Danach werden die konstanten Ansatzfunktionen durch Blasenfunktionen mit gleichem Integral ersetzt. Daher folgt die L^2-Stabilität

$$\|\pi_h^{(2)}\mathbf{v}\|_{L^2(\Omega)} \le c\|\mathbf{v}\|_{L^2(\Omega)}. \tag{14.11}$$

Die gleichmäßige Beschränktheit von Π_h ergibt sich aus (14.10), (14.11) und

$$\mathbf{v} - \Pi_h \mathbf{v} = \mathbf{v} - \big(\pi_h^{(1)}\mathbf{v} + \pi_h^{(2)}(\mathbf{v} - \pi_h^{(1)}\mathbf{v})\big) = (I - \pi_h^{(2)})(\mathbf{v} - \pi_h^{(1)}\mathbf{v}).$$

Denn mit Hilfe der inversen Abschätzung erhalten wir

$$\begin{aligned}\|\Pi_h \mathbf{v}\|_{H^1(\Omega)} &\le \big\|\pi_h^{(1)}\mathbf{v}\big\|_{H^1(\Omega)} + h^{-1}\big\|\pi_h^{(2)}(I - \pi_h^{(1)})\mathbf{v}\big\|_{L^2(\Omega)} \\ &\le c\Big\{\|\mathbf{v}\|_{H^1(\Omega)} + h^{-1}\big\|(I - \pi_h^{(1)})\mathbf{v}\big\|_{L^2(\Omega)}\Big\} \\ &\le c\|\mathbf{v}\|_{H^1(\Omega)}.\end{aligned}$$

Fortins Kriterium liefert schließlich die Behauptung. □

Korollar 14.5 Sei Ω ein konvexes Polygongebiet. Dann erfüllt das MINI-Element die Fehlerabschätzung

$$\|\mathbf{v} - \mathbf{v}_h\|_{H^1(\Omega)} + \|p - p_h\|_{L^2(\Omega)} \le ch\big\{\|\mathbf{v}\|_{H^2(\Omega)} + \|p\|_{H^1(\Omega)}\big\}.$$

Beweis. Wegen

$$\inf_{\mathbf{w}_h \in U_h} \|\mathbf{v} - \mathbf{w}_h\|_{H^1(\Omega)} \le ch\|\mathbf{v}\|_{H^2(\Omega)}, \qquad \inf_{q_h \in v_h} \|p - q_h\|_{L^2(\Omega)} \le ch\|p\|_{H^1(\Omega)}$$

ergibt sich das Behauptete unmittelbar aus Korollar 13.7. □

14.5 Taylor-Hood-Element

Bei dem oft benutzten Taylor-Hood-Element werden wie bei den instabilen Elementen für den Fluss Polynome höheren Grades als für den Druck herangezogen. Allerdings wird der Druck stetig angesetzt:

$$\begin{aligned}U_h &:= \big\{\mathbf{v} \in [C(\overline{\Omega})]^2 : \mathbf{v}|_\Gamma = \mathbf{0} \text{ und } \mathbf{v}|_T \in [\mathcal{P}_2]^2 \text{ für alle } T \in \mathcal{T}_h\big\}, \\ V_h &:= \big\{p \in C(\overline{\Omega}) : p|_T \in \mathcal{P}_1 \text{ für alle } T \in \mathcal{T}_h\big\}.\end{aligned}$$

Dabei sei $\mathcal{T}_h$ eine Zerlegung von Ω in Dreiecke. Auf Vierecken verwendet man entsprechend $\mathcal{Q}_2$-$\mathcal{Q}_1$-Elemente anstelle von $\mathcal{P}_2$-$\mathcal{P}_1$-Elementen.

Satz 14.6 Sei Ω ein konvexes Polygongebiet. Dann erfüllt das Taylor-Hood-Element die LBB-Bedingung mit einem $\beta > 0$.

Beweis. Ziel ist es, zu gegebem $p_h \in V_h$ ein $\mathbf{v}_h = \mathbf{v}_h(p_h) \in U_h$ zu konstruieren, für das die LBB-Bedingung erfüllt ist. Wir vollziehen den Beweis in drei Schritten.

(*i.*) Zu jeder Kante e im Inneren von Ω bezeichne $\mathbf{t}_e$ den Tangentenvekor und $\mathbf{n}_e$ den Normalenvektor. In jedem Kantenmittelpunkt $\mathbf{m}_e$ setzen wir

$$\langle \mathbf{v}_h(\mathbf{m}_e), \mathbf{t}_e\rangle := \langle \mathbf{t}_e, (\nabla p_h)(\mathbf{m}_e)\rangle, \quad \langle \mathbf{v}_h(\mathbf{m}_e), \mathbf{n}_e\rangle := 0.$$

Man beachte hierbei, dass die Tangentialableitung des Drucks entlang der Kante e wohldefiniert ist. Ferner setzen wir $\mathbf{v}_h(\mathbf{x}) := \mathbf{0}$ in den Eckpunkten aller Dreiecke und an den Kantenmittelpunkten auf Γ.
Nach Konstruktion gilt einerseits

$$\|\mathbf{v}_h\|_{L^2(\Omega)} \leq |p_h|_{H^1(\Omega)}. \tag{14.12}$$

Weil $\langle \mathbf{v}_h, \nabla p_h \rangle|_T$ quadratisch und $(\nabla p_h)|_T$ konstant ist, gilt andererseits

$$\begin{aligned}
\int_T \langle \mathbf{v}_h, \nabla p_h \rangle \,\mathrm{d}\mathbf{x} &= \frac{|T|}{3} \sum_{i=1}^{3} \langle \mathbf{v}_h(\mathbf{m}_{e_i}), (\nabla p_h)(\mathbf{m}_{e_i}) \rangle \\
&= \frac{|T|}{3} \sum_{i=1}^{3} \langle \mathbf{t}_{e_i}, (\nabla p_h)(\mathbf{m}_{e_i}) \rangle^2 \\
&\geq \frac{|T|}{3} c_\kappa \big\| (\nabla p_h)|_T \big\|^2 \\
&= \frac{c_\kappa}{3} \int_T \|\nabla p_h\|^2 \,\mathrm{d}\mathbf{x}.
\end{aligned}$$

Aufsummation über alle $T \in \mathcal{T}_h$ ergibt schliesslich

$$b(\mathbf{v}_h, p_h) \geq \frac{c_\kappa}{3} \sum_{T \in \mathcal{T}_h} \int_T \|\nabla p_h\|^2 \,\mathrm{d}\mathbf{x} = \frac{c_\kappa}{3} |p_h|^2_{H^1(\Omega)}. \tag{14.13}$$

(*ii.*) Zu $K > 0$ bezeichne

$$V_h^K := \big\{ p_h \in V_h : \|p_h\|_{L^2(\Omega)} \leq Kh |p_h|_{H^1(\Omega)} \big\}.$$

Für $p_h \in V_h^K$ gilt wegen (14.13) mit dem in Schritt (*i.*) konstruierten $\mathbf{v}_h = \mathbf{v}_h(p_h)$

$$\frac{b(\mathbf{v}_h, p_h)}{\|\mathbf{v}_h\|_{H^1(\Omega)}} \geq \frac{c_\kappa}{3} \frac{|p_h|^2_{H^1(\Omega)}}{\|\mathbf{v}_h\|_{H^1(\Omega)}} \geq \frac{c_\kappa}{3Kh} \frac{|p_h|_{H^1(\Omega)} \|p_h\|_{L^2(\Omega)}}{\|\mathbf{v}_h\|_{H^1(\Omega)}}.$$

Aus (14.12) und der inversen Abschätzung folgt daher die LBB-Bedingung für alle Funktionen aus V_h^K:

$$\frac{b(\mathbf{v}_h, p_h)}{\|\mathbf{v}_h\|_{H^1(\Omega)}} \geq \frac{c_\kappa}{3Kh} \frac{\|\mathbf{v}_h\|_{L^2(\Omega)} \|p_h\|_{L^2(\Omega)}}{\|\mathbf{v}_h\|_{H^1(\Omega)}} \geq \frac{c_\kappa}{3 c_{inv} K} \|p_h\|_{L^2(\Omega)}.$$

(*iii.*) Sei nun $p_h \notin V_h^K$. Wir wählen ein $\mathbf{v} \in [H_0^1(\Omega)]^2$ derart, dass $\|\mathbf{v}\|_{H^1(\Omega)} = 1$ und

$$b(\mathbf{v}, p_h) \geq \beta \|p_h\|_{L^2(\Omega)}$$

für ein $\beta > 0$ erfüllt ist. Dies ist aufgrund der Gültigkeit der kontinuierlichen LBB-Bedingung (vergleiche Satz 14.3) möglich.
Ferner bezeichne $\mathbf{v}_h$ die Clément-Approximation von $\mathbf{v}$, die nach Satz 9.1 der Fehlerabschätzung

$$\|\mathbf{v} - \mathbf{v}_h\|_{L^2(\Omega)} \leq ch \|\mathbf{v}\|_{H^1(\Omega)}$$

genügt und H^1-stabil ist

$$\|\mathbf{v}_h\|_{H^1(\Omega)} \leq \|\mathbf{v} - \mathbf{v}_h\|_{H^1(\Omega)} + \|\mathbf{v}\|_{H^1(\Omega)} \leq c \|\mathbf{v}\|_{H^1(\Omega)}.$$

