Wissenschaftliches Rechnen II

Vorlesungsskript SS 2012

Mario Bebendorf

Inhaltsverzeichnis

Vorwort

Dieses Skript fasst den Inhalt der von mir im Sommersemester 2012 an der Universität Bonn gehaltenen Vorlesung *Wissenschaftliches Rechnen II* des sechsten Semesters im Bachelorstudiengang Mathematik zusammen. Korrekturvorschläge sind willkommen.

Bonn, 18. Juli 2012

Einleitung

Die Mathematik stellt eine wichtige Grundlage für viele Anwendungsbereiche des täglichen Lebens dar. Ingenieure, Logistikexperten und Ökonomen profitieren in gleicher Weise von mathematischen Methoden und Modellen. Jedoch kann nur ein Bruchteil der auftretenden Probleme analytisch gelöst werden, der Großteil ist mit Papier und Bleistift nicht zu bewältigen. Aus diesem Grund nutzt man zur Umsetzung der immer komplexer werdenden Verfahren den Computer als effizientes Hilfsmittel. Der Hörer dieser Vorlesung lernt grundlegende Konzepte, Algorithmen und Methoden zur numerischen Lösung partieller Differentialgleichungen kennen. Er soll am Ende in der Lage sein, mit Hilfe der erworbenen Kenntnisse selbständig numerische Methoden problemorientiert zu entwickeln, zu analysieren und programmtechnisch umzusetzen.

Literaturangaben:

- K. Atkinson, W. Han: Theoretical Numerical Analysis, Springer
- D. Braess: Finite Elemente – Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie, Springer
- C. Großmann und H.-G. Roos: Numerische Behandlung partieller Differentialgleichungen, Teubner
- V. Girault und P.-A. Raviart: Finite Element Methods for Navier-Stokes Equations. Springer, 1986
- W. Hackbusch: Theorie und Numerik elliptischer Differentialgleichungen, Teubner
- B. Smith, P. Bjørstadt, und W. Gropp: Domain decomposition, Cambridge University Press, 2004
- A. Toselli und O. Widlund: Domain Decomposition Methods – Algorithms and Theory, Springer, 2005

3 Mehrgitterverfahren

In diesem Kapitel stellen wir eines der etabliertesten Verfahren zur Lösung von diskreten Variationsproblemen

$$u_h \in V_h : \quad a(u_h, v_h) = \ell(v_h) \quad \text{für alle } v_h \in V_h \subset V$$

mit symmetrischer, koerziver und stetiger Bilinearform $a : V \times V \to \mathbb{R}$ vor. Mehrgitterverfahren basieren auf der Glättungseigenschaft klassischer Iterationsverfahren.

3.1 Glättungseigenschaften von Iterationsverfahren

Mit der Zerlegung $A = A_L + A_D + A_R$, wobei A_L, A_R strikte untere bzw. obere Dreiecksmatrizen und A_D die Diagonale von A bezeichnen, haben wir in der *Algorithmischen Mathematik II* die folgenden klassischen Iterationsverfahren zur Lösung von linearen Gleichungssystemen $Ax = b$ kennen gelernt.

- **Jacobi-Verfahren** oder **Gesamtschritt-Verfahren**

$$A_D x^{(k+1)} = -(A_L + A_R)x^{(k)} + b, \quad k = 0, 1, 2, \ldots,$$

- **Gauß-Seidel-Verfahren** oder **Einzelschritt-Verfahren**

$$(A_D + A_L)x^{(k+1)} = -A_R x^{(k)} + b, \quad k = 0, 1, 2, \ldots,$$

- **Richardson-Verfahren**

$$x^{(k+1)} = x^{(k)} + \alpha(b - Ax^{(k)}), \quad k = 0, 1, 2, \ldots.$$

Diese Verfahren lassen sich auch in der Form

$$x^{(k+1)} = Tx^{(k)} + c, \quad k = 0, 1, 2, \ldots,$$

schreiben mit den Iterationsmatrizen $T_J := -A_D^{-1}(A_L + A_R)$, $T_{\text{GS}} := -(A_D + A_L)^{-1}A_R$ und $T_R := I - \alpha A$. Weil x die Fixpunktgleichung $x = Tx + c$ erfüllt, gilt dann für den Fehler

$$x - x^{(k+1)} = T(x - x^{(k)}), \quad k = 0, 1, 2, \ldots.$$

Diese Verfahren haben im Vergleich zu den moderneren Krylov-Raum-Methoden ein schlechteres Konvergenzverhalten. Ihre heutige Bedeutung erlangen sie aber durch einen anderen Effekt, den sog. **Glättungseffekt**.

Beispiel 3.1. Wir betrachten wieder die Poisson-Gleichung $-\Delta u = f$ in $\Omega = (0,1)^2$ mit homogenen Randbedingungen $u = 0$ auf $\partial\Omega$. Dabei sei $\overline{\Omega}$ mit einer Courant-Triangulierung (siehe Bspl. 2.59) der Gitterweite h zerlegt. Die Diskretisierung der zugehörigen Bilinearform

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

durch stückweise lineare Finite-Elemente führt auf lineare Gleichungssysteme $Ax = b$ mit dem 5-Punkte-Stern

$$\begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}$$

Die Eigenwerte von A sind (vgl. Übungsblatt 12)

$$\lambda_{k\ell} = 4\left(\sin^2\frac{k\pi h}{2} + \sin^2\frac{\ell\pi h}{2}\right), \quad k,\ell = 1,\dots,n := 1/h - 1.$$

Die zugehörigen Eigenfunktionen $v_{k\ell}$ sind gegeben durch die Knotenwerte der Funktionen

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

Daher ergibt sich für die Eigenwerte der Iterationsmatrix $T_R = I - \alpha A$ des Richardson-Verfahrens für $\alpha = 1/8$

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

Für den betragsmäßig größten Eigenwert μ_{11} folgt

$$\rho(I - A/8) = |\mu_{11}| \approx 1 - \frac{\pi^2 h^2}{4}.$$

Die zeigt, dass sich die Konvergenz des Richardson-Verfahrens für $h \to 0$ verlangsamt.

Auf der anderen Seite bilden die Eigenvektoren $v_{k\ell}$ der symmetrischen Matrix A eine Basis des $\mathbb{R}^{n^2}$. Für die Eigenwerte der Iterationsmatrix $I - A/8$ des Richardson-Verfahrens zu Eigenvektoren $\{v_{k\ell},\, k \ge (n+1)/2$ oder $\ell \ge (n+1)/2\}$ mit einem hochfrequenten Anteil in mindestens einer Richtung gilt

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

Mit der Basis-Darstellung

$$x - x^{(k)} = \sum_{k,\ell=1}^{n} \beta_{k\ell} v_{k\ell}$$

reduziert sich wegen

$$x - x^{(k+1)} = \sum_{k,\ell=1}^{n} \beta_{k\ell}(I - \frac{1}{8}A)v_{k\ell} = \sum_{k,\ell=1}^{n} \beta_{k\ell}\mu_{k\ell}v_{k\ell}$$

der hochfrequente Anteil des Fehlers pro Richardson-Iterationsschritt um mindestens den Faktor 1/2. Die Reduktion des hochfrequenten Anteils bedeutet eine Glättung des Fehlers. Für den niederfrequenten Anteil gilt eine solche Reduktion um einen konstanten Faktor nicht, was letztlich die langsame Konvergenz des Richardson-Verfahrens zur Folge hat.

Um die Glättungseigenschaft mathematisch zu pr¨azisieren, führen wir die folgenden diskreten Normen ein.

Definition 3.2. *Sei A positiv-definit mit $A = VDV^T$. Dabei sei V eine orthogonale Matrix und D eine Diagonalmatrix. Für $s \in \mathbb{R}$ und $x \in \mathbb{R}^n$ definieren wir*

$$|||x|||_s = \sqrt{x^T A^s x}, \quad A^s := VD^sV^T.$$

Bemerkung. Die Norm $|||\cdot|||_s$ hebt die Komponenten in Richtung der zu den größeren Eigenwerten λ von A gehörenden Eigenvektoren v, $\|v\|_2 = 1$, um so mehr hervor, je größer s ist. Dies sind wegen

$$\lambda = v^T Av = a(Jv, Jv) = \|\nabla Jv\|^2_{L^2(\Omega)}$$

beim Poisson-Problem gerade die hochfrequenten Eigenvektoren.

Wir fassen einige Eigenschaften dieser Normen zusammen.

Lemma 3.3. *Es gilt $|||x|||_0 = \|x\|_2$ und für alle $r = \frac{1}{2}(s+t)$*

$$|x^T A^r y| \le |||x|||_s |||y|||_t$$

und somit $|||x|||_r \le |||x|||_s^{1/2} |||x|||_t^{1/2}$ für alle $x, y \in \mathbb{R}^n$.

Beweis. Die verallgemeinerte Cauchy-Schwarzsche Ungleichung folgt aus

$$|x^T A^r y| = |x^T A^{t/2} A^{s/2} y| \le \|A^{s/2}x\|_2 \|A^{t/2}y\|_2 = |||x|||_s |||y|||_t.$$

□

Im folgenden Satz präzisieren wir die Glättungseigenschaft des Richardson-Verfahrens.

Satz 3.4. *Sei A positiv-definit und $T_R = I - \alpha A$ mit $0 < \alpha \le 1/\lambda_{\max}(A)$. Dann gilt für $s \ge 0$ und $t > 0$ mit $c := [t/(2\alpha e)]^{t/2}$*

$$|||T_R^k x|||_{s+t} \le c\, k^{-t/2} |||x|||_s, \quad x \in \mathbb{R}^n.$$

Ferner gilt $|||T_R x|||_s \le |||x|||_s$ für alle $x \in \mathbb{R}^n$.

Beweis. Es bezeichne (λ_i, v_i), $i = 1, \ldots, n$, die Eigenpaare von A. Für $x = \sum_{i=1}^n \beta_i v_i$ erhalten wir

$$T_R^k x = \sum_{i=1}^n \beta_i (1 - \alpha\lambda_i)^k v_i$$

und hieraus

$$\begin{aligned} |||T_R^k x|||_{s+t}^2 = x^T T_R^k A^{s+t} T_R^k x &= \sum_{i=1}^n (1-\alpha\lambda_i)^{2k} \lambda_i^{s+t} \beta_i^2 \le \max_{i=1,\ldots,n} \lambda_i^t (1-\alpha\lambda_i)^{2k} \sum_{i=1}^n \lambda_i^s \beta_i^2 \\ &= \frac{1}{\alpha^t} \max_{i=1,\ldots,n} (\alpha\lambda_i)^t (1-\alpha\lambda_i)^{2k} |||x|||_s^2. \end{aligned}$$

Nach Voraussetzung ist $0 \le \alpha\lambda_i \le 1$ für alle $i = 1, \dots, n$. Daher folgt der zweite Teil der Behauptung, und für den ersten Teil erhalten wir

$$\max_{i=1,\dots,n} (\alpha\lambda_i)^t (1-\alpha\lambda_i)^{2k} \le \max_{0\le\xi\le 1} \xi^t (1-\xi)^{2k}.$$

Die Funktion $f(\xi) := \xi(1-\xi)^{2k/t}$ nimmt wegen $f'(\xi) = (1-\xi)^{2k/t-1}(1-(2k+t)\xi/t)$ ihr Maximum in $\xi_{\max} := t/(2k+t) \in [0,1]$ an. Daher folgt

$$\max_{i=1,\dots,n} \alpha\lambda_i (1-\alpha\lambda_i)^{2k/t} \le f(\xi_{\max}) = \frac{t}{2k+t}\left(\frac{2k}{2k+t}\right)^{2k/t} \le \frac{t}{2k}\frac{1}{(1+t/(2k))^{2k/t}} \le \frac{t}{2ke}$$

und hieraus

$$|||T_R^k x|||_{s+t}^2 \le \left(\frac{t}{2\alpha k e}\right)^t |||x|||_s^2.$$

□

Ist $A = A_h$ eine Finite-Element-Steifigkeitsmatrix des Laplace-Operators, so wissen wir aus dem Beweis zu Satz 2.67, dass $\lambda_{\max}(A_h) \sim h^{d-2}$. Für $s = 0$ und $t = 2$ bzw. $s = t = 1$ folgt somit aus Satz 3.4

$$|||T_R^k x|||_2 \le \frac{c}{k} h^{d-2} |||x|||_0, \qquad |||T_R^k x|||_2 \le \frac{c}{\sqrt{k}} h^{d/2-1} |||x|||_1 \tag{3.1}$$

für alle $x \in \mathbb{R}^n$.

3.2 Zweigitterverfahren

Wir betrachten zwei Finite-Elemente-Räume V_h und V_H mit $V_H \subset V_h$ und Dimensionen $n = \dim V_h$, $N = \dim V_H$. Wie beim BPX-Verfahren definieren wir Prolongations- und Restriktionsoperatoren I_H^h und I_h^H. Ein Iterationsschritt des **Zweigitterverfahrens** zur Lösung von $A_h x_h = b_h$ mit der Glättung S_h setzt sich dann wie folgt zusammen.

1. *Vorglättung*: Setze $x_h^{\text{pre},0} := x_h^{\text{alt}}$ und führe ν_1 Glättungschritte durch:

$$x_h^{\text{pre},k+1} = S_h x_h^{\text{pre},k} + s_h, \quad k = 0, 1, \dots, \nu_1 - 1.$$

2. *Restriktion*: Restringiere das Residuum auf das gröbere Gitter

$$r_H := I_h^H (b_h - A_h x_h^{\text{pre},\nu_1}).$$

3. *Grob-Gitter-Korrektur*: Löse die **Defektgleichung** auf dem gröberen Gitter

$$A_H e_H = r_H$$

und korrigiere $x_h^{\text{post},0} := x_h^{\text{pre},\nu_1} + I_H^h e_H$.

4. *Nachglättung*: Führe ν_2 Glättungschritte

$$x_h^{\text{post},k+1} = S_h x_h^{\text{post},k} + s_h, \quad k = 0, 1, \dots, \nu_2 - 1.$$

durch und setzte $x_h^{\text{neu}} := x_h^{\text{post},\nu_2}$.

Durch diesen Algorithmus ist wiederum ein Iterationsverfahren der Form

$$x_h^{\text{neu}} = T_{2G} x_h^{\text{alt}} + t_h$$

mit Iterationsmatrix

$$T_{2G} := S_h^{\nu_2}(I - I_H^h A_H^{-1} I_h^H A_h) S_h^{\nu_1}$$

und einem $t_h \in V_h$ gegeben. Bezeichnet x_h die exakte Lösung von $A_h x_h = b_h$, so gilt insbesondere für den Fehler nach einem Schritt des Zweigitterverfahrens

$$x_h - x_h^{\text{neu}} = T_{2G}(x_h - x_h^{\text{alt}}).$$

Für die Konvergenzanalyse benötigen wir die folgende Approximationseigenschaft. Bei der Bildung der Normen $|||v|||_s$ identifizieren wir $v \in V_h$ wieder mit dem Koeffizientenvektor in der nodalen Basis. Wir verwenden die folgende offensichtliche Normäquivalenz (vgl. Lemma 2.65 und Lemma 2.66 bzw. die Äquivalenz von $\|\cdot\|_{H^1(\Omega)}$ und Energienorm)

$$c_1 h^{d/2} |||x|||_0 \le \|J_h x\|_{L^2(\Omega)} \le C_1 h^{d/2} |||x|||_0, \qquad c_2 |||x|||_1 \le \|J_h x\|_{H^1(\Omega)} \le C_2 |||x|||_1 \tag{3.2}$$

für alle $x \in \mathbb{R}^n$.

Lemma 3.5. *Zu $v \in V_h$ sei $\mathcal{P}_h v \in V_h$ die Galerkin-Projektion, d.h. die Lösung von $a(\mathcal{P}_h v, w) = a(v, w)$ für alle $w \in V_h$. Ferner sei Ω konvex oder besitze einen glatten Rand. Dann gilt*

$$\|v - \mathcal{P}_h v\|_{L^2(\Omega)} \le c\, h \, \|v - \mathcal{P}_h v\|_{H^1(\Omega)}, \qquad \|v - \mathcal{P}_h v\|_{H^1(\Omega)} \le c\, h^{1-d/2} \, |||v|||_2.$$

Beweis. Wegen des Regularitätssatzes 2.40 ist die duale Lösung u_φ im Aubin-Nitsche-Lemma H^2-regulär. Die erste Aussage folgt wie in der Bemerkung nach Satz 2.64. Ferner gilt nach Lemma 3.3 und (3.2) wie beim Céa-Lemma

$$\begin{aligned} c_K \|v - \mathcal{P}_h v\|_{H^1(\Omega)}^2 &\le a(v - \mathcal{P}_h v, v - \mathcal{P}_h v) = a(v - \mathcal{P}_h v, v) = v^T A_h (v - \mathcal{P}_h v) \\ &\le |||v - \mathcal{P}_h v|||_0 \, |||v|||_2 \le c_1 \, h^{-d/2} \, \|v - \mathcal{P}_h v\|_{L^2(\Omega)} \, |||v|||_2 \\ &\le c_1 c \, h^{1-d/2} \, \|v - \mathcal{P}_h v\|_{H^1(\Omega)} \, |||v|||_2. \end{aligned}$$

Division liefert die Behauptung. □

Das folgende Lemma zeigt, dass e_H aus der Defektgleichung die Galerkin-Projektion von $x_h - x_h^{\text{pre},\nu_1}$ auf V_H ist.

Lemma 3.6. *Für alle $v_H \in V_H$ gilt $a(J_H e_H, v_H) = a(J_h(x_h - x_h^{\text{pre},\nu_1}), v_H)$.*

Beweis. Für $y \in \mathbb{R}^N$ gilt nach Lemma 2.72

$$\begin{aligned} a(J_H e_H, J_H y) &= y^T A_H e_H = y^T I_h^H (b_h - A_h x_h^{\text{pre},\nu_1}) = (I_H^h y)^T A_h (x_h - x_h^{\text{pre},\nu_1}) \\ &= a(J_h(x_h - x_h^{\text{pre},\nu_1}), J_h I_H^h y) = a(J_h(x_h - x_h^{\text{pre},\nu_1}), J_H y). \end{aligned}$$

□

Der folgende Satz zeigt die Konvergenz des Zweigitterverfahrens bei einer genügend großen Anzahl von Vorglättungsschritten ν_1.

Satz 3.7 (Konvergenzsatz). *Es gelten die Voraussetzungen von Lemma 3.5. Dann gilt für das Zweigitterverfahren bei Richardson-Glättung mit* $0 < \alpha \leq 1/\lambda_{\max}(A_h)$

$$|||x_h - x_h^{\text{neu}}|||_0 \leq \frac{c}{\nu_1}|||x_h - x_h^{\text{alt}}|||_0, \qquad |||x_h - x_h^{\text{neu}}|||_1 \leq \frac{c}{\sqrt{\nu_1}}|||x_h - x_h^{\text{alt}}|||_1.$$

Beweis. Für die Glättung mittels Richardson-Verfahren gilt

$$x_h - x_h^{\text{pre},\nu_1} = T_R^{\nu_1}(x_h - x_h^{\text{alt}})$$

und somit nach (3.1)

$$|||x_h - x_h^{\text{pre},\nu_1}|||_2 \leq \frac{c}{\nu_1}h^{d-2}|||x_h - x_h^{\text{alt}}|||_0, \qquad |||x_h - x_h^{\text{pre},\nu_1}|||_2 \leq \frac{c}{\sqrt{\nu_1}}h^{d/2-1}|||x_h - x_h^{\text{alt}}|||_1.$$

Anwendung von Lemma 3.5 auf $v := x_h - x_h^{\text{pre},\nu_1}$ liefert wegen Lemma 3.6

$$\begin{aligned}\|J_h(x_h - x_h^{\text{post},0})\|_{L^2(\Omega)} &= \|J_h(x_h - x_h^{\text{pre},\nu_1} - I_H^h e_H)\|_{L^2(\Omega)} = \|J_h(x_h - x_h^{\text{pre},\nu_1}) - J_H e_H\|_{L^2(\Omega)} \\ &\leq c\,h\|J_h(x_h - x_h^{\text{pre},\nu_1}) - J_H e_H\|_{H^1(\Omega)} \leq c\,h^{2-d/2}|||x_h - x_h^{\text{pre},\nu_1}|||_2.\end{aligned}$$

Die Nachglättung verschlechtert die Approximation nicht, d.h. nach Satz 3.4 gilt für alle $x \in \mathbb{R}^n$, dass $|||T_R x|||_s \leq |||x|||_s$. Aus (3.2) folgt somit

$$\begin{aligned}|||x_h - x_h^{\text{neu}}|||_0 = |||x_h - x_h^{\text{post},\nu_2}|||_0 &\leq |||x_h - x_h^{\text{post},0}|||_0 \leq c\,h^{-d/2}\|J_h(x_h - x_h^{\text{post},0})\|_{L^2(\Omega)} \\ &\leq c'h^{2-d}|||x_h - x_h^{\text{pre},\nu_1}|||_2 \leq \frac{c''}{\nu_1}|||x_h - x_h^{\text{alt}}|||_0\end{aligned}$$

bzw. nach dem zweiten Teil von Lemma 3.5

$$\begin{aligned}|||x_h - x_h^{\text{neu}}|||_1 = |||x_h - x_h^{\text{post},\nu_2}|||_1 &\leq |||x_h - x_h^{\text{post},0}|||_1 \leq c\|J_h(x_h - x_h^{\text{post},0})\|_{H^1(\Omega)} \\ &\leq c'h^{1-d/2}|||x_h - x_h^{\text{pre},\nu_1}|||_2 \leq \frac{c''}{\sqrt{\nu_1}}|||x_h - x_h^{\text{alt}}|||_1.\end{aligned}$$

□

3.3 Mehrgitterverfahren

Es sei

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

eine geschachtelte Folge von Finite-Elemente-Räumen. Diese seien durch uniformes Verfeinern des Grobgitterraums V_0 erzeugt. Beim Zweigitterverfahren taucht das Problem auf, die Defektgleichung $A_H e_H = r_H$ auf dem gröberen Gitter zu lösen. Die Matrix A_H ist zwar kleiner dimensioniert als A_h, wird aber in der Regel noch immer großdimensioniert sein. Allerdings ist sie vom gleichen Typ wie das Ausgangsproblem $A_h x_h = b_h$, so dass die Idee des Zweigitterverfahrens rekursiv auf die Defektgleichung angewendet werden kann, um die Komplexität zu reduzieren. Ein Iterationsschritt des **Mehrgitterverfahrens** MGM_ℓ zur Lösung von $A_\ell x_\ell = b_\ell$ mit der Glättung S_ℓ auf dem ℓ-ten Gitter setzt sich dann wie folgt zusammen.

1. *Vorglättung*: Setze $x_\ell^{\mathrm{pre},0} := x_\ell^{\mathrm{alt}}$ und führe ν_1 Glättungschritte durch:
$$x_\ell^{\mathrm{pre},k+1} = S_\ell x_\ell^{\mathrm{pre},k} + s_\ell, \quad k = 0, 1, \ldots, \nu_1 - 1.$$

2. *Restriktion*: Restringiere das Residuum auf das nächst gröbere Gitter
$$r_{\ell-1} := I_\ell^{\ell-1}(b_\ell - A_\ell x_\ell^{\mathrm{pre},\nu_1}).$$

3. *Grob-Gitter-Korrektur*: Ist $\ell = 1$, so löse die Defektgleichung $A_0 e_0 = r_0$ exakt. Im Fall $\ell > 1$ wende μ Schritte des Mehrgitterverfahrens $\mathrm{MGM}_{\ell-1}$ auf $A_{\ell-1}e_{\ell-1} = r_{\ell-1}$ mit Startwert 0 an. Anschließend korrigiere
$$x_\ell^{\mathrm{post},0} := x_\ell^{\mathrm{pre},\nu_1} + I_{\ell-1}^{\ell} e_{\ell-1}.$$

4. *Nachglättung*: Führe ν_2 Glättungschritte
$$x_\ell^{\mathrm{post},k+1} = S_\ell x_\ell^{\mathrm{post},k} + s_\ell, \quad k = 0, 1, \ldots, \nu_2 - 1.$$
durch und setzte $x_\ell^{\mathrm{neu}} := x_\ell^{\mathrm{post},\nu_2}$.

Bemerkung.

(a) Die Defektgleichung entsteht aus dem niederfrequenten Rest der Vorglättung. Beim Mehrgitterverfahren wird hierauf auf dem nächst gröberen Gitter wieder eine Glättung angewendet. Diese kann den Fehler weiter reduzieren, weil die verbliebenen niederfrequenten Anteile des Fehlers auf dem gröberen Gitter hochfrequenter erscheinen.

(b) In der Praxis verwendet man nur $\mu = 1$ (**V-Zyklus**) und $\mu = 2$ (**W-Zyklus**). Diese Bezeichnungen leiten sich aus dem schematischen Ablauf der Iteration (siehe Abb. 3.1) ab.

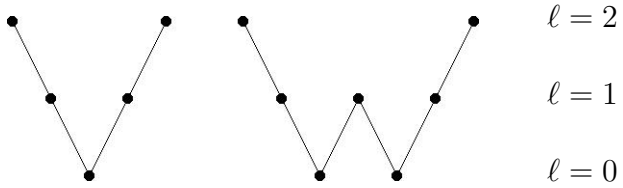


Abbildung 3.1: Schematischer Ablauf V- und W-Zyklus bei drei Gittern.

3.3.1 Konvergenzbeweis für den W-Zyklus

Im Gegensatz zum Zweigitterverfahren wird die Defektgleichung auf dem göberen Gitter nicht exakt gelöst, sondern durch rekursive Anwendung des Mehrgitterverfahrens behandelt. Für die Konvergenzanalyse fassen wir das Mehrgitterverfahren als gestörtes Zweigitterverfahren auf. Ziel ist es, eine Abschätzung der Art
$$|||x_\ell - x_\ell^{\mathrm{neu}}|||_s \leq \rho_\ell \, |||x_\ell - x_\ell^{\mathrm{alt}}|||_s, \quad s = 0, 1,$$
für die ℓ-te Ebene herzuleiten. Dabei setzen wir die Konvergenzrate ρ_1 für das Zweigitterverfahren bzgl. der Norm $|||\cdot|||_s$ als gegeben voraus; vgl. Satz 3.7.

Lemma 3.8. *Für $s = 0, 1$ und $\ell \geq 2$ gilt $\rho_\ell \leq \rho_1 + \rho_{\ell-1}^{\mu}(\rho_1 + 1)$.*

Beweis. Es bezeichne $\hat{x}_\ell^{\text{post},0}$ das Ergebnis der exakten Grob-Gitter-Korrektur. Dann gilt nach Satz 3.7

$$|||x_\ell - \hat{x}_\ell^{\text{post},0}|||_s \leq \rho_1 |||x_\ell - x_\ell^{\text{alt}}|||_s. \tag{3.3}$$

Weil nach Satz 3.4 die Richardson-Glättung den Fehler nicht verschlechtert, erhält man hieraus mit der Dreiecksungleichung

$$|||\hat{x}_\ell^{\text{post},0} - x_\ell^{\text{pre},\nu_1}|||_s \leq |||x_\ell - \hat{x}_\ell^{\text{post},0}|||_s + |||x_\ell - x_\ell^{\text{pre},\nu_1}|||_s \leq (\rho_1 + 1)\, |||x_\ell - x_\ell^{\text{alt}}|||_s. \tag{3.4}$$

Die linke Seite gibt die Größe der Grob-Gitter-Korrektur bei exakter Rechnung an. Die tatsächliche Korrektur unterscheidet sich davon um den Fehler auf der Ebene $\ell - 1$. Nach Induktionsvoraussetzung gilt

$$|||\hat{x}_\ell^{\text{post},0} - x_\ell^{\text{post},0}|||_s \leq \rho_{\ell-1}^{\mu} |||\hat{x}_\ell^{\text{post},0} - x_\ell^{\text{pre},\nu_1}|||_s. \tag{3.5}$$

Setzt man (3.4) in (3.5) ein, so folgt mit der Dreiecksungleichung und (3.3)

$$|||x_\ell - x_\ell^{\text{post},0}|||_s \leq [\rho_1 + \rho_{\ell-1}^{\mu}(\rho_1 + 1)]\, |||x_\ell - x_\ell^{\text{alt}}|||_s.$$

Weil die Nachglättung den Fehler nicht verschlechtert, gilt $|||x_\ell - x_\ell^{\text{neu}}|||_s \leq |||x_\ell - x_\ell^{\text{post},0}|||_s$ und somit die Behauptung. □

Im folgenden Satz zeigen wir mit der Abschätzung für ρ_ℓ die Konvergenz des W-Zyklus bei hinreichend vielen Vorglättungsschritten.

Satz 3.9 (Konvergenz des W-Zyklus). *Für die Zweigitterrate sei $\rho_1 \leq 1/5$. Dann gilt im Fall des W-Zyklus*

$$|||x_\ell - x_\ell^{\text{neu}}|||_s \leq \frac{5}{3}\rho_1 |||x_\ell - x_\ell^{\text{alt}}|||_s, \quad s = 0, 1.$$

Beweis. Für $\ell = 1$ ist die Aussage wahr. Angenommen, sie gilt für $\ell - 1$. Dann folgt nach Lemma 3.8

$$\rho_\ell \leq \rho_1 + \rho_{\ell-1}^2(\rho_1 + 1) \leq \rho_1 + \frac{1}{3}\frac{5}{3}\rho_1\left(\frac{1}{5} + 1\right) = \frac{5}{3}\rho_1.$$

□

ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_∞
0.1	0.1110	0.1136	0.1142	0.1143	0.1144
0.2	0.2480	0.2738	0.2900	0.3009	0.3333
0.3	0.4170	0.5261	0.6598	0.8659	∞

Tabelle 3.1: Konvergenzraten ρ_ℓ des W-Zyklus gemäß Lemma 3.8.

Im Fall der Energienorm kann man die in Lemma 3.8 gezeigte Abschätzung mit Orthogonalitätsargumenten verbessern.

Lemma 3.10. *Bzgl. der Energienorm ($s = 1$) gilt $\rho_\ell^2 \le \rho_1^2 + \rho_{\ell-1}^{2\mu}(1-\rho_1^2)$ für $\ell \ge 2$.*

Beweis. In der Energienorm ist die exakte Grob-Gitter-Korrektur $\hat{x}_\ell^{\text{post},0}$ nach Lemma 3.6 die orthogonale Projektion von x_ℓ auf $x_\ell^{\text{pre},\nu_1} + V_{\ell-1}$. Damit ist der Fehler $x_\ell - \hat{x}_\ell^{\text{post},0}$ orthogonal zu $V_{\ell-1}$ also insbesondere zu $\hat{x}_\ell^{\text{post},0} - x_\ell^{\text{pre},\nu_1}$ und zu $\hat{x}_\ell^{\text{post},0} - x_\ell^{\text{post},0}$. Daher kann (3.4) ersetzt werden durch

$$|||\hat{x}_\ell^{\text{post},0} - x_\ell^{\text{pre},\nu_1}|||_1^2 = |||x_\ell - x_\ell^{\text{pre},\nu_1}|||_1^2 - |||x_\ell - \hat{x}_\ell^{\text{post},0}|||_1^2.$$

Aus der Orthogonalität und (3.5) folgt

$$\begin{aligned} |||x_\ell - x_\ell^{\text{post},0}|||_1^2 &= |||x_\ell - \hat{x}_\ell^{\text{post},0}|||_1^2 + |||\hat{x}_\ell^{\text{post},0} - x_\ell^{\text{post},0}|||_1^2 && \text{(3.6a)} \\ &\le |||x_\ell - \hat{x}_\ell^{\text{post},0}|||_1^2 + \rho_{\ell-1}^{2\mu}|||\hat{x}_\ell^{\text{post},0} - x_\ell^{\text{pre},\nu_1}|||_1^2 && \text{(3.6b)} \\ &= (1-\rho_{\ell-1}^{2\mu})|||x_\ell - \hat{x}_\ell^{\text{post},0}|||_1^2 + \rho_{\ell-1}^{2\mu}|||x_\ell - x_\ell^{\text{pre},\nu_1}|||_1^2. && \text{(3.6c)} \end{aligned}$$

Aus (3.3) und $|||x_\ell - x_\ell^{\text{pre},\nu_1}|||_1 \le |||x_\ell - x_\ell^{\text{alt}}|||_1$ ergibt sich

$$|||x_\ell - x_\ell^{\text{post},0}|||_1^2 \le [(1-\rho_{\ell-1}^{2\mu})\rho_1^2 + \rho_{\ell-1}^{2\mu}]\,|||x_\ell - x_\ell^{\text{alt}}|||_1^2.$$

□

Satz 3.11. *Für die Zweigitterrate sei bzgl. der Energienorm $\rho_1 \le 1/2$. Dann gilt im Fall des W-Zyklus*

$$|||x_\ell - x_\ell^{\text{neu}}|||_1 \le \frac{6}{5}\rho_1 |||x_\ell - x_\ell^{\text{alt}}|||_1.$$

Beweis. Für $\ell = 1$ ist die Aussage wahr. Angenommen, sie gilt für $\ell - 1$. Dann folgt nach Lemma 3.10

$$\rho_\ell^2 \le \rho_1^2 + (\frac{6}{5}\rho_1)^4(1-\rho_1^2) = \rho_1^2\left[1 + \left(\frac{6}{5}\right)^4 \rho_1^2(1-\rho_1^2)\right] \le \rho_1^2\left[1 + \left(\frac{6}{5}\right)^4 \frac{1}{4}\frac{3}{4}\right] \le \frac{36}{25}\rho_1^2.$$

□

ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_∞
0.1	0.1005	0.1005	0.1005	0.1005	0.1005
0.2	0.2038	0.2041	0.2041	0.2041	0.2041
0.3	0.3120	0.3141	0.3144	0.3145	0.3145
0.4	0.4260	0.4332	0.4354	0.4361	0.4364
0.5	0.5449	0.5622	0.5700	0.5738	0.5774
0.6	0.6655	0.6968	0.7148	0.7260	0.7500
0.7	0.7826	0.8254	0.8525	0.8714	0.9802
0.8	0.8874	0.9291	0.9530	0.9680	1.0000

Tabelle 3.2: Konvergenzraten ρ_ℓ des W-Zyklus gemäß Lemma 3.10.

3.3.2 Konvergenzbeweis für den V-Zyklus

Wir haben bisher nur den W-Zyklus ($\mu = 2$) betrachtet. Zwar gelten die rekursiven Abschätzugen für ρ_ℓ auch im Fall $\mu = 1$, sie liefern allerdings keine ausreichenden Schranken. Im Folgenden werden wir die Beweistechnik verfeinern, um auch für den V-Zyklus Konvergenz zu beweisen. Ferner werden wir sehen, dass bereits ein Glättungsschritt ausreicht.

Für den Beweis benötigen wir zwei Lemmata. Dazu führen wir ein Maß für die Glattheit von Finite-Elemente-Funktionen $v \in V_h$ ein:

$$\beta(v) := \begin{cases} 1 - \lambda_{\max}^{-1}(A_h)\frac{|||v|||_2^2}{|||v|||_1^2}, & v \neq 0, \\ 0, & v = 0. \end{cases}$$

Offenbar gilt $0 \leq \beta < 1$. Glatte Funktionen v_h liefern einen Wert β nahe 1 und Funktionen mit großen oszillierenden Anteilen ein kleines β; vgl. die Bemerkung nach Definition 3.2.

Anders als der bisher verwendete Satz 3.4 zur Glättungseigenschaft des Richardson-Verfahrens macht das folgende Lemma Aussagen zwischen gleichen Normen.

Lemma 3.12. *Sei A positiv-definit und $T_R = I - \alpha A$. Dann gilt für $\alpha \geq 1/\lambda_{\max}(A)$*

$$|||T_R^k x|||_1 \leq \beta^k |||x|||_1, \quad x \in \mathbb{R}^n,$$

mit $\beta := \beta(T_R^k x)$.

Beweis. Sei $x = \sum_{i=1}^n \alpha_i v_i$ bzgl. der Orthonormalbasis $v_1, \ldots, v_n$ aus Eigenvektoren von A zerlegt, und es sei $\mu_i := 1 - \alpha\lambda_i$. Dann folgt aus der Hölderschen Ungleichung

$$\sum_{i=1}^n \lambda_i \mu_i^{2k} |\alpha_i|^2 \leq \left(\sum_{i=1}^n \lambda_i \mu_i^{2k+1} |\alpha_i|^2 \right)^{\frac{2k}{2k+1}} \left(\sum_{i=1}^n \lambda_i |\alpha_i|^2 \right)^{\frac{1}{2k+1}}.$$

Wegen $|||x|||_1^2 = \sum_{i=1}^n \lambda_i |\alpha_i|^2$ ist dies äquivalent mit

$$|||T_R^k x|||_1^{2k+1} \leq |||T_R^{k+1/2} x|||_1^{2k} \, |||x|||_1.$$

Mit $y := T_R^k x$ erhält man nach Division durch $|||y|||_1^{2k}$

$$|||T_R^k x|||_1 \leq \left(\frac{|||T_R^{1/2} y|||_1}{|||y|||_1} \right)^{2k} |||x|||_1.$$

Weil T_R symmetrisch und mit A vertauschbar ist, folgt aus $\alpha \geq 1/\lambda_{\max}(A)$

$$|||T_R^{1/2} y|||_1^2 = (T_R^{1/2} y)^T A T_R^{1/2} y = y^T A T_R y = y^T A y - \alpha y^T A^2 y \leq \beta(J_h y) \, |||y|||_1^2.$$

□

Auch die Genauigkeit der Grob-Gitter-Korrektur läßt sich mit Hilfe von β abschätzen.