Es sei angemerkt, dass die Clément-Approximation so modifiziert werden kann, dass auch $\mathbf{v}_h$ Nullrandbedingungen erfüllt. Insbesondere weisen wir darauf hin, dass die Konstante c nicht von p_h und damit nicht von K abhängt.
Wir haben nun

$$\begin{aligned} b(\mathbf{v}_h, p_h) &= b(\mathbf{v}, p_h) - b(\mathbf{v} - \mathbf{v}_h, p_h) \\ &\geq \beta \|p_h\|_{L^2(\Omega)} + (\mathbf{v} - \mathbf{v}_h, \nabla p_h)_{L^2(\Omega)} \\ &\geq \beta \|p_h\|_{L^2(\Omega)} - \|\mathbf{v} - \mathbf{v}_h\|_{L^2(\Omega)} |p_h|_{H^1(\Omega)} \\ &\geq \beta \|p_h\|_{L^2(\Omega)} - ch \underbrace{\|\mathbf{v}\|_{H^1(\Omega)}}_{=1} |p_h|_{H^1(\Omega)}. \end{aligned}$$

Wegen $p_h \notin V_h^K$, ist

$$|p_h|_{H^1(\Omega)} < \frac{1}{Kh} \|p_h\|_{L^2(\Omega)}$$

und es ergibt sich

$$b(\mathbf{v}_h, p_h) \geq \left(\beta - \frac{c}{K} \right) \|p_h\|_{L^2(\Omega)}.$$

Damit ist auch im Fall $p_h \notin V_h^K$ die LBB-Bedingung gezeigt, falls K hinreichend groß gewählt wurde. □

Korollar 14.7 Das Taylor-Hood-Element erfüllt die Fehlerabschätzung

$$\|\mathbf{v} - \mathbf{v}_h\|_{H^1(\Omega)} + \|p - p_h\|_{L^2(\Omega)} \leq ch^k \big\{ \|\mathbf{v}\|_{H^{k+1}(\Omega)} + \|p\|_{H^k(\Omega)} \big\}$$

mit $k \in \{1, 2\}$, sofern $\mathbf{v} \in [H^{k+1}(\Omega)]^d$ und $p \in H^k(\Omega)$ gilt.

Beweis. Für $k \in \{1, 2\}$ gelten die Approximationsabschätzungen

$$\inf_{\mathbf{w}_h \in U_h} \|\mathbf{v} - \mathbf{w}_h\|_{H^1(\Omega)} \leq ch^k \|\mathbf{v}\|_{H^{k+1}(\Omega)}, \quad \inf_{q_h \in V_h} \|p - q_h\|_{L^2(\Omega)} \leq ch^k \|p\|_{H^k(\Omega)},$$

vorausgesetzt die Normen auf der rechten Seite sind wohldefiniert. Damit folgt die Behauptung sofort aus Korollar 13.7. □

Bemerkung Man kann den Fluss auch linear auf einer einmal uniform verfeinerten Triangulierung ansetzen, anstelle ihn mit quadratischen Elementen zu approximieren. Auch dieses Element wird oftmals als Taylor-Hood-Element bezeichnet. △

15. Eigenwertprobleme

15.1 Motivation

Die klassische Formulierung eines Eigenwertproblems lautet

$$\mathcal{L}u = \lambda u \text{ in } \Omega, \quad u = 0 \text{ auf } \Gamma. \tag{15.1}$$

Dabei ist $\mathcal{L}$ ein elliptischer Differentialoperator zweiter Ordnung. Eine Lösung $u \not\equiv 0$ von (15.1) heißt *Eigenfunktion*, das zugehörige λ *Eigenwert*.

Beispiel 15.1 Gemäß Kapitel 1 lautet die Wellengleichung

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = f \text{ in } (0,T) \times \Omega, \quad u = 0 \text{ auf } (0,T) \times \Gamma,$$

wobei zum Zeitpunkt $t = 0$ Anfangsbedingungen vorgegeben sind:

$$u(0,\cdot) = g, \ \frac{\partial u}{\partial t}(0,\cdot) = h \text{ in } \Omega.$$

Unter der Vorraussetzung, dass $f = 0$ ist, führt der Separationsansatz

$$u(t,\mathbf{x}) = v(t)w(\mathbf{x})$$

auf die Gleichung

$$v''(t)w(\mathbf{x}) = v(t)\Delta w(\mathbf{x}).$$

Dies bedeutet, dass

$$\frac{v''(t)}{v(t)} = \frac{\Delta w(\mathbf{x})}{w(\mathbf{x})} := -\lambda \equiv const$$

gelten muss, was einerseits die Gleichung

$$v''(t) + \lambda v(t) = 0$$

und andererseits die Gleichung

$$-\Delta w = \lambda w \text{ in } \Omega, \quad w = 0 \text{ auf } \Gamma$$

impliziert. Die Funktion v besitzt demanch die Form

$$v(t) = A\cos(\sqrt{\lambda}t) + B\sin(\sqrt{\lambda}t), \quad A, B \in \mathbb{R},$$

während w durch ein Eigenwertproblem bestimmt ist. △

Wie in Kapitel 3 kann man die klassische Darstellung (15.1) des Eigenwertproblems durch eine äquivalente Variationsformulierung ersetzen, wobei eine geeignete Bilinearform $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ an die Stelle von $\mathcal{L}$ tritt:

$$\text{suche } u \in V, \text{ so dass } a(u,v) = \lambda(u,v)_{L^2(\Omega)} \text{ für alle } v \in V. \tag{15.2}$$

15.2 Spektraltheorie

Wir wollen einen abstrakten Rahmen für das Eigenwertproblem (15.2) entwickeln. Sei H ein separabler 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. Dabei sei die Einbettung $V \hookrightarrow H$ stetig, das heißt, es ist $\|v\|_H \le c\|v\|_V$ für alle $v \in V$. Ferner sei $a : V \times V \to \mathbb{R}$ eine stetige und V-elliptische Bilinearform.

Es bezeichne $\mathcal{S} : H \to V$ den Lösungsoperator des Problems

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

das heißt, $u = \mathcal{S}f$ ist dessen Lösung. Offensichtlich ist $\mathcal{S}$ linear und beschränkt

$$\|\mathcal{S}f\|_V \le c\|f\|_H,$$

denn es gilt

$$\|\mathcal{S}f\|_V^2 = \|\mathcal{S}f\|_V\|u\|_V = \|u\|_V^2 \le \frac{1}{c_E}a(u,u) = \frac{1}{c_E}(f,u) \le \frac{1}{c_E}\|f\|_H\|u\|_H \le \frac{c}{c_E}\|f\|_H\|u\|_V.$$

Deshalb ist das Eigenwertproblem äquivalent zu

$$\text{suche } (\lambda,u) \in \mathbb{C} \times H, \text{ so dass } u = \lambda\mathcal{S}u. \tag{15.3}$$

Lemma 15.2 Die Einbettung $V \hookrightarrow H$ sei kompakt und $a : V \times V \to \mathbb{R}$ symmetrisch und V-elliptisch. Dann ist der Lösungsoperator $\mathcal{S} : V \to V$ kompakt, symmetrisch bezüglich des Innenprodukts $a(\cdot,\cdot)$,

$$a(\mathcal{S}u,v) = a(u,\mathcal{S}v) \quad \text{für alle } u,v \in V,$$

und positiv

$$a(\mathcal{S}u,u) > 0 \quad \text{für alle } 0 \neq u \in V.$$

Beweis. Wir zeigen zunächst, dass $\mathcal{S} : V \to V$ ein kompakter Operator ist. Hierzu sei $B \subset V$ eine beschränkte Menge. Da die Einbettung $V \hookrightarrow H$ kompakt ist, ist $B \subset H$ eine kompakte Menge. Aufgrund der Stetigkeit von $\mathcal{S} : H \to V$ ist $\mathcal{S}(B) \subset V$ eine kompakte Menge. Daher bildet $\mathcal{S} : V \to V$ beschränkte Mengen auf kompakte Mengen ab, ist folglich also ein kompakter Operator.

Aus der Definition von $\mathcal{S}$ folgt, dass

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

Die Symmetrie der Bilinearform $a(\cdot,\cdot)$ impliziert daher

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

Der letzte Teil der Behauptung ergibt sich schließlich aus

$$a(\mathcal{S}u,u) = (u,u) = \|u\|_H^2 > 0 \quad \text{für alle } 0 \neq u \in V.$$

□

Satz 15.3 (Spektralsatz) Es sei H ein separabler Hilbert-Raum und $V \subset H$ dicht. Die Einbettung $V \hookrightarrow H$ sei kompakt und $a : V \times V \to \mathbb{R}$ symmetrisch und V-elliptisch. Dann existieren abzählbar viele reelle Eigenwerte

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k \leq \cdots$$

mit zugehörigen Eigenvektoren $\{u_k\} \subset V$ zum Eigenwertproblem

$$\text{suche } (\lambda, u) \in \mathbb{C} \times H, \text{ so dass } a(u, v) = \lambda(u, v) \text{ für alle } v \in V. \tag{15.4}$$

Dabei ist $+\infty$ der einzige Häufungspunkt von der Folge $\{\lambda_k\}$. Die Eigenvektoren $\{u_k\}$ bilden eine Orthonormalbasis von H, während $\{\lambda_k^{-1/2} u_k\}$ eine Orthonormalbasis von V bezüglich des Innenprodukts $a(\cdot, \cdot)$ bilden.