Lemma 3.13. *Für die exakte Grob-Gitter-Korrektur gilt*

$$|||x_\ell - \hat{x}_\ell^{\text{post},0}|||_1 \leq \min\{1, c\sqrt{1 - \beta(x_\ell - x_\ell^{\text{pre},\nu_1})}\} \, |||x_\ell - x_\ell^{\text{pre},\nu_1}|||_1.$$

Beweis. Wie im Beweis von Lemma 3.10 gesehen ist $\hat{x}_\ell^{\text{post},0}$ die orthogonale Projektion von x_ℓ auf $x_\ell^{\text{pre},\nu_1} + V_{\ell-1}$. Nach Lemma 3.5 gilt $|||x_\ell - \hat{x}_\ell^{\text{post},0}|||_1 \le c\, h^{1-d/2}\, |||x_\ell - x_\ell^{\text{pre},\nu_1}|||_2$. Mit Hilfe von $\lambda_{\max}(A_h) \le c\, h^{d-2}$ folgt

$$|||x_\ell - \hat{x}_\ell^{\text{post},0}|||_1 \le c\, \lambda_{\max}^{-1/2}(A_h)\, |||x_\ell - x_\ell^{\text{pre},\nu_1}|||_2.$$

Die Definition von β und die Tatsache, dass die Energienorm des Fehlers durch die Grob-Gitter-Korrektur nicht vergrößert wird, liefern die Behauptung. □

Satz 3.14. *Es sei $\alpha \ge 1/\lambda_{\max}(A_h)$. Dann gilt für das Mehrgitterverfahren MGM_ℓ mit V- oder W-Zyklus*

$$|||x_\ell - x_\ell^{\text{neu}}|||_1 \le \rho_\ell\, |||x_\ell - x_\ell^{\text{alt}}|||_1, \quad \ell = 0, 1, 2, \ldots,$$

mit

$$\rho_\ell \le \rho_\infty := \sqrt{\frac{c^2}{c^2 + 2\nu_1}} \tag{3.7}$$

und der von ℓ und ν_1 unabhängigen Konstanten c aus Lemma 3.13.

Beweis. Wir zeigen zunächst die rekursive Abschätzung

$$\rho_\ell^2 \le \max_{0\le\beta\le 1} \beta^{2\nu_1}[\rho_{\ell-1}^{2\mu} + (1 - \rho_{\ell-1}^{2\mu})M], \quad M := \min\{1, c^2(1-\beta)\}, \tag{3.8}$$

mit der Konstanten c aus Lemma 3.13. Mit $\beta = \beta(x_\ell - x_\ell^{\text{pre},\nu_1})$ gilt nach Lemma 3.13 und Lemma 3.12

$$|||x_\ell - \hat{x}_\ell^{\text{post},0}|||_1^2 \le M\, |||x_\ell - x_\ell^{\text{pre},\nu_1}|||_1^2 \le \beta^{2\nu_1} M\, |||x_\ell - x_\ell^{\text{alt}}|||_1^2.$$

Diese Abschätzung in (3.6) eingesetzt ergibt (3.8).

Mit $\rho_0 := 0$ ist (3.7) für $\ell = 0$ wahr. Sei diese Abschätzung für $\ell - 1$ gezeigt. Dann folgt aus (3.8)

$$\begin{aligned}
\rho_\ell^2 &\le \max_{0\le\beta\le 1} \beta^{2\nu_1}[\rho_{\ell-1}^{2\mu}(1-M) + M] \le \max_{0\le\beta\le 1} \beta^{2\nu_1}[\frac{c^2}{c^2+2\nu_1}(1-M) + M] \\
&= \max_{0\le\beta\le 1} \beta^{2\nu_1}[\frac{c^2}{c^2+2\nu_1} + (1 - \frac{c^2}{c^2+2\nu_1})M] \\
&\le \max_{0\le\beta\le 1} \beta^{2\nu_1}[\frac{c^2}{c^2+2\nu_1} + (1 - \frac{c^2}{c^2+2\nu_1})c^2(1-\beta)] \\
&= \frac{c^2}{c^2+2\nu_1} \max_{0\le\beta\le 1} \beta^{2\nu_1}[1 + 2\nu_1(1-\beta)] = \frac{c^2}{c^2+2\nu_1}.
\end{aligned}$$

□

Bemerkung. Für alle $\nu_1 \ge 1$ erhält man $\rho_\infty < 1$. Die Kontraktionszahl sinkt dabei wie $\nu_1^{-1/2}$. Nutzt man auch $\nu_2 = \nu_1$ Nachglättungsschritte, so kann ein Verhalten wie ν_1^{-1} gezeigt werden.

Insgesamt benötigen wir also $k = \log_{\rho_\infty} \varepsilon$ Schritte des Mehrgitterverfahrens MGM_ℓ, um eine vorgegebene Genauigkeit $\varepsilon > 0$ bei der Lösung von $A_\ell x_\ell = b_\ell$ zu erreichen. Im Folgenden analysieren wir noch die Kosten pro Iterationsschritt.

3.3.3 Komplexität des Mehrgitterverfahrens

Für die Abschätzung der Komplexität des Mehrgitterverfahrens gehen wir davon aus, dass die Anwendung sowohl des Glätters S_ℓ, der Restriktion $I_\ell^{\ell-1}$ und der Prolongation $I_{\ell-1}^{\ell}$ in linearer Komplexität durchführbar ist. Daher ist der Aufwand in der Ebene ℓ durch

$$c(\nu_1 + \nu_2 + 1)n_\ell, \quad n_\ell := \dim V_\ell, \tag{3.9}$$

abschätzbar. Bei Zerlegungen im $\mathbb{R}^d$ dürfen wir davon ausgehen, dass sich die Anzahl der Unbekannten n_ℓ mit jeder Vergröberung um den Faktor 2^d reduziert. Die Summation der Ausdrücke (3.9) ergibt im Fall des V-Zyklus

$$\begin{aligned} c(\nu_1 + \nu_2 + 1)(n_\ell + n_{\ell-1} + n_{\ell-2} + \ldots) &= c(\nu_1 + \nu_2 + 1)n_\ell(1 + 2^{-d} + 2^{-2d} + \ldots) \\ &\leq \frac{2^d}{2^d - 1} c(\nu_1 + \nu_2 + 1)n_\ell \end{aligned}$$

und beim W-Zyklus

$$\begin{aligned} c(\nu_1 + \nu_2 + 1)(n_\ell + 2n_{\ell-1} + 4n_{\ell-2} + \ldots) &= c(\nu_1 + \nu_2 + 1)n_\ell(1 + 2^{-(d-1)} + 2^{-2(d-1)} + \ldots) \\ &\leq \frac{2^{d-1}}{2^{d-1} - 1} c(\nu_1 + \nu_2 + 1)n_\ell. \end{aligned}$$

Der Aufwand für einen Schritt des Mehrgitterverfahrens MGM_ℓ ist somit von linearer Ordnung n_ℓ. Dabei durften wir den Aufwand für die exakte Lösung der Gleichungssysteme auf dem gröbsten Gitter vernachlässigen, falls für die Anzahl der Stufen L gilt $L \sim |\log_2 h|$.

3.3.4 Geschachtelte Iteration

Da wir das Mehrgitterverfahren als Iterationsverfahren formuliert haben, benötigen wir gute Startwerte x_ℓ^{alt}, damit die Iteration zur Lösung von $A_\ell x_\ell = b_\ell$ möglichst schnell konvergiert. Motivation der folgenden **geschachtelten Iteration** NI_ℓ ist, dass Approximationen auf einem Gitter der Stufe $\ell - 1$ gute Startwerte für das ℓ-te Gitter darstellen.

1. Ist $\ell = 0$, so berechne $v_0 = x_0 = A_0^{-1} b_0$;
2. Im Fall $\ell > 0$ bestimme $v_{\ell-1}$ per geschachtelter Iteration $\mathrm{NI}_{\ell-1}$ für $A_{\ell-1}x_{\ell-1} = b_{\ell-1}$. Verwende $x_\ell^{\mathrm{alt}} := I_{\ell-1}^{\ell} v_{\ell-1}$ als Startwert für $m \geq 1$ Schritte des Mehrgitterverfahrens MGM_ℓ und setze $v_\ell := x_\ell^{\mathrm{neu}}$.

Bemerkung. Anstelle der m Schritte des Mehrgitterverfahrens MGM_ℓ in der geschachtelten Iteration kann natürlich auch das CG-Verfahren mit passender Vorkonditionierung verwendet werden.

Im folgenden Satz untersuchen wir die Genauigkeit des durch die geschachtelte Iteration NI_ℓ gewonnenen Startwerts v_ℓ. Danach reduziert die geschachtelte Iteration den Fehler $v_\ell - x_\ell$ auf die Größe des Diskretisierungsfehlers. Dies ist ausreichend für die Konvergenz der Finite-Elemente-Methode und vermeidet unnötige Iterationen.

Satz 3.15. *Für die Finite-Elemente-Approximation $u_h \in V_h$ gelte $\|u - u_h\|_{H^1(\Omega)} \leq c\, h^\alpha$ mit einem $\alpha > 0$. Ferner sei die Konvergenzrate ρ der m Mehrgitterschritte bzgl. der Norm $|||\cdot|||_1$ kleiner als $2^{-\alpha}$. Dann gilt*

$$|||v_\ell - x_\ell|||_1 \leq c' \frac{\rho}{1 - 2^\alpha \rho}\, h_\ell^\alpha.$$

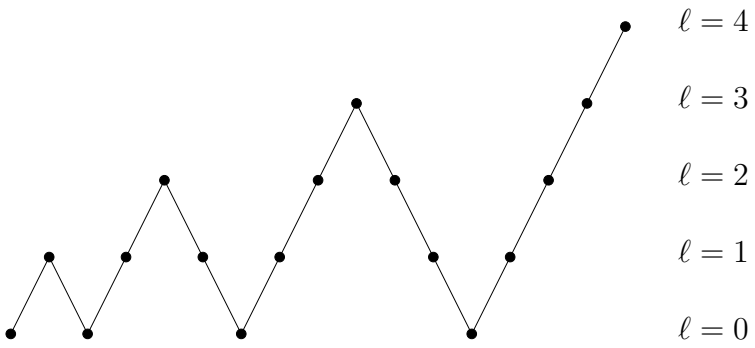


Abbildung 3.2: Geschachtelte Iteration bei V-Zyklus und fünf Gittern.

Beweis. Wir zeigen, dass

$$|||v_\ell - x_\ell|||_1 \leq (2^\alpha + 1)c_2 c \frac{\rho}{1 - 2^\alpha \rho} h_\ell^\alpha, \quad \ell \geq 0,$$

mit der Konstanten c_2 aus (3.2). Für $\ell = 0$ ist die Behautung wegen $v_0 = x_0$ klar. Sie sei wahr für $\ell - 1$. Wegen $h_{\ell-1} = 2h_\ell$ folgt aus der Vorraussetzung an den Diskretisierungsfehler

$$\|u - u_{\ell-1}\|_{H^1(\Omega)} \leq c\,(2h_\ell)^\alpha, \quad \|u - u_\ell\|_{H^1(\Omega)} \leq c\,h_\ell^\alpha.$$

Die Dreiecksungleichung liefert wegen $x_\ell^{\text{alt}} = I_{\ell-1}^\ell v_{\ell-1}$, $u_\ell = J_{h_\ell} x_\ell$ und $|||I_{\ell-1}^\ell x|||_1 = |||x|||_1$ nach (3.2)

$$\begin{aligned} |||x_\ell^{\text{alt}} - x_\ell|||_1 &\leq |||v_{\ell-1} - x_{\ell-1}|||_1 + c_2\|u_{\ell-1} - u\|_{H^1(\Omega)} + c_2\|u - u_\ell\|_{H^1(\Omega)} \\ &\leq (2^\alpha + 1)c_2 c \frac{\rho}{1 - 2^\alpha \rho} (2h_\ell)^\alpha + (2^\alpha + 1)c_2 c h_\ell^\alpha = \frac{(2^\alpha + 1)c_2 c}{1 - 2^\alpha \rho} h_\ell^\alpha. \end{aligned}$$

Das Mehrgitterverfahren verkleinert den Fehler um den Faktor ρ, d.h. es ist

$$|||x_\ell^{\text{neu}} - x_\ell|||_1 \leq \rho |||x_\ell^{\text{alt}} - x_\ell|||_1 \leq (2^\alpha + 1)c_2 c \frac{\rho}{1 - 2^\alpha \rho} h_\ell^\alpha.$$

□

Die mit Hilfe der geschachtelten Iteration berechnete Approximation v_ℓ kann wiederum als Startwert für weitere Schritte des Mehrgitterverfahrens verwendet werden. In diesem Fall kann die Abschätzung aus dem letzten Satz (falls nötig) mit Hilfe von Satz 3.14 verbessert werden.

Komplexität der geschachtelten Iteration

Wegen der linearen Komplexität des Mehrgitterverfahrens MGM_ℓ und des Ansteigens von n_ℓ von einer Ebenen zur nächsten um den Faktor 2^d gilt für den Aufwand der geschachtelten Iteration

$$c(n_\ell + n_{\ell-1} + n_{\ell-2} + \ldots) \leq c \frac{2^d}{2^d - 1} n_\ell.$$

4 Nichtkonforme und gemischte Methoden

4.1 Nichtkonforme Methoden

Bei einer Diskretisierung mittels konformer Elemente wurde dem kontinuierlichen Problem

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

durch Wahl eines endlichdimensionalen Unterraums $V_h \subset V$ ein diskretes Problem

$$u_h \in V_h: \quad a(u_h, v_h) = \ell(v_h) \quad \text{für alle } v_h \in V_h$$

zugeordnet. Beispielsweise in den folgenden Situationen stellt sich die konforme Diskretisierung aber als unzweckmäßig heraus:

- bei Verwendung einer Quadraturformel können die Bilinearform a und das Funktional ℓ nur näherungsweise bestimmt werden;
- bei Gleichungen höherer Ordnung kann die Wahl eines Funktionenraums $V_h \subset V$ kompliziert werden;
- krummlinige Ränder des Berechnungsgebietes Ω lassen im Allgemeinen keine exakte Darstellung der Randbedingungen zu.

Sind die Bilinearform a und das lineare Funktional ℓ etwa durch numerische Integration gestört, so ist (4.1) durch das endlichdimensionale Problem

$$u_h \in V_h: \quad a_h(u_h, v_h) = \ell_h(v_h) \quad \text{für alle } v_h \in V_h \tag{4.2}$$

ersetzt. Dabei sei $\ell_h \in V_h'$ und $a_h : V_h \times V_h \to \mathbb{R}$ auf dem Hilbert-Raum V_h stetig und **gleichgradig koerziv**, d.h. es existiert eine von h unabhängige Konstante $c_K > 0$ mit

$$a_h(v_h, v_h) \geq c_K \|v_h\|_{V_h}^2 \quad \text{für alle } v_h \in V_h.$$

Nach Satz 2.4 ist (4.2) eindeutig lösbar. Neben dem FE-Approximationsfehler entsteht bei (4.2) aber auch ein Konsistenzfehler. Den Einfluss dieser Störungen auf die Lösung beschreibt das folgende Lemma von Strang. Wir nehmen dazu zunächst $V_h \subset V$ und $\|\cdot\|_{V_h} = \|\cdot\|_V$ an.

Lemma 4.1 (erstes Lemma von Strang). *Es sei $u \in V$ eine Lösung von (4.1). Die Bilinearform $a_h : V_h \times V_h \to \mathbb{R}$ sei gleichgradig koerziv und $a : V \times V \to \mathbb{R}$ stetig. Dann gilt für die Lösung $u_h \in V_h \subset V$ von (4.2)*

$$\|u - u_h\|_V \le \inf_{v_h \in V_h} \left\{ \left(1 + \frac{c_S}{c_K}\right) \|u - v_h\|_V + c_K^{-1} \sup_{0 \neq w_h \in V_h} \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|_V} \right\} + c_K^{-1} \sup_{0 \neq w_h \in V_h} \frac{|\ell(w_h) - \ell_h(w_h)|}{\|w_h\|_V}.$$

Beweis. Für $v_h \in V_h$ und $w_h := u_h - v_h$ sehen wir wegen der gleichgradigen Koerzivität von a_h

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

Division durch $\|w_h\|_V$ liefert

$$\begin{aligned} c_K \|u_h - v_h\|_V &\le \sup_{0 \neq w_h \in V_h} \left\{ \frac{|a(u - v_h, w_h)|}{\|w_h\|_V} + \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|_V} + \frac{|\ell(w_h) - \ell_h(w_h)|}{\|w_h\|_V} \right\} \\ &\le c_S \|u - v_h\|_V + \sup_{0 \neq w_h \in V_h} \left\{ \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|_V} + \frac{|\ell(w_h) - \ell_h(w_h)|}{\|w_h\|_V} \right\}. \end{aligned}$$

Mit der Dreiecksungleichung $\|u - u_h\|_V \le \|u - v_h\|_V + \|u_h - v_h\|_V$ folgt die Behauptung. □

Beispiel 4.2 (Mass-Lumping). Es sei V_h der Raum der stetigen, stückweise linearen Funktionen $\mathcal{S}^{1,0}(\mathcal{T}_h)$. Wir wählen die Quadraturformel

$$Q(f) = \sum_{k=1}^{n} \mu(D_k) f(p_k),$$

wobei p_k die Eckpunkte der Dreiecke $\tau \in \mathcal{T}_h$ und D_k eine Umgebung von p_k bezeichnen, so dass $\{D_k,\ k = 1, \dots, n\}$ eine Zerlegung von $\Omega \subset \mathbb{R}^d$ bildet, sog. *duale Vernetzung*.

Indem man $\ell(v) = \int_\Omega f v \, \mathrm{d}x$ durch die Quadraturformel $\ell_h(v) := Q(fv)$ ersetzt, erhält man bei Lipschitz-stetigem f

$$|\ell(v_h) - \ell_h(v_h)| \le c h^{d+1} \sum_{\tau \in \mathcal{T}_h} \|\nabla v_h\|_\tau.$$

Diese spezielle Substitution der Funktionale ℓ durch Näherungsformeln ℓ_h wird als **mass-lumping** bezeichnet. Sie eignet sich besonders zur Diskretisierung von Randwertproblemen der Form

$$\begin{aligned} -\Delta u + c(x) u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{auf } \partial\Omega. \end{aligned}$$

Die zugehörige Bilinearform

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x + \int_\Omega cuv \, \mathrm{d}x$$

wird dann durch

$$a_h(u_h, v_h) = \int_\Omega \nabla u_h \cdot \nabla v_h \,\mathrm{d}x + \sum_{k=1}^{n} c(p_k)\mu(D_k)u_h(p_k)v_h(p_k)$$

ersetzt. Wird a_h auf die Lagrange-Funktionen φ_i, φ_j angewendet, so erhält man

$$a_h(\varphi_j, \varphi_i) = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i \,\mathrm{d}x + c(p_i)\mu(D_i)\delta_{ij}.$$

Bei Verwendung dieser Technik nimmt also insbesondere die Masse-Matrix $M \in \mathbb{R}^{n\times n}$ mit Einträgen $m_{ij} = (\varphi_j, \varphi_i)_{L^2}$ Diagonalgestalt an.

Finite-Element-Methoden, für die $V_h \not\subset V$ gilt, werden als **nichtkonform** bezeichnet. Im Gegensatz zu konformen Methoden gilt im Allgemeinen nicht, dass $\|\cdot\|_V$ auf V_h definiert ist. Daher verwendet man gitterabhängige Normen $\|\cdot\|_{V_h}$, die auf $V_h + V$ definiert seien; vgl. Satz 2.64. Ebenso setzen wir voraus, dass die Bilinearform a_h für Funktionen in $V_h + V$ erklärt ist.

Die Differenz $u - u_h$ der Lösung u von (4.1) und der Lösung $u_h \in V_h$ von (4.2) wird im folgenden zum Céa-Lemma analogen Lemma in der Norm $\|\cdot\|_{V_h}$ des Raums V_h abgeschätzt.

Lemma 4.3 (zweites Lemma von Strang). *Sei a_h gleichgradig koerziv und stetig. Dann gilt*

$$\|u - u_h\|_{V_h} \le \left(1 + \frac{c_S}{c_K}\right) \inf_{v_h\in V_h} \|u - v_h\|_{V_h} + c_K^{-1} \sup_{0\neq w_h\in V_h} \frac{|a_h(u, w_h) - \ell_h(w_h)|}{\|w_h\|_{V_h}}.$$

Beweis. Sei $v_h \in V_h$. Aus der gleichgradigen Koerzivität folgt

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

Mit der Bezeichnung $w_h := u_h - v_h$ erhalten wir nach Division durch $\|w_h\|_{V_h}$

$$c_K\|u_h - v_h\|_{V_h} \le c_S\|u - v_h\|_{V_h} + \frac{|a_h(u, w_h) - \ell_h(w_h)|}{\|w_h\|_{V_h}}.$$

Wie im ersten Lemma erhält man die Behauptung mittels Dreiecksungleichung. □

Bei nichtkonformer Diskretisierung erhält man im Aubin-Nitsche-Lemma zwei zusätzliche Terme.

Lemma 4.4 (Aubin-Nitsche). *Seien die Hilbert-Räume W und V wie in Lemma 2.45. Ferner gelte $V_h \subset W$ und die auf $V + V_h$ definierte stetige Bilinearform a_h stimme mit a auf $V \times V$ überein. Dann gilt*

$$\begin{aligned} \|u - u_h\|_W \le \sup_{0\neq\varphi\in W} \frac{1}{\|\varphi\|_W} \Big\{ & c_S\|u - u_h\|_{V_h}\|u_\varphi - u_{\varphi,h}\|_{V_h} + |a_h(u - u_h, u_\varphi) - (u - u_h, \varphi)_W| \\ & + |a_h(u, u_\varphi - u_{\varphi,h}) - \ell(u_\varphi) + \ell_h(u_{\varphi,h})| \Big\}. \end{aligned}$$

Dabei bezeichnen $u_\varphi \in V$ und $u_{\varphi,h} \in V_h$ für jedes $\varphi \in W$ die Lösungen der dualen Probleme $a(w, u_\varphi) = (w, \varphi)_W$, $w \in V$, bzw. $a_h(w, u_{\varphi,h}) = (w, \varphi)_W$, $w \in V_h$.

Beweis. Wegen der Einbettung von V in W und $V_h \subset W$ gilt $u - u_h \in W$. Die Norm von $u - u_h$ läßt sich mittels der Dualnorm

$$\|u - u_h\|_W = \sup_{0 \neq \varphi \in W} \frac{(u - u_h, \varphi)_W}{\|\varphi\|_W}$$

ausdrücken. Nach Definition von u_h, u_φ und $u_{\varphi,h}$ gilt für jedes $\varphi \in W$

$$\begin{aligned}(u - u_h, \varphi)_W &= a_h(u, u_\varphi) - a_h(u_h, u_{\varphi,h}) \\ &= a_h(u - u_h, u_\varphi - u_{\varphi,h}) + a_h(u_h, u_\varphi - u_{\varphi,h}) + a_h(u - u_h, u_{\varphi,h}).\end{aligned}$$

Mit

$$\begin{aligned}a_h(u_h, u_\varphi - u_{\varphi,h}) &= a_h(u_h - u, u_\varphi) + a_h(u, u_\varphi) - a_h(u_h, u_{\varphi,h}) \\ &= a_h(u_h - u, u_\varphi) + (u - u_h, \varphi)_W\end{aligned}$$

und

$$\begin{aligned}a_h(u - u_h, u_{\varphi,h}) &= a_h(u, u_{\varphi,h} - u_\varphi) + a_h(u, u_\varphi) - a_h(u_h, u_{\varphi,h}) \\ &= a_h(u, u_{\varphi,h} - u_\varphi) + \ell(u_\varphi) - \ell_h(u_{\varphi,h})\end{aligned}$$

folgt

$$\begin{aligned}(u - u_h, \varphi)_W = a_h(u - u_h, u_\varphi - u_{\varphi,h}) &- \{a_h(u - u_h, u_\varphi) - (u - u_h, \varphi)_W\} \\ &- \{a_h(u, u_\varphi - u_{\varphi,h}) - \ell(u_\varphi) + \ell_h(u_{\varphi,h})\}.\end{aligned}$$

Aus der Stetigkeit von a_h folgt die Behauptung. □

4.1.1 Das Crouzeix-Raviart-Element

Wir wenden diese Ergebnisse auf das einfachste nichtkonforme Element, das **Crouzeix-Raviart-Element**, im Zusammenhang mit der Lösung der Poisson-Gleichung mit homogenen Dirichlet-Randbedingungen an. Wir gehen von einem polygonalen Gebiet $\Omega \subset \mathbb{R}^2$ aus. Der Ansatzraum ist dann

$$\begin{aligned}V_h = \{v \in L^2(\Omega) : v|_\tau \text{ ist linear für jedes } \tau \in \mathcal{T}_h, \\ v \text{ ist stetig in den Kantenmittelpunkten}\}.\end{aligned}$$

Bei Nullrandbedingungen wird zusätzlich gefordert, dass $v = 0$ in den Mittelpunkten der Kanten auf $\partial\Omega$ gilt. Jedes $v \in V_h$ ist eindeutig bestimmt durch seine Werte in den Mit-

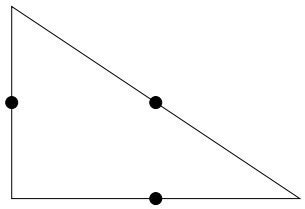


Abbildung 4.1: Crouzeix-Raviart-Element

telpunkten m_1, m_2, m_3 der Kanten. Bezeichnen z_1, z_2, z_3 die Eckpunkte von $\tau \in \mathcal{T}_h$ und

$\varphi_1, \varphi_2, \varphi_3$ die zugehörigen nodalen Basisfunktionen, so überzeugt man sich leicht davon, dass

$$v|_\tau = \sum_{i=1}^{3} v(m_i)\,\psi_i,$$

wobei $\psi_i := \varphi_{i+1} + \varphi_{i+2} - \varphi_i$. Dabei nehmen wir an, dass die i-te Kante dem i-ten Eckpunkt gegenüber liegt und alle Knotenindizes modulo 3 zu verstehen sind.

Die Funktionen $v \in V_h$ sind aber nicht stetig. Es gilt zwar $V_h \subset L^2(\Omega)$ aber $V_h \not\subset H^1(\Omega)$. Daher definieren wir

$$a_h(u,v) = \sum_{\tau\in\mathcal{T}_h} \int_\tau \nabla u \cdot \nabla v \,\mathrm{d}x, \quad \ell_h(v) = \int_\Omega f v \,\mathrm{d}x \quad \text{für } u, v \in H_0^1(\Omega) + V_h$$

und

$$\|v\|_{V_h} := \left(\sum_{\tau\in\mathcal{T}_h} |v|^2_{H^1(\tau)} \right)^{1/2}.$$

Im Hinblick auf das zweite Lemma von Strang geben wir die folgende Aussage an.

Lemma 4.5. *Sei $u \in H_0^1(\Omega) \cap H^2(\Omega)$ die Lösung von (4.1) und $L_u(w) := a_h(u,w) - \ell_h(w)$. Ist $\mathcal{T}_h$ nicht-entartet, so gilt für alle $w \in V_h$*

$$|L_u(w)| \le c\,h\,|u|_{H^2(\Omega)}\,\|w\|_{V_h}.$$

Beweis. Sei $w \in V_h$. Dann gilt

$$\begin{aligned} L_u(w) = a_h(u,w) - \ell_h(w) &= \sum_{\tau\in\mathcal{T}_h} \int_\tau \nabla u \cdot \nabla w \,\mathrm{d}x - \int_\Omega f w \,\mathrm{d}x \\ &= \sum_{\tau\in\mathcal{T}_h} \left(\int_{\partial\tau} w\,\partial_\nu u \,\mathrm{d}s - \int_\tau w \Delta u \,\mathrm{d}x \right) - \int_\Omega f w \,\mathrm{d}x \\ &= \sum_{\tau\in\mathcal{T}_h} \int_{\partial\tau} w\,\partial_\nu u \,\mathrm{d}s - \int_\Omega (\Delta u + f) w \,\mathrm{d}x = \sum_{\tau\in\mathcal{T}_h} \int_{\partial\tau} w\,\partial_\nu u \,\mathrm{d}s. \end{aligned}$$

Da $\partial_\nu u$ nur das Vorzeichen wechselt, je nachdem von welchem Element man eine innere Kante betrachtet, gilt $\sum_{\tau\in\mathcal{T}_h} \int_\tau \partial_\nu u \,\mathrm{d}s = 0$ und somit

$$L_u(w) = \sum_{\tau\in\mathcal{T}_h} \sum_{e\in\partial\tau} \int_e (w - w_e)\,\partial_\nu u \,\mathrm{d}s.$$

Dabei ist w zwar nicht über e stetig, es gilt aber für $e = \tau_1 \cap \tau_2$

$$w_e := \frac{1}{|e|} \int_e w|_{\tau_1} \,\mathrm{d}s = \frac{1}{|e|} \int_e w|_{\tau_2} \,\mathrm{d}s.$$

Da für den nodalen Interpolanten $\mathfrak{I}u \in V_h$ gilt, dass $\partial_\nu \mathfrak{I}u$ auf jeder Kante e konstant ist, gilt insbesondere

$$\int_e (w - w_e) \partial_\nu \mathfrak{I}u \,\mathrm{d}s = 0$$

und somit nach der Cauchy-Schwarzschen Ungleichung

$$\begin{aligned}|L_u(w)| &= |\sum_{\tau\in\mathcal{T}_h}\sum_{e\in\partial\tau}\int_e (w-w_e)\,\partial_\nu(u-\mathfrak{I}u)\,\mathrm{d}s| \\ &\leq \sum_{\tau\in\mathcal{T}_h}\sum_{e\in\partial\tau}\|w-w_e\|_{L^2(e)}\|\nabla(u-\mathfrak{I}u)\|_{L^2(e)}.\end{aligned}$$

Für $u\in H^2(\hat\tau)$ erhalten wir nach dem Spursatz 2.22 und Lemma 2.61 auf dem Referenzelement $\hat\tau$

$$\|\nabla(u-\mathfrak{I}u)\|_{L^2(\partial\hat\tau)} \leq c\|u-\mathfrak{I}u\|_{H^2(\hat\tau)} \leq c'|u|_{H^2(\hat\tau)}.$$

Mit der Transformationsformel folgt daraus für $\tau\in\mathcal{T}_h$

$$\|\nabla(u-\mathfrak{I}u)\|_{L^2(\partial\tau)} \leq ch^{1/2}|u|_{H^2(\tau)}.$$

Ebenso folgt mit dem Bramble-Hilbert-Lemma für jede Kante $e\in\partial\hat\tau$ und $w\in\Pi_1$

$$\|w-w_e\|_{L^2(e)} \leq c|w|_{H^1(\hat\tau)},$$

weil die linke Seite für konstante Funktionen verschwindet, und daraus mit der Transformationsformel für $e\in\partial\tau$ und $w\in V_h$

$$\|w-w_e\|_{L^2(e)} \leq ch^{1/2}|w|_{H^1(\tau)}.$$

Hieraus erhält man mit Hilfe der Cauchy-Schwarzschen Ungleichung für euklidische Skalarprodukte

$$\begin{aligned}|L_u(w)| &\leq \sum_{\tau\in\mathcal{T}_h} ch|u|_{H^2(\tau)}|w|_{H^1(\tau)} \leq ch\left(\sum_{\tau\in\mathcal{T}_h}|u|^2_{H^2(\tau)}\right)^{1/2}\left(\sum_{\tau\in\mathcal{T}_h}|w|^2_{H^1(\tau)}\right)^{1/2} \\ &\leq ch|u|_{H^2(\Omega)}\|w\|_{V_h}.\end{aligned}$$

□

Bemerkung. Inspiziert man den Beweis des letzten Lemmas, so erkennt man, dass die Aussage auch für $w\in V_h+H^1_0(\Omega)$ gilt.

Beachtet man nun, dass die stückweise linearen konformen Elemente in V_h enthalten sind, liefert Satz 2.64 und das zweite Lemma von Strang

$$\|u-u_h\|_{V_h} \leq c\,h\,|u|_{H^2(\Omega)} \leq c\,h\,\|f\|_{L^2(\Omega)}. \tag{4.3}$$

Um den Fehler in der L^2-Norm zu messen, wenden wir das Aubin-Nitsche Lemma 4.4 an.

Satz 4.6. *Sei $\mathcal{T}_h$ ein Familie nicht-entarteter Triangulierungen von $\Omega\subset\mathbb{R}^2$. Ferner seien $u\in H^2(\Omega)$ und $f\in L^2(\Omega)$. Dann gilt für den Diskretisierungsfehler des Poisson-Problems bei Crouzeix-Raviart-Elementen*

$$\|u-u_h\|_{L^2(\Omega)} + h\|u-u_h\|_{V_h} \leq c\,h^2|u|_{H^2(\Omega)}.$$

Beweis. Um das Aubin-Nitsche Lemma anzuwenden, setzen wir $V = H_0^1(\Omega)$ und $W = L^2(\Omega)$. Die Differenz $u_\varphi - u_{\varphi,h}$ der Lösungen der dualen Probleme schätzen wir mit (4.3) ab:

$$\|u_\varphi - u_{\varphi,h}\|_{V_h} \le ch\|\varphi\|_{L^2(\Omega)}.$$

Ferner folgt für die beiden Zusatzterme im Aubin-Nitsche Lemma

$$\begin{aligned}|a_h(u-u_h,u_\varphi) - (u-u_h,\varphi)_W| = |L_{u_\varphi}(u-u_h)| &\le ch|u_\varphi|_{H^2(\Omega)}\|u-u_h\|_{V_h}\\ &\le c'h\|\varphi\|_{L^2(\Omega)}\|u-u_h\|_{V_h}.\end{aligned}$$

und

$$|a_h(u,u_\varphi - u_{\varphi,h}) - (f,u_\varphi - u_{\varphi,h})_W| = |L_u(u_\varphi - u_{\varphi,h})| \le ch|u|_{H^2(\Omega)}\|u_\varphi - u_{\varphi,h}\|_{V_h}.$$

Daher erhalten wir

$$\|u-u_h\|_{L^2(\Omega)} \le ch\|u-u_h\|_{V_h} + c'h^2|u|_{H^2(\Omega)}$$

und mit (4.3) die Behauptung. □

4.1.2 Polygonale Approximation krummliniger Ränder

Bisher haben wir immer angenommen, dass $\Omega \subset \mathbb{R}^2$ polygonal ist. Bei nicht-polygonalen Gebieten Ω benötigt man krummlinige Elemente in der Zerlegung $\mathcal{T}_h$ von Ω.

Im Folgenden beschränken wir uns auf lineare Dreieckselemente

$$\begin{aligned}V_h := \{v \in C(\Omega) : v|_\tau \text{ ist linear für jedes } \tau \in \mathcal{T}_h,\\ v(x) = 0 \text{ für jeden Knoten } x \in \partial\Omega \text{ der Zerlegung}\}\end{aligned}$$

bei Gebieten $\Omega \subset \mathbb{R}^2$. Dann gilt $V_h \not\subset H_0^1(\Omega)$. Wegen $V_h \subset H^1(\Omega)$ benötigt man aber keine gitterabhängigen Normen und kann $a_h = a$ und $\ell_h = h$ setzen.

Lemma 4.7. *Sei Ω ein C^2-Gebiet und $\mathcal{T}_h$ eine Familie nicht-entarteter Zerlegungen. Dann gilt*

$$\|v_h\|_{L^2(\partial\Omega)} \le c\,h^{3/2}|v_h|_{H^1(\Omega)} \quad \textit{für alle } v_h \in V_h.$$

Beweis. Sei $\tau \in \mathcal{T}_h$ ein Element mit krummlinigem Rand $\Gamma_\tau := \tau \cap \partial\Omega$. Wir zeigen

$$\int_{\partial\tau} v_h^2 \,\mathrm{d}s \le ch_\tau^3 \int_\tau |\nabla v_h|^2 \,\mathrm{d}x.$$

Die Behauptung folgt hieraus durch Summation über alle Dreiecke $\tau \in \mathcal{T}_h$.

Wir dürfen annehmen, dass die beiden Knoten in Γ_τ auf der x_1-Achse liegen. Die Koordinaten dieser beiden Punkte seien $(\xi,0)$ und $(\xi',0)$ und der Rand Γ_τ werde durch $x_2 = \phi(x_1)$ beschrieben. Wegen $\phi(\xi) = \phi(\xi') = 0$, $|\xi - \xi_1'| \le h_\tau$ gilt für $\xi \le x_1 \le \xi'$

$$|\phi(x_1)| \le c\,h_\tau^2, \quad c := \max_{\xi \le x_1 \le \xi'} |\phi''(x_1)|. \tag{4.4}$$

Da $v_h \in V_h$ in τ linear ist und somit entlang der x_1-Achse verschwindet, gilt $v_h(x_1,x_2) = bx_2$ und damit $|\nabla v_h| = b$. Wegen $|\phi'| \le c$ folgt

$$\int_{\Gamma_\tau} v_h^2 \,\mathrm{d}s = \int_\xi^{\xi'} |b\phi(x_1)|^2 \sqrt{1+|\phi'(x_1)|^2}\,\mathrm{d}x_1 \le cb^2h_\tau^4 \int_\xi^{\xi'} 1\,\mathrm{d}x_1 = cb^2h_\tau^5.$$

Da die Fläche von τ nach unten durch die des Innenkreises abgeschätzt werden kann, gilt

$$\pi \frac{h_\tau^2}{c_E^2} b^2 \leq \int_\tau |\nabla v_h|^2 \, \mathrm{d}x.$$

Der Vergleich der beiden letzten Abschätzungen ergibt die Behauptung. □

Sei $u \in H_0^1(\Omega)$ die Lösung des Poisson-Problems und $u_h \in V_h$ die zugehörige schwache Lösung, d.h.

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

Wir schätzen den Diskretisierungsfehler in der H^1-Norm ab.

Satz 4.8. *Sei Ω ein C^2-Gebiet und $f \in L^2(\Omega)$. Dann gilt bei nicht-entarteter Zerlegung $\mathcal{T}_h$*

$$\|u - u_h\|_{H^1(\Omega)} \leq c\, h\, \|u\|_{H^2(\Omega)} \leq c'\, h\, \|f\|_{L^2(\Omega)}.$$

Beweis. Wegen der Glattheit des Randes ist $u \in H^2(\Omega) \cap H_0^1(\Omega)$. Mit partieller Integration ergibt sich für $v_h \in V_h \subset H^1(\Omega)$

$$(f, v_h)_{L^2(\Omega)} = (-\Delta u, v_h)_{L^2(\Omega)} = a(u, v_h) - \int_{\partial\Omega} v_h \partial_\nu u \, \mathrm{d}s.$$

Mit der Cauchy-Schwarzschen Ungleichung, dem Spursatz und Lemma 4.7 folgt

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

Die Behauptung folgt aus dem zweiten Lemma von Strang. Dabei ist zu beachten, dass die erhöhte Konvergenzordnung der letzten Abschätzung vom Approximationsfehlers

$$\inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)} \sim h$$

dominiert wird. □

Ersetzt man die krummlinigen Kanten in $\mathcal{T}_h$ durch Sehnen, so erhält man polygonale Approximationen Ω_h an das ursprüngliche Gebiet Ω. Man beachte, dass dabei wegen (4.4) nur eine Fläche $\tau' := \tau \cap (\Omega \setminus \Omega_h)$ mit

$$\mu(\tau') \leq c\, h\, \mu(\tau) \tag{4.5}$$

abgeschnitten wird. Daher bleibt die Abschätzung aus dem letzten Satz richtig, wenn a durch

$$a_h(u, v) := \int_{\Omega_h} \nabla u \cdot \nabla v \, \mathrm{d}x$$

ersetzt wird. Es ist nämlich $|a_h(u, v_h) - a(u, v_h)| \leq \|u\|_{H^1(\Omega)} \|v_h\|_{H^1(\Omega \setminus \Omega_h)}$, und weil ∇v_h für $v_h \in V_h$ auf jedem Element τ konstant ist, folgt aus (4.5)

$$\|v_h\|_{H^1(\Omega \setminus \Omega_h)} \leq c\, h\, \|v_h\|_{H^1(\Omega)}.$$

Da Lemma 4.7 nicht für $v_h \in V_h + H_0^1(\Omega)$ gilt, können wir das Aubin-Nitsche-Lemma 4.4 nicht anwenden, um Abschätzungen für die L^2-Norm zu beweisen. Wir müssen daher im Fall der Randapproximation einen gesonderten Beweis führen.

Satz 4.9. *Unter den Voraussetzungen von Satz 4.8 gilt* $\|u - u_h\|_{L^2(\Omega)} \le c\, h^{3/2}\, \|u\|_{H^2(\Omega)}$.

Beweis. Zu $w := u - u_h$ sei φ die Lösung von

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

Weil Ω glatt ist, ist $\varphi \in H^2(\Omega) \cap H_0^1(\Omega)$ und $\|\varphi\|_{H^2(\Omega)} \le c\|w\|_{L^2(\Omega)}$. Im Gegensatz zu Umformungen bei konformen Elementen erhalten wir wegen $w \notin H_0^1(\Omega)$ aus der Greenschen Formel Randterme

$$\|w\|^2_{L^2(\Omega)} = (w, -\Delta\varphi)_{L^2(\Omega)} = a(w,\varphi) - \int_{\partial\Omega} w\, \partial_\nu\varphi \,\mathrm{d}s.$$

Da für beliebige $v_h \in V_h$ wegen $\varphi \in H_0^1(\Omega)$ gilt

$$\begin{aligned} a(w, v_h) = a(u, v_h) - a(u_h, v_h) &= (-\Delta u, v_h)_{L^2(\Omega)} + \int_{\partial\Omega} v_h\, \partial_\nu u \,\mathrm{d}s - (f, v_h)_{L^2(\Omega)} \\ &= \int_{\partial\Omega} v_h\, \partial_\nu u \,\mathrm{d}s = \int_{\partial\Omega} (\varphi - v_h)\, \partial_\nu u \,\mathrm{d}s, \end{aligned}$$

folgt

$$\|w\|^2_{L^2(\Omega)} = a(w, \varphi - v_h) - \int_{\partial\Omega} (\varphi - v_h)\, \partial_\nu u \,\mathrm{d}s - \int_{\partial\Omega} w\, \partial_\nu\varphi \,\mathrm{d}s. \tag{4.6}$$

Für die Wahl $v_h = \mathfrak{I}_h\varphi$ erhalten wir nach Satz 4.8

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

Um den zweiten Term in (4.6) abzuschätzen, benötigen wir noch die Approximationsaussage $\|\varphi - \mathfrak{I}_h\varphi\|_{L^2(\partial\Omega)} \le c\, h^{3/2} \|\varphi\|_{H^2(\Omega)}$, die mit dem Spursatz und der Transformationsformel gezeigt werden kann. Die Anwendung des Spursatzes auf ∇u liefert

$$\begin{aligned} |(\partial_\nu u, \varphi - v_h)_{L^2(\partial\Omega)}| &\le \|\nabla u\|_{L^2(\partial\Omega)} \|\varphi - v_h\|_{L^2(\partial\Omega)} \le c\, h^{3/2} \|u\|_{H^2(\Omega)} \|\varphi\|_{H^2(\Omega)} \\ &\le c'\, h^{3/2} \|u\|_{H^2(\Omega)} \|w\|_{L^2(\Omega)}. \end{aligned}$$

Den letzten Term in (4.6) schätzen wir mit Hilfe von Lemma 4.7, Satz 4.8 und dem Spursatz ab:

$$\begin{aligned} |(w, \partial_\nu\varphi)_{L^2(\partial\Omega)}| &\le \|u - u_h\|_{L^2(\partial\Omega)} \|\nabla\varphi\|_{L^2(\partial\Omega)} = \|u_h\|_{L^2(\partial\Omega)} \|\nabla\varphi\|_{L^2(\partial\Omega)} \\ &\le c\, h^{3/2} \|u_h\|_{H^1(\Omega)} \|\varphi\|_{H^2(\Omega)} \le c\, h^{3/2} \{\|u\|_{H^1(\Omega)} + \|u - u_h\|_{H^1(\Omega)}\} \|w\|_{L^2(\Omega)} \\ &\le c' h^{3/2} \|u\|_{H^2(\Omega)} \|w\|_{L^2(\Omega)}. \end{aligned}$$

Zusammen haben wir

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

woraus die Behauptung folgt. □

4.2 Gemischte Finite Elemente

Wir betrachten exemplarisch die Poissonsche Differentialgleichung mit homogenen Dirichlet-Randbedingungen

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

Dieses Problem kann als einfaches Modell zur Beschreibung der Auslenkung einer eingespannten Membran bei äußerer Last f interpretiert werden. Die bisher betrachteten numerischen Methoden liefern Approximationen u_h an die Auslenkung u. Für technische Anwendungen sind aber die Spannungskräfte ∇u mindestens genauso wichtig. Zwar können wir Approximationen an die Spannung durch Differentiation der Approximation u_h gewinnen, diese wird aber i.A. um eine h-Potenz schlechter approximiert werden. Bei den sog. **gemischten Finite-Elemente-Methoden** führt man $\sigma := \nabla u \in \mathbb{R}^d$ als neue Variable ein, d.h. man betrachtet das System erster Ordnung für $[\sigma, u]$

$$\begin{aligned} -\operatorname{div}\sigma &= f \text{ in } \Omega, \\ -\nabla u + \sigma &= 0 \text{ in } \Omega, \\ u &= 0 \text{ auf } \partial\Omega. \end{aligned}$$

Die direkte Approximation erlaubt eine höhere Genauigkeit als der indirekte Zugang über Differentiation von u_h. Um hierzu eine Variationsformulierung aufzustellen, integrieren wir die erste Gleichung partiell

$$(\sigma, \nabla v)_{L^2(\Omega)} = (f, v)_{L^2(\Omega)} \quad \text{für alle } v \in H^1_0(\Omega), \tag{4.7a}$$

$$-(\nabla u, \tau)_{L^2(\Omega)} + (\sigma, \tau)_{L^2(\Omega)} = 0 \quad \text{für alle } \tau \in [L^2(\Omega)]^d. \tag{4.7b}$$

Die Lösung suchen wir im Produktraum $X := [L^2(\Omega)]^d \times H^1_0(\Omega)$, der mit der Norm

$$\|[\sigma, u]\|_X := \|\sigma\|_{L^2(\Omega)} + |u|_{H^1(\Omega)}$$

ausgestattet sei. Äquivalent zu (4.7) haben wir das Variationsproblem

$$[\sigma, u] \in X: \quad A([\sigma, u], [\tau, v]) = \ell([\tau, v]), \quad [\tau, v] \in X,$$

mit der Bilinearform $A : X \times X \to \mathbb{R}$

$$A([\sigma, u], [\tau, v]) := (\sigma, \nabla v)_{L^2(\Omega)} - (\nabla u, \tau)_{L^2(\Omega)} + (\sigma, \tau)_{L^2(\Omega)}$$

und der Linearform $\ell : X \to \mathbb{R}$ definiert durch $\ell([\tau, v]) = (f, v)_{L^2(\Omega)}$. Wegen

$$\begin{aligned} |A([\sigma, u], [\tau, v])| &\le \|\sigma\|_{L^2(\Omega)}|v|_{H^1(\Omega)} + |u|_{H^1(\Omega)}\|\tau\|_{L^2(\Omega)} + \|\sigma\|_{L^2(\Omega)}\|\tau\|_{L^2(\Omega)} \\ &\le \{\|\sigma\|_{L^2(\Omega)} + |u|_{H^1(\Omega)}\}\{\|\tau\|_{L^2(\Omega)} + |v|_{H^1(\Omega)}\} \\ &= \|[\sigma, u]\|_X \|[\tau, v]\|_X \end{aligned}$$

ist A stetig. Die Bilinearform ist jedoch nicht koerziv, weil

$$A([\tau, v], [\tau, v]) = (\tau, \nabla v)_{L^2(\Omega)} - (\nabla v, \tau)_{L^2(\Omega)} + \|\tau\|^2_{L^2(\Omega)} = \|\tau\|^2_{L^2(\Omega)}$$

nicht von v abhängt. Daher lassen sich die bisher entwickelten Techniken nicht anwenden.

4.2.1 Die inf-sup Bedingung

Im Folgenden stellen wir eine Lösbarkeitstheorie für allgemeine Variationsprobleme

$$u \in V: \quad a(u,w) = \ell(w), \quad w \in W, \tag{4.8}$$

mit der Bilinearform $a : V \times W \to \mathbb{R}$ und den Hilbert-Räumen V, W vor. Dazu verwenden wir den aus der Funktionalanalysis bekannten Satz vom abgeschlossenen Bild (engl. *closed range theorem*).

Sei dazu $T : X \to Y$ ein linearer Operator zwischen den Banach-Räumen X und Y. Dann ist der zu T **adjungierte Operator** $T' : Y' \to X'$ definiert durch

$$(T'\varphi)(x) = \varphi(Tx), \quad x \in X \text{ und } \varphi \in Y'.$$

Ferner definiert man die **Polare** des abgeschlossenen Unterraums $U \subset Y'$

$$U^0 := \{y \in Y : \varphi(y) = 0 \text{ für alle } \varphi \in U\}.$$

Satz 4.10 (Satz vom abgeschlossenen Bild). *Sei $T : X \to Y$ linear und stetig zwischen den Banach-Räumen X und Y. Dann sind folgende Aussagen äquivalent:*

(i) Das Bild $\operatorname{Ran} T$ *ist abgeschlossen in Y.*

(ii) Es gilt $\operatorname{Ran} T = (\operatorname{Ker} T')^0$.

Mit Hilfe dieser Aussage können wir die Lösbarkeit von (4.8) untersuchen. Dazu definieren wir durch

$$(Av)(w) = a(v,w), \quad v \in V,\ w \in W,$$

einen linearen Operator $A : V \to W'$. Für den zu A adjungierten Operator A' gilt dann $(A'w)(v) = a(v,w)$, $v \in V$, $w \in W$.

Satz 4.11. *Das Problem (4.8) besitzt für jedes $\ell \in W'$ eine eindeutige Lösung $u \in V$, falls die folgenden Bedingungen erfüllt sind:*

(i) Die Bilinearform a ist stetig, d.h. $|a(v,w)| \le c_S \|v\|_V \|w\|_W$ für alle $v \in V$, $w \in W$;

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

$$\inf_{0 \neq v \in V} \sup_{0 \neq w \in W} \frac{a(v,w)}{\|v\|_V \|w\|_W} \ge c_E > 0; \tag{4.9}$$

(iii) Zu jedem $0 \neq w \in W$ existiert ein $v \in V$ mit $a(v,w) \neq 0$.

In diesem Fall ist $\operatorname{Ran} A = W'$ *und für die Lösung u gilt $\|u\|_V \le c_E^{-1} \|\ell\|_{W'}$.*

Beweis. Die Stetigkeit von a überträgt sich auf A. Ferner ist A injektiv, weil aus $Av_1 = Av_2$ folgt $\sup_{0 \neq w \in W} a(v_1 - v_2, w) = 0$ und nach (4.9) $v_1 - v_2 = 0$. Zu jedem $\varphi \in \operatorname{Ran} A$ existiert

dann ein eindeutiges Inverses $A^{-1}\varphi$. Nach (4.9) folgt $\|Av\|_{W'} \geq c_E\|v\|_V$ für alle $v \in V$ und hiermit die Stetigkeit von A^{-1} auf $\operatorname{Ran} A$:

$$\|A^{-1}\varphi\|_V \leq \frac{1}{c_E}\|\varphi\|_{W'} \quad \text{für alle } \varphi \in \operatorname{Ran} A.$$

Wegen der Stetigkeit von A und A^{-1} ist $\operatorname{Ran} A$ abgeschlossen in W'. Aus Satz 4.10 folgt $\operatorname{Ran} A = (\operatorname{Ker} A')^0$ und nach Voraussetzung ist wegen der Reflexivität[1] von Hilbert-Räumen

$$\operatorname{Ker} A' = \{w \in W : a(v,w) = 0 \text{ für alle } v \in V\} = \{0\}.$$

Daher ist $\operatorname{Ran} A = \{0\}^0 = W'$, was die Surjektivität von A zeigt. Für beliebiges $\ell \in W'$ erhalten wir daher die Lösung $u := A^{-1}\ell$. Diese ist wegen der Injektivität von A eindeutig. Die Stabilitätsaussage folgt aus der Stetigkeit von A^{-1}. □

Die inf-sup-Bedingung ist also äquivalent mit der Stetigkeit von A und A^{-1}.

Korollar 4.12. *Ist a stetig, so sind folgende Aussagen äquivalent.*

(i) Es gilt die inf-sup-Bedingung (4.9);

(ii) Der Operator $A : V \to (\operatorname{Ker} A')^0$ ist ein Isomorphismus mit $\|Av\|_{W'} \geq c_E\|v\|_V$, $v \in V$;

(iii) Der Operator $A' : (\operatorname{Ker} A')^\perp \to V'$ ist ein Isomorphismus mit $\|A'w\|_{V'} \geq c_E\|w\|_W$, $w \in (\operatorname{Ker} A')^\perp$.

Beweis. Aus dem Beweis von Satz 4.11 sieht man, dass aus (i) die Aussage (ii) folgt.

Gilt (ii), so folgt aus $\|Av\|_{W'} \geq c_E\|v\|_V$ für alle $v \in V$ die Bedingung (4.9). Ferner ist zu jedem $q \in (\operatorname{Ker} A')^\perp$ durch $w \mapsto (w,q)_W$ ein Funktional $\varphi \in (\operatorname{Ker} A')^0$ mit $\|\varphi\|_{W'} = \|q\|_W$ erklärt. Da A ein Isomorphismus ist, gibt es ein $p \in V$ mit $a(p,w) = (w,q)_W$ für alle $w \in W$. Somit folgt $\|q\|_W = \|\varphi\|_{W'} = \|Ap\|_{W'} \geq c_E\|p\|_V$ und mit $a(p,q) = \|q\|_W^2$

$$\sup_{0\neq v\in V} \frac{a(v,q)}{\|v\|_V} \geq \frac{a(p,q)}{\|p\|_V} = \frac{\|q\|_W^2}{\|p\|_V} \geq c_E\|q\|_W.$$

Wegen $\|Av\|_{W'} > 0$ für alle $v \neq 0$ erfüllt $A' : (\operatorname{Ker} A')^\perp \to V'$ die drei Voraussetzungen von Satz 4.11, und wir erhalten (iii). Ist umgekehrt $A' : (\operatorname{Ker} A')^\perp \to V'$ ein Isomorphismus, so gilt für gegebenes $v \in V$

$$\begin{aligned}\|v\|_V &= \sup_{0\neq\varphi\in V'} \frac{\varphi(v)}{\|\varphi\|_{V'}} = \sup_{0\neq w\in(\operatorname{Ker} A')^\perp} \frac{(A'w)(v)}{\|A'w\|_{V'}} \\ &= \sup_{0\neq w\in(\operatorname{Ker} A)^\perp} \frac{a(v,w)}{\|A'w\|_{V'}} \leq c_E^{-1} \sup_{0\neq w\in(\operatorname{Ker} A)^\perp} \frac{a(v,w)}{\|w\|_W}\end{aligned}$$

und somit (i). □

[1] Ein Banach-Raum V heißt **reflexiv**, falls $V'' = V$.

Bemerkung. Satz 4.11 ist eine Verallgemeinerung des Satzes von Lax-Milgram, weil im Fall $V = W$ und einer koerziven Bilinearform die inf-sup-Bedingung

$$\inf_{0\neq v\in V}\sup_{0\neq w\in V}\frac{a(v,w)}{\|v\|_V\|w\|_V} \geq \inf_{0\neq v\in V}\frac{a(v,v)}{\|v\|_V^2} \geq c_K$$

gilt. Die Bedingung (iii) in Satz 4.11 ergibt sich, indem man $v = w$ wählt und die Koerzivität verwendet.

Zu (4.8) kann auch eine zum Céa-Lemma analoge Konvergenzaussage für die Lösung des diskreten Problems

$$u_h \in V_h: \quad a(u_h, w_h) = \ell(w_h), \quad w_h \in W_h, \tag{4.10}$$

bei konformer Diskretisierung $V_h \subset V$ und $W_h \subset W$ bewiesen werden.

Satz 4.13. *Die Bedingungen aus Satz 4.11 seien sowohl für die Hilbert-Räume V, W als auch für die konformen Finite-Element-Räume $V_h \subset V$ und $W_h \subset W$ erfüllt. Dann gilt für die eindeutig bestimmte Lösung $u_h \in V_h$*

$$\|u - u_h\|_V \leq \left(1 + \frac{c_S}{c_E}\right)\inf_{v_h\in V_h}\|u - v_h\|_V.$$

Beweis. Für $\varphi(w) := a(u - v_h, w)$ gilt $\|\varphi\|_{W'} \leq c_S\|u - v_h\|_V$. Ferner gilt für $A_h : V_h \to W_h'$ definiert durch $(A_h v_h)(w_h) := a(v_h, w_h)$ für $v_h \in V_h$ und $w_h \in W_h$, dass $\|A_h^{-1}\| \leq 1/c_E$. Aus der Galerkin-Orthogonalität erhält man

$$\varphi(w_h) = a(u - v_h, w_h) = a(u_h - v_h, w_h) = (A_h(u_h - v_h))(w_h), \quad w_h \in W_h.$$

Also folgt $\varphi|_{W_h} = A_h(u_h - v_h)$ oder äquivalent $u_h - v_h = A_h^{-1}\varphi|_{W_h}$. Hieraus folgt

$$\|u_h - v_h\|_V \leq \frac{1}{c_E}\|\varphi\|_{W'} \leq \frac{c_S}{c_E}\|u - v_h\|_V$$

und mit der Dreiecksungleichung die Behauptung. □

4.3 Sattelpunktprobleme und restringierte Variationsprobleme

Setzt man $a(v, w) = (v, w)_{L^2(\Omega)}$, $b(v, q) = -(v, \nabla q)_{L^2(\Omega)}$ und $V = [L^2(\Omega)]^d$, $W = H_0^1(\Omega)$, so ist (4.7) von der Form: suche $(u, p) \in V \times W$

$$a(u, v) + b(v, p) = f(v) \quad \text{für alle } v \in V, \tag{4.11a}$$
$$b(u, q) = g(q) \quad \text{für alle } q \in W. \tag{4.11b}$$

Dabei sind $f \in V'$, $g \in W'$ und $a : V \times V \to \mathbb{R}$, $b : V \times W \to \mathbb{R}$ stetige Bilinearformen mit Stetigkeitskonstanten c_S bzw. c_S'. Probleme dieser Struktur werden als **Sattelpunktprobleme** oder **gemischte Variationsprobleme** bezeichnet.

Die Lösbarkeit von Sattelpunktproblemen untersuchen wir im Zusammenhang mit **restringierten Variationsgleichungen**

$$u \in G: \quad a(u,z) = f(z) \quad \text{für alle } z \in Z. \tag{4.12}$$

Dabei ist durch die Menge

$$G := \{v \in V : b(v,q) = g(q) \text{ für alle } q \in W\}$$

die Nebenbedingung vorgegeben und $Z := \{v \in V : b(v,q) = 0 \text{ für alle } q \in W\}$.

Satz 4.14. *Es sei a auf Z koerziv. Gilt $G \neq \emptyset$, so besitzt (4.12) eine eindeutige Lösung $u \in G$. Für diese gilt*

$$\|u\|_V \leq \frac{1}{c_K}\|f\|_{V'} + \left(\frac{c_S}{c_K} + 1\right) \sup_{v \in G} \|v\|_V.$$

Beweis. Sei $v \in G$. Wegen der Bilinearität von b gilt $v + z \in G$ für alle $z \in Z$. Andererseits läßt sich $u \in G$ auch durch

$$u = v + \tilde{z} \tag{4.13}$$

mit einem $\tilde{z} \in Z$ darstellen. Die Aufgabe (4.12) ist damit äquivalent zur Bestimmung eines $\tilde{z} \in Z$ mit

$$a(\tilde{z}, z) = f(z) - a(v,z) \quad \text{für alle } z \in Z. \tag{4.14}$$

Wegen der Voraussetzung kann das Lemma von Lax-Milgram angewendet werden. Also ist (4.14) und damit auch (4.12) eindeutig lösbar. Aus (4.14) folgt für $z = \tilde{z}$

$$\|\tilde{z}\|_V \leq \frac{1}{c_K}(\|f\|_{V'} + c_S\|v\|_V). \tag{4.15}$$

Mit (4.13) und der Dreiecksgleichung erhält man hieraus die Behauptung. □

Beispiel 4.15. Wir betrachten das Poissonsche Randwertproblem

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

Mit $V = H^1(\Omega)$ und $W = H^{-1/2}(\partial\Omega)$ lässt sich

$$a(v,w) = \int_\Omega \nabla v \cdot \nabla w \,\mathrm{d}x, \quad f(v) = \int_\Omega fv \,\mathrm{d}x,$$
$$b(v,q) = -\int_{\partial\Omega} vq \,\mathrm{d}s, \qquad g(q) = -\int_{\partial\Omega} gq \,\mathrm{d}s$$

wählen. Man erhält damit

$$G = \{v \in H^1(\Omega) : \int_{\partial\Omega} vq \,\mathrm{d}s = \int_{\partial\Omega} gq \,\mathrm{d}s \text{ für alle } q \in H^{-1/2}(\partial\Omega)\}$$

sowie $Z = H_0^1(\Omega)$. Die Dirichletschen Randbedigungen lassen sich auch in der Form $u \in V$, $\gamma_0 u = g$ mit dem Spur-Operator $\gamma_0 : V \to H^{1/2}(\partial\Omega)$ schreiben. Im vorliegenden Fall ist die Bedingung $G \neq \emptyset$ zu der häufig angegebenen Forderung $g \in H^{1/2}(\partial\Omega)$ äquivalent.

Wir untersuchen nun den Zusammenhang der Lösbarkeit von (4.11) bzw. (4.12).

Lemma 4.16. *Ist $(u,p) \in V \times W$ eine Lösung des Sattelpunktproblems (4.11), dann löst u die restringierte Variationsgleichung (4.12).*

Beweis. Der zweite Teil von (4.11) ist äquivalent zu $u \in G$. Für $z \in Z \subset V$ gilt wegen $p \in W$, dass $b(z,p) = 0$. Mit dem ersten Teil von (4.11) liefert dies

$$a(u,z) = f(z) \quad \text{für alle } z \in Z.$$

Folglich ist u eine Lösung von (4.12). □

Im Folgenden untersuchen wir die Umkehrung dieses Lemmas, d.h. wir werden zeigen, dass zu jeder Lösung u von (4.12) ein $p \in W$ existiert, so dass (u,p) das Sattelpunktproblem (4.11) löst. Das orthogonale Komplement $Z^\perp$ zu Z bildet einen abgeschlossenen Unterraum von V. Daher läßt sich der Raum V als direkte Summe

$$V = Z \oplus Z^\perp$$

darstellen. Wegen (4.12) löst $(u,p) \in V \times W$ genau dann das System (4.11), falls

$$b(v,p) = f(v) - a(u,v) \quad \text{für alle } v \in Z^\perp. \tag{4.16}$$

Für die Lösbarkeit der letzten Gleichung bzgl. p setzen wir die folgende nach Ladyshenskaja, Babuška und Brezzi benannte **LBB-Bedingung** für b voraus.

Lemma 4.17. *Die Bilinearform b erfülle die LBB-Bedingung*

$$\sup_{0 \neq v \in V} \frac{b(v,q)}{\|v\|_V} \geq c_{\mathrm{LBB}} \|q\|_W \quad \textit{für alle } q \in Y^\perp, \tag{4.17}$$

wobei $Y := \{q \in W : b(v,q) = 0 \text{ für alle } v \in V\}$. Dann besitzt (4.16) eine Lösung $p \in W$.

Beweis. Wegen der Bilinearität und Stetigkeit von b wird durch $(Bv)(q) := b(v,q)$ ein stetiger linearer Operator $B : V \to W'$ definiert. Aus Korollar 4.12 (iii) angewendet auf $A = B'$ und $V = Y^\perp$ folgt wegen der LBB-Bedingung

$$\|Bv\|_{W'} \geq c_{\mathrm{LBB}} \|v\|_V \quad \text{für alle } v \in Z^\perp. \tag{4.18}$$

Es bezeichne $\mathcal{R} : W' \to W$ den Rieszschen Darstellungsoperator, und wir definieren eine symmetrische, stetige Bilinearform $d : V \times V \to \mathbb{R}$ durch

$$d(w,v) := (Bv)(\mathcal{R}Bw) = (\mathcal{R}Bv, \mathcal{R}Bw)_W \quad \text{für alle } v,w \in V.$$

Mit (4.18) erhält man

$$d(v,v) = \|\mathcal{R}Bv\|_W^2 = \|Bv\|_{W'}^2 \geq c_{\mathrm{LBB}}^2 \|v\|_V^2 \quad \text{für alle } v \in Z^\perp.$$

Somit ist d $Z^\perp$-koerziv, und nach dem Lemma von Lax-Milgram gibt es ein $y \in Z^\perp$ mit

$$d(y,v) = f(v) - a(u,v) \quad \text{für alle } v \in Z^\perp.$$

Das durch $p := \mathcal{R}By \in W$ definierte Element löst (4.16). □

Wir fassen die Ergebnisse zusammen.

Satz 4.18. *Es sei a auf Z koerziv. Für b gelte die LBB-Bedingung (4.17). Ist $G \neq \emptyset$, so besitzt das Sattelpunktproblem (4.11) mindestens eine Lösung $(u,p) \in V \times W$. Dabei ist die erste Komponente $u \in V$ eindeutig bestimmt, und es gelten die Abschätzungen*

$$\|u\|_V \leq \frac{1}{c_K}\|f\|_{V'} + \frac{1}{c_{\mathrm{LBB}}}\left(\frac{c_S}{c_K}+1\right)\|g\|_{W'}$$

und

$$\inf_{y\in Y}\|p+y\|_W \leq \frac{1}{c_{\mathrm{LBB}}}\left(\frac{c_S}{c_K}+1\right)\left\{\|f\|_{V'} + \frac{c_S}{c_{\mathrm{LBB}}}\|g\|_{W'}\right\}.$$

Beweis. Nach Satz 4.14 besitzt die restringierte Variationsgleichung (4.12) eine eindeutige Lösung $u \in G$. Mit Lemma 4.17 folgt die Existenz einer Komponente $p \in W$, so dass $(u,p) \in V\times W$ das Sattelpunktproblem (4.11) löst. Weil nach Lemma 4.16 die V-Komponente jeder Lösung von (4.11) auch (4.12) löst, ist die Komponente u eindeutig bestimmt.

Das Element $u \in V$ wird durch $u = v + \tilde{z}$ mit $\tilde{z} \in Z$, $v \in Z^\perp$ eindeutig dargestellt. Die Bilinearität von b, (4.11) und (4.18) liefern

$$\|g\|_{W'} = \|Bu\|_{W'} = \|Bv\|_{W'} \geq c_{\mathrm{LBB}}\|v\|_V.$$

Somit gilt $\|v\|_V \leq c_{\mathrm{LBB}}^{-1}\|g\|_{W'}$. Weil mit $u \in G$ auch $v \in G$ ist, erhält man mit (4.15) die Abschätzung

$$\|u\|_V \leq \|v\|_V + \|\tilde{z}\|_V \leq \frac{1}{c_K}\|f\|_{V'} + \frac{1}{c_{\mathrm{LBB}}}\left(\frac{c_S}{c_K}+1\right)\|g\|_{W'}.$$

Wir untersuchen nun die zweite Komponente $p \in W$. Aus (4.16) folgt

$$|b(v,p)| \leq |f(v)| + |a(u,v)| \leq (\|f\|_{V'} + c_S\|u\|_V)\|v\|_V \quad \text{für alle } v \in Z^\perp$$

und somit

$$\sup_{0\neq v\in V}\frac{b(v,p)}{\|v\|_V} \leq \|f\|_{V'} + c_S\|u\|_V.$$

Weil ein $y \in Y$ existiert mit $p+y \in Y^\perp$, folgt mit (4.17)

$$c_{\mathrm{LBB}}\|p+y\|_W \leq \sup_{0\neq v\in V}\frac{b(v,p+y)}{\|v\|_V} \leq \|f\|_{V'} + c_S\|u\|_V.$$

Unter Beachtung der ersten Abschätzung folgt somit die gesamte Behauptung. □

Bemerkung. Gilt $Y = \{0\}$, so erkennt man aus dem Beweis von Lemma 4.17, dass die W-Komponente und somit die gesamte Lösung $(u,p) \in V \times W$ von (4.11) eindeutig ist. Die beiden Abschätzungen des vorangehenden Satzes stellen dann Stabilitätsabschätzungen zum Einfluss der Störungen in f und g auf die Lösung (u,p) dar.

Sattelpunkte und Optimalitätskriterien

Ist die Bilinearform a symmetrisch und gilt $a(v,v) > 0$ für alle $0 \neq v \in Z$, so bildet die restringierte Variationsgleichung (4.12) eine notwendige und hinreichende Bedingung dafür, dass $u \in G$ das Variationsproblem

$$\text{minimiere} \quad J(v) := \frac{1}{2}a(v,v) - f(v) \quad \text{über } v \in G \tag{4.19}$$

löst; siehe Übungen. Wie in Abschnitt 4.1 der *Einführung in die Numerik* definieren wir das zu diesem Problem gehörende Lagrange-Funktional

$$L(v,q) := J(v) + b(v,q) - g(q) \quad \text{für alle } v \in V,\ q \in W.$$

Mit Hilfe des Lagrange-Funktionals L kann ein hinreichendes Optimalitätskriterium für (4.19) in Sattelpunktform angegeben werden. Dabei heißt $(u,p) \in V \times W$ **Sattelpunkt** von L, falls

$$L(u,q) \le L(u,p) \le L(v,p) \quad \text{für alle } v \in V,\ q \in W. \tag{4.20}$$

Lemma 4.19. *Es sei $(u,p) \in V \times W$ ein Sattelpunkt von L. Dann löst die zugehörige Komponente $u \in V$ das Variationsproblem (4.19).*

Beweis. Wir zeigen zunächst, dass $u \in G$ gilt. Aus dem linken Teil von (4.20) folgt

$$b(u,q) - g(q) \le b(u,p) - g(p) \quad \text{für alle } q \in W. \tag{4.21}$$

Weil W ein linearer Raum ist, hat man $q := p + y \in W$ für alle $y \in W$. Unter Beachtung der Linearität von $b(u,\cdot)$ und g liefert damit (4.21) die Abschätzung

$$b(u,y) - g(y) \le 0 \quad \text{für alle } y \in W.$$

Mit $y \in W$ ist auch $-y \in W$. Also folgt $b(u,y) - g(y) = 0$ für alle $y \in W$ und hieraus $u \in G$. Mit dem rechten Teil von (4.20) erhält man für alle $v \in G$

$$\begin{aligned} J(u) &= J(u) + b(u,p) - g(p) = L(u,p) \\ &\le L(v,p) = J(v) + b(v,p) - g(p) = J(v). \end{aligned}$$

Damit löst u das Variationsproblem (4.19). □

In Satz 2.1 haben wir im Fall einer symmetrischen Bilinearform gezeigt, dass das Minimum des Energiefunktionals J mit der Lösung des Variationsproblems übereinstimmt. Im folgenden Satz zeigen wir, dass die Lösung des Sattelpunktproblems (4.11) mit dem Sattelpunkt (also dem Minimum für $q = p$ und dem Maximum für $v = u$) übereinstimmt.

Satz 4.20. *$(u,p) \in V \times W$ ist genau dann Lösung des Sattelpunktproblems (4.11), wenn (u,p) ein Sattelpunkt ist.*

Beweis. siehe Übung. □

4.3.1 Konforme Approximation gemischter Variationsgleichungen

Wir betrachten nun eine konforme Finite-Elemente-Diskretisierung der gemischten Variationsgleichung (4.11). Entsprechend seien $V_h \subset V$ und $W_h \subset W$ gewählt. Gesucht wird eine Lösung $(u_h, p_h) \in V_h \times W_h$ der **diskreten gemischten Variationsgleichung**

$$a(u_h, v_h) + b(v_h, p_h) = f(v_h) \quad \text{für alle } v_h \in V_h, \tag{4.22a}$$
$$b(u_h, q_h) = g(q_h) \quad \text{für alle } q_h \in W_h. \tag{4.22b}$$

Diese Finite-Elemente-Diskretisierung von (4.11) und damit letztlich von (4.12) wird **Methode der gemischten finiten Elemente** genannt. Die Lösbarkeit wie auch die Stabilität der Lösungen von (4.22) untersuchen wir analog zum stetigen Fall. Es sei

$$G_h := \{v_h \in V_h : b(v_h, q_h) = g(q_h) \text{ für alle } q_h \in W_h\}$$

sowie $Z_h := \{v_h \in V_h : b(v_h, q_h) = 0 \text{ für alle } q_h \in W_h\}$. Wir weisen darauf hin, dass trotz der getroffenen Voraussetzung $V_h \subset V$ im Allgemeinen gilt

$$G_h \not\subset G \quad \text{und} \quad Z_h \not\subset Z.$$

Damit folgt aus der Z-Koerzivität von a nicht automatisch deren Z_h-Koerzivität. Wir setzen daher voraus, dass $c_K > 0$ existiert mit

$$c_K \|v_h\|_V^2 \le a(v_h, v_h) \quad \text{für alle } v_h \in Z_h. \tag{4.23}$$

Ferner gebe es eine Konstante $c_{\mathrm{LBB}} > 0$ derart, dass

$$\sup_{0 \neq v_h \in V_h} \frac{b(v_h, q_h)}{\|v_h\|_V} \ge c_{\mathrm{LBB}} \|q_h\|_W \quad \text{für alle } q_h \in Y_h^{\perp} \tag{4.24}$$

mit $Y_h := \{q_h \in W_h : b(v_h, q_h) = 0 \text{ für alle } v_h \in V_h\}$ gilt. Unter den getroffenen Voraussetzungen lässt sich Satz 4.18 unmittelbar auf das diskrete Problem (4.22) übertragen. Wir erhalten

Satz 4.21. *Es sei $G_h \neq \varnothing$, und a erfülle die Bedingungen (4.23) und (4.24). Dann besitzt (4.22) mindestens eine Lösung $(u_h, p_h) \in V_h \times W_h$. Die erste Komponente $u_h \in V_h$ ist dabei eindeutig bestimmt, und es gelten die Abschätzungen*

$$\|u_h\|_V \le \frac{1}{c_K}\|f\|_{V'} + \frac{1}{c_{\mathrm{LBB}}}\left(\frac{c_S}{c_K} + 1\right)\|g\|_{W'}$$

und

$$\inf_{y_h \in Y_h} \|p_h + y_h\|_W \le \frac{1}{c_{\mathrm{LBB}}}\left(\frac{c_S}{c_K} + 1\right)\left\{\|f\|_{V'} + \frac{c_S}{c_{\mathrm{LBB}}}\|g\|_{W'}\right\}.$$

Der Einfachheit halber untersuchen wir das Konvergenzverhalten der Methode der gemischten finiten Elemente für den Fall $Y_h = \{0\}$. In diesem Fall kann die Konvergenzanalyse von Galerkin-Verfahren angewendet werden.