Beweis. Anstelle des Eigenwertproblems (15.4) betrachten wir (15.3), beziehungsweise

$$\text{suche } (\mu, u) \in \mathbb{C} \times H, \text{ so dass } \mathcal{S}u = \mu u.$$

Da $\mathcal{S} : V \to V$ symmetrisch und kompakt ist, folgt aus dem Spektralsatz für symmetrische und kompakte Operatoren, dass eine abzählbare Folge von positive Eigenwerten $\{\mu_k\}$ mit zugehörigen Eigenvektoren $\{v_k\} \subset V$ existieren. Die Eigenwerte $\{\mu_k\}$ häufen sich höchstens in der 0, die Eigenvektoren $\{v_k\}$ bilden eine Orthonormalbasis von V bezüglich des Innenprodukts $a(\cdot, \cdot)$.
Wir setzen $\lambda_k := 1/\mu_k$ für alle $k \in \mathbb{N}$ und erhalten

$$a(v_k, w) = \lambda_k a(\mathcal{S}v_k, w) = \lambda_k (v_k, w) \quad \text{für alle } w \in V.$$

Daraus ergibt sich für $u_k := \sqrt{\lambda_k} v_k$

$$(u_k, u_\ell) = \frac{1}{\lambda_k} a(u_k, u_\ell) = \sqrt{\frac{\lambda_\ell}{\lambda_k}} a(v_k, v_\ell) = \delta_{k,\ell}.$$

Folglich ist $\{u_k\}$ orthonormal in H bezüglich des Innenprodukts $(\cdot, \cdot)$.
Wir wollen nun zeigen, dass $\{u_k\}$ auch eine Basis von H ist. Dazu sei das Gegenteil angenommen. Dann existiert ein $0 \neq f \in H$, so dass $(f, u_k) = 0$ für alle $k \in \mathbb{N}$, beziehungsweise

$$(f, v_k) = 0, \quad k = 1, 2, \ldots.$$

Da $\{v_k\}$ aber eine Orthonormalbasis in V bezüglich des Innenprodukts $a(\cdot, \cdot)$ ist, ergibt sich

$$\lim_{n \to \infty} \left\| w - \sum_{k=1}^{n} a(w, v_k) v_k \right\|_H = 0 \quad \text{für alle } w \in V.$$

Weil $V \subset H$ folgt

$$|(f, w)| \leq \left| \sum_{k=1}^{n} a(w, v_k) \underbrace{(f, v_k)}_{=0} \right| + \|f\|_H \left\| w - \sum_{k=1}^{n} a(w, v_k) v_k \right\|_H \overset{n \to \infty}{\longrightarrow} 0$$

für alle $w \in V$. Da $V \subset H$ dicht ist, schließen wir $(f, v) = 0$ für alle $v \in H$, speziell also $\|f\|_H = 0$ beziehungsweise $f = 0$ in H, was zu einem Widerspruch führt. ∎

15.3 Min-Max-Prinzip

Betrachte den *Raleigh-Quotient*

$$R(v) = \frac{a(v,v)}{(v,v)} \quad \text{für alle } v \in V.$$

Die Eigenpaare $\{(\lambda_k, u_k)\}$ von (15.4) erfüllen gerade

$$R(u_k) = \lambda_k \quad k = 1, 2, \ldots.$$

Sei $0 \neq v = \sum_{k=1}^{\infty} \alpha_k u_k \in V$. Dann folgt

$$R(v) = \frac{\sum_{k=1}^{\infty} \lambda_k \alpha_k^2}{\sum_{k=1}^{\infty} \alpha_k^2} \geq \lambda_1$$

und $\lambda_1 = \min_{0 \neq v \in V} R(v)$. Sei

$$U_m := \operatorname{span}\{u_k : k = 1, 2, \ldots, m\} \subset V$$

und

$$\begin{aligned} U_m^{\perp} :=& \{v \in V : a(v,w) = 0 \text{ für alle } w \in U_m\} \\ =& \{v \in V : a(v,u_k) = 0 \text{ für alle } k = 1, 2, \ldots, m\}. \end{aligned}$$

Falls $v = \sum_{k=1}^{\infty} \alpha_k u_k \in U_{m-1}^{\perp}$, dann folgt $\alpha_k = 0$ für alle $k = 1, 2, \ldots, m-1$ und

$$R(v) = \frac{\sum_{k=m}^{\infty} \lambda_k \alpha_k^2}{\sum_{k=m}^{\infty} \alpha_k^2} \geq \lambda_m.$$

Insbesondere ist

$$\lambda_m = \min_{0 \neq v \in U_{m-1}^{\perp}} R(v). \tag{15.5}$$

Allgemein haben wir

Satz 15.4 (Min-Max-Prinzip von Courant-Fischer) Unter den Annahmen von Satz 15.3 gilt

$$\lambda_m = \min_{\substack{V_m \subset V \\ \dim V_m = m}} \max_{0 \neq v \in V_m} R(v).$$

Beweis. Sei $V_m \subset V$ mit $\dim V_m = m$ beliebig und $0 \neq v \in V_m$ so gewählt, dass $(v, u_k) = 0$ für alle $k = 1, 2, \ldots, m-1$, das heißt, $v \in V_m \cap U_{m-1}^{\perp}$. Gleichung (15.5) impliziert $R(v) \geq \lambda_m$ und da v beliebig war gilt folglich

$$\max_{0 \neq v \in V_m} R(v) \geq \lambda_m.$$

Im Fall $V_m = U_m$ gilt sogar Gleichheit. Denn aus $0 \neq v = \sum_{k=1}^{m} \alpha_k u_k$ folgt

$$R(v) = \frac{\sum_{k=1}^{m} \lambda_k \alpha_k^2}{\sum_{k=1}^{m} \alpha_k^2} \leq \lambda_m.$$

Wegen $R(v_m) = \lambda_m$ erhalten wir daher

$$\max_{0 \neq v \in U_m} R(v) = \lambda_m.$$

□

15.4 Finite-Element-Approximation

Sei $V_N \subset V$ ein endlichdimensionaler Teilraum. Die Galerkin-Diskretisierung von (15.4) lautet:

$$\text{suche } (\lambda, u) \in \mathbb{C} \times V_N, \text{ so dass } a(u,v) = \lambda(u,v) \text{ für alle } v \in V_N. \tag{15.6}$$

Satz 15.5 Unter den Annahmen von Satz 15.3 existieren N diskrete Eigenwerte $\{\lambda_{N,k}\} \subset \mathbb{R}^N$ von (15.6) und eine Orthonormalbasis $\{u_{N,k}\}$ von V_N in H, so dass

$$a(u_{N,k}, v_N) = \lambda_{N,k}(u_{N,k}, v_N) \quad \text{für alle } v_N \in V_N.$$

Beweis. Sei $V_N = \operatorname{span}\{\varphi_k : k = 1, 2, \ldots, N\}$ und $u_N = \sum_{k=1}^N x_k \varphi_k$. Dann ist (15.6) äquivalent zu

$$\sum_{k=1}^N x_k a(\varphi_k, \varphi_\ell) = \lambda \sum_{k=1}^N x_k (\varphi_k, \varphi_\ell), \quad \ell = 1, 2, \ldots, N.$$

Sei $\mathbf{A} = [a(\varphi_k, \varphi_\ell)]_{k,\ell} \in \mathbb{R}^{N\times N}$ die Steiffigkeitsmatrix und $\mathbf{M} = [(\varphi_k, \varphi_\ell)]_{k,\ell} \in \mathbb{R}^{N\times N}$ die Massenmatrix. Dann ist (15.6) also äquivalent zum verallgemeinerten Eigenwertproblem

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{M}\mathbf{x}.$$

Mit der Cholesky-Zerlegung $\mathbf{L}\mathbf{L}^T = \mathbf{M}$ ist dieses Problem äquivalent zu

$$\widetilde{\mathbf{A}}\widetilde{\mathbf{x}} := \mathbf{L}^{-1}\mathbf{A}\mathbf{L}^{-T}\widetilde{\mathbf{x}} = \lambda\widetilde{\mathbf{x}}, \quad \widetilde{\mathbf{x}} = \mathbf{L}^T\mathbf{x}.$$

Dabei ist $\widetilde{\mathbf{A}}$ symmetrisch und positiv definit, weshalb N Eigenwerte

$$0 < \lambda_{N,1} \le \lambda_{N,2} \le \cdots \le \lambda_{N,N}$$

mit zugehörigen Eigenvektoren $\{\widetilde{\mathbf{x}}_k\}_{k=1}^N$ existieren, welche eine Orthonormalbasis des $\mathbb{R}^N$ bilden. Die entsprechenden Eigenfunktionen $\{u_{N,k}\}_{k=1}^N \subset V_N$ sind orthonormal in H wegen

$$(u_{N,k}, u_{N,\ell}) = \mathbf{x}_k^T \mathbf{M}\mathbf{x}_\ell = \mathbf{x}_k^T \mathbf{L}\mathbf{L}^T \mathbf{x}_\ell = \widetilde{\mathbf{x}}_k^T \widetilde{\mathbf{x}}_\ell = \delta_{k,\ell}.$$

□

15.5 Konvergenz der Eigenwerte

Lemma 15.6 Es bezeichne $P_N : V \to V_N$ die Galerkin-Projektion, definiert durch

$$a(u - P_N u, v_N) = 0 \quad \text{für alle } v_N \in V_N, \tag{15.7}$$

und sei

$$\sigma_{N,m} := \inf_{0 \neq v \in U_m} \frac{\|P_N v\|_H}{\|v\|_H}.$$

Dann gilt

$$\lambda_m \le \lambda_{N,m} \le \frac{1}{\sigma_{N,m}^2} \lambda_m \quad \text{für alle } m = 1, 2, \ldots, N,$$

vorausgesetzt es ist $\sigma_{N,m} > 0$.

Beweis. Aus $\sigma_{N,m} > 0$ folgt $\dim(P_N U_m) = m$. Denn falls nicht, dann ist der Projektor $P_N : U_m \to P_N U_m$ nicht bijektiv und es würde ein $0 \neq v \in U_m$ existieren mit $P_N v = 0$. Dies widerspricht jedoch $\sigma_{N,m} > 0$.
Es folgt

$$\lambda_{N,m} \leq \max_{0 \neq v \in P_N U_m} R(v) = \max_{0 \neq v \in U_m} \frac{a(P_N v, P_N v)}{\|P_N v\|_H^2} = \max_{\substack{v \in U_m \\ \|v\|_H = 1}} \frac{a(P_N v, P_N v)}{\|P_N v\|_H^2}.$$

Da P_N die orthogonale Projektion von v auf V_N bezüglich des Innenprodukts $a(\cdot, \cdot)$ ist, folgt

$$a(P_N v, P_N v) \leq a(v, v).$$

Wegen

$$\lambda_m = \max_{0 \neq v \in U_m} R(v) = \max_{0 \neq v \in U_m} \frac{a(v, v)}{\|v\|_H^2},$$

folgt hieraus

$$\lambda_{N,m} \leq \sup_{\substack{v \in U_m \\ \|v\|_H = 1}} \frac{a(v, v)}{\|P_N v\|_H^2} \leq \lambda_m \cdot \sup_{\substack{v \in U_m \\ \|v\|_H = 1}} \frac{1}{\|P_N v\|_H^2},$$

das ist die obere Schranke für $\lambda_{N,m}$.
Die untere Schranke $\lambda_m \leq \lambda_{N,m}$ ergibt sich sofort aus dem Min-Max-Prinzip. □

Lemma 15.7 Für jedes $m \geq 1$ existiert eine Konstante $c = c(m) > 0$, so dass für jedes $V_N \subset V$ mit $N \geq m$ gilt

$$\sigma_{N,m}^2 \geq 1 - c(m) \sup_{\substack{v \in U_m \\ \|v\|_H = 1}} \|v - P_N v\|_V^2.$$

Beweis. Sei $w = \sum_{k=1}^m \alpha_k u_k \in U_m$ mit $\|w\|_H^2 = \sum_{k=1}^m \alpha_k^2 = 1$. Dann gilt

$$\begin{aligned} 1 - \|P_N w\|_H^2 &= (w, w) - (P_N w, P_N w) \\ &= (w - P_N w, w + P_N w) \\ &= -\|w - P_N w\|_H^2 + 2(w - P_N w, w), \end{aligned}$$

dies bedeutet,

$$\|P_N w\|_H^2 \geq 1 - 2(w - P_N w, w).$$

Wegen $a(u_k, v) = \lambda_k (u_k, v)$ für alle $v \in V$ ergibt sich

$$(w - P_N w, w) = \sum_{k=1}^m \alpha_k (w - P_N w, u_k) = \sum_{k=1}^m \frac{\alpha_k}{\lambda_k} a(w - P_N w, u_k).$$

Weil $a(w - P_N w, v) = 0$ ist für alle $v \in V_N$, folgt weiter

$$(w - P_N w, w) = \sum_{k=1}^m \frac{\alpha_k}{\lambda_k} a(w - P_N w, u_k - \underbrace{P_N u_k}_{\in V_N}).$$

Die Stetigkeit der Bilinearform $a(\cdot,\cdot)$ liefert

$$\begin{aligned}|(w - P_N w, w)| &\le c_S \|w - P_N w\|_V \left(\sum_{k=1}^{m} \frac{|\alpha_k|}{\lambda_k} \|u_k - P_N u_k\|_V \right) \\ &\le c_S \|w - P_N w\|_V \left(\sum_{k=1}^{m} \frac{\alpha_k^2}{\lambda_k^2} \right)^{1/2} \left(\sum_{k=1}^{m} \|u_k - P_N u_k\|_V^2 \right)^{1/2} \\ &\le c_S \frac{\sqrt{m}}{\lambda_1} \|w - P_N w\|_V \sup_{\substack{v \in U_m \\ \|v\|_H = 1}} \|v - P_N v\|_V.\end{aligned}$$

Damit erhalten wir die Behauptung mit $c(m) = 2c_S\sqrt{m}/\lambda_1$. □

Satz 15.8 (Konvergenz approximativer Eigenwerte) Es gelten die Annahmen von Satz 15.3 und sei $\lim_{N\to\infty} \inf_{v_N \in V_N} \|v - v_N\|_V = 0$ für alle $v \in V$. Dann existiert zu jedem $m \in \mathbb{N}$ ein $N_0 > 0$ derart, dass

$$0 \le \lambda_{N,m} - \lambda_m \le c(m) \sup_{\substack{v \in U_m \\ \|v\|_H = 1}} \inf_{v_N \in V_N} \|v - v_N\|_V^2$$

für alle $N \ge N_0$.

Beweis. Die Galerkin-Projektion erfüllt aufgrund des Céa-Lemmas

$$\|v - P_N v\|_V \le \frac{c_S}{c_E} \inf_{v_N \in V_N} \|v - v_N\|_V \overset{N\to\infty}{\longrightarrow} 0$$

für alle $v \in V$. Für jedes $v = \sum_{k=1}^{m} \alpha_k u_k \in U_m$ mit $\|v\|_H = \sum_{k=1}^{m} \alpha_k^2 = 1$ gilt außerdem

$$\|v - P_N v\|_V \le \sum_{k=1}^{m} |\alpha_k| \|u_k - P_N u_k\|_V \le \left(\underbrace{\sum_{k=1}^{m} \alpha_k^2}_{=1} \right)^{1/2} \left(\sum_{k=1}^{m} \|u_k - P_N u_k\|_V^2 \right)^{1/2}.$$

Die Kombination beider Abschätzungen ergibt

$$\lim_{N\to\infty} \sup_{\substack{v \in U_m \\ \|v\|_H = 1}} \|v - P_N v\|_V = 0.$$

Hieraus folgt mit Hilfe von Lemma 15.7, dass ein $N \ge N_0$ existiert, so dass $\sigma_{N,m} \ge 1/2$ ist. Lemmata 15.6 und 15.7 liefern schließlich

$$\lambda_m \le \lambda_{N,m} \le \left(1 + c(m) \sup_{\substack{v \in U_m \\ \|v\|_H = 1}} \|v - P_N v\|_V^2 \right) \lambda_m$$

mit einer Konstante $c(m) > 0$. □

15.6 Konvergenz der Eigenfunktionen

Wir zeigen die Konvergenz der Eigenfunktionen nur im Fall, dass λ_m ein einfacher Eigenwert ist. Unter den Voraussetzungen von Satz 15.8 gilt dann $\lambda_{N,k} \neq \lambda_m$ für alle $k \neq m$ und $N \geq N_0$. Deshalb ist die Größe

$$\rho_{N,m} := \max_{\substack{1\leq k\leq N\\ k\neq m}} \frac{\lambda_{N,m}}{|\lambda_{N,k} - \lambda_m|}$$

wohldefiniert.

Lemma 15.9 Sei λ_m ein einfacher Eigenwert. Dann existiert ein N_0 derart, dass mit geeigneten $u_{N,m} \in V_N$ gilt

$$\|u_m - u_{N,m}\|_H \leq 2(1 + \rho_{N,m})\|u_m - P_N u_m\|_H$$

für alle $N \geq N_0$.