Satz 4.22. *Die Bilinearformen a und b erfüllen die Voraussetzung der Sätze 4.18 und 4.21. Ferner gelte*

$$\sup_{0 \neq v_h \in V_h} \frac{b(v_h, q_h)}{\|v_h\|_V} \ge \tilde{c}_{\mathrm{LBB}} \|q_h\|_W \quad \textit{für alle } q_h \in W_h. \tag{4.25}$$

Dann gilt für die Differenz der Lösung (u_h, p_h) von (4.22) und der Lösung (u, p) von (4.11)

$$\|u - u_h\|_V + \|p - p_h\|_W \le c\left\{\inf_{v_h \in V_h}\|u - v_h\|_V + \inf_{q_h \in W_h}\|p - q_h\|_W\right\}$$

mit einer Konstanten $c > 0$.

Beweis. Die Bedingung (4.25) sichert die Gültigkeit von $Y_h = \{0\}$. Zunächst erhalten wir für beliebige $\tilde{v}_h \in V_h$, $\tilde{q}_h \in W_h$, dass

$$\begin{aligned} a(u_h - \tilde{v}_h, v_h) + b(v_h, p_h - \tilde{q}_h) &= a(u - \tilde{v}_h, v_h) + b(v_h, p - \tilde{q}_h) \quad \text{für alle } v_h \in V_h, \\ b(u_h - \tilde{v}_h, q_h) &= b(u - \tilde{v}_h, q_h) \quad \text{für alle } q_h \in W_h. \end{aligned}$$

Unter Beachtung von $Y_h = \{0\}$ folgt hieraus mit Satz 4.21

$$\|u_h - \tilde{v}_h\| \leq \left\{ \frac{c_S}{c_K} + \frac{c_S'}{\tilde{c}_{\mathrm{LBB}}} \left(\frac{c_S}{c_K} + 1 \right) \right\} \|u - \tilde{v}_h\| + \frac{c_S'}{c_K} \|p - \tilde{q}_h\|$$

und

$$\|p_h - \tilde{q}_h\| \leq \frac{1}{\tilde{c}_{\mathrm{LBB}}} \left(\frac{c_S}{c_K} + 1 \right) \left\{ c_S' \|p - \tilde{q}_h\| + c_S \left(1 + \frac{c_S'}{c_{\mathrm{LBB}}} \right) \|u - \tilde{v}_h\| \right\}.$$

Nutzt man $\|u - u_h\| \leq \|u - \tilde{v}_h\| + \|u_h - \tilde{v}_h\|$ bzw. $\|p - p_h\| \leq \|p - \tilde{q}_h\| + \|p_h - \tilde{q}_h\|$, so erhält man hieraus die Behauptung. □

Die Gültigkeit der diskreten LBB-Bedingung (4.25) direkt zu überprüfen, ist oft schwierig. Das folgende Kriterium gibt eine einfache, hinreichende Bedingung an.

Lemma 4.23 (Kriterium von Fortin). *Die Bilinearform b genüge der LBB-Bedingung (4.17). Sei $V_h \times W_h \subset V \times W$ eine Familie von Teilräumen, so dass Projektoren $\Pi_h : V \to V_h$ existieren mit*

$$b(v - \Pi_h v, q_h) = 0 \quad \textit{für alle } v \in V \textit{ und } q_h \in W_h$$

und $\|\Pi_h v\|_V \leq c_\Pi \|v\|_V$ für alle $v \in V$ mit einer h-unabhängigen Konstanten $c_\Pi > 0$. Dann erfüllt b auch die diskrete LBB-Bedingung (4.25).

Beweis. Nach Voraussetzung gilt wegen $\Pi_h v \in V_h$ für alle $q_h \in W_h \subset W$

$$\begin{aligned} c_{\mathrm{LBB}} \|q_h\|_W &\leq \sup_{0 \neq v \in V} \frac{b(v, q_h)}{\|v\|_V} = \sup_{0 \neq v \in V} \frac{b(\Pi_h v, q_h)}{\|v\|_V} \leq c_\Pi \sup_{0 \neq v \in V} \frac{b(\Pi_h v, q_h)}{\|\Pi_h v\|_V} \\ &\leq c_\Pi \sup_{0 \neq v \in V_h} \frac{b(v, q_h)}{\|v\|_V}. \end{aligned}$$

Hieraus folgt die Behauptung mit $\tilde{c}_{\mathrm{LBB}} := c_{\mathrm{LBB}} / c_\Pi$. □

4.4 Lösbarkeit der gemischten Formulierung des Poisson-Problems

Wir haben das Kapitel mit einer gemischten Formulierung (4.7) des Poisson-Problems begonnen. Nach den allgemeinen Aussagen zu Existenz und Eindeutigkeit solcher Probleme, soll nun geprüft werden, ob (4.7) und eine weitere Formulierung den Voraussetzungen dieser Theorie genügt. Wir betrachten als zunächst die sog. **primal-gemischte Formulierung** des Poisson-Problems

$$(u, v)_{L^2(\Omega)} - (v, \nabla p)_{L^2(\Omega)} = 0 \qquad \text{für alle } v \in [L^2(\Omega)]^d, \tag{4.26a}$$

$$(u, \nabla q)_{L^2(\Omega)} = (f, q)_{L^2(\Omega)} \qquad \text{für alle } q \in H_0^1(\Omega). \tag{4.26b}$$

Satz 4.24. *Zur primal-gemischten Formulierung (4.26) existiert für jedes $f \in H^{-1}(\Omega)$ eine eindeutige Lösung $(u,p) \in [L^2(\Omega)]^d \times H_0^1(\Omega)$.*

Beweis. Wir setzen $V = [L^2(\Omega)]^d$, $W = H_0^1(\Omega)$ (normiert mit $|\cdot|_{H^1(\Omega)}$) und

$$a(v,w) := (v,w)_{L^2(\Omega)}, \quad b(v,q) := -(v,\nabla q)_{L^2(\Omega)}.$$

Dann ist a auf ganz V koerziv. Um die LBB-Bedingung zu zeigen, wählen wir $v := \nabla q \in V$. Für $q \neq 0$ gilt $v = \nabla q \neq 0$. Denn wäre $v = 0$, so wäre q konstant, was wegen $q \in H_0^1(\Omega)$ impliziert, dass $q = 0$. Hieraus folgt also

$$\inf_{0\neq q\in W} \sup_{0\neq v\in V} \frac{b(v,q)}{\|v\|_V \|q\|_W} \geq \inf_{0\neq q\in W} \frac{b(\nabla q, q)}{\|\nabla q\|_V \|q\|_W} = \inf_{0\neq q\in W} \frac{(\nabla q, \nabla q)_{L^2(\Omega)}}{|q|^2_{H^1(\Omega)}} = 1.$$

Folglich sind die Voraussetzungen von Satz 4.18 erfüllt, und die Behautung folgt. □

Führen wir den Hilbert-Raum

$$H(\operatorname{div};\Omega) = \{v \in [L^2(\Omega)]^d : \operatorname{div} v \in L^2(\Omega)\}$$

mit dem Skalarprodukt $(v,w)_{H(\operatorname{div};\Omega)} := (v,w)_{L^2(\Omega)} + (\operatorname{div} v, \operatorname{div} w)_{L^2(\Omega)}$ ein, so kann alternativ zur primal-gemischten Formulierung (4.26) unter Verwendung von

$$-(v,\nabla q)_{L^2(\Omega)} = (\operatorname{div} v, q)_{L^2(\Omega)} \quad \text{für } v \in H(\operatorname{div};\Omega) \text{ und } q \in H_0^1(\Omega)$$

die **dual-gemischte Formulierung** des Poisson-Problems betrachtet werden: suche $(u,p) \in H(\operatorname{div};\Omega) \times L^2(\Omega)$, so dass

$$(u,v)_{L^2(\Omega)} + (\operatorname{div} v, p)_{L^2(\Omega)} = 0 \qquad \text{für alle } v \in H(\operatorname{div};\Omega), \tag{4.27a}$$
$$(\operatorname{div} u, q)_{L^2(\Omega)} = -(f,q)_{L^2(\Omega)} \qquad \text{für alle } q \in L^2(\Omega). \tag{4.27b}$$

Dabei fällt auf, dass die Dirichlet-Bedingung $p|_{\partial\Omega} = 0$ nicht explizit auftritt. Der folgende Satz zeigt aber, dass sie implizit in der Formulierung enthalten ist.

Satz 4.25. *Zur dual-gemischten Formulierung (4.27) existiert für jedes $f \in H^{-1}(\Omega)$ eine eindeutige Lösung $(u,p) \in H(\operatorname{div};\Omega) \times L^2(\Omega)$. Für p gilt sogar $p \in H_0^1(\Omega)$.*

Beweis. Das Problem (4.27) besitzt Sattelpunktform für die Hilbert-Räume $V = H(\operatorname{div};\Omega)$ und $W = L^2(\Omega)$ und die Bilinearformen

$$a(v,w) := (v,w)_{L^2(\Omega)}, \quad b(v,q) := (\operatorname{div} v, q)_{L^2(\Omega)}.$$

Dabei ist der Kern von b

$$Z := \{v \in V : (\operatorname{div} v, q)_{L^2(\Omega)} = 0 \text{ für alle } q \in L^2(\Omega)\}$$

gerade der Raum der divergenzfreien Funktionen. Die Bilinearform a ist auf $V \times V$ stetig und auf Z elliptisch, weil für $v \in Z$ gilt

$$a(v,v) = \|v\|^2_{L^2(\Omega)} = \|v\|^2_{L^2(\Omega)} + \|\operatorname{div} v\|^2_{L^2(\Omega)} = \|v\|^2_{H(\operatorname{div};\Omega)}.$$

Es ist noch die LBB-Bedingung für b zu zeigen. Sei hierzu $0 \neq q \in L^2(\Omega)$ beliebig. Wegen der Dichtheit von $C_0^\infty(\Omega)$ in $L^2(\Omega)$ existiert $0 \neq \tilde{q} \in C_0^\infty(\Omega)$ mit

$$\|q - \tilde{q}\|_{L^2(\Omega)}^2 \leq \frac{1}{2}\|q\|_{L^2(\Omega)}^2.$$

Mit diesem $\tilde{q}$ sei v mit den Komponenten

$$v_1(x_1, \ldots, x_d) := \int_{-\infty}^{x_1} \tilde{q}(t, x_2, \ldots, x_d)\,\mathrm{d}t$$

und $v_i = 0$, $i \geq 2$, definiert. Dann gilt punktweise

$$\operatorname{div} v = \frac{\partial v}{\partial x_1} = \tilde{q}.$$

Daher ist $v \in V \backslash \{0\}$, und wie im Beweis der Poincaré-Friedrichsschen Ungleichung (Satz 2.22) sieht man $\|v\|_{L^2(\Omega)} \leq c\|\tilde{q}\|_{L^2(\Omega)}$, woraus

$$\|v\|_{H(\operatorname{div};\Omega)}^2 = \|v\|_{L^2(\Omega)}^2 + \|\operatorname{div} v\|_{L^2(\Omega)}^2 \leq (c^2+1)\|\tilde{q}\|_{L^2(\Omega)}^2$$

folgt. Wegen

$$\begin{aligned}(\tilde{q}, q)_{L^2(\Omega)} &= \frac{1}{2}\left(\|\tilde{q}\|_{L^2(\Omega)}^2 + \|q\|_{L^2(\Omega)}^2 - \|q - \tilde{q}\|_{L^2(\Omega)}^2\right) \geq \frac{1}{4}\left(\|\tilde{q}\|_{L^2(\Omega)}^2 + \|q\|_{L^2(\Omega)}^2\right) \\ &\geq \frac{1}{2}\|\tilde{q}\|_{L^2(\Omega)}\|q\|_{L^2(\Omega)}\end{aligned}$$

erhält man

$$\frac{b(v,q)}{\|v\|_{H(\operatorname{div};\Omega)}} \geq \frac{(\operatorname{div} v, q)_{L^2(\Omega)}}{\sqrt{c^2+1}\,\|\tilde{q}\|_{L^2(\Omega)}} = \frac{(\tilde{q}, q)_{L^2(\Omega)}}{\sqrt{c^2+1}\,\|\tilde{q}\|_{L^2(\Omega)}} \geq \frac{\|q\|_{L^2(\Omega)}}{2\sqrt{c^2+1}}.$$

Daher existiert nach Satz 4.18 eine eindeutige Lösung $(u, p) \in H(\operatorname{div};\Omega) \times L^2(\Omega)$.

Wir zeigen noch, dass sogar $p \in H_0^1(\Omega)$ gilt. Wegen $[C_0^\infty(\Omega)]^d \subset V$ gilt nach (4.27)

$$(u, \varphi)_{L^2(\Omega)} = -(p, \operatorname{div}\varphi)_{L^2(\Omega)} \quad \text{für alle } \varphi \in [C_0^\infty(\Omega)]^d.$$

Nach Definition der schwachen Ableitung folgt hieraus $u = \nabla p$ und somit $p \in H^1(\Omega)$. Ferner erhält man aus (4.27) für $\varphi \in [C^\infty(\Omega)]^d \subset V$

$$0 = (u, \varphi)_{L^2(\Omega)} + (\operatorname{div}\varphi, p)_{L^2(\Omega)} = (\nabla p, \varphi)_{L^2(\Omega)} + (\operatorname{div}\varphi, p)_{L^2(\Omega)} = \int_{\partial\Omega} \varphi \cdot \nu\, p\,\mathrm{d}s.$$

Also ist $p \in H_0^1(\Omega)$. □

4.4.1 Das Raviart-Thomas-Element

Im Folgenden stellen wir eine passende Diskretisierung von $V = H(\operatorname{div};\Omega)$ in zwei Dimensionen vor. Dazu sei

$$V_h := \{v_h \in [L^2(\Omega)]^2 : v_h(x,y)|_\tau = \begin{bmatrix} a_\tau \\ b_\tau \end{bmatrix} + c_\tau \begin{bmatrix} x \\ y \end{bmatrix} \text{ mit } a_\tau, b_\tau, c_\tau \in \mathbb{R}$$
$$\text{und } v_h \cdot \nu \text{ ist stetig an den Elementgrenzen}\}.$$

Wir zeigen in den Übungen, dass die Stetigkeit der Normalkomponenten von $v_h \in V_h$ die Konformität $V_h \subset H(\mathrm{div}; \Omega)$ impliziert.

Raviart-Thomas-Funktionen $v_h \in V_h$ erfüllen offenbar $v_h|_\tau \in [\Pi_1]^2$ und besitzen drei Freiheitsgrade pro Element. Die Normalkomponente von Raviart-Thomas-Funktionen ist auf jeder Dreieckskante e konstant. Um dies zu sehen, sei e gegeben durch $\{\alpha + t\beta,\, t \in [0,1]\}$. Dann ist $\nu_e = [-\beta_2, \beta_1]^T$ und

$$\nu_e \cdot v_h(\alpha + t\beta) = \nu_e \cdot \left(\begin{bmatrix} a_\tau \\ b_\tau \end{bmatrix} + c_\tau(\alpha + t\beta) \right) = \nu_e \cdot \left(\begin{bmatrix} a_\tau \\ b_\tau \end{bmatrix} + c_\tau \alpha \right).$$

Die drei Freiheitsgrade von v_h auf jedem τ können durch die Normalkomponenten $\nu_e \cdot v_h$ zu

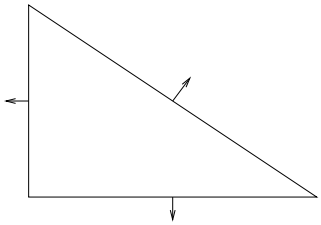


Abbildung 4.2: Raviart-Thomas-Element

den drei Kanten $e \in \partial\tau$ repräsentiert werden. Dazu bestimmt man die beiden Komponenten von $v_h(x_i)$ in jedem der drei Eckpunkte x_i, $i = 1, 2, 3$, von τ aus den Normalkomponenten der jeweiligen beiden anliegenden Kanten. Zu den sechs Werten $v_h(x_i)$, $i = 1, 2, 3$, existiert ein eindeutiges lineares Polynom $v_h|_\tau \in [\Pi_1]^2$. Als Element von $[\Pi_1]^2$ sind die Normalkomponenten von $v_h|_\tau$ auf den Kanten linear. Sie sind sogar konstant, weil sie nach Konstruktion an den beiden Endpunkten übereinstimmen.

Den Finite-Element-Fehler können wir nach Satz 4.22 durch den Approximationsfehler abschätzen. Diesen wiederum schätzen wir wie bisher immer durch Interpolation ab. Dazu definieren wir $\mathfrak{I}_\tau : H^1(\tau) \to V_h|_\tau$ durch

$$\int_e (v - \mathfrak{I}_\tau v) \cdot \nu_e \, \mathrm{d}s = 0 \quad \text{für alle } e \in \partial\tau,$$

und für eine Triangulierung $\mathcal{T}_h$ von Ω definieren wir $\mathfrak{I}_h : H^1(\Omega) \to V_h$ durch

$$(\mathfrak{I}_h v)|_\tau = \mathfrak{I}_\tau(v|_\tau) \quad \text{für alle } \tau \in \mathcal{T}_h.$$

Die Interpolierende wird also durch die Kantenmittelwerte der Normalkomponenten von $v \in H^1(\Omega)$ definiert. Mit dem Gaußschen Divergenzsatz erhält man

$$\int_\tau \mathrm{div}(v - \mathfrak{I}_\tau v) \, \mathrm{d}x = \sum_{e \in \partial\tau} \int_e (v - \mathfrak{I}_\tau v) \cdot \nu_e \, \mathrm{d}s = 0. \tag{4.28}$$

Weil $\mathfrak{I}_\tau v$ linear und somit $\mathrm{div}\, \mathfrak{I}_\tau v$ konstant ist, haben wir die Eigenschaft

$$\mathrm{div}(\mathfrak{I}_h v) = P_0 \, \mathrm{div}\, v \tag{4.29}$$

gezeigt. Hierbei bezeichnet $P_0 : L^2(\Omega) \to \Pi_0$ die L^2-Projektion auf Π_0.

Um die diskrete LBB-Bedingung zu zeigen, benötigen wir folgendes Lemma. Im Folgenden sei

$$W_h := \{q \in L^2(\Omega) : q|_\tau \in \Pi_0 \text{ für alle } \tau \in \mathcal{T}_h\}.$$

Lemma 4.26. *Die Abbildung* $\operatorname{div} : V_h \to W_h$ *ist surjektiv und für jedes* $q_h \in W_h$ *genügt das Urbild* $v_h \in V_h$ *der Abschätzung*

$$\|v_h\|_{H(\operatorname{div};\Omega)} \leq c\|q_h\|_{L^2(\Omega)}. \tag{4.30}$$

Beweis. Sei $q_h \in W_h$ vorgegeben. Wir wählen ein konvexes, polygonales Gebiet $\tilde{\Omega} \supset \Omega$ und setzen q_h trivial auf $\tilde{\Omega}$ fort. Dann existiert $u \in H^2(\tilde{\Omega}) \cap H_0^1(\tilde{\Omega})$ mit $\Delta u = q_h$. Für $v := \nabla u \in H^1(\Omega)$ folgt für $v_h := \mathfrak{I}_h v$ aus (4.28) und dem Divergenzsatz von Gauß

$$\int_\tau \operatorname{div} v_h \,\mathrm{d}x = \int_\tau \operatorname{div} v \,\mathrm{d}x = \int_\tau q_h \,\mathrm{d}x. \tag{4.31}$$

Weil $\operatorname{div} v_h$ und q_h auf τ konstant sind, folgt $\operatorname{div} v_h = q_h$.

Durch Transformation auf das Referenzelement zeigt man schnell, dass

$$\|\mathfrak{I}_\tau v\|_{L^2(\tau)} \leq c\|v\|_{L^2(\tau)} \quad \text{für alle } \tau \in \mathcal{T}_h.$$

Hieraus folgt

$$\begin{aligned}\|v_h\|^2_{H(\operatorname{div};\Omega)} &= \|\operatorname{div} v_h\|^2_{L^2(\Omega)} + \|v_h\|^2_{L^2(\Omega)} = \|q_h\|^2_{L^2(\Omega)} + \sum_{\tau\in\mathcal{T}_h} \|\mathfrak{I}_\tau v\|^2_{L^2(\tau)} \\ &\leq \|q_h\|^2_{L^2(\Omega)} + c^2\|v\|^2_{L^2(\Omega)}.\end{aligned}$$

Wegen

$$\|v\|^2_{L^2(\Omega)} = |u|^2_{H^1(\Omega)} \leq |u|^2_{H^1(\tilde{\Omega})} \leq c'\|q_h\|^2_{L^2(\tilde{\Omega})} = c'\|q_h\|^2_{L^2(\Omega)}$$

folgt hieraus die Behauptung. □

Nach Satz 4.25 ist die dual-gemischte Formulierung des Poisson-Problems in $H(\operatorname{div};\Omega) \times L^2(\Omega)$ eindeutig lösbar. Es ist noch zu prüfen, ob die eindeutige Lösbarkeit auch für den diskreten Ansatzraum $V_h \times W_h$ gilt.

Satz 4.27. *Die Diskretisierung mit Raviart-Thomas-Elementen liefert für die dual-gemischte Formulierung des Poisson-Problems zu jeder rechten Seite* $f \in H^{-1}(\Omega)$ *eine eindeutige Lösung* $(u_h, p_h) \in V_h \times W_h$.

Beweis. Für $v_h \in Z_h = \{v_h \in V_h : (\operatorname{div} v_h, q_h)_{L^2(\Omega)} = 0 \text{ für alle } q_h \in W_h\}$ gilt $\int_\tau \operatorname{div} v_h \,\mathrm{d}x = 0$ für alle $\tau \in \mathcal{T}_h$, und wegen $(\operatorname{div} v_h)|_\tau \in \Pi_0$ folgt hieraus $\operatorname{div} v_h = 0$ in τ. Aus $V_h \subset V$ folgt $\operatorname{div} v_h = 0$. Daher ist $\|v_h\|^2_{H(\operatorname{div};\Omega)} = \|v_h\|^2_{L^2(\Omega)} = a(v_h, v_h)$ und somit a auf Z_h koerziv. Es bleibt, die LBB-Bedingung für b zu zeigen. Nach dem Kriterium von Fortin (Lemma 4.23) müssen wir einen Projektor $\Pi_h : V \to V_h$ mit h-unabhängiger Stetigkeitskonstanten und

$$b(v - \Pi_h v, q_h) = (\operatorname{div}(v - \Pi_h v), q_h)_{L^2(\Omega)} = 0 \quad \text{für alle } v \in V \text{ und } q_h \in W_h$$

konstruieren. Die letzte Bedingung wird für $\Pi_h := \mathfrak{I}_h$ wegen (4.28) erfüllt. Wegen (4.30) erhält man die Stetigkeit von Π_h unter Verwendung von (4.31) und $q_h|_\tau \in \Pi_0$ aus

$$\|q_h\|_{L^2(\tau)} \leq |\tau|^{1/2}|q_h|_\tau| \leq |\tau|^{-1/2}\|\operatorname{div} v\|_{L^1(\tau)} \leq \|\operatorname{div} v\|_{L^2(\tau)} \leq \|v\|_{H(\operatorname{div};\tau)}$$

und der Summation über $\tau \in \mathcal{T}_h$. □

Nachdem die Lösbarkeit des kontinuierlichen und des diskreten Problems geklärt ist, sollen nun Fehlerabschätzungen hergeleitet werden. Dafür benötigen wir eine Aussage zur Approximationseigenschaft in V_h.

Lemma 4.28. *Sei $\mathcal{T}_h$ eine nicht-entartete Triangulierung von Ω. Dann gilt für $v \in H^1(\Omega)$*

$$\|v - \mathfrak{I}_h v\|_{H(\mathrm{div};\Omega)} \leq ch|v|_{H^1(\Omega)} + \inf_{q_h \in W_h} \|\mathrm{div}\, v - q_h\|_{L^2(\Omega)}.$$

Beweis. Sei $\tau \in \mathcal{T}_h$. Nach dem Spur-Satz ist das Funktional $v \mapsto \int_e v \cdot \nu_e \, \mathrm{d}s$, $e \in \partial\tau$, auf $[H^1(\tau)]^2$ beschränkt. Wegen $[\Pi_0]^2 \subset V_h$ gilt außerdem $\mathfrak{I}_\tau v = v$ für $v \in [\Pi_0]^2$. Das Bramble-Hilbert-Lemma und die Transformation auf Referenzelemente liefern

$$\|v - \mathfrak{I}_h v\|_{L^2(\Omega)} \leq ch|v|_{H^1(\Omega)}.$$

Die Abschätzung für $\|\mathrm{div}(v - \mathfrak{I}_h v)\|_{L^2(\Omega)}$ folgt aus (4.29). □

Wegen

$$\inf_{q_h \in W_h} \|p - q_h\|_{L^2(\Omega)} \leq ch\|p\|_{H^1(\Omega)}$$

erhält man somit aus Satz 4.22 die Fehlerabschätzung

$$\|u - u_h\|_{H(\mathrm{div};\Omega)} + \|p - p_h\|_{L^2(\Omega)} \leq c \left(h|u|_{H^1(\Omega)} + h\|p\|_{H^1(\Omega)} + \inf_{f_h \in W_h} \|f - f_h\|_{L^2(\Omega)} \right).$$

4.5 Lösung der diskreten Probleme

Aus der Diskretisierung von Sattelpunktproblemen (4.11) erhält man lineare Gleichungssysteme der Form

$$\begin{bmatrix} A & B \\ B^T & -C \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}. \tag{4.32}$$

Dabei ist wegen der Z_h-Koerzivität (4.23) der Bilinearform a die Matrix $A \in \mathbb{R}^{m \times m}$ mit

$$a_{ij} := a(\varphi_j, \varphi_i), \quad i, j = 1, \ldots, m,$$

(symmetrisch) positiv-definit. Hierbei gelte

$$V_h = \mathrm{span}\{\varphi_1, \ldots, \varphi_m\} \quad \text{und} \quad W_h = \mathrm{span}\{\psi_1, \ldots, \psi_n\}.$$

Aus der diskreten LBB-Bedingung (4.25) folgt, dass $B \in \mathbb{R}^{m \times n}$ mit

$$b_{ij} = b(\varphi_i, \psi_j), \quad i = 1, \ldots, m, \ j = 1, \ldots, n,$$

den Rang n hat. Als Verallgemeinerung der bisherigen Sattelpunkt-Struktur betrachten wir ferner eine (symmetrisch) positiv-semidefinite Matrix $C \in \mathbb{R}^{n \times n}$. Die Systemmatrix

$$\hat{A} := \begin{bmatrix} A & B \\ B^T & -C \end{bmatrix}$$

ist daher zwar symmetrisch, jedoch wegen

$$\begin{bmatrix} x \\ 0 \end{bmatrix}^T \hat{A} \begin{bmatrix} x \\ 0 \end{bmatrix} = x^T A x > 0, \quad x \neq 0, \qquad \begin{bmatrix} 0 \\ y \end{bmatrix}^T \hat{A} \begin{bmatrix} 0 \\ y \end{bmatrix} = -y^T C y \leq 0,$$

indefinit.

Lemma 4.29. *Das Gleichungssystem (4.32) ist eindeutig lösbar, und das Schur-Komplement $S := B^T A^{-1} B + C$ ist positiv-definit.*

Beweis. Die Matrix $S = B^T A^{-1} B + C \in \mathbb{R}^{n \times n}$ ist offensichtlich positiv-definit. Aus der Zerlegung

$$\begin{bmatrix} I & 0 \\ -B^T A^{-1} & I \end{bmatrix} \begin{bmatrix} A & B \\ B^T & -C \end{bmatrix} = \begin{bmatrix} A & B \\ 0 & -D \end{bmatrix}$$

sieht man, dass die Systemmatrix $\hat{A}$ regulär ist. □

Für solche indefiniten Probleme haben wir das MINRES-Verfahren angesprochen. Wir stellen im Folgenden zwei weitere Methoden vor, die auf die Sattelpunktform zugeschnitten sind.

4.5.1 Das Uzawa-Verfahren

Das folgende klassische Iterationsverfahren zur Lösung von (4.32) für die Wahl $C = 0$ wird als **Uzawa-Verfahren** bezeichnet. Sei $(x_0, y_0) \in \mathbb{R}^{m+n}$. Dann setze für $k = 0, 1, \ldots$

$$\begin{aligned} x_{k+1} &= A^{-1}(f - By_k), \\ y_{k+1} &= y_k + \omega(B^T x_{k+1} - g). \end{aligned}$$

Durch Elimination von x_{k+1} erhält man

$$y_{k+1} = y_k + \omega(B^T A^{-1}(f - By_k) - g),$$

was der Richardson-Iteration angewendet auf

$$Sy = B^T A^{-1} f - g$$

entspricht. Aus Beispiel 12.10 (AlMa II) ergibt sich daher sofort

Satz 4.30. *Das Uzawa-Verfahren konvergiert genau dann, wenn $\omega \in (0, 2/\lambda_{\max}(S))$. Der optimale Relaxationsparameter ist*

$$\omega_{\text{opt}} := \frac{2}{\lambda_{\min}(S) + \lambda_{\max}(S)}.$$

Im Uzawa-Verfahren muss in jedem Schritt das Gleichungssystem

$$Ax_{k+1} = f - By_k \tag{4.33}$$

gelöst werden. Dies entspricht der Minimierung des Funktionals

$$h_k(x) := \frac{1}{2} x^T Ax - x^T(f - By_k).$$

Um den Aufwand zu reduzieren, bestimmt man x_{k+1} nicht als exakte Lösung von (4.33), sondern verwendet ein iteratives Verfahren (z.B. das CG-Verfahren). Beim **Arrow-Hurwicz-Verfahren** geht man einen Schritt mit Schrittweite ε in Richtung des Gradienten $Ax_k - (f - By_k)$ von h_k in x_k: Sei $(x_0, y_0) \in \mathbb{R}^{m+n}$. Für $k = 0, 1, \ldots$ setze

$$\begin{aligned} x_{k+1} &= x_k + \varepsilon(f - Ax_k - By_k), \\ y_{k+1} &= y_k + \omega(B^T x_{k+1} - g). \end{aligned}$$

4.5.2 Das Bramble-Pasciak-CG-Verfahren

Obwohl $\hat{A}$ indefinit ist, lässt sich ein Skalarprodukt angeben, bzgl. dessen diese Matrix nach geeigneter Transformation positiv-definit ist. Daher können die besseren Konvergenzeigenschaften des CG-Verfahrens genutzt werden.

Es sei $A_0 \in \mathbb{R}^{m\times m}$ eine positiv-definite Matrix mit

$$0 < x^T A_0 x < x^T A x, \quad x \neq 0.$$

Multiplikation von (4.32) von links mit

$$T := \begin{bmatrix} A_0^{-1} & 0 \\ B^T A_0^{-1} & -I \end{bmatrix} \in \mathbb{R}^{(m+n)\times(m+n)}$$

ergibt

$$T\hat{A}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} A_0^{-1}A & A_0^{-1}B \\ B^T A_0^{-1}(A - A_0) & B^T A_0^{-1}B + C \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} A_0^{-1}f \\ B^T A_0^{-1}f - g \end{bmatrix}. \tag{4.34}$$

Satz 4.31. *Die Systemmatrix $A' := T\hat{A}$ ist positiv-definit bzgl. des Skalarproduktes $(x,y)_M := x^T My$ mit*

$$M := \begin{bmatrix} A - A_0 & 0 \\ 0 & I \end{bmatrix}.$$

Beweis. Nach Voraussetzung ist M positiv-definit. Daher definiert $(\cdot,\cdot)_M$ tatsächlich ein Skalarprodukt. Wegen

$$\begin{aligned}\left(A'\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix}\right)_M &= \left(\begin{bmatrix} A_0^{-1}(Ax + By) \\ B^T A_0^{-1}(A - A_0)x + (B^T A_0^{-1}B + C)y \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix}\right)_M \\ &= (Ax + By)^T A_0^{-1}(A - A_0)u + (B^T A_0^{-1}(A - A_0)x + (B^T A_0^{-1}B + C)y)^T v \\ &= x^T(AA_0^{-1}A - A)u + y^T B^T A_0^{-1}(A - A_0)u + x^T(A - A_0)A_0^{-1}Bv + \\ &\quad + y^T(B^T A_0^{-1}B + C)v\end{aligned}$$

und

$$\begin{aligned}\left(\begin{bmatrix} x \\ y \end{bmatrix}, A'\begin{bmatrix} u \\ v \end{bmatrix}\right)_M &= \left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} A_0^{-1}(Au + Bv) \\ B^T A_0^{-1}(A - A_0)u + (B^T A_0^{-1}B + C)v \end{bmatrix}\right)_M \\ &= x^T(A - A_0)A_0^{-1}(Au + Bv) + y^T(B^T A_0^{-1}(A - A_0)u + (B^T A_0^{-1}B + C)v) \\ &= x^T(AA_0^{-1}A - A)u + x^T(A - A_0)A_0^{-1}Bv + y^T B^T A_0^{-1}(A - A_0)u + \\ &\quad + y^T(B^T A_0^{-1}B + C)v\end{aligned}$$

ist A' selbstadjungiert bzgl. $(\cdot,\cdot)_M$. Insbesondere ist

$$\left(A'\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix}\right)_M = ((A - A_0)x + By)^T A_0^{-1}((A - A_0)x + By) + x^T(A - A_0)x + y^T Cy.$$

Mit der Zerlegung

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_0 \\ 0 \end{bmatrix} + \begin{bmatrix} x_1 \\ y \end{bmatrix},$$

wobei x_1 die eindeutige Lösung von $Ax_1 = -By$ bezeichnet, ergibt sich

$$x_1^T A x_1 = y^T B^T A^{-1} B y.$$

Hieraus erhält man

$$\left(A' \begin{bmatrix} x_0 \\ 0 \end{bmatrix}, \begin{bmatrix} x_0 \\ 0 \end{bmatrix}\right)_M = x_0^T (A - A_0) A_0^{-1} (A - A_0) x_0 \geq \underbrace{\lambda_{\min}((A - A_0) A_0^{-1} (A - A_0))}_{=:c_1 > 0} \|x_0\|_2^2$$

und

$$\begin{aligned}\left(A' \begin{bmatrix} x_1 \\ y \end{bmatrix}, \begin{bmatrix} x_1 \\ y \end{bmatrix}\right)_M &= x_1^T A_0 x_1 + x_1^T (A - A_0) x_1 + y^T C y = \frac{1}{2} x_1^T A x_1 + y^T (\frac{1}{2} B^T A^{-1} B + C) y \\ &\geq \underbrace{\frac{1}{2} \lambda_{\min}(A)}_{=:c_2 > 0} \|x_1\|_2^2 + \underbrace{\frac{1}{2} \lambda_{\min}(B^T A^{-1} B)}_{=:c_3 > 0} \|y\|_2^2.\end{aligned}$$

Aus

$$\left(A' \begin{bmatrix} x_1 \\ y \end{bmatrix}, \begin{bmatrix} x_0 \\ 0 \end{bmatrix}\right)_M = (Ax_1 + By) A_0^{-1} (A - A_0) x_0 = 0$$

folgt

$$\left(A' \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix}\right)_M = \left(A' \begin{bmatrix} x_0 \\ 0 \end{bmatrix}, \begin{bmatrix} x_0 \\ 0 \end{bmatrix}\right)_M + \left(A' \begin{bmatrix} x_1 \\ y \end{bmatrix}, \begin{bmatrix} x_1 \\ y \end{bmatrix}\right)_M \geq c_1 \|x_0\|_2^2 + c_2 \|x_1\|_2^2 + c_3 \|y\|_2^2.$$

Mit $\|x\|_2^2 = \|x_0 + x_1\|_2^2 \leq 2\|x_0\|_2^2 + 2\|x_1\|_2^2$ ergibt sich wegen

$$\left(A' \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix}\right)_M \geq \frac{1}{2} \min\{c_1, c_2\} \|x\|_2^2 + c_3 \|y\|_2^2$$

die Behauptung. □

Das CG-Verfahren für (4.34) und dem Skalarprodukt $(\cdot, \cdot)_M$ kann so formuliert werden, dass nur Matrix-Vektor-Produkte mit A, B, C und die Anwendung von A_0^{-1} realisiert werden müssen.