Beweis. Es bezeichne $P_N : H \to V_N$ die Galerkin-Projektion auf V_N. Sei $v_{N,m}$ die H-orthogonale Projektion von $P_N u_m$ auf die lineare Hülle von $\{u_{N,m}\}$, das heißt

$$v_{N,m} = (P_N u_m, u_{N,m})u_{N,m}. \tag{15.8}$$

Da $\{u_{N,k}\}_{k=1}^N$ eine Orthonormalbasis in V_N bezüglich des Innenprodukts $(\cdot,\cdot)$ ist, folgt

$$\|P_N u_m - v_{N,m}\|_H^2 = \left\| \sum_{\substack{k=1\\ k\neq m}}^{N} (P_N u_m, u_{N,k})u_{N,k} \right\|_H^2 = \sum_{\substack{k=1\\ k\neq m}}^{N} (P_N u_m, u_{N,k})_H^2. \tag{15.9}$$

Aufgrund der Definition von P_N (vergleiche (15.7)) gilt

$$\begin{aligned}(P_N u_m, u_{N,k}) &= \frac{1}{\lambda_{N,k}} a(P_N u_m, u_{N,k}) \\ &= \frac{1}{\lambda_{N,k}} a(u_m, u_{N,k}) \\ &= \frac{\lambda_m}{\lambda_{N,k}} (u_m, u_{N,k})\end{aligned}$$

und daher

$$(\lambda_{N,k} - \lambda_m)(P_N u_m, u_{N,k}) = \lambda_m (u_m - P_N u_m, u_{N,k}).$$

Diese Gleichung eingesetzt in (15.9) ergibt

$$\begin{aligned}\|P_N u_m - v_{N,m}\|_H^2 &\leq \rho_{N,m}^2 \sum_{\substack{1\leq k\leq N\\ k\neq m}} (u_m - P_N u_m, u_{N,k})^2 \\ &\leq \rho_{N,m}^2 \sum_{k=1}^{N} (u_m - P_N u_m, u_{N,k})^2 \\ &\leq \rho_{N,m}^2 \|u_m - P_N u_m\|_H^2. \end{aligned} \tag{15.10}$$

Wegen (15.8) ist

$$\|v_{N,m} - u_{N,m}\|_H = \big\|\big\{(P_N u_m, u_{N,m}) - 1\big\}u_{N,m}\big\|_H = |(P_N u_m, u_{N,m}) - 1| \underbrace{\|u_{N,m}\|_H}_{=1} \tag{15.11}$$

und

$$\underbrace{\|u_m\|_H}_{=1} - \|u_m - v_{N,m}\|_H \le \underbrace{\|v_{N,m}\|_H}_{=|(P_N u_m, u_{N,m})|} \le \underbrace{\|u_m\|_H}_{=1} + \|u_m - v_{N,m}\|_H,$$

das heißt,

$$|1 - (P_N u_m, u_{N,m})| \le \|u_m - v_{N,m}\|_H. \tag{15.12}$$

Die Kombination von (15.11) und (15.12) ergibt

$$\|v_{N,m} - u_{N,m}\|_H \le \|u_m - v_{N,m}\|_H \le \|u_m - P_N u_m\| + \|P_N u_m - v_{N,m}\|_H.$$

Damit erhalten wir

$$\begin{aligned}\|u_m - u_{N,m}\|_H &\le \|u_m - P_N u_m\|_H + \|P_n u_m - v_{N,m}\|_H + \|v_{N,m} - u_{N,m}\|_H \\ &\le 2\|u_m - P_N u_m\| + 2\|P_N u_m - v_{N,m}\|_H.\end{aligned}$$

Abschätzung (15.10) liefert schließlich die Behauptung. □

Lemma 15.10 Es gilt die Identität

$$a(u_{N,m} - u_m, u_{N,m} - u_m) = \lambda_m \|u_{N,m} - u_m\|_H^2 + \lambda_{N,m} - \lambda_m.$$

Beweis. Das Behauptete folgt aus

$$\begin{aligned}a(u_{N,m} - u_m, u_{N,m} - u_m) &= a(u_{N,m}, u_{N,m}) + a(u_m, u_m) - 2a(u_{N,m}, u_m) \\ &= \lambda_{N,m} + \lambda_m - 2\lambda_m(u_{N,m}, u_m) \\ &= \lambda_{N,m} - \lambda_m + 2\lambda_m\big\{1 - (u_{N,m}, u_m)\big\}\end{aligned}$$

und

$$\|u_{N,m} - u_m\|_H^2 = \|u_{N,m}\|_H^2 + \|u_m\|_H^2 - 2(u_{N,m}, u_m) = 2\big\{1 - (u_{N,m}, u_m)\big\}.$$

□

Satz 15.11 (Konvergenz approximativer Eigenfunktionen) Es gelten die Annahmen von Satz 15.3 und sei $\lim_{N\to\infty} \inf_{v_N \in V_N} \|v - v_N\|_V = 0$ für alle $v \in V$. Ist λ_m ein einfacher Eigenwert, dann existiert ein $N_0 > 0$ derart, dass für alle $N \ge N_0$ $\lambda_{N,m}$ ebenfalls ein einfacher Eigenwert ist. Weiterhin existiert eine von N unabhängige Konstante $c > 0$, so dass die Fehlerabschätzungen

$$\|u_m - u_{N,m}\|_V \le c \sup_{\substack{v \in U_m \\ \|v\|_H = 1}} \inf_{v_N \in V_N} \|v - v_N\|_V$$

und

$$\|u_m - u_{N,m}\|_H \le c\|u_m - P_N u_m\|_H$$

gelten.

Beweis. Aus Satz 15.8 ergibt sich, dass der Eigenwert $\lambda_{N,m}$ einfach ist. Ferner ist $\rho_{N,m}$ unabhängig von N beschränkt, so dass die zweite Abschätzung aus (15.9) folgt. Zusammen mit Satz 15.8 und Lemma 15.10 folgt hieraus dann die erste Abschätzung

$$\begin{aligned}\|u_m - u_{N,m}\|_V^2 &\le \frac{1}{c_E} a(u_{N,m} - u_m, u_{N,m} - u_m) \\ &= \frac{1}{c_E}\Big\{\lambda_m \|u_{N,m} - u_m\|_H^2 + \lambda_{N,m} - \lambda_m\Big\} \\ &\le \frac{1}{c_E}\Big\{c\lambda_m \|u_m - P_N u_m\|_H^2 + c(m) \sup_{\substack{v \in U_m \\ \|v\|_H = 1}} \inf_{v_N \in V_N} \|v - v_N\|_V^2\Big\},\end{aligned}$$

denn es ist

$$\|u_m - P_N u_m\|_H^2 \le c\|u_m - P_N u_m\|_V^2 \le c \inf_{v_N \in V_N} \|u_m - v_N\|_V^2.$$

□

Bemerkung Für einen Differentialoperator zweiter Ordnung ist $V = H_0^1(\Omega)$ und $H = L^2(\Omega)$. Bei einer Galerkin-Diskretisierung mit linearen Elementen konvergieren die Eigenwerte quadratisch in h. Hingegen konvergieren die Eigenfunktionen in der Energienorm nur linear, während sie in $L^2(\Omega)$ ebenfalls quadratisch konvergieren. Man beachte jedoch, dass die Konstanten dabei von der Nummer m des Eigenwerts abhängen. △

16. Lineare Elastizität

16.1 Herleitung

In der Elastizitätstheorie werden Verformungen und Spannungen von Körpern betrachtet. Ausgegangen wird von einem elastischen Körper, der mit einem beschänkten und zusammenhängenden Gebiet $\Omega \subset \mathbb{R}^d$ identifiziert werde. Wirken auf diesen Körper *Oberflächenkräfte* $\mathbf{g} : \Gamma_N \to \mathbb{R}^d$ und *Volumenkräfte* $\mathbf{f} : \Omega \to \mathbb{R}^d$, so erzeugen diese einen *Spannungszustand* im Innern des Körpers. Dieser wird beschrieben durch den *Spannungstensor*

$$\boldsymbol{\sigma} = [\sigma_{i,j}] : \Omega \to \mathbb{R}^{d\times d}_{\text{sym}}.$$

Die Spannung erzeugt eine *Deformation* $\boldsymbol{\varphi} : \Omega \to \mathbb{R}^d$ 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

$$\|\boldsymbol{\varphi}(\mathbf{x}+\mathbf{y}) - \boldsymbol{\varphi}(\mathbf{x})\|^2 = \|\nabla\boldsymbol{\varphi}(\mathbf{x})\mathbf{y}\|^2 + \mathcal{O}(\|\mathbf{y}\|^3) = \mathbf{y}^T\big(\nabla\boldsymbol{\varphi}(\mathbf{x})\big)^T\nabla\boldsymbol{\varphi}(\mathbf{x})\mathbf{y} + \mathcal{O}(\|\mathbf{y}\|^3).$$

Für die lokale Änderung von Längen ist also der *Cauchy-Greensche Verzerrungstensor* ausschlaggebend:

$$\mathbf{C} = \big(\nabla\boldsymbol{\varphi}(\mathbf{x})\big)^T\nabla\boldsymbol{\varphi}(\mathbf{x}).$$

Die durch

$$\mathbf{E} = \frac{1}{2}(\mathbf{C} - \mathbf{I}) \tag{16.1}$$

definierte Abweichung von der Identität bezeichnet man als *Verzerrung*.

In der Praxis ist die *Verschiebung* $\mathbf{u} : \Omega \to \mathbb{R}^d$ wichtiger als die Deformation. Sie ist gegeben durch

$$\mathbf{u}(\mathbf{x}) = \boldsymbol{\varphi}(\mathbf{x}) - \mathbf{x}, \quad \mathbf{x} \in \Omega, \tag{16.2}$$

Durch Einsetzen von (16.2) in (16.1) erhält man

$$\mathbf{E} = \frac{1}{2}\big[\nabla\mathbf{u} + (\nabla\mathbf{u})^T\big] + \frac{1}{2}(\nabla\mathbf{u})^T\nabla\mathbf{u}.$$

Bei kleinen Verzerrungen kann der letzte Term vernachlässigt werden und so entsteht der linearisierte *Verzerrungstensor*

$$\boldsymbol{\varepsilon} = [\varepsilon_{i,j}] : \Omega \to \mathbb{R}^{d\times d}_{\text{sym}},$$

welcher gegeben ist durch

$$\boldsymbol{\varepsilon}(\mathbf{u}) := \frac{1}{2}\big[\nabla\mathbf{u} + (\nabla\mathbf{u})^T\big]. \tag{16.3}$$

Gemäß des Hookeschen Gesetzes wird der Zusammenhang zwischen den Verzerrungen $\boldsymbol{\varepsilon}$ und den Spannungen $\boldsymbol{\sigma}$ durch den (konstanten) *Elastizitätstensor* $\mathbf{A} : \mathbb{R}^{d\times d}_{\text{sym}} \to \mathbb{R}^{d\times d}_{\text{sym}}$ beschrieben:

$$\boldsymbol{\sigma} = \mathbf{A}\boldsymbol{\varepsilon} := \frac{E\nu}{(1+\nu)(1-2\nu)}\,\mathrm{tr}\,\big(\boldsymbol{\varepsilon}\big)\mathbf{I} + \frac{E}{1+\nu}\boldsymbol{\varepsilon}.$$

Hier ist $E > 0$ das *Elastizitätsmodul* und $0 < \nu < 1/2$ die *Poisson-Zahl*.

Beispiel 16.1 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

$$|\lambda - (1-\nu)| \le \nu \quad \text{oder} \quad \lambda = 1-2\nu$$

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

$$|\lambda - 1| \le 2\nu \quad \text{oder} \quad \lambda = 1+\nu.$$

△

Im Gleichgewichtszustand gilt im Körper

$$-\operatorname{div}\big(\boldsymbol{\sigma}(\mathbf{u})\big) = \mathbf{f} \quad \text{in } \Omega, \tag{16.4}$$

wobei die Divergenz einer matrixwertigen Funktion $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_d]^T$ gegeben ist durch

$$\operatorname{div} \mathbf{S} = \big[\operatorname{div} \mathbf{s}_1, \operatorname{div} \mathbf{s}_2, \ldots, \operatorname{div} \mathbf{s}_d\big]^T.$$

Zusätzlich zu (16.4) gelten Dirichlet- bzw. Verschiebungsrandbedingungen

$$\mathbf{u} = \mathbf{0} \quad \text{auf } \Gamma_D \subset \partial\Omega \tag{16.5}$$

oder Neumann- bzw. Spannungsrandbedingungen

$$\boldsymbol{\sigma}(\mathbf{u})\mathbf{n} = \mathbf{g} \quad \text{auf } \Gamma_N = \partial\Omega \backslash \overline{\Gamma}_D. \tag{16.6}$$

Wir wollen abschließend noch eine äquivalente Schreibweise der linken Seite von (16.4) herleiten. Mit Hilfe der *Lamé-Konstanten*

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}, \quad \mu = \frac{E}{2(1+\nu)} \tag{16.7}$$

und (16.3) folgt der direkte Zusammenhang zwischen den Verschiebungen $\mathbf{u}$ und dem Spannungstensor $\boldsymbol{\sigma}$:

$$\boldsymbol{\sigma}(\mathbf{u}) = \lambda \operatorname{div}(\mathbf{u})\mathbf{I} + \mu\big[\nabla\mathbf{u} + (\nabla\mathbf{u})^T\big]. \tag{16.8}$$

Demnach haben wir die Beziehung

$$\operatorname{div}\big(\boldsymbol{\sigma}(\mathbf{u})\big) = \mu\Delta\mathbf{u} + (\lambda+\mu)\nabla(\operatorname{div}\mathbf{u})$$

mit dem Laplace-Operator für vektorwertige Funktionen:

$$\Delta\mathbf{u} = \frac{\partial}{\partial x_1^2}\mathbf{u} + \frac{\partial}{\partial x_2^2}\mathbf{u} + \cdots + \frac{\partial}{\partial x_d^2}\mathbf{u}.$$

16.2 Variationsfomulierung

Um die Variationsformulierung zu erhalten, müssen wir (16.4) mit einer Testfunktion $\mathbf{v} : \Omega \to \mathbb{R}^d$ multiplizieren und über das Gebiet Ω integrieren:

$$-\int_\Omega \langle \operatorname{div}\big(\boldsymbol{\sigma}(\mathbf{u})\big), \mathbf{v}\rangle \,\mathrm{d}\mathbf{x} = \int_\Omega \langle \mathbf{f}, \mathbf{v}\rangle \,\mathrm{d}\mathbf{x}. \tag{16.9}$$

Die partielle Integration der linken Seite via komponentenweisem Anwenden des Gaußschen Integralsatzes liefert

$$-\int_\Omega \langle \operatorname{div}\big(\boldsymbol{\sigma}(\mathbf{u})\big), \mathbf{v}\rangle \,\mathrm{d}\mathbf{x} = \int_\Omega \boldsymbol{\sigma}(\mathbf{u}) : \nabla\mathbf{v} \,\mathrm{d}\mathbf{x} - \int_{\partial\Omega} \langle \boldsymbol{\sigma}(\mathbf{u})\mathbf{n}, \mathbf{v}\rangle \,\mathrm{d}\sigma. \tag{16.10}$$

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 (16.5) erfüllen alle kinematisch zulässigen Verschiebungen die Bedingung $\mathbf{v} = \mathbf{0}$ auf Γ_D. Demnach lautet der Energieraum

$$V := \Big\{ \mathbf{v} \in [H^1(\Omega)]^d : \mathbf{v}\big|_{\Gamma_D} = \mathbf{0} \Big\}.$$

Definieren wir also die Bilinearform $a : V \times V \to \mathbb{R}$ durch

$$a(\mathbf{u}, \mathbf{v}) := \int_\Omega \boldsymbol{\sigma}(\mathbf{u}) : \nabla \mathbf{v} \, \mathrm{d}\mathbf{x}$$

und die Linearform $\ell : V \to \mathbb{R}$ durch

$$\ell(\mathbf{v}) := \int_\Omega \langle \mathbf{f}, \mathbf{v} \rangle \, \mathrm{d}\mathbf{x} + \int_{\Gamma_N} \langle \mathbf{g}, \mathbf{v} \rangle \, \mathrm{d}\sigma,$$

so erhalten wir wegen (16.6) offensichtlich die Variationsformulierung:

$$\text{suche } \mathbf{u} \in V, \text{ so dass} \quad a(\mathbf{u}, \mathbf{v}) = \ell(\mathbf{v}) \quad \text{für alle } \mathbf{v} \in V. \tag{16.11}$$

Hierbei können wir wegen der Symmetrie von $\boldsymbol{\sigma}(\mathbf{u})$ und (16.8) die Bilinearform $a(\cdot, \cdot)$ auch umschreiben gemäß

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) &= \int_\Omega \boldsymbol{\sigma}(\mathbf{u}) : \frac{1}{2}\big[\nabla \mathbf{v} + (\nabla \mathbf{v})^T\big] \, \mathrm{d}\mathbf{x} = \int_\Omega \mathbf{A}\boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}) \, \mathrm{d}\mathbf{x} \\ &= 2\mu \int_\Omega \boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}) \, \mathrm{d}\mathbf{x} + \lambda \int_\Omega \operatorname{div}(\mathbf{u}) \operatorname{div}(\mathbf{v}) \, \mathrm{d}\mathbf{x}. \end{aligned} \tag{16.12}$$

Hieraus ist sofort ersichtlich, dass die Bilinearform $a(\cdot, \cdot)$ symmetrisch ist. Mit Hilfe der Cauchy-Schwarzschen Ungleichung weist man ausserdem leicht

$$\begin{aligned} |a(\mathbf{u}, \mathbf{v})| &\le \frac{1}{2}\mu \underbrace{\big\|\nabla \mathbf{u} + (\nabla \mathbf{u})^T\big\|_{L^2(\Omega)}}_{\le 2\|\nabla \mathbf{u}\|_{L^2(\Omega)}} \underbrace{\big\|\nabla \mathbf{v} + (\nabla \mathbf{v})^T\big\|_{L^2(\Omega)}}_{\le 2\|\nabla \mathbf{v}\|_{L^2(\Omega)}} + d\lambda \|\nabla \mathbf{u}\|_{L^2(\Omega)} \|\nabla \mathbf{v}\|_{L^2(\Omega)} \\ &\le (2\mu + d\lambda) |\mathbf{u}|_{H^1(\Omega)} |\mathbf{v}|_{H^1(\Omega)} \end{aligned}$$

nach, das heißt die Bilinearform $a(\cdot, \cdot)$ ist stetig. Auch das Funktional $\ell(\cdot)$ ist stetig:

$$|\ell(\mathbf{v})| \le \|\mathbf{f}\|_{L^2(\Omega)} \|\mathbf{v}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\Gamma_N)} \|\mathbf{v}\|_{L^2(\Gamma_N)} \le c \|\mathbf{v}\|_{H^1(\Omega)}.$$

Die Existenz und die Eindeutigkeit der Lösung der Variationsformulierung (16.11) folgt deshalb aus dem Satz von Lax-Milgram (Satz 3.11), vorausgesetzt $a(\cdot, \cdot)$ ist elliptisch.