Bemerkung. Ist A_0 spektraläquivalent zu A, d.h. gilt

$$\alpha\, x^T A x \leq x^T A_0 x \leq \beta\, x^T A x \quad \text{für alle } x \in \mathbb{R}^m$$

mit von m und n unabhängigen Konstanten $0 < \alpha \leq \beta$, so ist die Systemmatrix $MT\hat{A}$ spektraläquivalent zu

$$\begin{bmatrix} I & 0 \\ 0 & S \end{bmatrix}.$$

Daher benötigt man nur noch einen geeigneten Vorkonditionierer für das Schurkomplement S von A in $\hat{A}$.

5 Die Stokessche Gleichung

Zur Beschreibung der stationären Strömung einer inkompressiblen und extrem zähen Flüssigkeit (z.B. Honig) in einem beschränkten Gebiet $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, verwendet man das System der **Stokesschen Gleichungen**

$$-\Delta u + \nabla p = f \quad \text{in } \Omega, \tag{5.1a}$$
$$\operatorname{div} u = 0 \quad \text{in } \Omega, \tag{5.1b}$$
$$u = u_0 \quad \text{auf } \partial\Omega. \tag{5.1c}$$

Dabei bezeichnet $u : \Omega \to \mathbb{R}^d$ das gesuchte Geschwindigkeitsfeld und $p : \Omega \to \mathbb{R}$ den gesuchten Druck. Im Gegensatz zu weniger viskosen Fluiden treten hier keine mikroskaligen Wirbel auf. Das System ist linear und daher relativ einfach zu analysieren. Dennoch zeigen sich grundsätzliche numerische Schwierigkeiten.

Die Divergenz-Freiheit (5.1b) sichert die Massenerhaltung. Nach dem Gaußschen Divergenzsatz gilt nämlich für beliebiges $\omega \subset \Omega$

$$\int_{\partial\omega} u \cdot \nu \, \mathrm{d}s = \int_{\omega} \operatorname{div} u \, \mathrm{d}x = 0.$$

Durch den Rand $\partial\omega$ strömt also im Mittel so viel Flüssigkeit heraus wie einströmt. Insbesondere muss für die Dirichlet-Daten u_0 die Kompatibilitätsbedingung

$$\int_{\partial\Omega} u_0 \cdot \nu \, \mathrm{d}s = 0$$

gelten. Ferner fällt auf, dass der Druck p nicht eindeutig sein kann, weil nur dessen Gradient auftritt. Daher benötigt man eine weitere Normierungsbedingung. Eine mögliche Wahl ist

$$\int_{\Omega} p \, \mathrm{d}x = 0.$$

5.1 Variationsformulierung

Mittels partieller Integration erhält man

$$-(\Delta u, \phi)_{L^2(\Omega)} = (\nabla u, \nabla \phi)_{L^2(\Omega)} - (\nu \cdot \nabla u, \phi)_{L^2(\partial\Omega)},$$
$$(\nabla p, \phi)_{L^2(\Omega)} = -(p, \operatorname{div} \phi)_{L^2(\Omega)} + (\nu\, p, \phi)_{L^2(\partial\Omega)}.$$

Multipliziert man (5.1a) mit $\phi \in [C_0^\infty(\Omega)]^d$, so folgt hieraus

$$(\nabla u, \nabla \phi)_{L^2(\Omega)} - (\operatorname{div} \phi, p)_{L^2(\Omega)} = (f, \phi)_{L^2(\Omega)}.$$

Wie üblich dürfen wir von homogenen Randbedingungen $u_0 = 0$ ausgehen. Mit den Bilinearformen $a : [H_0^1(\Omega)]^d \times [H_0^1(\Omega)]^d \to \mathbb{R}$,

$$a(v, w) := (\nabla v, \nabla w)_{L^2(\Omega)} = \sum_{i,j=1}^{d} \int_{\Omega} \partial_j v_i \, \partial_j w_i \, \mathrm{d}x,$$

und $b : [H_0^1(\Omega)]^d \times L_0^2(\Omega) \to \mathbb{R}$,

$$b(v,q) := -(\operatorname{div} v, q)_{L^2(\Omega)},$$

wobei $L_0^2(\Omega) := \{q \in L^2(\Omega) : \int_\Omega q \,\mathrm{d}x = 0\}$ bezeichnet, erhält man die Variationsformulierung der Stokesschen Gleichung: suche $u \in [H_0^1(\Omega)]^d$ und $p \in L_0^2(\Omega)$, so dass

$$\begin{aligned} a(u,v) + b(v,p) &= (f,v)_{L^2(\Omega)}, \quad && v \in [H_0^1(\Omega)]^d, && \text{(5.2a)} \\ b(u,q) &= 0, && q \in L_0^2(\Omega). && \text{(5.2b)} \end{aligned}$$

Man beachte, dass es sich hierbei um $d+1$ Gleichungen handelt. Speziell für $d = 2$ erhalten wir ausgeschrieben

$$\begin{aligned} \int_\Omega \partial_1 u_1 \, \partial_1 v_1 + \partial_2 u_1 \, \partial_2 v_1 \,\mathrm{d}x - \int_\Omega p \, \partial_1 v_1 &= \int_\Omega f_1 v_1 \,\mathrm{d}x, \quad && v_1 \in H_0^1(\Omega), \\ \int_\Omega \partial_1 u_1 \, \partial_2 v_2 + \partial_2 u_2 \, \partial_2 v_2 \,\mathrm{d}x - \int_\Omega p \, \partial_2 v_2 &= \int_\Omega f_2 v_2 \,\mathrm{d}x, \quad && v_2 \in H_0^1(\Omega), \\ \int_\Omega (\partial_1 u_1 + \partial_2 u_2) q \,\mathrm{d}x &= 0, && q \in L_0^2(\Omega). \end{aligned}$$

Satz 5.1. *Für rechte Seiten $f \in [L^2(\Omega)]^d$ sind schwache Lösungen $(u,p) \in [H_0^1(\Omega)]^d \times L_0^2(\Omega)$ von (5.2) mit $u \in [C^2(\Omega)]^d$ und $p \in C^1(\Omega)$ klassische Lösungen von (5.1).*

Beweis. Sei $q := \operatorname{div} u \in L^2(\Omega)$. Dann gilt mit einer Konstante c, dass $q - c \in L_0^2(\Omega)$. Wegen (5.2b) gilt

$$\|\operatorname{div} u\|_{L^2(\Omega)}^2 = (\operatorname{div} u, q)_{L^2(\Omega)} = (\operatorname{div} u, c)_{L^2(\Omega)} = c \int_\Omega \operatorname{div} u \,\mathrm{d}x.$$

Wegen $u|_{\partial\Omega} = 0$ schließen wir aus dem Gaußschen Integralsatz

$$\int_\Omega \operatorname{div} u \,\mathrm{d}x = \int_{\partial\Omega} u \cdot \nu \,\mathrm{d}s = 0.$$

Also ist $\|\operatorname{div} u\|_{L^2(\Omega)} = 0$, und aus $u \in [C^2(\Omega)]^d$ folgt daher die Divergenz-Freiheit in jedem Punkt $x \in \Omega$.

Die erste Gleichung (5.1a) erhalten wir durch Zurückführen auf das Poisson-Problem. Für die schwache Lösung $u \in [H_0^1(\Omega)]^d$ gilt

$$(\nabla u, \nabla v)_{L^2(\Omega)} = g(v), \quad g(v) := (f,v)_{L^2(\Omega)} + (p, \operatorname{div} v)_{L^2(\Omega)}.$$

Somit ist u die klassische Lösung des Poisson-Problems

$$-\Delta u = g = f - \nabla p \quad \text{in } \Omega$$

mit homogenen Dirichlet-Randbedingungen. Dies ist aber gerade die gesuchte Gleichung. □

Bei (5.2) handelt es sich offenbar um ein Sattelpunktproblem mit einer auf ganz $[H_0^1(\Omega)]^d$ koerziven Bilinearform a. Um die Existenz von Lösungen von (5.2) zu garantieren, muss nach Satz 4.18 nur die LBB-Bedinung für die Bilinearform b nachgewiesen werden. Dies erfolgt über eine Abschätzung deren Beweis den Rahmen dieser Vorlesung sprengen würde.

Lemma 5.2. *Für beschränkte Lipschitz-Gebiete Ω gilt*

(i) Das Bild der linearen Abbildung $\nabla : L^2(\Omega) \to [H^{-1}(\Omega)]^d$ ist abgeschlossen.

(ii) Es existiert eine Konstante $c = c(\Omega)$ mit

$$\begin{aligned} \|p\|_{L^2(\Omega)} &\le c(\|\nabla p\|_{H^{-1}(\Omega)} + \|p\|_{H^{-1}(\Omega)}) \quad \text{für alle } p \in L^2(\Omega), \\ \|p\|_{L^2(\Omega)} &\le c\|\nabla p\|_{H^{-1}(\Omega)} \quad \text{für alle } p \in L^2_0(\Omega). \end{aligned} \tag{5.3}$$

Satz 5.3. *In beschränkten Lipschitz-Gebieten Ω besitzt das Stokes-Problem (5.2) für beliebiges $f \in ([H^1_0(\Omega)]^d)'$ eine eindeutige Lösung $(u, p) \in [H^1_0(\Omega)]^d \times L^2_0(\Omega)$.*

Beweis. Für Satz 4.18 müssen wir nur noch die LBB-Bedingung für b nachweisen. Für $p \in L^2_0(\Omega)$ folgt aus (5.3)

$$\|\nabla p\|_{H^{-1}(\Omega)} \ge \frac{1}{c}\|p\|_{L^2(\Omega)}.$$

Nach Definition der $H^{-1}(\Omega)$-Norm gibt es ein $v \in [H^1_0(\Omega)]^d$ mit $\|v\|_{H^1(\Omega)} = 1$ und

$$(v, \nabla p)_{L^2(\Omega)} \ge \frac{1}{2}\|v\|_{H^1(\Omega)}\|\nabla p\|_{H^{-1}(\Omega)} \ge \frac{1}{2c}\|p\|_{L^2(\Omega)}.$$

Wegen $b(v, p) = -(\operatorname{div} v, p)_{L^2(\Omega)} = (v, \nabla p)_{L^2(\Omega)}$ folgt

$$\frac{b(v, p)}{\|v\|_{H^1(\Omega)}} = (v, \nabla p)_{L^2(\Omega)} \ge \frac{1}{2c}\|p\|_{L^2(\Omega)}$$

und damit die LBB-Bedingung. □

5.2 Die diskrete LBB-Bedingung für Stokes

Wie wir bereits erwähnt haben, folgt aus der LBB-Bedingung (4.17) nicht automatisch die diskrete LBB-Bedingung (4.25) für diskrete Räume $V_h \times W_h$. Im Fall der Stokes-Gleichung besteht die Möglichkeit, dass ein nicht-trivialer Druck $q \in W_h$ im Kern von b existiert, d.h. es gilt

$$(\operatorname{div} v, q) = 0 \quad \text{für alle } v \in V_h.$$

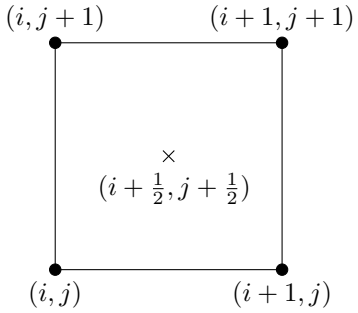
In diesem Fall bezeichnet man die Diskretisierung $V_h \times W_h$ als instabil. Eine schwächere Form der Instabilität liegt vor, wenn zwar kein $q \in W_h \setminus \{0\}$ im Kern von b liegt aber die LBB-Konstante von der Gitterweite h abhängt.

Instabile Stokes-Elemente

Ein wegen seiner Einfachheit beliebtes Element ist das sog. Q_1/P_0-Rechteckelement. Der Fluss wird hierbei mit stückweise bilinearen, der Druck mit stückweise konstanten Funktionen approximiert:

$$\begin{aligned} V_h &:= \{v \in [C(\overline{\Omega})]^2 : v|_{\partial\Omega} = 0 \text{ und } v|_\tau \text{ bilinear für alle } \tau \in \mathcal{T}_h\}, \\ W_h &:= \{q \in L^2_0(\Omega) : q|_\tau \in \Pi_0 \text{ für alle } \tau \in \mathcal{T}_h\}. \end{aligned}$$

Diese Diskretisierung ist instabil, weil der Kern von $B' : W_h \to V_h'$ nicht-trivial ist. Wir numerieren die Knoten im Element τ_{ij} wie im Bild dargestellt.



Weil auf jedem Element $\operatorname{div} v$ linear und q konstant ist, ergibt sich

$$\begin{aligned}\int_{\tau_{ij}} q \operatorname{div} v \,\mathrm{d}x &= h^2 q_{(i+\frac{1}{2},j+\frac{1}{2})} \operatorname{div} v_{(i+\frac{1}{2},j+\frac{1}{2})} \\ &= h^2 q_{(i+\frac{1}{2},j+\frac{1}{2})} \frac{1}{2h} \big(v_{1,(i+1,j+1)} + v_{1,(i+1,j)} - v_{1,(i,j+1)} - v_{1,(i,j)} + \\ &\quad + v_{2,(i+1,j+1)} + v_{2,(i,j+1)} - v_{2,(i+1,j)} - v_{2,(i,j)} \big) .\end{aligned}$$

Summation über die Elemente und Sortieren der Terme nach den Gitterpunkten liefern (vgl. auch die Formel der partiellen Summation vor Satz 1.16)

$$\int_\Omega q \operatorname{div} v \,\mathrm{d}x = h^2 \sum_{i,j} \left[v_{1,(i,j)} (\nabla_1 q)_{ij} + v_{2,(i,j)} (\nabla_2 q)_{ij} \right]$$

mit den Differenzenquotienten

$$\begin{aligned}(\nabla_1 q)_{ij} &:= \frac{1}{2h} \left(q_{(i+\frac{1}{2},j+\frac{1}{2})} + q_{(i+\frac{1}{2},j-\frac{1}{2})} - q_{(i-\frac{1}{2},j+\frac{1}{2})} - q_{(i-\frac{1}{2},j-\frac{1}{2})} \right) , \\ (\nabla_2 q)_{ij} &:= \frac{1}{2h} \left(q_{(i+\frac{1}{2},j+\frac{1}{2})} + q_{(i-\frac{1}{2},j+\frac{1}{2})} - q_{(i+\frac{1}{2},j-\frac{1}{2})} - q_{(i-\frac{1}{2},j-\frac{1}{2})} \right) .\end{aligned}$$

Wegen $v \in [H_0^1(\Omega)]^d$ erstreckt sich die Summation nur über innere Knoten. Es ist $q \in \operatorname{Ker} B_h'$, wenn

$$\int_\Omega q \operatorname{div} v \,\mathrm{d}x = 0 \quad \text{für alle } v \in V_h$$

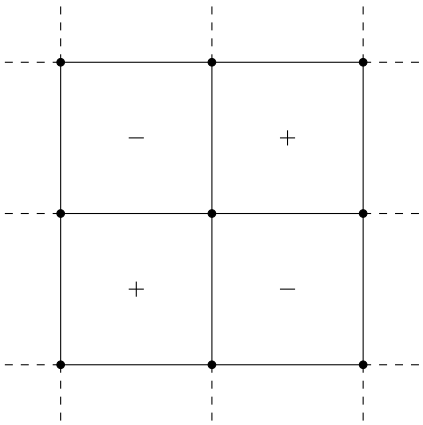
gilt, also $(\nabla_1 q)_{ij}$ und $(\nabla_2 q)_{ij}$ an allen inneren Knoten verschwindet. Die tritt ein, falls

$$q_{(i+\frac{1}{2},j+\frac{1}{2})} = q_{(i-\frac{1}{2},j-\frac{1}{2})}, \quad q_{(i+\frac{1}{2},j-\frac{1}{2})} = q_{(i-\frac{1}{2},j+\frac{1}{2})}.$$

Um diese beiden Gleichungen zu erfüllen, muss q nicht konstant sein. Es genügt

$$q_{(i+\frac{1}{2},j+\frac{1}{2})} = \begin{cases} a, & i+j \text{ gerade}, \\ b, & i+j \text{ ungerade}, \end{cases}$$

mit Zahlen a, b, so dass $\int_\Omega q \,\mathrm{d}x = 0$. Insbesondere haben a und b entgegengesetzte Vorzeichen, d.h. es bildet sich ein Schachbrett-Muster. Deshalb spricht man von **Schachbrettinstabilität**.



Bemerkung. Eine denkbare Abhilfe für die Problematik wäre die Einschränkung von V_h auf das orthogonale Komplement von $\operatorname{Ker} B_h'$. In diesem Fall ist der Kern zwar trivial, die Konstante in (4.24) hängt allerdings von h ab.

Das Taylor-Hood-Element

Bei dem oft verwendeten **Taylor-Hood-Element** werden wie bei den instabilen Q_1/P_0-Elementen für die Geschwindigkeit Polynome höheren Grades als als für den Druck herangezogen. Der Druck wird allerdings stetig angesetzt:

$$V_h := \{v \in [C(\overline{\Omega})]^d : v|_{\partial\Omega} = 0 \text{ und } v|_\tau \in [\Pi_2]^d \text{ für alle } \tau \in \mathcal{T}_h\},$$
$$W_h := \{q \in C(\overline{\Omega}) \cap L_0^2(\Omega) : q|_\tau \in \Pi_1 \text{ für alle } \tau \in \mathcal{T}_h\}.$$

Satz 5.4. *Sei $\mathcal{T}_h$ eine quasi-uniforme Triangulierung des polygonalen Gebiets $\Omega \subset \mathbb{R}^2$. Dann erfüllt das Taylor-Hood-Element die LBB-Bedingung (4.25).*

Beweis. Wir zeigen, dass zu jedem $q_h \in W_h$ ein $0 \neq v_h \in V_h$ existiert, so dass

$$\frac{b(v_h, q_h)}{\|v_h\|_{H^1(\Omega)}} \geq \tilde{c}_{\mathrm{LBB}} \|q_h\|_{L^2(\Omega)}.$$

Der Beweis wird in drei Schritten geführt.

(i) Zu jeder Kante e im Inneren von Ω bezeichne t_e den Tangentialvektor und ν_e die Normale. In jedem Kantenmittelpunkt m_e setzen wir

$$t_e \cdot v_h(m_e) = t_e \cdot (\nabla q_h)(m_e), \quad \nu_e \cdot v_h(m_e) = 0.$$

Ferner sei $v_h(x) := 0$ in den Eckpunkten aller Dreiecke und Kantenmittelpunkten auf $\partial\Omega$. Nach Konstruktion gilt dann

$$\|v_h\|_{L^2(\Omega)} \leq |q_h|_{H^1(\Omega)}. \tag{5.4}$$

Weil $(v_h \cdot \nabla q_h)|_\tau$ quadratisch und $(\nabla q_h)|_\tau$ konstant ist, gilt außerdem (vgl. Übungsaufgaben)

$$\int_\tau v_h \cdot \nabla q_h \,\mathrm{d}x = \frac{|\tau|}{3} \sum_{i=1}^{3} v_h(m_{e_i}) \cdot (\nabla q_h)|_\tau = \frac{|\tau|}{3} \sum_{i=1}^{3} [t_{e_i} \cdot (\nabla q_h)|_\tau]^2$$
$$\geq \frac{|\tau|}{3} c \|(\nabla q_h)|_\tau\|_2^2 = \frac{c}{3} \int_\tau \|(\nabla q_h)|_\tau\|_2^2 \,\mathrm{d}x$$

mit einer von der Entartetheit des Gitters $\mathcal{T}_h$ abhängigen Konstanten $c > 0$. Summation über alle $\tau \in \mathcal{T}_h$ ergibt

$$b(v_h, q_h) \geq \frac{c}{3} \sum_{\tau \in \mathcal{T}_h} \int_\tau \|(\nabla q_h)|_\tau\|_2^2 \, \mathrm{d}x = \frac{c}{3} |q_h|_{H^1(\Omega)}^2.$$

(ii) Zu $K > 0$ bezeichne

$$W_h^K := \{q_h \in W_h : \|q_h\|_{L^2(\Omega)} \leq Kh|q_h|_{H^1(\Omega)}\}.$$

Für $q_h \in W_h^K$ gilt mit dem in Schritt (i) konstruierten v_h

$$\frac{b(v_h, q_h)}{\|v_h\|_{H^1(\Omega)}} \geq \frac{c}{3} \frac{|q_h|_{H^1(\Omega)}^2}{\|v_h\|_{H^1(\Omega)}} \geq \frac{c}{3Kh} \frac{|q_h|_{H^1(\Omega)} \|q_h\|_{L^2(\Omega)}}{\|v_h\|_{H^1(\Omega)}} \geq \frac{c}{3Kc_{\mathrm{inv}}} \|q_h\|_{L^2(\Omega)}.$$

Dabei haben wir (5.4) und die inversen Ungleichung $\|v_h\|_{H^1(\Omega)} \leq c_{\mathrm{inv}} h^{-1} \|v_h\|_{L^2(\Omega)}$ (siehe Abschnitt 2.6, WissRech I) verwendet.

(iii) Sei nun $q_h \notin W_h^K$. Wegen der Gültigkeit der kontinuierlichen LBB-Bedingung (4.17) existiert $v \in [H_0^1(\Omega)]^2$ mit $\|v\|_{H^1(\Omega)} = 1$ und

$$b(v, q_h) \geq c_{\mathrm{LBB}} \|q_h\|_{L^2(\Omega)}.$$

Ferner bezeichne $0 \neq v_h \in V_h$ die Clément-Approximation von v, die nach Satz 2.83 (WissRech I) der Abschätzung

$$\|v - v_h\|_{L^2(\Omega)} \leq ch\|v\|_{H^1(\Omega)}$$

genügt. Dabei hängt die Konstante nicht von q_h und somit nicht von K ab. Insgesamt folgt

$$\begin{aligned} b(v_h, q_h) &= b(v, q_h) + b(v_h - v, q_h) \geq c_{\mathrm{LBB}} \|q_h\|_{L^2(\Omega)} + (v_h - v, \nabla q_h)_{L^2(\Omega)} \\ &\geq c_{\mathrm{LBB}} \|q_h\|_{L^2(\Omega)} - \|v - v_h\|_{L^2(\Omega)} |q_h|_{H^1(\Omega)} \\ &\geq c_{\mathrm{LBB}} \|q_h\|_{L^2(\Omega)} - ch\|v\|_{H^1(\Omega)} |q_h|_{H^1(\Omega)}. \end{aligned}$$

Wegen $q_h \notin W_h^K$ gilt $|q_h|_{H^1(\Omega)} < \|q_h\|_{L^2(\Omega)}/(Kh)$. Ferner ist

$$\|v_h\|_{H^1(\Omega)} \leq \|v\|_{H^1(\Omega)} + \|v - v_h\|_{L^2(\Omega)} \leq (1 + ch)\|v\|_{H^1(\Omega)},$$

und mit $\|v\|_{H^1(\Omega)} = 1$ folgt

$$\frac{b(v_h, q_h)}{\|v_h\|_{H^1(\Omega)}} \geq \frac{c_{\mathrm{LBB}} - c/K}{1 + ch} \|q_h\|_{L^2(\Omega)}.$$

Für hinreichend große K ist somit die LBB-Bedingung auch im Fall $q_h \notin W_h^K$ gezeigt. □

Satz 5.5. *Für das Taylor-Hood-Element gilt die Fehlerabschätzung*

$$\|u - u_h\|_{H^1(\Omega)} + \|p - p_h\|_{L^2(\Omega)} \leq ch^k \left\{\|u\|_{H^{k+1}(\Omega)} + \|p\|_{H^k(\Omega)}\right\}$$

mit $k \in \{1, 2\}$, falls $u \in H^{k+1}(\Omega)$ und $p \in H^k(\Omega)$ gilt.

Beweis. Wegen der Approximationsaussagen

$$\inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)} \leq ch^k \|u\|_{H^{k+1}(\Omega)}, \qquad \inf_{q_h \in W_h} \|p - q_h\|_{L^2(\Omega)} \leq ch^k \|p\|_{H^k(\Omega)}$$

für $k \in \{1, 2\}$ folgt die Behauptung aus Satz 4.22. □

Bemerkung. Anstatt quadratische Ansatzfunktionen für die Geschwindigkeit zu verwenden, können auch stückweise lineare Funktionen bei verfeinerter Triangulierung zugelassen werden:

$$\begin{aligned} V_h &:= \{v \in [C(\overline{\Omega})]^d : v|_{\partial\Omega} = 0 \text{ und } v|_\tau \in [\Pi_1]^d \text{ für alle } \tau \in \mathcal{T}_{h/2}\}, \\ W_h &:= \{q \in C(\overline{\Omega}) \cap L_0^2(\Omega) : q|_\tau \in \Pi_1 \text{ für alle } \tau \in \mathcal{T}_h\}. \end{aligned}$$

Dies liefert ebenfalls ein stabiles Element, und die Anzahl der Freiheitsgrade ist dieselbe wie beim Taylor-Hood-Element. Diese Variante bezeichnet man ebenfalls als Taylor-Hood-Element.

Das MINI-Element

Ein Nachteil beim Taylor-Hood-Element ist, dass sich die Werte für Druck und Geschwindigkeit auf verschieden feine Triangulierungen beziehen. Statt unterschiedlichen Triangulierungen ist das sog. **MINI-Element** ein P_1/P_1-Element mit durch Blasenfunktionen

$$b_\tau(x) := 9\lambda_{\tau,1}\lambda_{\tau,2}\lambda_{\tau,3},$$

angereichertem Geschwindigkeitsraum, d.h.

$$\begin{aligned} V_h &:= \{v \in [C(\overline{\Omega})]^2 : v|_{\partial\Omega} = 0 \text{ und } v|_\tau \in [\Pi_1 \oplus \operatorname{span}\{b_\tau\}]^2 \text{ für alle } \tau \in \mathcal{T}_h\}, \\ &= [\mathcal{S}_0^{1,0}(\mathcal{T}_h) \oplus B_3]^2 \end{aligned}$$

mit $B_3 := \{v \in C(\overline{\Omega}) : v|_\tau \in \operatorname{span}\{b_\tau\} \text{ für alle } \tau \in \mathcal{T}_h\}$ und

$$W_h := \{q \in C(\overline{\Omega}) \cap L_0^2(\Omega) : q|_\tau \in \Pi_1 \text{ für alle } \tau \in \mathcal{T}_h\}.$$

Dabei bezeichnet $(\lambda_{\tau,1}, \lambda_{\tau,2}, \lambda_{\tau,3})$ die baryzentrischen Koordinaten von x auf dem Dreieck $\tau \in \mathcal{T}_h$. Diese fallen im Einheitsdreieck mit x_1, x_2 und $1 - x_1 - x_2$ zusammen. Man beachte, dass $b_\tau|_{\partial\tau} = 0$.

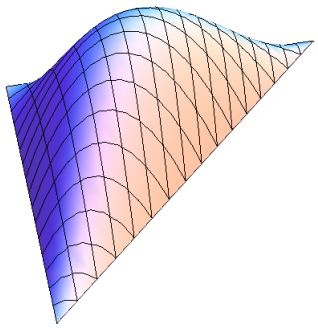


Abbildung 5.1: Blasenfunktion für das Einheitsdreieck.

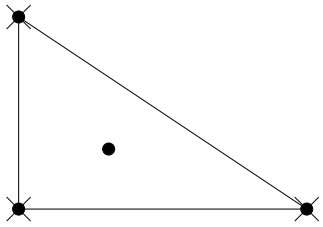


Abbildung 5.2: MINI-Element (v ist an den Punkten, q an den Kreuzen vorgegeben).

Satz 5.6. *Sei Ω ein polygonales Gebiet im $\mathbb{R}^2$. Dann erfüllt das MINI-Element die LBB-Bedingung.*

Beweis. Wir weisen die Gültigkeit des Kriteriums von Fortin nach, d.h. wir zeigen die Existenz einer Projektion $\Pi_h : [H_0^1(\Omega)]^2 \to V_h$ mit der Orthogonalitätseigenschaft

$$(\operatorname{div}(v - \Pi_h v), q_h)_{L^2(\Omega)} = 0 \quad \text{für alle } v \in [H_0^1(\Omega)]^2 \text{ und } q_h \in W_h$$

und der Beschränktheit mit einer h-unabhängigen Konstanten $c_\Pi > 0$. Sei dazu $\Pi_h^{(0)} : H_0^1(\Omega) \to \mathcal{S}_0^{1,0}(\mathcal{T}_h)$ der Clément-Operator. Für diesen gilt

$$\|v - \Pi_h^{(0)} v\|_{L^2(\Omega)} \le c_1 h \|v\|_{H^1(\Omega)}.$$

und $\|\Pi_h^{(0)} v\|_{H^1(\Omega)} \le c_2 \|v\|_{H^1(\Omega)}$. Wir definieren ferner $\Pi_h^{(1)} : L^2(\Omega) \to B_3$ durch

$$\Pi_h^{(1)} v = \sum_{\tau \in \mathcal{T}_h} \beta_\tau b_\tau \quad \text{mit} \quad \beta_\tau := \left(\int_\tau b_\tau \,\mathrm{d}x \right)^{-1} \int_\tau v \,\mathrm{d}x.$$

Die Projektion $\Pi_h^{(1)}$ kann man somit als zweistufigen Prozess interpretieren. Zunächst erfolgt die L^2-Projektion auf die stückweise konstanten Funktionen. Anschließend werden die Konstanten durch Blasenfunktionen mit gleichem Integral ersetzt. Hieraus folgt

$$\|\Pi_h^{(1)} v\|_{L^2(\Omega)} \le c_3 \|v\|_{L^2(\Omega)}.$$

Wir setzen nun $\Pi_h v := \Pi_h^{(0)} v + \Pi_h^{(1)}(v - \Pi_h^{(0)} v)$. Da $(\Pi_h^{(1)} v)|_\tau$ den Mittelwert von $v|_\tau$ erhält, gilt mit $\tilde{v} := v - \Pi_h^{(0)} v$

$$\int_\tau v - \Pi_h v \,\mathrm{d}x = \int_\tau \tilde{v} - \Pi_h^{(1)} \tilde{v} \,\mathrm{d}x = 0 \quad \text{für alle } \tau \in \mathcal{T}_h.$$

Hieraus erhält man mit partieller Integration

$$\begin{aligned}(\operatorname{div}(v - \Pi_h v), q_h)_{L^2(\Omega)} &= \sum_{\tau \in \mathcal{T}_h} (\operatorname{div}(v - \Pi_h v), q_h)_{L^2(\tau)} \\ &= \sum_{\tau \in \mathcal{T}_h} (v - \Pi_h v, \nabla q_h)_{L^2(\tau)} - \int_{\partial\tau} (v - \Pi_h v) \cdot \nu \, q_h \,\mathrm{d}s.\end{aligned}$$

Die Integrale über die inneren Kanten verschwinden wegen des wechselnden Vorzeichens der Normalen. Die Integrale entlang der äußeren Kanten verschwinden wegen $v|_{\partial\Omega} = 0 = (\Pi_h v)|_{\partial\Omega}$. Weil ∇q_h stückweise konstant ist, ergibt sich insgesamt

$$b(v - \Pi_h v, q_h) = (\operatorname{div}(v - \Pi_h v), q_h)_{L^2(\Omega)} = \sum_{\tau \in \mathcal{T}_h} (\nabla q_h)|_\tau \int_\tau v - \Pi_h v \, \mathrm{d}x = 0.$$

Wir dehnen nun die Definition von Π_h auf vektorwertige Funktionen $v \in [H_0^1(\Omega)]^2$ durch komponentenweise Anwendung von Π_h aus. Mit Hilfe der inversen Abschätzung erhalten wir aus der Stetigkeit von $\Pi_h^{(0)}$ und $\Pi_h^{(1)}$

$$\begin{aligned}
\|\Pi_h v\|_{H^1(\Omega)} &\leq \|\Pi_h^{(0)} v\|_{H^1(\Omega)} + c_{\mathrm{inv}} h^{-1} \|\Pi_h^{(1)}(v - \Pi_h^{(0)} v)\|_{L^2(\Omega)} \\
&\leq c_2 \|v\|_{H^1(\Omega)} + c_3 c_{\mathrm{inv}} h^{-1} \|v - \Pi_h^{(0)} v\|_{L^2(\Omega)} \\
&\leq (c_2 + c_1 c_3 c_{\mathrm{inv}}) \|v\|_{H^1(\Omega)}.
\end{aligned}$$

Lemma 4.23 liefert schließlich die Behauptung. □

Satz 5.7. *Für das MINI-Element gilt die Fehlerabschätzung*

$$\|u - u_h\|_{H^1(\Omega)} + \|p - p_h\|_{L^2(\Omega)} \leq ch \left\{ \|u\|_{H^2(\Omega)} + \|p\|_{H^1(\Omega)} \right\},$$

falls $u \in H^2(\Omega)$ und $p \in H^1(\Omega)$ gilt.

Beweis. Wegen der Approximationsaussagen

$$\inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)} \leq ch \|u\|_{H^2(\Omega)}, \qquad \inf_{q_h \in W_h} \|p - q_h\|_{L^2(\Omega)} \leq ch \|p\|_{H^1(\Omega)}$$

folgt die Behauptung aus Satz 4.22. □

5.2.1 Lösung der diskreten Probleme

Die Variationsformulierung der Stokesschen Gleichung führt als Sattelpunktproblem auf lineare Gleichungssysteme der Form (4.32). Zur Lösung dieser hatten wir das Uzawa- und das Bramble-Pasciak-CG-Verfahren vorgestellt. Das Uzawa-Verfahren war äquivalent zur Richardson-Iteration für das Schur-Komplement, und beim Bramble-Pasciak-CG-Verfahren ist die Kondition des Schur-Komplements für die Konvergenz ausschlaggebend.

Das folgende Lemma zeigt, dass die Kondition des Schur-Komplements $S = B^T A B$ im Fall der Stokesschen Gleichung beschränkt ist. Dazu sei V_h ein m-dimensionaler und W_h ein n-dimensionaler Raum zur Diskretisierung von V bzw. W. Wir wissen bereits, dass die Massenmatrix $M \in \mathbb{R}^{n \times n}$ zu W_h gut konditioniert ist.

Lemma 5.8. *Es gelte die diskrete LBB-Bedingung (4.25). Dann gilt für den maximalen und den minimalen Eigenwert von $M^{-1}S$*

$$0 < \tilde{c}_{\mathrm{LBB}}^2 \leq \lambda_{\min} \leq \lambda_{\max} \leq d.$$

Beweis. Der kleinste Eigenwert von $M^{-1}S$ lässt sich beschränken durch

$$\begin{aligned}\lambda_{\min}^2 &\geq \min_{y\neq 0} \frac{\|M^{-1}S\|_M^2}{\|y\|_M^2} = \min_{y\neq 0} \frac{(Sy, M^{-1}Sy)}{\|y\|_M^2} = \min_{y\neq 0} \frac{(Sy, M^{-1}Sy)}{(Sy,y)} \frac{(Sy,y)}{\|y\|_M^2} \\ &= \lambda_{\min} \min_{y\neq 0} \frac{(Sy,y)}{\|y\|_M^2}.\end{aligned}$$

Für den Zähler gilt mit $x := A^{-1}By$ wegen $\|x\|_A = \max_{z\neq 0}(x,z)_A/\|z\|_A$

$$\begin{aligned}(Sy,y) = (B^TA^{-1}By, y) = (A^{-1}By, By) = (x, Ax) = \|x\|_A^2 \\ \geq \max_{z\neq 0} \frac{(Ax,z)^2}{\|z\|_A^2} = \max_{z\neq 0} \frac{(By,z)^2}{\|z\|_A^2}.\end{aligned}$$

Wir erhalten zusammen

$$\lambda_{\min} \geq \min_{y\neq 0} \max_{z\neq 0} \frac{(By,z)^2}{\|z\|_A^2\|y\|_M^2} = \left(\min_{y\neq 0} \max_{z\neq 0} \frac{(y, B^Tz)}{\|z\|_A\|y\|_M}\right)^2.$$

Für $v_h = J_h z$ und $q_h = J_h y$ gilt

$$\|z\|_A = \|\nabla v_h\|_{L^2(\Omega)}, \quad \|y\|_M = \|q_h\|_{L^2(\Omega)} \quad \text{und} \quad (y, B^Tz) = (\operatorname{div} v_h, q_h)_{L^2(\Omega)}.$$

Daher erhalten wir mit der diskreten LBB-Bedingung (4.25)

$$\lambda_{\min} \geq \left(\min_{q_h\in W_h} \max_{v_h\in V_h} \frac{(\operatorname{div} v_h, q_h)_{L^2(\Omega)}}{\|\nabla v_h\|_{L^2(\Omega)}\|q_h\|_{L^2(\Omega)}}\right)^2 \geq \tilde{c}_{\mathrm{LBB}}^2.$$

Analog erhält man

$$\lambda_{\max} \leq \left(\max_{q_h\in W_h} \max_{v_h\in V_h} \frac{(\operatorname{div} v_h, q_h)_{L^2(\Omega)}}{\|\nabla v_h\|_{L^2(\Omega)}\|q_h\|_{L^2(\Omega)}}\right)^2 \leq \left(\max_{v_h\in V_h} \frac{\|\operatorname{div} v_h\|_{L^2(\Omega)}}{\|\nabla v_h\|_{L^2(\Omega)}}\right)^2.$$

Die Behauptung folgt wegen der Cauchy-Schwarzschen Ungleichung aus

$$\|\operatorname{div} v_h\|_{L^2(\Omega)}^2 = \int_\Omega \left(\sum_{i=1}^d \partial_i v_h\right)^2 \mathrm{d}x \leq \int_\Omega d\sum_{i=1}^d (\partial_i v_h)^2 \,\mathrm{d}x \leq d\|\nabla v_h\|_{L^2(\Omega)}^2.$$

□

6 Eigenwertprobleme

Wie bei Matrizen bezeichnet man für elliptische Differentialoperatoren zweiter Ordnung $\mathcal{L}$

$$\mathcal{L}u = \lambda u \text{ in } \Omega, \quad u = 0 \text{ auf } \partial\Omega, \tag{6.1}$$

als **Eigenwertproblem**. Jede Lösung $u \neq 0$ heißt **Eigenfunktion**, und das zugehörige λ wird als **Eigenwert** bezeichnet.