16.3 Elliptizitätsabschätzung

Um die Elliptizität der Bilinearform $a(\cdot, \cdot)$ im Fall eines Dirichlet-Randwertproblems nachzuweisen, benötigen wir folgendes Lemma.

Lemma 16.2 (erste Kornsche Ungleichung) Für alle $\mathbf{v} \in [H_0^1(\Omega)]^d$ gilt

$$\int_\Omega \boldsymbol{\varepsilon}(\mathbf{v}) : \boldsymbol{\varepsilon}(\mathbf{v}) \, \mathrm{d}\mathbf{x} \ge \frac{1}{2} |\mathbf{v}|^2_{H^1(\Omega)}.$$

Beweis. Für $\boldsymbol{\varphi} \in [C_0^\infty(\Omega)]^d$ folgt

$$\begin{aligned}\int_\Omega \boldsymbol{\varepsilon}(\boldsymbol{\varphi}) : \boldsymbol{\varepsilon}(\boldsymbol{\varphi})\,\mathrm{d}\mathbf{x} &= \frac{1}{4}\int_\Omega \left[\nabla\boldsymbol{\varphi} + (\nabla\boldsymbol{\varphi})^T\right] : \left[\nabla\boldsymbol{\varphi} + (\nabla\boldsymbol{\varphi})^T\right]\mathrm{d}\mathbf{x} \\ &= \frac{1}{2}\int_\Omega \nabla\boldsymbol{\varphi} : \nabla\boldsymbol{\varphi}\,\mathrm{d}\mathbf{x} + \frac{1}{2}\int_\Omega \nabla\boldsymbol{\varphi} : (\nabla\boldsymbol{\varphi})^T\,\mathrm{d}\mathbf{x}.\end{aligned}$$

Zweimaliges partielles integrieren ergibt

$$\begin{aligned}\int_\Omega \nabla\boldsymbol{\varphi} : (\nabla\boldsymbol{\varphi})^T\,\mathrm{d}\mathbf{x} &= \sum_{i,j=1}^d \int_\Omega \frac{\partial\varphi_i}{\partial x_j}\frac{\partial\varphi_j}{\partial x_i}\,\mathrm{d}\mathbf{x} = \sum_{i,j=1}^d \int_\Omega \frac{\partial\varphi_i}{\partial x_i}\frac{\partial\varphi_j}{\partial x_j}\,\mathrm{d}\mathbf{x} = \int_\Omega \left[\sum_{i=1}^d \frac{\partial\varphi_i}{\partial x_i}\right]^2 \mathrm{d}\mathbf{x} \\ &= \int_\Omega \big(\operatorname{div}(\boldsymbol{\varphi})\big)^2\,\mathrm{d}\mathbf{x} \geq 0.\end{aligned}$$

Daher erhalten wir

$$\int_\Omega \boldsymbol{\varepsilon}(\boldsymbol{\varphi}) : \boldsymbol{\varepsilon}(\boldsymbol{\varphi})\,\mathrm{d}\mathbf{x} \geq \frac{1}{2}\int_\Omega \nabla\boldsymbol{\varphi} : \nabla\boldsymbol{\varphi}\,\mathrm{d}\mathbf{x} = \frac{1}{2}|\boldsymbol{\varphi}|^2_{H^1(\Omega)}.$$

Diese Abschätzung impliziert die Behauptung aufgrund der Dichtheit von $[C_0^\infty(\Omega)]^d$ in $[H_0^1(\Omega)]^d$. □

Aus (16.12) folgern wir

$$a(\mathbf{v},\mathbf{v}) \geq 2\mu\int_\Omega \boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{u})\,\mathrm{d}\mathbf{x} \geq \mu|\mathbf{v}|^2_{H^1(\Omega)}.$$

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

$$a(\mathbf{v},\mathbf{v}) \geq c_E\|\mathbf{v}\|^2_{H^1(\Omega)} \quad \text{für alle } \mathbf{v} \in [H_0^1(\Omega)]^d. \tag{16.13}$$

Bemerkung Die Elliptizitätsabschätzung (16.13) 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. △

16.4 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

$$\mathcal{R} := \left\{\mathbf{v}(\mathbf{x}) = \mathbf{B}\mathbf{x} + \mathbf{d} : \mathbf{B} = -\mathbf{B}^T,\ \mathbf{B} \in \mathbb{R}^{d\times d},\ \mathbf{d} \in \mathbb{R}^d\right\} \subset [L^2(\Omega)]^d \tag{16.14}$$

charakterisiert, denn es gilt:

Lemma 16.3 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 $\mathbf{v} \in \mathcal{R}$. □

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

$$\mathcal{R} = \operatorname{span}\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} x_2 \\ -x_1 \end{bmatrix} \right\},$$

während in $d = 3$ Raumdimensionen $\dim \mathcal{R} = 6$ ist und

$$\mathcal{R} = \operatorname{span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} x_2 \\ -x_1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ x_3 \\ -x_2 \end{bmatrix}, \begin{bmatrix} -x_3 \\ 0 \\ x_1 \end{bmatrix} \right\}.$$

△

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 16.4 (zweite Kornsche Ungleichung) Für alle $\mathbf{v} \in [H^1(\Omega)]^d \cap \mathcal{R}^\perp$ gilt

$$\int_\Omega \boldsymbol{\varepsilon}(\mathbf{v}) : \boldsymbol{\varepsilon}(\mathbf{v}) \, \mathrm{d}\mathbf{x} \geq c \|\mathbf{v}\|^2_{H^1(\Omega)}.$$

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". □

Weil insbesondere $\operatorname{div}(\mathbf{v}) = 0$ gilt für alle $\mathbf{v} \in \mathcal{R}$, schließen wir aus Lemma 16.3, dass

$$a(\mathbf{v}, \mathbf{v}) = 0 \quad \text{für alle } \mathbf{v} \in \mathcal{R}. \tag{16.15}$$

Deshalb benötigen wir im Fall eines reinen Neumann-Randwertproblems eine Kompatibilitätsbedingung an die rechte Seite:

$$\ell(\mathbf{v}) = \int_\Omega \langle \mathbf{f}, \mathbf{v} \rangle \, \mathrm{d}\mathbf{x} + \int_{\Gamma_N} \langle \mathbf{g}, \mathbf{v} \rangle \, \mathrm{d}\sigma = 0 \quad \text{für alle } \mathbf{v} \in \mathcal{R}. \tag{16.16}$$

Definieren wir nun

$$V := \big\{\mathbf{v} \in [H^1(\Omega)]^d : \mathbf{v} \perp \mathcal{R}\big\},$$

so folgt aus der zweiten Kornschen Ungleichung die V-Elliptizität der Bilinearform $a(\cdot,\cdot)$ auch im Fall eines reinen Neumann-Randwertproblems:

$$a(\mathbf{v},\mathbf{v}) \geq 2\mu \int_\Omega \boldsymbol{\varepsilon}(\mathbf{v}) : \boldsymbol{\varepsilon}(\mathbf{v})\,\mathrm{d}\mathbf{x} \geq 2\mu c \|\mathbf{v}\|^2_{H^1(\Omega)} \quad \text{für alle } \mathbf{v} \in V.$$

Deshalb besitzt das Variationsproblem (16.11) unter der Voraussetzung (16.16) eine eindeutige Lösung in V.