Beispiel 6.1. Auf die Wellengleichung

$$\partial_t^2 u - \Delta u = f \text{ in } (0,T) \times \Omega, \quad u = 0 \text{ auf } (0,T) \times \partial\Omega,$$

wobei zum Zeitpunkt $t = 0$ Anfangsbedingungen

$$u(0,\cdot) = g, \quad \partial_t u(0,\cdot) = h \quad \text{in } \Omega$$

gestellt sind, kann für $f = 0$ der Separationsansatz $u(x,t) = v(t)w(x)$ angewendet werden. Dies liefert

$$v''(t)w(x) = v(t)\Delta w(x) \iff \frac{v''(t)}{v(t)} = \frac{\Delta w(x)}{w(x)} = -\lambda = \text{const.}$$

Beide Gleichungen $-v'' = \lambda v$ und $-\Delta w = \lambda w$ sind Eigenwertprobleme. Die erste Gleichung ist eine gewöhnliche Differentialgleichung, deren Lösungen bekannt sind, d.h. v besitzt die Form

$$v(t) = \alpha \cos(\sqrt{\lambda}\, t) + \beta \sin(\sqrt{\lambda}\, t), \quad \alpha, \beta \in \mathbb{R},$$

während w im Allgemeinen nicht explizit bestimmt werden kann.

Zur numerischen Behandlung des Randwertproblems (6.1) ersetzen wir es durch das Variationsproblem

$$(\lambda, u) \in \mathbb{C} \times V: \quad a(u,v) = \lambda\,(u,v)_{L^2(\Omega)} \quad \text{für alle } v \in V \tag{6.2}$$

mit der Bilinearform $a : V \times V \to \mathbb{R}$ definiert durch $a(v,w) := (\mathcal{L}v, w)_{L^2(\Omega)}$.

6.1 Spektraltheorie

Wir untersuchen das Eigenwertproblem in einer allgemeinen Situation. Sei $(W, (\cdot,\cdot)_W)$ ein Hilbert-Raum und $V \subset W$ ein Unterraum mit $(V, (\cdot,\cdot)_V) \hookrightarrow (W, (\cdot,\cdot)_W)$. Ferner sei $a : V \times V \to \mathbb{R}$ eine stetige und V-koerzive Bilinearform.

Bemerkung. Für Dirichlet-Probleme von Differentialgleichungen zweiter Ordnung ist $V = H_0^1(\Omega)$ und $W = L^2(\Omega)$.

Es bezeichne $\mathcal{S} : W \to V$ den Lösungsoperator des Problems

$$u \in V : \quad a(u, v) = (f, v)_W \quad \text{für alle } v \in V$$

zu vorgegebenem $f \in W$, d.h. $u = \mathcal{S}f$. Offensichtlich ist $\mathcal{S}$ linear und beschränkt, weil

$$\|\mathcal{S}f\|_V^2 = \|u\|_V^2 \leq \frac{1}{c_K} a(u, u) = \frac{1}{c_K}(f, u)_W \leq \frac{1}{c_K}\|f\|_W\|u\|_W \leq \frac{c}{c_K}\|f\|_W\|u\|_V.$$

Daher ist das Eigenwertproblem $a(u, v) = \lambda\,(u, v)_W$, $v \in V$, äquivalent zu

$$(\lambda, u) \in \mathbb{C} \times V : \quad u = \lambda\,\mathcal{S}u. \tag{6.3}$$

Lemma 6.2. *Die Einbettung $V \hookrightarrow W$ sei kompakt und $a : V \times V \to \mathbb{R}$ symmetrisch und V-koerziv. Dann ist der Lösungsoperator $\mathcal{S} : V \to V$ kompakt, selbstadjungiert bzgl. des Skalarproduktes a, d.h.*

$$a(\mathcal{S}v, w) = a(v, \mathcal{S}w) \quad \textit{für alle } v, w \in V,$$

und positiv, d.h.

$$a(\mathcal{S}v, v) > 0 \quad \textit{für alle } 0 \neq v \in V.$$

Beweis. Weil $\mathcal{S} : W \to V$ stetig und $V \hookrightarrow W$ kompakt ist, ist $\mathcal{S} : V \to V$ kompakt. Aus der Definition von $\mathcal{S}$ folgt

$$a(\mathcal{S}v, w) = (v, w)_W = (w, v)_W = a(\mathcal{S}w, v) \quad \text{für alle } v, w \in V.$$

Die Symmetrie der Bilinearform a impliziert daher $a(\mathcal{S}v, w) = a(v, \mathcal{S}w)$ für alle $v, w \in V$. Die letzte Aussage erhält man aus

$$a(\mathcal{S}v, v) = (v, v)_W = \|v\|_W^2 > 0 \quad \text{für alle } 0 \neq v \in V.$$

□

Satz 6.3 (Spektralsatz). *Es sei W ein Hilbert-Raum und $V \subset W$ dicht. Die Einbettung $V \hookrightarrow W$ sei kompakt und $a : V \times V \to \mathbb{R}$ symmetrisch und V-koerziv. Dann existieren abzählbar viele reelle Eigenwerte*

$$0 < \lambda_1 \leq \lambda_2 \leq \dots$$

mit zugehörigen Eigenvektoren $\{u_k\}_{k\in\mathbb{N}} \subset V$ zum Eigenwertproblem

$$(\lambda, u) \in \mathbb{C} \times V : \quad a(u, v) = \lambda\,(u, v)_W \quad \textit{für alle } v \in V. \tag{6.4}$$

Dabei ist $+\infty$ der einzige Häufungspunkt der Folge $\{\lambda_k\}_{k\in\mathbb{N}}$. Die Folge der Eigenvektoren $\{u_k\}$ bildet eine Orthonormalbasis von W, während $\{\lambda_k^{-1/2} u_k\}$ eine Orthonormalbasis von V bzgl. des Skalarproduktes a bildet.

Beweis. Anstelle des Eigenwertproblems (6.4) betrachten wir (6.3), d.h.

$$(\mu, u) \in \mathbb{C} \times V : \quad \mathcal{S}u = \mu u.$$

Da $\mathcal{S} : V \to V$ selbstadjungiert, positiv und kompakt ist, folgt aus dem Spektralsatz für solche Operatoren, dass eine abzählbare Folge von positiven Eigenwerten $\{\mu_k\}_{k\in\mathbb{N}}$ mit zugehörigen Eigenvektoren $\{v_k\}_{k\in\mathbb{N}} \subset V$ existiert. Die Eigenwerte $\{\mu_k\}_{k\in\mathbb{N}}$ häufen sich höchstens bei 0, die Eigenvektoren $\{v_k\}_{k\in\mathbb{N}}$ bilden eine Orthonormalbasis von V bzgl. des Skalarproduktes a.

Mit $\lambda_k := 1/\mu_k$ erhalten wir

$$a(v_k, v) = \lambda_k\, a(\mathcal{S}v_k, v) = \lambda_k\, (v_k, v)_W \quad \text{für alle } v \in V.$$

Daraus ergibt sich für $u_k := \sqrt{\lambda_k}\, v_k$

$$(u_k, u_\ell)_W = \frac{1}{\lambda_k} a(u_k, u_\ell) = \sqrt{\frac{\lambda_\ell}{\lambda_k}}\, a(v_k, v_\ell) = \delta_{k\ell}.$$

Daher ist $\{u_k\}$ orthonormal in W bzgl. des Skalarproduktes $(\cdot,\cdot)_W$. Wir zeigen, dass $\{u_k\}$ auch eine Basis von W ist. Sei dazu das Gegenteil angenommen, d.h. es existiert ein $0 \neq f \in W$ mit $(f, u_k)_W = 0$ für alle $k \in \mathbb{N}$. Da v_k aber eine Orthonormalbasis in V bzgl. des Skalarproduktes a ist, erhält man

$$\lim_{n\to\infty} \|v - \sum_{k=1}^{n} a(v, v_k)\, v_k\|_W = 0 \quad \text{für alle } v \in V.$$

Hieraus folgt für alle $v \in V$

$$|(f, v)_W| \leq |\sum_{k=1}^{n} a(v, v_k) \underbrace{(f, v_k)_W}_{=0}| + \|f\|_W \|v - \sum_{k=1}^{n} a(v, v_k)\, v_k\|_W \to 0 \quad \text{für } n \to \infty.$$

Aus der Dichtheit von V in W schließen wir $(f, v)_W = 0$ für alle $v \in W$ und somit $\|f\|_W = 0$, was den Widerspruch liefert. □

In der Vorlesung *Einführung in die Numerik* (Kapitel 3) haben wir den Satz von Courant-Fischer für Matrizen bewiesen. Dieser gilt auch allgemeiner. Wir betrachten wieder den **Rayleigh-Quotienten**

$$R(v) := \frac{a(v, v)}{(v, v)_W} \quad \text{für alle } v \in V.$$

Die Eigenpaare (λ_k, u_k) von (6.4) erfüllen offenbar $R(u_k) = \lambda_k$, $k \in \mathbb{N}$. 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 $U_m := \operatorname{span}\{u_k : 1 \leq k \leq m\} \subset V$ und

$$U_m^\perp = \{v \in V : a(v, w) = 0 \text{ für alle } w \in U_m\} = \{v \in V : a(v, u_k) = 0,\ 1 \leq k \leq m\}.$$

Für $v = \sum_{k=1}^{\infty} \alpha_k u_k \in U_{m-1}^\perp$ folgt $\alpha_k = 0$ für $1 \leq k < m$ und

$$R(v) = \frac{\sum_{k=m}^{\infty} \lambda_k \alpha_k^2}{\sum_{k=m}^{\infty} \alpha_k^2} \geq \lambda_m. \tag{6.5}$$

Insbesondere ist $\lambda_m = \min_{0\neq v\in U_{m-1}^\perp} R(v)$.

Satz 6.4 (Courant-Fischer). *Unter den Voraussetzungen von Satz 6.3 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 $0 \neq v \in V_m \cap U_{m-1}^{\perp}$. Wegen (6.5) folgt $R(v) \geq \lambda_m$ und somit

$$\max_{0 \neq v \in V_m} R(v) \geq \lambda_m.$$

Im Fall $V_m = U_m$ gilt sogar Gleichheit, weil 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 $R(u_m) = \lambda_m$ erhalten wir daher $\max_{0 \neq v \in U_m} R(v) = \lambda_m$. □

6.2 Finite-Elemente-Approximation

Sei $V_h \subset V$ ein n-dimensionaler Teilraum. Dann lautet die Galerkin-Approximation von (6.4)

$$(\lambda_h, u_h) \in \mathbb{C} \times V_h : \quad a(u_h, v_h) = \lambda_h \, (u_h, v_h)_W \quad \text{für alle } v \in V_h. \tag{6.6}$$

Satz 6.5. *Unter den Voraussetzungen von Satz 6.3 existiert eine Orthonormalbasis $u_{h,k}$ von V_h in W und Werte $\lambda_{h,k} > 0$, so dass $(\lambda_{h,k}, u_{h,k})$, $k = 1, \ldots, n$, das diskrete Eigenwertproblem (6.6) lösen.*

Beweis. Sei $V_h = \operatorname{span}\{\varphi_k,\, k = 1, \ldots, n\}$ und $u_h = J_h x = \sum_{k=1}^{n} x_k \varphi_k$. Dann ist (6.6) äquivalent zu

$$\sum_{k=1}^{n} x_k \, a(\varphi_k, \varphi_\ell) = \lambda_h \sum_{k=1}^{n} x_k \, (\varphi_k, \varphi_\ell)_W, \quad \ell = 1, \ldots, n \quad \Longleftrightarrow \quad A_h x = \lambda_h M_h x$$

mit $A_h = [a(\varphi_k, \varphi_\ell)]_{k,\ell=1,\ldots,n}$ und $M_h := [(\varphi_k, \varphi_\ell)_W]_{k,\ell=1,\ldots,n}$. Mit der Cholesky-Zerlegung der Massenmatrix $M_h = LL^T$ ist dieses Problem äquivalent zu

$$L^{-1} A_h L^{-T} \tilde{x} = \lambda_h \tilde{x}, \quad \tilde{x} := L^T x.$$

Weil $\tilde{A}_h := L^{-1} A_h L^{-T}$ positiv-definit ist, existieren n Eigenwerte

$$0 < \lambda_{h,1} \leq \lambda_{h,2} \leq \ldots \leq \lambda_{h,n}$$

mit zugehörigen Eigenvektoren $\tilde{x}_k \in \mathbb{R}^n$, die eine Orthonormalbasis des $\mathbb{R}^n$ bilden. Die entsprechenden Eigenfunktionen $u_{h,k} := J_h L^{-T} \tilde{x}_k \subset V_h$ sind wegen

$$(u_{h,k}, u_{h,\ell})_W = (L^{-T} \tilde{x}_k)^T M_h L^{-T} \tilde{x}_\ell = \tilde{x}_k^T L^{-1} M_h L^{-T} \tilde{x}_\ell = \tilde{x}_k^T \tilde{x}_\ell = \delta_{k\ell}$$

orthonormal in W. □

Wir werden im Folgenden die Konvergenz der diskreten Eigenwerte $\lambda_{h,k}$ gegen die exakten Eigenwerte λ_k von (6.4) untersuchen. Dazu benötigen wir zwei Lemmata.

Lemma 6.6. *Es bezeichne $P_h : V \to V_h$ die Galerkin-Projektion definiert durch*

$$a(u - P_h u, v_h) = 0 \quad \text{für alle } v_h \in V_h,$$

und es sei

$$\sigma_{h,m} := \min_{\substack{v \in U_m \\ \|v\|_W = 1}} \|P_h v\|_W.$$

Dann gilt

$$\lambda_m \le \lambda_{h,m} \le \sigma_{h,m}^{-2} \lambda_m, \quad m = 1, \dots, n,$$

falls $\sigma_{h,m} > 0$.

Beweis. Im Fall $\sigma_{h,m} > 0$ ist $\operatorname{Ker} P_h = \{0\}$ und somit $\dim P_h U_m = m$. Es folgt

$$\lambda_{h,m} \le \max_{0 \ne v \in P_h U_m} R(v) = \max_{0 \ne v \in U_m} \frac{a(P_h v, P_h v)}{\|P_h v\|_W^2} \le \max_{0 \ne v \in U_m} \frac{a(v,v)}{\|P_h v\|_W^2}.$$

Die letzte Abschätzung $a(P_h v, P_h v) \le a(v,v)$ gilt, weil $P_h v$ die orthogonale Projektion von v auf V_h bzgl. des Skalarproduktes a ist. Wegen

$$\lambda_m \ge \max_{0 \ne v \in U_m} R(v) = \max_{0 \ne v \in U_m} \frac{a(v,v)}{\|v\|_W^2}$$

erhält man hieraus

$$\lambda_{h,m} \le \max_{0 \ne v \in U_m} \frac{a(v,v)}{\|P_h v\|_W^2} \le \lambda_m \max_{0 \ne v \in U_m} \frac{\|v\|_W^2}{\|P_h v\|_W^2}.$$

Die untere Schranke $\lambda_m \le \lambda_{h,m}$ erhält man sofort aus dem Satz von Courant-Fischer. □

Lemma 6.7. *Für jedes $m \ge 1$ existiert eine Konstante $c = c(m) > 0$, so dass*

$$\sigma_{h,m}^2 \ge 1 - c(m) \max_{\substack{v \in U_m \\ \|v\|_W = 1}} \|v - P_h v\|_V^2.$$

Beweis. Für $w = \sum_{k=1}^m \alpha_k u_k \in U_m$ mit $\|w\|_W = 1$ gilt

$$\begin{aligned} 1 - \|P_h w\|_W^2 &= (w,w)_W - (P_h w, P_h w)_W = (w - P_h w, w + P_h w)_W \\ &= 2(w - P_h w, w)_W - \|w - P_h w\|_W^2 \le 2(w - P_h w, w)_W. \end{aligned}$$

Wegen $a(u_k, v) = \lambda_k (u_k, v)_W$ für alle $v \in V$ ergibt sich

$$(w - P_h w, w)_W = \sum_{k=1}^m \alpha_k (w - P_h w, u_k)_W = \sum_{k=1}^m \frac{\alpha_k}{\lambda_k} a(w - P_h w, u_k)$$

und mit $w - P_h w \perp V_h$ folgt weiter

$$(w - P_h w, w)_W = \sum_{k=1}^m \frac{\alpha_k}{\lambda_k} a(w - P_h w, u_k - P_h u_k).$$

Die Stetigkeit der Bilinearform liefert

$$\begin{aligned}|(w - P_h w, w)_W| &\leq c_S \|w - P_h w\|_V \left(\sum_{k=1}^{m} \frac{|\alpha_k|}{\lambda_k} \|u_k - P_h u_k\|_V \right) \\ &\leq c_S \|w - P_h w\|_V \left(\sum_{k=1}^{m} \frac{\alpha_k^2}{\lambda_k^2} \right)^{1/2} \left(\sum_{k=1}^{m} \|u_k - P_h u_k\|_V^2 \right)^{1/2} \\ &\leq c_S \frac{\sqrt{m}}{\lambda_1} \|w - P_h w\|_V \max_{\substack{v \in U_m \\ \|v\|_W = 1}} \|v - P_h v\|_V.\end{aligned}$$

Hieraus folgt die Behauptung mit $c(m) := 2c_S\sqrt{m}/\lambda_1$. □

Satz 6.8 (Konvergenz approximativer Eigenwerte). *Es gelten die Annahmen von Satz 6.3, und es seien $V_h \subset V$ mit $\dim V_h \geq m$ so gewählt, dass $\lim_{h\to 0} \inf_{v_h \in V_h} \|v - v_h\|_V = 0$ für alle $v \in V$. Dann existiert zu jedem $m \in \mathbb{N}$ ein $h_m > 0$ derart, dass*

$$0 \leq \lambda_{h,m} - \lambda_m \leq 2c(m) \max_{\substack{v \in U_m \\ \|v\|_W = 1}} \inf_{v_h \in V_h} \|v - v_h\|_V^2$$

für alle $0 < h \leq h_m$.

Beweis. Für jedes $v = \sum_{k=1}^{m} \alpha_k u_k \in U_m$ mit $\|v\|_W = 1$ gilt

$$\|v - P_h v\|_V \leq \sum_{k=1}^{m} |\alpha_k| \|u_k - P_h u_k\|_V \leq \underbrace{\left(\sum_{k=1}^{m} \alpha_k^2 \right)^{1/2}}_{=1} \left(\sum_{k=1}^{m} \|u_k - P_h u_k\|_V^2 \right)^{1/2}.$$

Für die Galerkin-Approximation gilt aufgrund des Céa-Lemmas für $k = 1, \ldots, m$

$$\|u_k - P_h u_k\|_V \leq \frac{c_S}{c_K} \inf_{v_h \in V_h} \|u_k - v_h\|_V$$

und somit

$$\max_{\substack{v \in U_h \\ \|v\|_W = 1}} \|v - P_h v\|_V \leq \sqrt{m} \frac{c_S}{c_K} \max_{k=1,\ldots,m} \inf_{v_h \in V_h} \|u_k - v_h\|_V \overset{h\to 0}{\longrightarrow} 0.$$

Hieraus folgt mit Hilfe von Lemma 6.7, dass ein $h_m > 0$ existiert mit $\sigma_{h,m}^2 \geq 1/2$ für $0 < h \leq h_m$. Lemma 6.6 und Lemma 6.7 liefern wegen $1/(1-x) \leq 1 + 2x$ für $x \leq 1/2$

$$\lambda_m \leq \lambda_{h,m} \leq \left(1 + 2c(m) \max_{\substack{v \in U_m \\ \|v\|_W = 1}} \|v - P_h v\|_V^2 \right) \lambda_m.$$

□

Konvergenz der Eigenfunktionen

Wir zeigen die Konvergenz der Eigenfunktionen $u_{h,m}$ gegen u_m für den Fall, dass λ_m ein einfacher Eigenwert ist. Unter den Voraussetzungen von Satz 6.8 gilt dann $\lambda_{h,k} \neq \lambda_m$ für alle $k \neq m$ und $0 < h \leq h_m$. Dann ist auch die Größe

$$\rho_{h,m} := \max_{\substack{1 \leq k \leq n \\ k \neq m}} \frac{\lambda_m}{|\lambda_{h,k} - \lambda_m|}$$

wohldefiniert. Diese Größe enthält die Information über den relativen Abstand der Eigenwerte. Je größer dieser ist, desto kleiner wird $\rho_{h,m}$ und desto besser wird folgende Abschätzung.

Lemma 6.9. *Sei λ_m ein einfacher Eigenwert. Dann existiert ein $h_m > 0$ derart, dass*

$$\|u_m - u_{h,m}\|_W \leq 2(1 + \rho_{h,m})\|u_m - P_h u_m\|_W$$

für alle $0 < h \leq h_m$.

Beweis. Sei $v_{h,m} := (P_h u_m, u_{h,m})_W u_{h,m}$ die W-orthogonale Projektion von $P_h u_m$ auf die lineare Hülle von $u_{h,m}$. Da $\{u_{h,k}\}_{k=1}^n$ eine Orthonormalbasis von V_h bzgl. des Skalarproduktes $(\cdot,\cdot)_W$ ist, folgt

$$\|P_h u_m - v_{h,m}\|_W^2 = \|\sum_{\substack{k=1 \\ k \neq m}}^{n} (P_h u_m, u_{h,k})_W u_{h,k}\|_W^2 = \sum_{\substack{k=1 \\ k \neq m}}^{n} (P_h u_m, u_{h,k})_W^2. \tag{6.7}$$

Wegen der Definition von P_h gilt

$$(P_h u_m, u_{h,k})_W = \frac{1}{\lambda_{h,k}} a(P_h u_m, u_{h,k}) = \frac{1}{\lambda_{h,k}} a(u_m, u_{h,k}) = \frac{\lambda_m}{\lambda_{h,k}} (u_m, u_{h,k})_W$$

und daher

$$(\lambda_{h,k} - \lambda_m)(P_h u_m, u_{h,k})_W = \lambda_m (u_m - P_h u_m, u_{h,k})_W.$$

Diese Gleichung in (6.7) eingesetzt ergibt

$$\begin{aligned} \|P_h u_m - v_{h,m}\|_W^2 &\leq \rho_{h,m}^2 \sum_{\substack{k=1 \\ k \neq m}}^{n} (u_m - P_h u_m, u_{h,k})_W^2 \leq \rho_{h,m}^2 \sum_{k=1}^{n} (u_m - P_h u_m, u_{h,k})_W^2 \\ &\leq \rho_{h,m}^2 \|u_m - P_h u_m\|_W^2. \end{aligned} \tag{6.8}$$

Aus der Definition von $v_{h,m}$ folgt

$$\|v_{h,m} - u_{h,m}\|_W = \|[(P_h u_m, u_{h,m})_W - 1] u_{h,m}\|_W = |(P_h u_m, u_{h,m})_W - 1| \underbrace{\|u_{h,m}\|_W}_{=1} \tag{6.9}$$

und

$$\underbrace{\|u_m\|_W}_{=1} - \|u_m - v_{h,m}\|_W \leq \underbrace{\|v_{h,m}\|_W}_{=|(P_h u_m, u_{h,m})_W|} \leq \underbrace{\|u_m\|_W}_{=1} + \|u_m - v_{h,m}\|_W,$$

d.h.

$$|(P_h u_m, u_{h,m})_W - 1| \leq \|u_m - v_{h,m}\|_W.$$

Die Kombination von (6.9) mit der letzten Abschätzung ergibt

$$\|v_{h,m} - u_{h,m}\|_W \leq \|u_m - v_{h,m}\|_W.$$

Damit erhalten wir

$$\begin{aligned}\|u_m - u_{h,m}\|_W &\leq \|u_m - v_{h,m}\|_W + \|v_{h,m} - u_{h,m}\|_W \leq 2\|u_m - v_{h,m}\|_W \\ &\leq 2\|u_m - P_h u_m\|_W + 2\|P_h u_m - v_{h,m}\|_W.\end{aligned}$$

Die Behauptung folgt aus (6.8). □

Lemma 6.10. *Es gilt*

$$a(u_{h,m} - u_m, u_{h,m} - u_m) = \lambda_m \|u_{h,m} - u_m\|_W^2 + \lambda_{h,m} - \lambda_m.$$

Beweis. Die Behauptung folgt aus

$$\begin{aligned}a(u_{h,m} - u_m, u_{h,m} - u_m) &= a(u_{h,m}, u_{h,m}) + a(u_m, u_m) - 2a(u_{h,m}, u_m) \\ &= \lambda_{h,m} + \lambda_m - 2\lambda_m (u_{h,m}, u_m)_W \\ &= \lambda_{h,m} - \lambda_m + 2\lambda_m [1 - (u_{h,m}, u_m)_W]\end{aligned}$$

und

$$\|u_{h,m} - u_m\|_W^2 = \|u_{h,m}\|_W^2 + \|u_m\|_W^2 - 2(u_{h,m}, u_m)_W = 2[1 - (u_{h,m}, u_m)_W].$$

□

Satz 6.11 (Konvergenz approximativer Eigenfunktionen). *Es gelten die Annahmen von Satz 6.3, und es seien $V_h \subset V$ mit $\dim V_h \geq m$ so gewählt, dass $\lim_{h\to 0} \inf_{v_h \in V_h} \|v - v_h\|_V = 0$ für alle $v \in V$. Ist λ_m ein einfacher Eigenwert, dann existiert ein $h_m > 0$ derart, dass für alle $0 < h \leq h_m$ der diskrete Eigenwert $\lambda_{h,m}$ ebenfalls einfach ist. Weiterhin existiert eine von h unabhängige Konstante $c = c(m) > 0$, so dass die Fehlerabschätzungen*

$$\|u_m - u_{h,m}\|_V \leq c \max_{\substack{v \in U_m \\ \|v\|_W = 1}} \inf_{v_h \in V_h} \|v - v_h\|_V$$

und

$$\|u_m - u_{h,m}\|_W \leq c \|u_m - P_h u_m\|_W$$

gelten.

Beweis. Aus Satz 6.8 ergibt sich, dass der Eigenwert $\lambda_{h,m}$ einfach ist. Ferner ist $\rho_{h,m}$ unabhängig von h beschränkt, so dass die zweite Abschätzung aus Lemma 6.9 folgt. Zusammen mit Satz 6.8 und Lemma 6.10 folgt hieraus die erste Abschätzung

$$\begin{aligned}\|u_m - u_{h,m}\|_V^2 &\leq \frac{1}{c_K} a(u_{h,m} - u_m, u_{h,m} - u_m) \\ &= \frac{1}{c_K} \left[\lambda_m \|u_{h,m} - u_m\|_W^2 + \lambda_{h,m} - \lambda_m\right] \\ &\leq \frac{1}{c_K} \left[c\lambda_m \|u_m - P_h u_m\|_W^2 + 2c(m) \max_{\substack{v \in U_m \\ \|v\|_W = 1}} \inf_{v_h \in V_h} \|v - v_h\|_V^2\right],\end{aligned}$$

denn es ist

$$\|u_m - P_h u_m\|_W^2 \leq c\|u_m - P_h u_m\|_V^2 \leq c' \inf_{v_h \in V_h} \|u_m - v_h\|_V^2.$$

□

Bemerkung. Bei Diskretisierung mit stückweise linearen Elementen konvergieren die Eigenwerte quadratisch in h. Die Eigenvektoren konvergieren in der Energienorm nur linear, während sie in $L^2(\Omega)$ ebenfalls quadratisch konvergieren. Man beachte, dass dabei die Konstanten vom Index m des Eigenwerts abhängen.

7 Gebietszerlegungsmethoden

In diesem Kapitel werden wir Methoden kennenlernen, mit deren Hilfe die Lösung von Randwertproblemen durch Aufteilung in Teilgebiete parallelisiert werden kann.

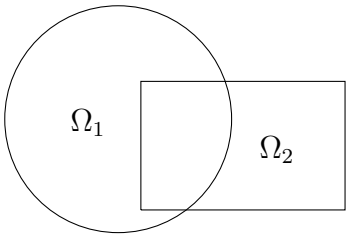
7.1 Die klassische Schwarz-Iteration

Von H.A. Schwarz wurde bereits im Jahr 1869 eine konstruktive Methode vorgestellt, um die Existenz von Lösungen Poissonscher Randwertaufgaben

$$-\Delta u = f \quad \text{in } \Omega := \Omega_1 \cup \Omega_2, \tag{7.1a}$$

$$u = 0 \quad \text{auf } \partial\Omega \tag{7.1b}$$

zu beweisen. Dabei sind Ω_1, Ω_2 zwei sich überlappende Gebiete. Zur damaligen Zeit war nur die Existenz von Lösungen auf Kreis- bzw. Rechteckgebieten bekannt.



Eine Lösung auf Ω findet man mit Hilfe des folgenden Algorithmus: Gegeben sei $u_2^0 : \Omega_2 \to \mathbb{R}$. Für $n = 0, 1, 2, \ldots$ löse auf Ω_1

$$\begin{aligned} -\Delta u_1^{n+1} &= f && \text{in } \Omega_1, \\ u_1^{n+1} &= u_2^n && \text{auf } \partial\Omega_1 \cap \Omega_2, \\ u_1^{n+1} &= 0 && \text{auf } \partial\Omega_1 \setminus \Omega_2 \end{aligned}$$

und entsprechend auf Ω_2

$$\begin{aligned} -\Delta u_2^{n+1} &= f && \text{in } \Omega_2, \\ u_2^{n+1} &= u_1^{n+1} && \text{auf } \partial\Omega_2 \cap \Omega_1, \\ u_2^{n+1} &= 0 && \text{auf } \partial\Omega_2 \setminus \Omega_1. \end{aligned} \tag{7.2}$$

Damit definiert man die Funktion

$$u^n := \begin{cases} u_1^n, & \text{in } \Omega_1 \setminus \Omega_2, \\ u_2^n, & \text{in } \Omega_2. \end{cases}$$

Offenbar ist dies gleichbedeutend mit folgender Iteration: bei gegebenem $u^0 : \Omega \to \mathbb{R}$, das auf $\partial\Omega$ verschwindet, löse für $n = 0, 1, 2, \ldots$

$$\begin{aligned} -\Delta u^{n+1/2} &= f && \text{in } \Omega_1, \\ u^{n+1/2} &= u^n && \text{in } \overline{\Omega} \setminus \Omega_1 \end{aligned}$$

und

$$\begin{aligned} -\Delta u^{n+1} &= f && \text{in } \Omega_2, \\ u^{n+1} &= u^{n+1/2} && \text{in } \overline{\Omega} \setminus \Omega_2. \end{aligned}$$

Wir werden diese Iteration nun in variationeller Form notieren. Für das Ausgangsproblem (7.1) verwenden wir $V := H_0^1(\Omega)$ und die Bilinearform

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x.$$

Ferner seien $V_k := H_0^1(\Omega_k)$ und $P_k : V_k \to V$, $k = 1, 2$, bezeichne jeweils die Fortsetzung durch Null. Dann gilt $V = P_1 V_1 + P_2 V_2$. Außerdem definieren wir die lokalen Bilinearformen

$$a_k(u, v) = \int_{\Omega_k} \nabla u \cdot \nabla v \, \mathrm{d}x, \quad u, v \in V_k.$$

Mit $u^0 \in V$ bestimme für $n = 0, 1, 2, \ldots$ die Lösung $v^{n+1/2} \in V_1$ von

$$a_1(u^n|_{\Omega_1} + v^{n+1/2}, v) = (f, v)_{L^2(\Omega_1)} \quad \text{für alle } v \in V_1 \tag{7.3}$$

und setze $u^{n+1/2} = u^n + P_1 v^{n+1/2}$. Mit diesem $u^{n+1/2} \in V$ bestimme $v^{n+1} \in V_2$, so dass

$$a_2(u^{n+1/2}|_{\Omega_2} + v^{n+1}, v) = (f, v)_{L^2(\Omega_2)} \quad \text{für alle } v \in V_2 \tag{7.4}$$

und setze $u^{n+1} = u^{n+1/2} + P_2 v^{n+1} \in V$.

Aus dieser Formulierung lässt sich eine Darstellung der Schwarz-Iteration durch zwei orthogonale Projektoren $\mathcal{P}_k := P_k \hat{\mathcal{P}}_k$, $k = 1, 2$, wobei $\hat{\mathcal{P}}_k : V \to V_k$ durch

$$a_k(\hat{\mathcal{P}}_k u, v) = a(u, P_k v) \quad \text{für alle } v \in V_k.$$

definiert ist, herleiten; vgl. auch Übungsblatt 13 der Vorlesung *Wissenschaftliches Rechnen I*.

Lemma 7.1. *Es sei $e^n := u^n - u$. Dann gilt*

$$e^{n+1} = (I - \mathcal{P}_2)(I - \mathcal{P}_1)e^n, \quad n = 0, 1, 2, \ldots.$$

Beweis. Wegen (7.3) gilt

$$a_1((u^{n+1/2} - u^n)|_{\Omega_1}, v) = (f, v)_{L^2(\Omega_1)} - a_1(u^n|_{\Omega_1}, v) = a(u - u^n, P_1 v) \quad \text{für alle } v \in V_1$$

und somit $(u^{n+1/2} - u^n)|_{\Omega_1} = \hat{\mathcal{P}}_1(u - u^n)$. Weil $u^{n+1/2} - u^n$ auf $\Omega \setminus \Omega_1$ verschwindet, erhalten wir

$$u^{n+1/2} - u^n = P_1(u^{n+1/2} - u^n)|_{\Omega_1} = \mathcal{P}_1(u - u^n).$$

und hieraus

$$u^{n+1/2} - u = (I - \mathcal{P}_1)(u^n - u).$$

Analog erhält man aus (7.4), dass

$$u^{n+1} - u = (I - \mathcal{P}_2)(u^{n+1/2} - u) = (I - \mathcal{P}_2)(I - \mathcal{P}_1)(u^n - u).$$

□

Bemerkung.

(a) Wegen der Fehlerdarstellung des letzten Lemmas spricht von einer **multiplikativen Schwarz-Methode**.

(b) Wegen

$$(I - \mathcal{P}_2)(I - \mathcal{P}_1) = I - \mathcal{P}_1 - \mathcal{P}_2 + \mathcal{P}_2\mathcal{P}_1 \tag{7.5}$$

kann die Schwarz-Iteration auch als Richardson-Iteration zur Lösung des Problems

$$\mathcal{P}_{\text{mu}} u = g$$

mit $\mathcal{P}_{\text{mu}} := \mathcal{P}_1 + \mathcal{P}_2 - \mathcal{P}_2\mathcal{P}_1$ und passender rechter Seite $g \in V$ gesehen werden.

(c) Anstelle der kontinuierlichen Räume V, V_k können auch Diskretisierungen $V := \mathcal{S}_0^{1,0}(\mathcal{T}_h)$ bzw. $V_k = \mathcal{S}_0^{1,0}(\mathcal{T}_k)$, $k = 1, 2$, verwendet werden, wobei $\mathcal{T}_k$ eine Zerlegung von Ω_k ist und $\mathcal{T}_h := \mathcal{T}_1 \cup \mathcal{T}_2$.

7.2 Die multiplikative Schwarz-Methode

Offenbar kann die Schwarz-Iteration auf K Teilgebiete Ω_i, $i = 1, \ldots, K$, verallgemeinert werden, indem alle Teilgebiete in einer vorgebenen Reihenfolge abgearbeitet werden und jeweils eine Lösung unter Beachtung der jeweils aktuellen Randwerte ausgerechnet wird. Dabei nehmen wir an, dass Ω_i aus Mengen $\Omega_i' \subset \Omega_i$, $i = 1, \ldots, K$, enstehen, so dass

$$\overline{\Omega} = \bigcup_{i=1}^{K} \overline{\Omega_i'}, \quad \Omega_k' \cap \Omega_\ell' = \emptyset, \quad k \neq \ell, \tag{7.6}$$

und die Breite der Überlappung mindestens βH mit

$$H := \max_{k=1,\ldots,K} \operatorname{diam} \Omega_k'$$

und einem $0 < \beta < 1$ beträgt. $\{\Omega_k'\}$ kann somit als grobes Gitter von Ω gedeutet werden. Für die lokalen Räume V_k wählen wir

$$V_k := \mathcal{S}_0^{1,0}(\mathcal{T}_k), \quad k = 1, \ldots, K,$$

wobei die Zerlegung $\mathcal{T}_k$ von Ω_k so gewählt sei, dass $\mathcal{T}_h = \bigcup_{k=1}^{K} \mathcal{T}_k$. Wie bisher sei $P_k : V_k \to V_h$ die Fortsetzung durch 0.

Dann erhält man die Rekursionsformel für den Fehler e^n der multiplikativen Schwarz-Methode

$$e^{n+1} = (I - \mathcal{P}_K)(I - \mathcal{P}_{K-1}) \cdot \ldots \cdot (I - \mathcal{P}_1) e^n, \quad n = 0, 1, 2, \ldots,$$

und somit

$$\mathcal{P}_{\text{mu}} := I - (I - \mathcal{P}_K)(I - \mathcal{P}_{K-1}) \cdot \ldots \cdot (I - \mathcal{P}_1).$$

7.2.1 Abstrakte Konvergenzanalyse

In Abschnitt 2.7 von *Wissenschaftliches Rechnen I* haben wir in einem allgemeinen Zugang die additive Schwarz-Methode untersucht. Aus (7.5) erkennt man, dass sich die multiplikative Schwarz-Methode (wegen des Terms $\mathcal{P}_2\mathcal{P}_1$) nicht als additive Methode umschreiben lässt. Allerdings können wir die multiplikative Methode unter den gleichen allgemeinen Voraussetzungen (ohne den Bezug zur Gebietszerlegung) wie die bei der additiven Methode verwendeten analysieren.

Seien Einbettungen $P_k : V_k \to V$ mit der Zerlegung

$$V = \sum_{k=0}^{K} P_k V_k, \quad V \ni v = \sum_{k=0}^{K} P_k v^k, \tag{7.7}$$

wobei $v^0 \in V_0, \ldots, v^K \in V_K$, definiert. Dabei hat der 0-te Raum eine Sonderstellung; er wird z.B. als Grobgitterraum gewählt. Auf den Räumen V_k betrachten wir symmetrische, koerzive und stetige Bilinearformen

$$b_k : V_k \times V_k \to \mathbb{R}.$$

Nach dem Satz von Lax-Milgram existiert zu $u \in V$ die eindeutige Lösung $\hat{\mathcal{P}}_k u \in V_k$ von

$$b_k(\hat{\mathcal{P}}_k u, v) = a(u, P_k v) \quad \text{für alle } v \in V_k.$$

Durch $\mathcal{P}_k : V \to P_k V_k \subset V$, $u \mapsto P_k \hat{\mathcal{P}}_k u$, definieren wir die sog. **Schwarz-Projektoren**. Bei der Analyse der additiven Schwarz-Methode haben wir die folgenden Voraussetzungen verwendet.

(i) **Stabile Zerlegung**: Es existiert $c_0 > 0$, so dass mit der Zerlegung (7.7) gilt

$$\sum_{k=0}^{K} b_k(v^k, v^k) \le c_0^2 \, a(v, v) \quad \text{für alle } v \in V.$$

(ii) **Verschärfte Cauchy-Schwarz-Ungleichung**: Es existieren Konstanten $0 \le \epsilon_{\ell k} \le 1$, $\ell, k = 1, \ldots, K$, mit

$$|a(P_\ell v, P_k w)|^2 \le \epsilon_{\ell k}^2 \, a(P_\ell v, P_\ell v) \, a(P_k w, P_k w) \quad \text{für alle } v \in V_\ell, \ w \in V_k.$$

(iii) **Lokale Stabilität**: Es gibt $\omega > 0$ mit

$$a(P_k v, P_k v) \le \omega \, b_k(v, v) \quad \text{für alle } v \in \operatorname{Ran} \hat{\mathcal{P}}_k \subset V_k, \quad k = 0, \ldots, K.$$

Im Folgenden untersuchen wir die Konvergenz der multiplikativen Schwarz-Methode und damit insbesondere die Konvergenz der klassischen Schwarz-Iteration.

Bemerkung. Im Fall der Schwarz-Iteration reduzieren sich die Annahmen (i)–(iii) auf die verschärfte Cauchy-Schwarz-Ungleichung, weil in diesem Fall

$$b_k(v, w) = a_k(v, w) = a(P_k v, P_k w), \quad v, w \in V_k,$$

gilt und somit (i) und (iii) mit $\omega = 1$ offenbar richtig sind.

Wir benötigen zunächst ein technisches Lemma.

Lemma 7.2. *Unter den Annahmen (ii) und (iii) gilt für $\ell, k = 1, \dots, K$ und $v, w \in V$*

$$a(\mathcal{P}_k v, w) \le \sqrt{a(\mathcal{P}_k v, v)}\,\sqrt{a(\mathcal{P}_k w, w)},$$
$$a(\mathcal{P}_\ell v, \mathcal{P}_k w) \le \omega \epsilon_{\ell k} \sqrt{a(\mathcal{P}_\ell v, v)}\,\sqrt{a(\mathcal{P}_k w, w)}.$$

Beweis. Nach Lemma 2.68 ist $\mathcal{P}_k$ selbstadjungiert. Die erste Abschätzung folgt aus der Definition von $\hat{\mathcal{P}}_k$ und der Cauchy-Schwarz-Ungleichung:

$$\begin{aligned} a(\mathcal{P}_k v, w) &= a(v, \mathcal{P}_k w) = a(v, P_k \hat{\mathcal{P}}_k w) = b_k(\hat{\mathcal{P}}_k v, \hat{\mathcal{P}}_k w) \\ &\le \sqrt{b_k(\hat{\mathcal{P}}_k v, \hat{\mathcal{P}}_k v)}\,\sqrt{b_k(\hat{\mathcal{P}}_k w, \hat{\mathcal{P}}_k w)} = \sqrt{a(v, \mathcal{P}_k v)}\,\sqrt{a(w, \mathcal{P}_k w)}. \end{aligned}$$

Die zweite Abschätzung erhält man durch Anwendung von (ii) und (iii) auf $\hat{\mathcal{P}}_\ell v$ und $\hat{\mathcal{P}}_k w$

$$\begin{aligned} a(\mathcal{P}_\ell v, \mathcal{P}_k w) &\le \epsilon_{\ell k} \sqrt{a(\mathcal{P}_\ell v, \mathcal{P}_\ell v)}\,\sqrt{a(\mathcal{P}_k w, \mathcal{P}_k w)} \le \omega \epsilon_{\ell k} \sqrt{b_\ell(\hat{\mathcal{P}}_\ell v, \hat{\mathcal{P}}_\ell v)}\,\sqrt{b_k(\hat{\mathcal{P}}_k w, \hat{\mathcal{P}}_k w)} \\ &= \omega \epsilon_{\ell k} \sqrt{a(v, \mathcal{P}_\ell v)}\,\sqrt{a(w, \mathcal{P}_k w)}. \end{aligned}$$

□

Der folgende Satz zeigt die lineare Konvergenz $\|e^n\|_a \to 0$ für $n \to \infty$. Ferner erhält man für den multiplikativen Schwarz-Projektor $\|\mathcal{P}_{\mathrm{mu}}\|_a < 2$.

Satz 7.3. *Unter den Annahmen (i)–(iii) mit $\omega \in (0, 2)$ gilt*

$$\|I - \mathcal{P}_{\mathrm{mu}}\|_a^2 \le 1 - \frac{2 - \omega}{c_0^2 (1 + 2\hat{\omega}^2 \rho^2(\mathcal{E}))} < 1,$$

wobei $\hat{\omega} := \max\{1, \omega\}$.

Beweis. Wir definieren $E_{-1} := I$ und für $k = 0, \dots, K$

$$E_k := (I - \mathcal{P}_k) \cdot \ldots \cdot (I - \mathcal{P}_0) \quad \text{sowie} \quad Q_k := 2\mathcal{P}_k - \mathcal{P}_k^2.$$

Wegen (iii) folgt (vgl. auch (2.28) aus dem Beweis von Lemma 2.71)

$$a^2(\mathcal{P}_k v, \mathcal{P}_k v) \le \omega^2 b_k^2(\hat{\mathcal{P}}_k v, \hat{\mathcal{P}}_k v) = \omega^2 a^2(v, \mathcal{P}_k v) \le \omega^2\, a(v, v)\, a(\mathcal{P}_k v, \mathcal{P}_k v). \tag{7.8}$$

und hieraus $\|\mathcal{P}_k\|_a \le \omega < 2$. Weil $Q_k = (2I - \mathcal{P}_k)\mathcal{P}_k \ge (2 - \omega)\mathcal{P}_k$, ist Q_k positiv-semidefinit. Mit dem zu E_k bzgl. a adjungierten Operator E_k^* gilt

$$E_{k-1}^* E_{k-1} - E_k^* E_k = E_{k-1}^* Q_k E_{k-1}, \quad k = 0, \dots, K.$$

Summation über k ergibt

$$I - E_K^* E_K = \sum_{k=0}^{K} E_{k-1}^* Q_k E_{k-1} \ge (2 - \omega) \sum_{k=0}^{K} E_{k-1}^* \mathcal{P}_k E_{k-1}. \tag{7.9}$$

Eine obere Schranke für $\|E_K\|_a$ erhält man, indem wir im Folgenden zeigen, dass die rechte Seite der letzten Abschätzung genügend positiv-definit ist. Zunächst erhält man aus der Definition von E_k

$$I - E_{k-1} = I - (I - \mathcal{P}_{k-1})E_{k-2} = I - E_{k-2} + \mathcal{P}_{k-1}E_{k-2} = \mathcal{P}_0 + \sum_{j=1}^{k-1} \mathcal{P}_j E_{j-1}.$$

Dies zeigt für $k > 0$

$$a(\mathcal{P}_k u, u) = a(\mathcal{P}_k u, E_{k-1}u) + a(\mathcal{P}_k u, \mathcal{P}_0 u) + \sum_{j=1}^{k-1} a(\mathcal{P}_k u, \mathcal{P}_j E_{j-1} u).$$

Mit Lemma 7.2 erhält man hieraus

$$a(\mathcal{P}_k u, u) \leq \sqrt{a(\mathcal{P}_k u, u)} \Bigg[\sqrt{a(\mathcal{P}_k E_{k-1} u, E_{k-1} u)} + \sqrt{a(\mathcal{P}_k \mathcal{P}_0 u, \mathcal{P}_0 u)} + \\ + \omega \sum_{j=1}^{k-1} \epsilon_{kj} \sqrt{a(\mathcal{P}_j E_{j-1} u, E_{j-1} u)} \Bigg].$$

Wegen $\epsilon_{kk} = 1$ können wir den ersten Term in den Summenausdruck integrieren. Division durch $\sqrt{a(\mathcal{P}_k u, u)}$ ergibt mit dem Vektor $c \in \mathbb{R}^K$ definiert durch

$$c_j = \sqrt{a(\mathcal{P}_j E_{j-1} u, E_{j-1} u)}, \quad j = 1, \ldots, K,$$

die Abschätzung

$$a(\mathcal{P}_k u, u) \leq 2a(\mathcal{P}_k \mathcal{P}_0 u, \mathcal{P}_0 u) + 2\hat{\omega}^2 (\mathcal{E} c)_k^2.$$

Summation über $k = 1, \ldots, K$ und Verwendung der Abschätzung

$$\sum_{k=1}^{K} a(\mathcal{P}_k v, v) \leq \omega \rho(\mathcal{E})\, a(v, v)$$

aus dem Beweis von Lemma 2.71 ergibt mit (7.8)

$$\begin{aligned} a(\mathcal{P}_{\text{ad}} u, u) &= a(\mathcal{P}_0 u, u) + \sum_{k=1}^{K} a(\mathcal{P}_k u, u) \\ &\leq a(\mathcal{P}_0 u, u) + 2 \sum_{k=1}^{K} a(\mathcal{P}_k \mathcal{P}_0 u, \mathcal{P}_0 u) + 2\hat{\omega}^2 \|\mathcal{E} c\|_2^2 \\ &\leq a(\mathcal{P}_0 u, u) + 2\omega\rho(\mathcal{E})\, a(\mathcal{P}_0 u, \mathcal{P}_0 u) + 2\hat{\omega}^2 \rho^2(\mathcal{E}) \|c\|_2^2 \\ &\leq (1 + 2\omega^2 \rho(\mathcal{E}))\, a(\mathcal{P}_0 u, u) + 2\hat{\omega}^2 \rho^2(\mathcal{E}) \|c\|_2^2 \\ &\leq (1 + 2\hat{\omega}^2 \rho^2(\mathcal{E})) \sum_{k=0}^{K} a(E_{k-1}^* \mathcal{P}_k E_{k-1} u, u). \end{aligned}$$

Mit (7.9) und der unteren Abschätzung $a(\mathcal{P}_{\text{ad}} u, u) \geq c_0^{-2} a(u, u)$ (dabei ist c_0 aus (i)) aus Lemma 2.70 folgt die Behauptung aus

$$c_0^{-2} a(u, u) \leq \frac{1 + 2\hat{\omega}^2 \rho^2(\mathcal{E})}{2 - \omega} a((I - E_K^* E_K) u, u).$$

□

7.2.2 Implementierung des multiplikativen Schwarz-Projektors

Die Schwarz-Projektoren $\mathcal{P}_{\rm ad}$ und $\mathcal{P}_{\rm mu}$ der additiven bzw. multiplikativen Schwarz-Methode können als Produkt CA eines passenden Vorkonditionierers C mit dem Ausgangsoperator A geschrieben werden. Weil nach Lemma 2.68 gilt

$$C_{\rm ad} = \sum_{k=0}^{K} P_k B_k^{-1} B_k^T,$$

ist die Anwendung von $C_{\rm ad}$ offensichtlich. Im Folgenden werden wir die Realisierung der Anwendung von $C_{\rm mu}$ im Fall der multiplikativen Schwarz-Methode beschreiben. Wegen

$$\mathcal{P}_{\rm mu} = I - (I - \mathcal{P}_K) \dots (I - \mathcal{P}_0) = C_{\rm mu} A$$

gilt im Fall $K = 0$ nach Lemma 2.68

$$y_0 := C_{\rm mu}^{(0)} x = \mathcal{P}_{\rm mu} A^{-1} x = [I - (I - \mathcal{P}_0)] A^{-1} x = \mathcal{P}_0 A^{-1} x = P_0 B_0^{-1} P_0^T x.$$

Angenommen, die Anwendung $y_{K-1} := C_{\rm mu}^{(K-1)} x$ von $C_{\rm mu}^{(K-1)}$ im Fall von K Gebieten wurde bereits durchgeführt. Dann gilt

$$\begin{aligned} y_K = C_{\rm mu}^{(K)} x &= A^{-1} x - (I - \mathcal{P}_K) \dots (I - \mathcal{P}_0) A^{-1} x \\ &= A^{-1} x - (I - \mathcal{P}_K) A^{-1} x + (I - \mathcal{P}_K) y_{K-1} = \mathcal{P}_K A^{-1} x + (I - \mathcal{P}_K) y_{K-1} \\ &= P_K B_K^{-1} P_K^T x + (I - P_K B_K^{-1} P_K^T A) y_{K-1}. \end{aligned}$$

Die Anwendung $y = C_{\rm mu} x$ von $C_{\rm mu}$ kann daher durch $y := P_0 B_0^{-1} P_0^T x$ und für $k = 1, \dots, K$

$$y := y + P_k B_k^{-1} P_k^T (x - Ay)$$

realisiert werden.

7.3 Die additive Schwarz-Methode

Wegen der Überlappung der beiden Gebiete Ω_1 und Ω_2 kann die Schwarz-Iteration nicht parallelisiert werden. Im Folgenden betrachten wir zwei Möglichkeiten, die gewünschte Parallelität zu erzielen.

7.3.1 Coloring

Ist Ω in viele Gebiete Ω_i, $i = 1, \dots, K$, zerlegt, so überlappen zwar benachbarte Gebiete, weiter entfernte Gebiete sind jedoch disjunkt. Die Nachbarschaftsbeziehung definiert einen Graphen mit Knotengrad höchstens q. Mit dem einfachen graphentheoretischen Werkzeug des sog. **Colorings** können Klassen $\mathcal{C}_i$, $i = 1, \dots, N_c$, von disjunkten Teilgebieten identifiziert werden, d.h. zwei Teilgebiete Ω_k und Ω_ℓ, $k, \ell \in \{1, \dots, K\}$, besitzen diesselbe Farbe, falls

$$a(P_k v_k, P_\ell v_\ell) = 0, \quad v_k \in V_k,\ v_\ell \in V_\ell. \tag{7.10}$$

Wegen

$$0 = a(\mathcal{P}_k v_k, P_\ell v_\ell) = b_\ell(\hat{\mathcal{P}}_\ell \mathcal{P}_k v_k, v_\ell) \quad \text{für alle } v_\ell \in V_\ell$$

bedeutet dies, dass $\mathcal{P}_\ell \mathcal{P}_k = 0$ und aus Symmetriegründen auch $\mathcal{P}_k \mathcal{P}_\ell = 0$ gilt. Daher folgt

$$I - \mathcal{P}_{\text{mu}} = \left(I - \sum_{i \in \mathcal{C}_{N_c}} \mathcal{P}_i \right) \dots \left(I - \sum_{i \in \mathcal{C}_1} \mathcal{P}_i \right) (I - \mathcal{P}_0)$$

Offensichtlich reduziert sich damit die Anzahl sequentieller Schritte von K auf $N_c + 1$. Ferner ist der Spekralradius von $\mathcal{E}$ durch den maximalen Knotengrad q beschränkt.

Lemma 7.4. *Es gilt $\rho(\mathcal{E}) \leq q + 1$.*

Beweis. Wegen (7.10) darf $\epsilon_{jk} = \epsilon_{kj} = 0$ gewählt weden. Daher existieren pro Zeile und Spalte von $\mathcal{E}$ höchstens $q + 1$ Einträge, die nicht verschwinden. Nach Lemma 2.76 gilt

$$\rho(\mathcal{E}) = \|\mathcal{E}\|_2 \leq \sqrt{q+1} \max_{i=1,\dots,K} \|\mathcal{E}_i\|_2 \leq q + 1,$$

weil jede Zeile $\mathcal{E}_i$ höchstens $q + 1$ Einträge ϵ_{ij} mit $0 < \epsilon_{ij} \leq 1$ enthält. □

7.3.2 Die additive Schwarz-Methode

Will man eine vollständige Parallelität erreichen, so muss $\mathcal{P}_k \mathcal{P}_\ell$ für alle $k, \ell = 1, \dots, K$ verschwinden. Der Term $\mathcal{P}_2 \mathcal{P}_1$ in (7.5) ist dafür verantwortlich, dass die beiden Aufgaben für die jeweiligen Teilgebiete nicht simultan bearbeitet werden können. An dieser Stelle setzt die additive Methode an. Hierbei ersetzt man u_1^{n+1} in (7.2) durch u_1^n und u^n durch

$$\tilde{u}^n(x) := \begin{cases} u_1^n(x), & x \in \Omega \setminus \overline{\Omega}_2, \\ u_2^n(x), & x \in \Omega \setminus \overline{\Omega}_1, \\ \frac{1}{2}\big(u_1^n(x) + u_2^n(x)\big), & x \in \Omega_1 \cap \Omega_2. \end{cases}$$

In diesem Fall erhält man die Rekursionsformel $\tilde{e}^{n+1} = (I - \mathcal{P}_{\text{ad}})\tilde{e}^n$ für den Fehler $\tilde{e}^n := \tilde{u}^n - u$ mit

$$\mathcal{P}_{\text{ad}} := \mathcal{P}_1 + \mathcal{P}_2.$$

Auch diese Methode ist auf K Gebiete verallgemeinerbar, indem man im $n+1$-ten Durchlauf durch die Gebiete die im n-ten Schritt berechneten Randwerte anstelle der innerhalb des $n+1$-ten aktualisierten Werte verwendet. Ähnlich wie beim Vergleich von Gauß-Seidel- und Jacobi-Verfahren konvergiert die multiplikative Methode im Vergleich zur additiven Methode schneller.

Schwarz-Methode mit Grobgitterkorrektur

Wir werden nun klären, welche Bedeutung der Index 0 in den Voraussetzungen (i)-(iii) der abstrakten Schwarz-Methode hat. Sei Ω_k ein inneres Teilgebiet, d.h. $\partial\Omega_k \cap \partial\Omega = \emptyset$. Wir betrachten den Vektor $e^{(k)}$ definiert durch

$$(e^{(k)})_j := \begin{cases} 1, & \text{Gitterpunkt } x_j \in \overline{\Omega}_k, \\ \text{beliebig}, & \text{sonst.} \end{cases}$$

Weil die Steifigkeitsmatrix A aus der Diskretisierung des Poisson-Problems entstanden ist, verschwindet die Zeilensumme für innere Punkte und somit ist $P_k^T A e^{(k)} = 0$. Also gilt auch für die im Gebiet Ω_k berechnete Korrektur

$$A_k^{-1} P_k^T A e^{(k)} = 0.$$

Daher wird keine Fehlerkorrektur durchgeführt, obwohl $e^{(k)} = 1$ in Ω_k. Dieser Effekt tritt allgemeiner für "glatte" Fehler auf und kann durch die im Folgenden beschriebene Grobgitterkorrektur behoben werden.

Sei $\mathcal{T}_H := \{\Omega_k'\}$ das in (7.6) eingeführte Grobgitter mit Gitterweite $H = \max_{k=1,\dots,K} \operatorname{diam} \Omega_k'$; siehe Abb. 7.1. Für $V_H := \mathcal{S}_0^{1,0}(\mathcal{T}_h)$ definieren wir den Prolongationsoperator $P_H : V_H \to V_h$

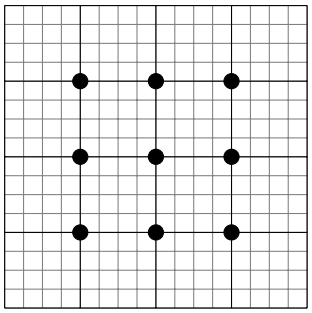


Abbildung 7.1: Grobgitter $\mathcal{T}_H$.

durch

$$P_H u := \mathfrak{I}_h[u]$$

definiert mit dem Interpolationsoperator $\mathfrak{I}_h : C(\overline{\Omega}) \to V_h$. Dieser interpoliert Grobgitterfunktionen auf dem feinen Gitter $\mathcal{T}_h$. Sei eine Näherungslösung $u^0 \in V_h$ für $a(u,v) = \ell(v)$, $v \in V_h$ geben. Dann wird die Grobgitterkorrektur $v_H \in V_H$ durch das Lösen des linearen Gleichungssystems (vgl. (7.3))

$$a(v_h, v) = \ell(v) - a(P_H^* u^0, v), \quad v \in V_H$$

bestimmt. Als verbesserte Näherung ergibt sich $u^1 := u^0 + P_H v_H$. Dies ist die gleiche Update-Formel wie bei der additiven bzw. multiplikativen Schwarz-Methode. Daher berücksichtigt man die Grobgitterkorrektur als 0-ten Index im abstrakten Setting:

$$P_0 := P_H, \quad V_0 := V_H, \quad b_0 := a.$$

7.3.3 Abschätzung der Kondition

Wir schätzen zunächst den größten Eigenwert von $\mathcal{P}_{\text{ad}}$ ab.

Lemma 7.5. *Es gilt $a(\mathcal{P}_{\text{ad}} v, v) \le (N_c + 1)\, a(v,v)$ für alle $v \in V_h$.*

Beweis. Nach Lemma 2.68 ist $\mathcal{P}_k$, $k = 0, \dots, K$, ein Projektor. Wir setzen $\mathcal{P}_i' := \sum_{k \in \mathcal{C}_i} \mathcal{P}_k$. Wegen

$$(\mathcal{P}_i')^2 = \sum_{k,\ell \in \mathcal{C}_i} \mathcal{P}_k \mathcal{P}_\ell = \sum_{k \in \mathcal{C}_i} \mathcal{P}_k^2 = \sum_{k \in \mathcal{C}_i} \mathcal{P}_k = \mathcal{P}_i'$$

ist auch $\mathcal{P}_i'$, $i = 1, \dots, N_c$, ein selbstadjungierter Projektor. Dann gilt wie in (7.8)

$$\sum_{k=0}^{K} a(\mathcal{P}_k v, v) = a(\mathcal{P}_0 v, v) + \sum_{i=1}^{N_c} a(\mathcal{P}_i' v, v) = a(\mathcal{P}_0 v, \mathcal{P}_0 v) + \sum_{i=1}^{N_c} a(\mathcal{P}_i' v, \mathcal{P}_i' v)$$
$$\leq (N_c + 1)\, a(v, v).$$

□

Bemerkung. Man kann auch zeigen, dass im letzten Lemma anstelle von N_c die maximale Anzahl $\hat{N}_c$ sich in einem Punkt überlappender Teilgebiete treten kann.

Den kleinsten Eigenwert schätzen wir mit Hilfe von Lemma 2.70 ab. Dazu muss nur die Annahme (i) gezeigt werden. Dazu werden wir die folgenden Hilfsmittel benötigen.

Lemma 7.6. *Es sei $Q_H : L^2(\Omega) \to V_H$ die L^2-Projektion auf V_H definiert durch*

$$(Q_H u, v_H)_{L^2(\Omega)} = (u, v_H)_{L^2(\Omega)}, \quad \text{für alle } u \in L^2(\Omega),\ v_H \in V_H.$$

Dann existiert $c > 0$, so dass für alle $u \in H_0^1(\Omega)$

$$|Q_H u|_{H^1(\Omega)} \leq c|u|_{H^1(\Omega)}, \quad \|u - Q_H u\|_{L^2(\Omega)} \leq cH|u|_{H^1(\Omega)}.$$

Lemma 7.7. *Es existiert eine **Partition der Eins** $\{\chi_i\}_{k=1}^K$ mit den Eigenschaften*

(i) $\chi_k \in C^\infty(\overline{\Omega})$,

(ii) $0 \leq \chi_k \leq 1$, $\chi_k(x) = 0$ *für* $x \in \Omega \setminus \Omega_k$,

(iii) $\sum_{k=1}^K \chi_k(x) = 1$ *für alle* $x \in \Omega$,

(iv) $\|\nabla \chi_k\|_{L^\infty(\Omega)} \leq c/(\beta H)$.

Für den Beweis des folgenden Satzes benötigen wir ferner den FE-Interpolationsoperator $\mathfrak{I}_h : C(\overline{\Omega}) \to V_h$ definiert durch

$$\mathfrak{I}_h[v](x) := \sum_{i \in I} v(x_i)\varphi_i(x), \quad v \in C(\overline{\Omega}),$$

wobei x_i der zugehörige Knoten zu φ_i ist. Für diesen gibt es für $s = 0, 1$ und eine Konstante $c > 0$ mit

$$|\mathfrak{I}_h[v]|_{H^s(\tau)} \leq c|v|_{H^s(\tau)} \quad \text{für alle } v \in H^s(\tau),\ \tau \in \mathcal{T}_h.$$

Satz 7.8. *Sei $v_h \in V_h \subset H_0^1(\Omega)$. Für die Zerlegung $v_h = \sum_{k=0}^K P_k v_h^k$ mit*

$$v_h^0 := Q_H v_h \in V_0 \quad \textit{und} \quad v_h^k := \mathfrak{I}_h[(v_h - v_h^0)\chi_k]|_{\Omega_k}, \quad k = 1, \dots, K,$$

gilt

$$\sum_{k=0}^{K} a(P_k v_h^k, P_k v_h^k) \leq c(1+1/\beta^2)\, a(v_h, v_h)$$

mit einer Konstanten $c > 0$, die unabhängig von H und h ist.

Beweis. Es gilt $v_h = \sum_{k=0}^{K} P_k v_h^k$, weil für $w_h := v_h - v_h^0$

$$\sum_{k=0}^{K} P_k v_h^k = \mathfrak{I}_h[v_h^0] + \sum_{k=1}^{K} \mathfrak{I}_h[w_h \chi_k] = \mathfrak{I}_h[v_h^0] + \mathfrak{I}_h[w_h \sum_{k=1}^{K} \chi_k] = \mathfrak{I}_h[v_h^0] + \mathfrak{I}_h[w_h] = v_h.$$

Sei $\tau \in \mathcal{T}_h$ ein Element und $k \in \{1, \ldots, K\}$ mit $\tau \subset \Omega_k'$. Aus der Taylorentwicklung um den Mittelpunkt x_τ von τ folgt

$$\chi_k(x) - \chi_k(x_\tau) = (x - x_\tau)^T \nabla \chi_k(x_\tau) + O(h^2) \quad \text{für alle } x \in \tau$$

und hieraus

$$\|\chi_k - \chi_k(x_\tau)\|_{L^\infty(\tau)} \leq c \frac{h}{\beta H}.$$

Hieraus erhält man mit Hilfe der inversen Abschätzung

$$\begin{aligned} |\mathfrak{I}_h[w_h(\chi_k - \chi_k(x_\tau)]|^2_{H^1(\tau)} &\leq c h^{-2} \|\mathfrak{I}_h[w_h(\chi_k - \chi_k(x_\tau))]\|^2_{L^2(\tau)} \\ &\leq c h^{-2} \|w_h\|^2_{L^2(\tau)} \|\chi_k - \chi_k(x_\tau)\|^2_{L^\infty(\tau)} \\ &\leq \frac{c}{\beta^2 H^2} \|w_h\|^2_{L^2(\tau)} \end{aligned}$$

und somit wegen $|\chi_k(x_\tau)| \leq 1$

$$\begin{aligned} |v_h^k|^2_{H^1(\tau)} &= |\mathfrak{I}_h[w_h \chi_k]|^2_{H^1(\tau)} \\ &\leq 2|\mathfrak{I}_h[w_h(\chi_k - \chi_k(x_\tau))]|^2_{H^1(\tau)} + 2|\mathfrak{I}_h[\chi_k(x_\tau) w_h|^2_{H^1(\tau)} \\ &\leq \frac{c}{\beta^2 H^2} \|w_h\|^2_{L^2(\tau)} + c|w_h|^2_{H^1(\tau)}. \end{aligned}$$

Summation über alle $k = 1, \ldots, K$ ergibt

$$\sum_{k=1}^{K} |v_h^k|^2_{H^1(\tau)} \leq N_c \left(\frac{c}{\beta^2 H^2} \|w_h\|^2_{L^2(\tau)} + c|w_h|^2_{H^1(\tau)} \right).$$

Hieraus erhält man nach Summation über $\tau \in \mathcal{T}_h$

$$\begin{aligned} \sum_{k=0}^{K} a(P_k v_h^k, P_k v_h^k) &= \sum_{k=0}^{K} |v_h^k|^2_{H^1(\Omega)} = |v_h^0|^2_{H^1(\Omega)} + \sum_{k=1}^{K} |v_h^k|^2_{H^1(\Omega)} \\ &\leq |Q_H v_h|^2_{H^1(\Omega)} + N_c \left(\frac{c}{\beta^2 H^2} \|v_h - Q_H v_h\|^2_{L^2(\Omega)} + c|v_h - Q_H v_h|^2_{H^1(\Omega)} \right) \\ &\leq c\, |v_h|^2_{H^1(\Omega)} + \frac{c N_c}{\beta^2 H^2} \|v_h - Q_H v_h\|^2_{L^2(\Omega)} \\ &\leq c(1 + 1/\beta^2)\, |v_h|^2_{H^1(\Omega)} = c(1+1/\beta^2)\, a(v_h, v_h). \end{aligned}$$

□

Aus Lemma 7.5 und Satz 7.8 folgt, dass die Kondition

$$\mathrm{cond}(\mathcal{P}_{\mathrm{ad}}) \leq c\left(1 + \frac{1}{\beta^2}\right)$$

nur durch den Überlappungsparameter β und nicht durch die Gitterweiten h, H noch die Anzahl der Teilgebiete K beschränkt ist.

Satz 7.8 zusammen mit Lemma 7.4 zeigen aber wegen Satz 7.3 auch die Beschänktheit der Norm von $I - \mathcal{P}_{\mathrm{mu}}$ und somit die Konvergenz der multiplikativen Methode.

8 Elliptische Variationsungleichungen

Bisher haben wir immer Variationsgleichungen betrachtet. Diese führen auf lineare Probleme. Die im Folgenden untersuchten *Variationsungleichungen* treten bei viele komplizierten physikalischen Prozessen auf und sind nicht-linear.

Beispiel 8.1 (Hindernis-Problem). Wir untersuchen die Auslenkung $u : \Omega \to \mathbb{R}$ einer elastischen Membran im Gleichgewicht, die

(i) am Rand $\partial\Omega$ des ebenen Gebietes $\Omega \subset \mathbb{R}^3$ fixiert, d.h. $u = 0$ auf $\partial\Omega$,

(ii) oberhalb eines Hindernisses der Höhe $\psi \in H^1(\Omega)$ mit $\psi \leq 0$ auf $\partial\Omega$ liegt und

(iii) einer vertikalen äußeren Kraft τg ausgesetzt ist, wobei τ die Spannung und $g \in H^{-1}(\Omega)$ eine gegebene Funktion bezeichnen.

Weil die Membran versucht,einen Zustand minimaler Energie anzunehmen, ist u Lösung des Minimierungsproblems

$$u \in K : \quad E(u) = \inf\{E(v) : v \in K\} \tag{8.1}$$

mit dem Energiefunktional

$$E(v) := \int_\Omega \frac{1}{2}|\nabla v|^2 - gv \,\mathrm{d}x$$

und der Menge der zulässigen Auslenkungen

$$K := \{v \in H_0^1(\Omega) : v \geq \psi \text{ in } \Omega\}.$$

Die Menge K ist nicht-leer, weil $\max\{0, \psi\}$ ein Element von K ist. Ferner überlegt man sich leicht, dass K abgeschlossen und konvex ist. Das Energiefunktional $E : H^1(\Omega) \to \mathbb{R}$ ist streng konvex und stetig auf K.

Im folgenden Lemma untersuchen wir die Lösbarkeit von (8.1). Für spätere Zwecke betrachten wir unterhalbstetige Funktionale.

Definition 8.2. *Sei V ein Banach-Raum und $K \subset V$. Eine Funktion $f : K \to \mathbb{R}$ heißt* ***unterhalbstetig****, falls aus $\{v_n\} \subset K$ und $v_n \to v \in K$ folgt*

$$f(v) \leq \liminf_{n\to\infty} f(v_n).$$

Eine Funktion f heißt ***schwach unterhalbstetig****, falls diese Ungleichung für alle $\{v_n\} \subset K$, die schwach gegen $v \in K$ konvergieren (d.h. $\lim_{n\to\infty} \ell(v_n - v) = 0$ für alle $\ell \in V'$), gilt.*

Lemma 8.3. *Sei V ein reflexiver Banach-Raum und $K \subset V$ schwach abgeschlossen. Ist $f : K \to \mathbb{R}$ schwach unterhalbstetig mit $\lim_{\|v\|\to\infty} f(v) = \infty$ und $f(v_0) < \infty$ für ein $v_0 \in V$, so besitzt das Minimierungsproblem*

$$u \in K : \quad f(u) \le f(v) \quad \text{für alle } v \in K$$

eine Lösung. Ist zusätzlich K konvex und f streng konvex, so ist das Minimum eindeutig.

Beweis. Wegen $f(v) \to \infty$ für $\|v\| \to \infty$ ist

$$K_0 := \{v \in K : f(v) \le f(v_0)\}$$

beschränkt. Weil K schwach abgeschlossen ist, folgt aus der schwachen Unterhalbstetigkeit, dass auch K_0 schwach abgeschlossen ist. Sei

$$\alpha := \inf_{v\in K_0} f(v).$$

Dann existiert eine Folge $\{u_n\} \subset K_0$ mit $\alpha = \lim_{n\to\infty} f(u_n)$. Weil K_0 beschränkt ist, ist auch die Folge $\{u_n\}$ beschränkt. Die Reflexivität von V impliziert die Existenz einer Teilfolge $\{u_{k_n}\}$, die gegen $u \in V$ schwach konvergiert. Weil K_0 schwach abgeschlossen ist, gilt $u \in K_0$ und

$$f(u) \le \liminf_{n\to\infty} f(u_{k_n}) = \alpha,$$

woraus $f(u) = \alpha$ folgt.

Angenommen, es existieren zwei Minima $u_1, u_2 \in K$ mit $f(u_1) = f(u_2)$. Weil K konvex ist, gilt $(u_1 + u_2)/2 \in K$. Die strenge Konvexität zeigt

$$f(\frac{u_1+u_2}{2}) < \frac{1}{2}[f(u_1) + f(u_2)] = f(u_1),$$

was den Widerspruch zeigt. □

Bemerkung. Schwache Abgeschlossenheit und schwache Unterhalbstetigkeit sind als Voraussetzung unhandlich. Wir geben hinreichende Bedingungen an.

(i) Ist K konvex und abgeschlossen, so ist K schwach abgeschlossen.