16.5 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 $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_R\}$ mit $R = \dim \mathcal{R}$ eine Basis von $\mathcal{R}$. Wir suchen nun $\mathbf{u} \in [H^1(\Omega)]^d$ und $\lambda_1, \lambda_2, \ldots, \lambda_R \in \mathbb{R}$, so dass

$$\begin{aligned} a(\mathbf{u},\mathbf{v}) + \sum_{k=1}^{R} \lambda_k \langle \mathbf{r}_k, \mathbf{v}\rangle &= \ell(\mathbf{v}) \qquad \text{für alle } \mathbf{v} \in [H^1(\Omega)]^d, \\ \langle \mathbf{r}_k, \mathbf{u}\rangle &= 0 \qquad \text{für alle } k = 1,2,\ldots,R. \end{aligned} \tag{16.17}$$

Wählen wir als Testfunktionen gerade die $\mathbf{r}_k$, $k = 1,2,\ldots,R$, so folgt wegen (16.15), (16.16) und der linearen Unabhängigkeit der Funktionen $\{\mathbf{r}_k\}$ sofort $\lambda_k = 0$ für alle $k = 1,2,\ldots,R$. Folglich können wir $\lambda_k = 0$ in der ersten Gleichung aus (16.17) durch die entsprechende zweite Gleichung aus (16.17) ersetzen und erhalten

$$a(\mathbf{u},\mathbf{v}) + \sum_{k=1}^{R} \langle \mathbf{r}_k, \mathbf{u}\rangle \langle \mathbf{r}_k, \mathbf{v}\rangle = \ell(\mathbf{v}) \quad \text{für alle } \mathbf{v} \in [H^1(\Omega)]^d. \tag{16.18}$$

Die modifizierte Bilinearform

$$a_{\mathrm{mod}}(\mathbf{u},\mathbf{v}) := a(\mathbf{u},\mathbf{v}) + \sum_{k=1}^{R} \langle \mathbf{r}_k, \mathbf{u}\rangle \langle \mathbf{r}_k, \mathbf{v}\rangle$$

ist nun elliptisch auf ganz $[H^1(\Omega)]^d$, da in Lemma 16.5 gezeigt wird, dass durch

$$|\!|\!|\mathbf{v}|\!|\!|^2 := |\mathbf{v}|^2_{H^1(\Omega)} + \sum_{k=1}^{R} |\langle \mathbf{r}_k, \mathbf{v}\rangle|^2$$

eine zur $[H^1(\Omega)]^d$-Norm äquivalente Norm definiert wird. Folglich ist das Variationsproblem (16.18) eindeutig lösbar in $[H^1(\Omega)]^d$.

Lemma 16.5 Es gilt

$$\underline{c}\|\mathbf{v}\|_{H^1(\Omega)} \leq |\!|\!|\mathbf{v}|\!|\!| \leq \overline{c}\|\mathbf{v}\|_{H^1(\Omega)}$$

für alle $\mathbf{v} \in [H^1(\Omega)]^d$.

Beweis. Wegen

$$\sum_{k=1}^{R} |\langle \mathbf{r}_k, \mathbf{v}\rangle|^2 \leq \sum_{k=1}^{R} \|\mathbf{r}_k\|^2_{L^2(\Omega)} \|\mathbf{v}\|^2_{L^2(\Omega)} \leq c\|\mathbf{v}\|^2_{H^1(\Omega)},$$

ist die Abschätzung nach oben trivial. Die Abschätzung nach untere 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

$$\lim_{n\to\infty} \sum_{k=1}^{R} |\langle \mathbf{r}_k, \mathbf{v}_n\rangle|^2 = 0 \quad \text{und} \quad \lim_{n\to\infty} |\mathbf{v}_n|_{H^1(\Omega)} = 0$$

gelten. Weil nach dem Rellichschen Auswahlsatz 5.9 die Einbettung $[H^1(\Omega)]^d \hookrightarrow [L^2(\Omega)]^d$ kompakt ist, existiert es 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

$$\|\mathbf{v}_n - \mathbf{v}_m\|^2_{H^1(\Omega)} = \|\mathbf{v}_n - \mathbf{v}_m\|^2_{L^2(\Omega)} + |\mathbf{v}_n - \mathbf{v}_m|^2_{H^1(\Omega)}$$

schließen wir, dass $\{\mathbf{v}_n\}$ eine Cauchy-Folge in $[H^1(\Omega)]^d$ ist. Wegen der Vollständigkeit des Raums existiert ein Grenzelement $\mathbf{v}^\star \in [H^1(\Omega)]^d$ mit $\|\mathbf{v}_n - \mathbf{v}^\star\|_{H^1(\Omega)} \to 0$ für $n \to \infty$. Aus Stetigkeitsgründen ergibt sich

$$\|\mathbf{v}^\star\|_{H^1(\Omega)} = 1 \quad \text{und} \quad |\!|\!|\mathbf{v}^\star|\!|\!| = 0.$$

Es folgt also $|\mathbf{v}^\star|_{H^1(\Omega)} = 0$, weshalb $\mathbf{v}^\star$ konstant sein muss. Wegen $\sum_{k=1}^{R} |\langle \mathbf{r}_k, \mathbf{v}^\star\rangle|^2 = 0$ muss aber $\mathbf{v}^\star = \mathbf{0}$ sein, was ein Widerspruch zu $\|\mathbf{v}^\star\|_{H^1(\Omega)} = 1$ darstellt. □

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. △

16.6 Finite-Element-Approximation

Zunächst erörtern wir die Galerkin-Diskretisierung der Variationsformulierung (16.11) 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:

$$V_h = \big\{\mathbf{v} \in [C(\Omega)]^d : \mathbf{v}|_T \in [\mathcal{P}_1]^d \text{ für alle } T \in \mathcal{T}_h \text{ und } \mathbf{v}|_{\Gamma_D} = \mathbf{0}\big\}.$$

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 (16.11) lautet nun:

$$\text{suche } \mathbf{u}_h \in V_h, \text{ so dass} \quad a(\mathbf{u}_h, \mathbf{v}_h) = \ell(\mathbf{v}_h) \quad \text{für alle } \mathbf{v}_h \in V_h. \tag{16.19}$$

Aufgrund von Stetigkeit und Elliptizität der Bilinearform $a(\cdot,\cdot)$ folgt sofort aus dem Céa-Lemma (Satz 4.1), dass

$$\|\mathbf{u} - \mathbf{u}_h\|_{H^k(\Omega)} \le c h^{2-k} \|\mathbf{u}\|_{H^2(\Omega)}, \quad k = 0, 1, \tag{16.20}$$

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 Polygongebiet 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 16.5 gezeigt wurde, behält die Fehleranschätzung (16.20) auch im Fall eines reinen Neumann-Randwertproblems ihre Gültigkeit, falls die modifizierte Bilinearform aus (16.18) zugrundegelegt wird.

Schließlich wollen wir die Bilinearform (16.12) auch noch explizit anzugeben. Dazu verwenden wir die unter Ingenieuren übliche *Voigtsche Notation*. Für

$$\widehat{\boldsymbol{\varepsilon}} := \begin{bmatrix} \varepsilon_{1,1} \\ \varepsilon_{2,2} \\ 2\varepsilon_{1,2} \end{bmatrix} = \begin{bmatrix} \frac{\partial u_1}{\partial x_1} \\ \frac{\partial u_2}{\partial x_2} \\ \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} \end{bmatrix}$$

gilt, ausgedrückt mit Hilfe der Lamé-Konstanten aus (16.7),

$$\widehat{\boldsymbol{\sigma}} := \begin{bmatrix} \sigma_{1,1} \\ \sigma_{2,2} \\ \sigma_{1,2} \end{bmatrix} = \begin{bmatrix} \lambda + 2\mu & \lambda & \\ \lambda & \lambda + 2\mu & \\ & & \mu \end{bmatrix} \widehat{\boldsymbol{\varepsilon}},$$

vergleiche Beispiel 16.1. 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

$$a(\mathbf{u}, \mathbf{v}) = \int_\Omega \begin{bmatrix} \frac{\partial u_1}{\partial x_1} \\ \frac{\partial u_2}{\partial x_2} \\ \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} \end{bmatrix}^T \begin{bmatrix} \lambda + 2\mu & \lambda & \\ \lambda & \lambda + 2\mu & \\ & & \mu \end{bmatrix} \begin{bmatrix} \frac{\partial v_1}{\partial x_1} \\ \frac{\partial v_2}{\partial x_2} \\ \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} \end{bmatrix} \mathrm{d}\mathbf{x}.$$

Analog ergibt sich im dreidimensionalen Fall für

$$\widehat{\boldsymbol{\varepsilon}} := \begin{bmatrix} \varepsilon_{1,1} \\ \varepsilon_{2,2} \\ \varepsilon_{3,3} \\ 2\varepsilon_{1,2} \\ 2\varepsilon_{1,3} \\ 2\varepsilon_{2,3} \end{bmatrix} = \begin{bmatrix} \frac{\partial u_1}{\partial x_1} \\ \frac{\partial u_2}{\partial x_2} \\ \frac{\partial u_3}{\partial x_3} \\ \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} \\ \frac{\partial u_1}{\partial x_1} + \frac{\partial u_3}{\partial x_3} \\ \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3} \end{bmatrix},$$

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

$$a(\mathbf{u},\mathbf{v}) = \int_\Omega \begin{bmatrix} \frac{\partial u_1}{\partial x_1} \\ \frac{\partial u_2}{\partial x_2} \\ \frac{\partial u_3}{\partial x_3} \\ \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} \\ \frac{\partial u_1}{\partial x_1} + \frac{\partial u_3}{\partial x_3} \\ \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3} \end{bmatrix}^T \begin{bmatrix} \lambda+2\mu & \lambda & \lambda & & & \\ \lambda & \lambda+2\mu & \lambda & & & \\ \lambda & \lambda & \lambda+2\mu & & & \\ & & & \mu & & \\ & & & & \mu & \\ & & & & & \mu \end{bmatrix} \begin{bmatrix} \frac{\partial v_1}{\partial x_1} \\ \frac{\partial v_2}{\partial x_2} \\ \frac{\partial v_3}{\partial x_3} \\ \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} \\ \frac{\partial u_1}{\partial x_1} + \frac{\partial u_3}{\partial x_3} \\ \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3} \end{bmatrix} \mathrm{d}\mathbf{x}.$$

Index