(ii) Ist f konvex und unterhalbstetig, so if f schwach unterhalbstetig.

Diese beiden Aussagen folgen aus dem Lemma von Mazur.

Die Lösung von (8.1) kann durch folgendes Lemma charakterisiert werden; vgl. den Projektionssatz aus *Algoritmische Mathematik II*.

Lemma 8.4. *Sei V ein Hilbert-Raum und $K \subset V$ konvex und $a : K \times K \to \mathbb{R}$ eine symmetrische Bilinearform und $\ell : K \to \mathbb{R}$ ein lineares Funktional. Genau dann ist $u \in K$ ein Minimum von $f(v) := \frac{1}{2}a(v, v) - \ell(v)$, wenn*

$$a(u, v - u) \ge \ell(v - u) \quad \text{für alle } v \in K.$$

Beweis. Für beliebige $v \in K$ ist $u + \lambda(v-u) \in K$, $\lambda \in [0,1]$. Daher nimmt die Funktion

$$\varphi(\lambda) := \frac{1}{2}a(u+\lambda(v-u), u+\lambda(v-u)) - \ell(u+\lambda(v-u))$$

für $\lambda = 0$ ihr Minimum an. Somit folgt

$$0 \le \varphi'(0) = a(u, v-u) - \ell(v-u).$$

Sei umgekehrt $a(u, v-u) \ge \ell(v-u)$ für alle $v \in K$. Dann gilt

$$\begin{aligned} f(v) = \frac{1}{2}a(v,v) - \ell(v) &= \frac{1}{2}a(u,u) + \frac{1}{2}a(v-u, v-u) + a(u, v-u) - \ell(v-u) - \ell(u) \\ &\ge \frac{1}{2}a(u,u) - \ell(u) = f(u). \end{aligned}$$

□

Wir erhalten also folgendes zu (8.1) äquivalentes Variationsproblem

$$u \in K: \quad \int_\Omega \nabla u \cdot \nabla(v-u)\,\mathrm{d}x \ge \int_\Omega g(v-u)\,\mathrm{d}x \quad \text{für alle } v \in K.$$

Bemerkung. Das letzte Lemma liefert insbesondere eine Charakterisierung der Bestapproximation $u \in K$ an $u_0 \in V$. Für $a(v,w) := (v,w)$, $\ell = 0$ und $K' := K - u_0$ erhält man nämlich für das Minimum $u' \in K'$ die Bedingung $(u', v'-u') \ge 0$ für alle $v \in K'$, d.h.

$$(u - u_0, v - u) \ge 0 \quad \text{für alle } v \in K.$$

8.1 Elliptische Variationsungleichungen erster Art

Sei V ein Hilbert-Raum, $a : V \times V \to \mathbb{R}$ eine stetige und V-koerzive Bilinearform und $\ell : V \to \mathbb{R}$ ein lineares, stetiges Funktional. Sei $K \subset V$. Eine **elliptische Variatiationsungleichung erster Art** besitzt die Form

$$u \in K: \quad a(u, v-u) \ge \ell(v-u) \quad \text{für alle } v \in K. \tag{8.2}$$

Ist a nicht symmetrisch, so kann diese Variationsungleichung nicht wie in Beispiel 8.1 als Minimierungsproblem interpretiert werden.

Ist K ein Unterraum von V, so ist (8.2) eine Variationsgleichung und somit eindeutig lösbar. Ist K eine konvexe Teilmenge, so hat man folgende Verallgemeinerung des Satzes von Lax-Milgram.

Satz 8.5. *Sei $\emptyset \ne K \subset V$ abgeschlossen und konvex. Dann besitzt (8.2) eine eindeutige Lösung $u \in K$.*

Beweis. Wir schreiben (8.2) als Fixpunkt-Problem. Dazu verwenden wir $L \in V$ mit $\|L\|_V = \|\ell\|_{V'}$ und

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

Wieder mit dem Rieszschen Darstellungssatz definieren wir für $u \in V$ die Abbildung $A : V \to V$ durch

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

Dann ist A linear und stetig mit $\|A\| \leq c_S$. Daher ist(8.2) für jedes $\theta > 0$ äquivalent mit

$$(u - [u - \theta(Au - L)], v - u) \geq 0 \quad \text{für alle } v \in K. \tag{8.3}$$

Nach der Bemerkung zu Lemma 8.4 ist damit u die orthogonale Projektion von $u - \theta(Au - L)$ auf K, d.h. es gilt

$$u = P_K(u - \theta(Au - L)).$$

Wir zeigen, dass der Operator auf der rechten Seite für $\theta \in (0, 2c_K/c_S^2)$ eine Kontraktion ist. Es gilt nämlich für $v_1, v_2 \in V$

$$\begin{aligned}
&\|P_K(v_1 - \theta(Av_1 - L)) - P_K(v_2 - \theta(Av_2 - L))\|^2 \\
&\qquad \leq \|(v_1 - \theta(Av_1 - L)) - (v_2 - \theta(Av_2 - L))\|^2 \\
&\qquad = \|(v_1 - v_2) - \theta A(v_1 - v_2)\|^2 \\
&\qquad = \|v_1 - v_2\|^2 - 2\theta a(v_1 - v_2, v_1 - v_2) + \theta^2 \|A(v_1 - v_2)\|^2 \\
&\qquad \leq (1 - 2\theta c_K + \theta^2 c_S^2) \, \|v_1 - v_2\|^2.
\end{aligned}$$

Nach dem Banachschen Fixpunktsatz besitzt (8.3) und somit (8.2) eine eindeutige Lösung. □

Die folgende Charakterisierung der Lösung von (8.2) zeigt einen Weg zur numerischen Lösung von (8.2) auf.

Satz 8.6. *Es gelten die Voraussetzungen von Satz 8.5 und a sei symmetrisch. Dann ist die Lösung $u \in K$ von (8.2) die eindeutige Bestapproximation in K von $w \in V$, d.h.*

$$\|u - w\|_a = \inf_{v \in K} \|v - w\|_a.$$

Dabei bezeichnet w die eindeutige Lösung des Variationsproblems

$$w \in V : \quad a(w, v) = \ell(v) \quad \textit{für alle } v \in V.$$

Beweis. Für alle $v \in K$ gilt

$$a(u, v - u) \geq \ell(v - u) = a(w, v - u)$$

und somit

$$a(u - w, v - u) \geq 0 \quad \text{für alle } v \in V.$$

Nach der Bemerkung zu Lemma 8.4 ist damit $u \in K$ die orthogonale Projektion von w auf K bzgl. des Skalarprodukts a. □

Zur Lösung von (8.2) kann man also zunächst ein Randwertproblem lösen und die Lösung w anschließend auf K projizieren.

8.1.1 Numerische Lösung

Sei $V_h \subset V$ ein FE-Raum und $\emptyset \neq K_h \subset V_h$ abgeschlossen und konvex. Dann ist die Finite-Elemente-Approximation von (8.2)

$$u_h \in K_h : \quad a(u_h, v_h - u_h) \geq \ell(v_h - u_h) \quad \text{für alle } v_h \in K_h. \tag{8.4}$$

Die Anwendung von Satz 8.5 auf dieses diskrete Problem liefert die Existenz eines eindeutigen $u_h \in K_h$. Für den Approximationsfehler hat man folgende Verallgemeinerung des Céa-Lemmas.

Satz 8.7. *Es existiert eine von h und u unabhängige Konstante $c > 0$, so dass*

$$\|u - u_h\|_V \leq c \Big\{ \inf_{v_h \in K_h} \big[\|u - v_h\|_V + |a(u, v_h - u) - \ell(v_h - u)|^{1/2} \big] + \inf_{v \in K} |a(u, v - u_h) - \ell(v - u_h)|^{1/2} \Big\}. \tag{8.5}$$

Beweis. Wegen (8.2) und (8.4) gilt

$$\begin{aligned} a(u, u) &\leq a(u, v) - \ell(v - u) \quad \text{für alle } v \in K, \\ a(u_h, u_h) &\leq a(u_h, v_h) - \ell(v_h - u_h) \quad \text{für alle } v_h \in K_h. \end{aligned}$$

Hieraus folgt mit der V-Koerzivität und der Stetigkeit von a für $v \in K$ und $v_h \in K_h$

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

Die Behauptung folgt mit $xy \leq \varepsilon x^2 + y^2/(4\varepsilon)$ für alle $x, y \in \mathbb{R}$, $\varepsilon > 0$, aus

$$a(u_h - u, v_h - u) \leq c_S \|u - u_h\|_V \|u - v_h\|_V \leq \frac{c_K}{2} \|u - u_h\|_V^2 + \frac{c_S}{2c_K} \|u - v_h\|_V^2.$$

□

Bemerkung. Im Fall $K_h \subset K$ gilt $u_h \in K$ und der dritte Term in der Fehlerabschätzung (8.5) verschwindet, d.h. es gilt

$$\|u - u_h\|_V \leq c \Big\{ \inf_{v_h \in K_h} \big[\|u - v_h\|_V + |a(u, v_h - u) - \ell(v_h - u)|^{1/2} \big] \Big\}.$$

Beispiel 8.8. Wir werden nun mit Hilfe von Satz 8.7 die Konvergenzordnung für das Hindernis-Problem bei linearen Elementen zeigen. Wir nehmen an, dass $u, \psi \in H^2(\Omega)$ und Ω ein Polygon ist, und verwenden lineare Elemente auf einer nicht-entarteten Triangulierung von Ω. Dann ist die diskrete zulässige Menge

$$K_h = \{ v_h \in H_0^1(\Omega) : v_h \text{ ist stückweise linear, } v_h(x) \geq \psi(x) \text{ für alle Gitterknoten } x \}.$$

Offenbar ist jede Funktion in K_h stetig, stückweise linear und verschwindet auf dem Rand. Allerdings dominiert sie die Hindernis-Funktion ψ nur in den inneren Knoten des Gitters. Daher gilt im Allgemeinen $K_h \not\subset K$. Für jedes $v \in H_0^1(\Omega)$ gilt

$$a(u,v) - \ell(v) = \int_\Omega \nabla u \cdot \nabla v - fv \,\mathrm{d}x = \int_\Omega (-\Delta u - f)v \,\mathrm{d}x.$$

Aus Satz 8.7 folgt

$$\begin{aligned}\|u-u_h\|_{H^1(\Omega)} \le c\Big\{ &\inf_{v_h\in K_h}\Big[\|u-v_h\|_{H^1(\Omega)} + \|-\Delta u - f\|_{L^2(\Omega)}^{1/2}\|u-v_h\|_{L^2(\Omega)}^{1/2}\Big] \\ &+ \|-\Delta u - f\|_{L^2(\Omega)}^{1/2} \inf_{v\in K}\|v-u_h\|_{L^2(\Omega)}^{1/2}\Big\}. \end{aligned} \tag{8.6}$$

Sei $\mathfrak{I}_h u$ der stückweise lineare Interpolant von u. Man überzeugt sich leicht davon, dass $\mathfrak{I}_h u \in K_h$. Dann gilt

$$\begin{aligned}&\inf_{v_h\in K_h}\Big[\|u-v_h\|_{H^1(\Omega)} + \|-\Delta u - f\|_{L^2(\Omega)}^{1/2}\|u-v_h\|_{L^2(\Omega)}^{1/2}\Big] \\ &\quad\le \|u-\mathfrak{I}_h u\|_{H^1(\Omega)} + \|-\Delta u - f\|_{L^2(\Omega)}^{1/2}\|u-\mathfrak{I}_h u\|_{L^2(\Omega)}^{1/2} \\ &\quad\le ch\Big[|u|_{H^2(\Omega)} + \|-\Delta u - f\|_{L^2(\Omega)}^{1/2}|u|_{H^2(\Omega)}^{1/2}\Big]\end{aligned}$$

Um den Ausdruck $\inf_{v\in K}\|v-u_h\|_{L^2(\Omega)}$ in (8.6) abzuschätzen, definiere

$$u_h^* := \max\{u_h, \psi\}.$$

Wegen $u_h, \psi \in H^1(\Omega)$ folgt $u_h^* \in H^1(\Omega)$ und offensichtlich $u_h^* \ge \psi$. Ferner gilt wegen $\psi \le 0$ auf $\partial\Omega$, dass $u_h^* = 0$ auf $\partial\Omega$ und somit $u_h^* \in K$. Sei

$$\Omega^* = \{x \in \Omega : u_h(x) < \psi(x)\}.$$

Dann gilt $u_h^* = u_h$ in $\Omega \setminus \Omega^*$ und daher

$$\inf_{v\in K}\|v-u_h\|_{L^2(\Omega)}^2 \le \|u_h^* - u_h\|_{L^2(\Omega)}^2 = \int_{\Omega^*} |u_h - \psi|^2 \,\mathrm{d}x.$$

Weil in jedem Gitterpunkt $u_h \ge \psi = \mathfrak{I}_h\psi$, gilt $u_h \ge \mathfrak{I}_h\psi$ auch in Ω. Daher folgt für $x \in \Omega^*$

$$0 < |u_h - \psi| = \psi - u_h \le \psi - \mathfrak{I}_h\psi \le |\psi - \mathfrak{I}_h\psi|.$$

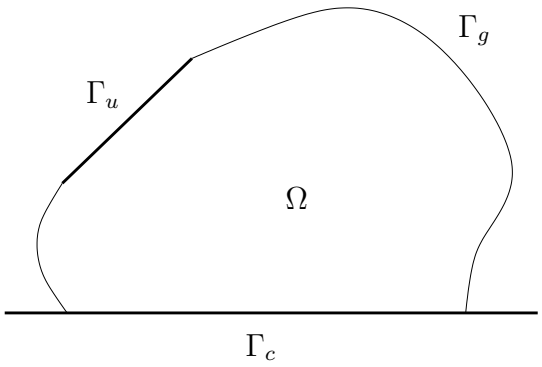
und hieraus

$$\int_{\Omega^*} |u_h-\psi|^2 \,\mathrm{d}x \le \int_{\Omega^*} |\psi - \mathfrak{I}_h\psi|^2 \,\mathrm{d}x \le \int_\Omega |\psi - \mathfrak{I}_h\psi|^2 \,\mathrm{d}x \le ch^4|\psi|_{H^2(\Omega)}^2,$$

was $\inf_{v\in K}\|v-u_h\|_{L^2(\Omega)}^{1/2} \le ch|\psi|_{H^2(\Omega)}^{1/2}$ liefert. Aus (8.6) folgt hieraus die Fehlerabschätzung

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

mit einer nur von $|u|_{H^2(\Omega)}$, $\|f\|_{L^2(\Omega)}$ und $|\psi|_{H^2(\Omega)}$ abhängigen Konstanten $c > 0$.



8.2 Elliptische Variationsungleichungen zweiter Art

Zur Erläuterung des zweiten Typs beginnen wir wieder mit einem Beispiel.

Beispiel 8.9 (Gleitreibungsproblem). Sei Ω ein elastischer Körper im Gleichgewicht auf einem festen Untergrund Γ_c, auf dem er gleitet. Der Rand $\partial\Omega$ wird als Lipschitz-stetig angenommen und besteht aus drei disjunkten Teilen Γ_u, Γ_g und Γ_c. An Γ_u wird Ω fixiert. Ω wird einer Volumenkraft mit Dichte $f \in [L^2(\Omega)]^d$ und der Oberflächenkraft $g \in [L^2(\Gamma_g)]^d$ an Γ_g ausgesetzt. Im restlichen Teil des Randes Γ_c besteht Gleitreibungskontakt. Gesucht ist das Verschiebungsfeld $u : \Omega \to \mathbb{R}^d$. Mit der Gleichgewichtsbedingung

$$-\operatorname{div}\sigma = f \quad \text{in } \Omega, \tag{8.7}$$

wobei $\sigma = [\sigma_{ij}]_{i,j=1}^d$ den symmetrischen **Cauchyschen Spannungstensor** bezeichnet, nehmen wir an, dass das Material linear-elastisch ist, d.h. es gilt

$$\sigma = C\varepsilon \iff \sigma_{ij} = \sum_{k,\ell=1}^{d} C_{ijk\ell}\varepsilon_{k\ell}(u) \quad \text{in } \Omega$$

mit dem Dehnungstensor $\varepsilon = [\varepsilon_{ij}]_{i,j=1}^d$

$$\varepsilon(u) = \frac{1}{2}(\nabla u + (\nabla u)^T)$$

und dem elastischen Tensor $C \in \mathbb{R}^{d^4}$ mit $C_{ijk\ell} \in L^\infty(\Omega)$,

$$C_{ijk\ell} = C_{jik\ell} = C_{k\ell ij}$$

und

$$E : CE \geq \alpha\|E\|^2 \quad \text{für alle symmetrischen } E \in \mathbb{R}^{d\times d},$$

wobei wir das innere Produkt

$$A : B = \sum_{i,j=1}^{d} a_{ij}b_{ij}$$

für Matrizen $A, B \in \mathbb{R}^{d\times d}$ verwenden.

Die Randbedingungen auf Γ_u und Γ_g sind

$$u = 0 \text{ on } \Gamma_u, \quad \sigma\nu = g \text{ on } \Gamma_g.$$

Um die Randbedingung auf dem restlichen Teil des Randes zu beschreiben, benötigen wir folgende Notation. Für ein Vektorfeld $u \in \mathbb{R}^d$ bezeichnen wir mit $u_\nu := u \cdot \nu \in \mathbb{R}$ die

Verschiebung in Normalenrichtung und mit $u_\tau := u - u_\nu \nu \in \mathbb{R}^d$ die Tangentialverschiebung. Dann gilt

$$u = u_\nu \nu + u_\tau$$

Für einen Spannungstensor $\sigma \in \mathbb{R}^{d\times d}$ definieren wir $\sigma_\nu = \nu^T \sigma \nu \in \mathbb{R}$ und $\sigma_\tau := \sigma\nu - \sigma_\nu \nu \in \mathbb{R}^d$. Hiermit gilt

$$\sigma\nu = \sigma_\nu \nu + \sigma_\tau.$$

Auf Γ_c stellen wir die folgenden vereinfachten Bedingungen der Gleitreibung

$$\sigma_\nu = -G, \tag{8.8a}$$
$$\|\sigma_\tau\| \le \mu_F G \tag{8.8b}$$

und $u_\tau = 0$, falls $\|\sigma_\tau\| < \mu_F G$, und $u_\tau = -\lambda\sigma_\tau$ mit einem $\lambda \ge 0$, falls $\|\sigma_\tau\| = \mu_F G$. Hierbei sind $G > 0$ und der Reibungskoeffizient $\mu_F > 0$ gegebene Funktionen $G, \mu_F \in L^\infty(\Gamma_c)$. Aus den beiden letzten Bedingungen erhält man

$$\sigma_\tau \cdot u_\tau = -\mu_F G \|u_\tau\| \quad \text{auf } \Gamma_c.$$

Im Folgenden leiten wir die schwache Formulierung des Gleitreibungsproblems her. Sei

$$V = \{v \in [H^1(\Omega)]^d : v = 0 \text{ auf } \Gamma_u\}.$$

Wir nehmen für die nächsten Rechnungen an, dass u genügend glatt ist. Multiplikation von (8.7) mit $v - u$, $v \in V$, ergibt

$$-\int_\Omega (v-u)\operatorname{div}\sigma \,\mathrm{d}x = \int_\Omega f \cdot (v-u)\,\mathrm{d}x.$$

Partielle Integration unter Verwendung der Randbedingung liefert

$$\begin{aligned}-\int_\Omega (v-u)\operatorname{div}\sigma \,\mathrm{d}x &= -\int_{\partial\Omega} \sigma\nu \cdot (v-u)\,\mathrm{d}s + \int_\Omega \sigma : \varepsilon(v-u)\,\mathrm{d}x \\ &= -\int_{\Gamma_g} g\cdot(v-u)\,\mathrm{d}s - \int_{\Gamma_c} \sigma\nu\cdot(v-u)\,\mathrm{d}s + \int_\Omega C\varepsilon(u) : \varepsilon(v-u)\,\mathrm{d}x.\end{aligned}$$

Die Zerlegung in Normal- und Tangentialkomponenten ergibt unter Verwendung von (8.9)

$$\begin{aligned}-\int_{\Gamma_c} \sigma\nu\cdot(v-u)\,\mathrm{d}s &= -\int_{\Gamma_c} \sigma_\nu(v_\nu - u_\nu) + \sigma_\tau\cdot(v_\tau - u_\tau)\,\mathrm{d}s \\ &= \int_{\Gamma_c} G(v_\nu - u_\nu)\,\mathrm{d}s + \int_{\Gamma_c} -\sigma_\tau\cdot v_\tau - \mu_F G\|u_\tau\|\,\mathrm{d}s \\ &\le \int_{\Gamma_c} G(v_\nu - u_\nu)\,\mathrm{d}s + \int_{\Gamma_c} \mu_F G(\|v_\tau\| - \|u_\tau\|)\,\mathrm{d}s.\end{aligned}$$

Zusammenfassend lautet die Variationsungleichung für das Verschiebungsfeld $u \in V$

$$\begin{aligned}&\int_\Omega C\varepsilon(u) : \varepsilon(u-v)\,\mathrm{d}x + \int_{\Gamma_c} \mu_F G\|v_\tau\|\,\mathrm{d}s - \int_{\Gamma_c} \mu_F G\|u_\tau\|\,\mathrm{d}s \\ &\ge \int_\Omega f\cdot(v-u)\,\mathrm{d}x + \int_{\Gamma_g} g\cdot(v-u)\,\mathrm{d}s - \int_{\Gamma_c} G(v_\nu - u_\nu)\,\mathrm{d}s\end{aligned}$$

für alle $v \in V$. Das zugehörige Minimierungsproblem lautet

$$u \in V: \quad E(u) = \inf\{E(v) : v \in V\},$$

bei dem das Energiefunktional

$$E(v) = \frac{1}{2}\int_\Omega C\varepsilon(u) : \varepsilon(v)\,\mathrm{d}x - \int_\Omega f \cdot v\,\mathrm{d}x - \int_{\Gamma_g} g \cdot v\,\mathrm{d}s + \int_{\Gamma_c} G(v_\nu + \mu_F\|v_\tau\|)\,\mathrm{d}s$$

nicht differenzierbar ist. Für den nicht-differenzierbaren Teil von E ist der Term, der den Reibungseffekt berücksichtigt, verantwortlich.

Das letzte Beispiel fällt in die Klasse der *elliptischen Variationsungleichungen zweiter Art*. Neben der Bilinearform a und dem linearen Funktional ℓ wie oben führen wir ein konvexes und unterhalbstetiges Funktional $j : V \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ ein. Dann bezeichnet man das Problem

$$u \in V: \quad a(u, v-u) + j(v) - j(u) \geq \ell(v-u) \quad \text{für alle } v \in V \tag{8.9}$$

als **elliptische Variationsungleichung zweiter Art**.

Um die Existenz von Lösungen von (8.9) zu zeigen, benötigen wir das folgende Lemma.

Lemma 8.10. *Sei V ein normierter Raum. Ist $j : V \to \overline{\mathbb{R}}$ konvex und unterhalbstetig mit $j(v_0) < \infty$ für ein $v_0 \in V$, so existieren $\ell_j \in V'$ und $c_j \in \mathbb{R}$ mit*

$$j(v) \geq \ell_j(v) + c_j \quad \textit{für alle } v \in V.$$

Beweis. Sei $a_0 \in \mathbb{R}$ mit $a_0 < j(v_0)$. Weil j unterhalbstetig ist, überprüft man leicht, dass die Menge

$$J := \{(v, a) \in V \times \mathbb{R} : j(v) \leq a\}$$

abgeschlossen in $V \times \mathbb{R}$ ist. Weil j konvex ist, ist J außerdem konvex. Weil die Menge $\{(v_0, a_0)\}$ kompakt und konvex und J konvex und abgeschlossen ist, folgt aus ihrer Disjunktheit nach dem aus der Funktionalanalysis bekannten *Trennungssatz* die Existenz eines nicht-trivialen Funktionals $\ell \in V'$ und $\alpha \in \mathbb{R}$, so dass

$$\ell(v_0) + \alpha\, a_0 < \ell(v) + \alpha\, a \quad \text{für alle } (v, a) \in J. \tag{8.10}$$

Hieraus erhält man für die Wahl $v = v_0$ und $a = j(v_0)$

$$\alpha(j(v_0) - a_0) > 0$$

und somit $\alpha > 0$. Division von (8.10) durch α liefert für $a = j(v)$

$$j(v) > -\ell(v)/\alpha + a_0 + \ell(v_0)/\alpha.$$

Hieraus folgt die Behauptung mit $\ell_j := -\ell/\alpha$ und $c_j := a_0 + \ell(v_0)/\alpha$. □

Satz 8.11. *Sei V ein Hilbert-Raum, $a : V \times V \to \mathbb{R}$ eine stetige, V-koerzive Bilinearform, $\ell \in V'$ und $j : V \to \overline{\mathbb{R}}$ ein konvexes, unterhalbstetiges Funktional mit $j(v_0) < \infty$ für ein $v_0 \in V$. Dann besitzt (8.9) eine eindeutige Lösung.*

Beweis. Wir zeigen zunächst die Eindeutigkeit. Jede Lösung u von (8.9) erfüllt

$$j(u) \le a(u, v_0 - u) + j(v_0) - \ell(v_0 - u) < \infty,$$

d.h. $j(u)$ ist eine reelle Zahl. Seien u_1, u_2 zwei Lösungen von (8.9). Dann gilt

$$\begin{aligned} a(u_1, u_2 - u_1) + j(u_2) - j(u_1) &\ge \ell(u_2 - u_1), \\ a(u_2, u_1 - u_2) + j(u_1) - j(u_2) &\ge \ell(u_1 - u_2). \end{aligned}$$

Addition dieser beiden Ungleichungen liefert

$$a(u_1 - u_2, u_1 - u_2) \le 0$$

und wegen der V-Koerzivität $u_1 = u_2$.

Für den Existenzbeweis betrachten wir zunächst den Fall, dass a symmetrisch ist. In diesem Fall ist (8.9) äquivalent zum Minimierungsproblem

$$u \in V: \quad E(u) = \inf\{E(v) : v \in V\}, \tag{8.11}$$

wobei $E(v) := \frac{1}{2}a(v, v) + j(v) - \ell(v)$. Aus Lemma 8.10 folgt die Existenz von $\ell_j \in V'$ und $c_j \in \mathbb{R}$ mit

$$j(v) \ge \ell_j(v) + c_j \quad \text{für alle } v \in V.$$

Wegen der gemachten Annahmen an a, j und ℓ sieht man, dass E konvex, unterhalbstetig mit $E(v_0) < \infty$ und

$$E(v) \to \infty \quad \text{für } \|v\| \to \infty$$

ist. Nach Lemma 8.3 besitzt das Minimierungsproblem (8.11) und damit (8.9) eine eindeutige Lösung.

Wir betrachten nun den allgemeinen Fall einer unsymmetrischen Bilinearform a. Wir transformieren das Problem (8.9) in ein Fixpunktproblem. Für jedes $\theta > 0$ ist (8.9) äquivalent mit

$$u \in V: \quad (u, v - u) + \theta j(v) - \theta j(u) \ge (u, v - u) - \theta a(u, v - u) + \theta\ell(v - u) \quad \forall v \in V.$$

Für gegebenes $u \in V$ betrachte das Problem

$$w \in V: \quad (w, v - w) + \theta j(v) - \theta j(w) \ge (u, v - w) - \theta a(u, v - w) + \theta\ell(v - w) \quad \forall v \in V.$$

Wegen der Symmetrie von $(\cdot, \cdot)$ existiert nach obiger Diskussion eine eindeutige Lösung $w := P_\theta u$ dieses Problems. Offenbar ist jeder Fixpunkt von P_θ eine Lösung von (8.9). Wir zeigen, dass P_θ für genügend kleine θ eine Kontraktion ist. Dann folgt aus dem Banachschen Fixpunktsatz die Existenz eines eindeutigen Fixpunkts. Für $u_1, u_2 \in V$ setze $w_1 := P_\theta u_1$ und $w_2 := P_\theta u_2$. Dann gilt

$$\begin{aligned} (w_1, w_2 - w_1) + \theta j(w_2) - \theta j(w_1) &\ge (u_1, w_2 - w_1) - \theta a(u_1, w_2 - w_1) + \theta\ell(w_2 - w_1), \\ (w_2, w_1 - w_2) + \theta j(w_1) - \theta j(w_2) &\ge (u_2, w_1 - w_2) - \theta a(u_2, w_1 - w_2) + \theta\ell(w_1 - w_2). \end{aligned}$$

Addition der beiden Ungleichungen liefert

$$\|w_1 - w_2\|^2 \le (u_1 - u_2, w_1 - w_2) - \theta a(u_1 - u_2, w_1 - w_2) = ((I - \theta A)(u_1 - u_2), w_1 - w_2),$$

wobei der Operator A durch $a(v,w) = (Av, w)$ für alle $v, w \in V$ definiert ist. Hieraus folgt

$$\|w_1 - w_2\| \leq \|(I - \theta A)(u_1 - u_2)\|.$$

Wegen

$$\begin{aligned}\|(I-\theta A)u\|^2 = \|u - \theta Au\|^2 &= \|u\|^2 - 2\theta a(u,u) + \theta^2\|Au\|^2 \\ &\leq (1 - 2\theta c_K + \theta^2 c_S^2)\|u\|^2\end{aligned}$$

ist P_θ für alle $\theta \in (0, 2c_K/c_S^2)$ eine Kontraktion. □

Beispiel 8.12. Mit der Wahl

$$\begin{aligned}V &= \{v \in [H^1(\Omega)]^d : v = 0 \text{ auf } \Gamma_u\}, \\ a(v,w) &= \sum_{i,j,k,\ell=1}^{d} \int_\Omega C_{ijk\ell} v_{ij} w_{k\ell}, \\ j(v) &= \int_{\Gamma_c} \mu_F G \|v_\tau\| \,\mathrm{d}s, \\ \ell(v) &= \int_\Omega f \cdot v \,\mathrm{d}x + \int_{\Gamma_g} g \cdot v \,\mathrm{d}s - \int_{\Gamma_c} G v_\nu \,\mathrm{d}s,\end{aligned}$$

sieht man, dass das Gleitreibungsproblem Beispiel 8.9 eine elliptisches Variationsungleichung zweiter Art ist. Die Voraussetzungen von Satz 8.11 überprüft man leicht. Daher besitzt dieses Problem eine eindeutige Lösung.

8.2.1 Numerische Lösung

Es gelten die Voraussetzungen von Satz 8.11. Dann besitzt die folgende elliptische Variationsungleichung eine eindeutige Lösung.

$$u \in V: \quad a(u, v-u) + j(v) - j(u) \geq \ell(v-u) \quad \text{für alle } v \in V. \tag{8.12}$$

Sei $V_h \subset V$ ein Finite-Elemente-Raum. Dann lautet die Finite-Elemente-Approximation von (8.12)

$$u_h \in V_h: \quad a(u_h, v_h - u_h) + j(v_h) - j(u_h) \geq \ell(v_h - u_h) \quad \text{für alle } v_h \in V_h. \tag{8.13}$$

Unter der Annahme, dass $v_0 \in V_h$ existiert mit $j(v_0) < \infty$, besitzt auch (8.13) eine eindeutige Lösung. Wir zeigen wieder eine Verallgemeinerung des Lemmas von Céa.

Satz 8.13. *Es gilt*

$$\|u - u_h\|_V \leq c \inf_{v_h \in V_h} \left\{\|u - v_h\|_V + |a(u, v_h - u) + j(v_h) - j(u) - \ell(v_h - u)|^{1/2}\right\}$$

mit einer von h und u unabhängigen Konstanten $c > 0$.

Beweis. Wir setzen $v = u_h$ in (8.12) und addieren diese Ungleichung zu (8.13). Daraus folgt

$$a(u, u_h - u) + a(u_h, v_h - u_h) + j(v_h) - j(u) \geq \ell(v_h - u) \quad \text{für alle } v_h \in V_h$$

und mit der V-Koerzivität und der Beschränktheit von a

$$\begin{aligned} c_K \|u - u_h\|_V^2 &\leq a(u - u_h, u - u_h) \\ &= -a(u, u_h - u) - a(u_h, v_h - u_h) + a(u_h - u, v_h - u) + a(u, v_h - u) \\ &\leq a(u_h - u, v_h - u) + a(u, v_h - u) + j(v_h) - j(u) - \ell(v_h - u) \\ &\leq c_S \|u - u_h\|_V \|u - v_h\|_V + a(u, v_h - u) + j(v_h) - j(u) - \ell(v_h - u) \\ &\leq \frac{c_K}{2} \|u - u_h\|_V^2 + c\|u - v_h\|_V^2 + a(u, v_h - u) + j(v_h) - j(u) - \ell(v_h - u), \end{aligned}$$

woraus die Behauptung folgt. □

Wir verwenden den letzten Satz, um für folgendes Modell-Problem eine Fehlerabschätzung herzuleiten. Auf dem Gebiet $\Omega \subset \mathbb{R}^2$ mit Lipschitz-Rand $\partial\Omega$ betrachte $V := H^1(\Omega)$,

$$a(v, w) := \int_\Omega \nabla v \cdot \nabla w + vw \,\mathrm{d}x, \quad \ell(v) := \int_\Omega fv \,\mathrm{d}x, \quad j(v) := g \int_{\partial\Omega} |v| \,\mathrm{d}s \tag{8.14}$$

mit gegebenen $f \in L^2(\Omega)$ und $g > 0$. Der Einfachheit halber sei Ω polygonal und

$$\partial\Omega = \bigcup_{i=1}^{K} \Gamma_i,$$

wobei jedes Γ_i ein Liniensegment bezeichne. Hierbei handelt es sich um eine vereinfachte Version des Gleitreibungsproblems Beispiel 8.9.

Satz 8.14. *Sei $\mathcal{T}_h$ eine nicht-entartete Zerlegung von Ω und $V_h = \mathcal{S}^{1,0}(\mathcal{T}_h)$. Gilt für die Lösung von (8.12), dass $u \in H^2(\Omega)$ und $u|_{\Gamma_i} \in H^2(\Gamma_i)$, so folgt für die Lösung von (8.13)*

$$\|u - u_h\| \leq c(u)\, u.$$

Beweis. Zunächst gilt

$$\begin{aligned} &a(u, v_h - u) + j(v_h) - j(u) - \ell(v_h - u) \\ &\quad = \int_{\partial\Omega} \partial_\nu u(v_h - u) + g(|v_h| - |u|) \,\mathrm{d}s + \int_\Omega (u - \Delta u - f)(v_h - u) \,\mathrm{d}x \\ &\quad \leq \left(\|\partial_\nu u\|_{L^2(\partial\Omega)} + g\sqrt{|\partial\Omega|} \right) \|v_h - u\|_{L^2(\partial\Omega)} + \|u - \Delta u - f\|_{L^2(\Omega)} \|v_h - u\|_{L^2(\Omega)}. \end{aligned}$$

Nach Satz 8.13 folgt

$$\|u - u_h\|_{H^1(\Omega)} \leq c(u) \inf_{v_h \in V_h} \left\{ \|u - v_h\|_{H^1(\Omega)} + \|v_h - u\|_{L^2(\partial\Omega)}^{1/2} + \|v_h - u\|_{L^2(\Omega)}^{1/2} \right\}.$$

Die Behauptung folgt nun wie in Beispiel 8.8. □

Ein grundlegendes Problem bei der Behandlung des diskreten Problems (8.13) ist die Behandlung des nicht-differenzierbaren Terms j. Wir stellen zwei Zugänge vor, um dieses Problem zu lösen.

Regularisierung

Eine mögliche Vorgehensweise ist die Regularisierung von j, d.h. statt j verwende

$$j_\varepsilon(x) := \int_{B_\varepsilon(0)} j(x-y)\eta_\varepsilon(y)\,\mathrm{d}y$$

mit dem Glättungskern $\eta_\varepsilon : B_\varepsilon(0) \to \mathbb{R}$,

$$\eta_\varepsilon(x) := \frac{\eta(x/\varepsilon)}{\varepsilon^d \|\eta\|_{L^1(\mathbb{R}^d)}},$$

wobei $\varepsilon > 0$ und

$$\eta(x) := \begin{cases} \exp(4/(4\|x\|_2^2 - 1)), & \|x\|_2 \le 1/2, \\ 0, & \|x\|_2 > 1/2. \end{cases}$$

Dann gilt $j_\varepsilon \in C^\infty(\Omega)$ und für Lipschitz-stetige j gilt

$$\begin{aligned} |j(x) - j_\varepsilon(x)| &= \Big| \int_{B_\varepsilon(0)} [j(x) - j(x-y)]\eta_\varepsilon(y)\,\mathrm{d}y \Big| \le \int_{B_\varepsilon(0)} |j(x) - j(x-y)|\eta_\varepsilon(y)\,\mathrm{d}y \\ &\le \int_{B_\varepsilon(0)} L\|y\|_2 \eta_\varepsilon(y)\,\mathrm{d}y \le L\varepsilon \int_{B_\varepsilon(0)} \eta_\varepsilon(y)\,\mathrm{d}y = L\varepsilon. \end{aligned}$$

Das regularisierte Problem

$$u_\varepsilon \in V: \quad a(u_\varepsilon, v - u_\varepsilon) + j_\varepsilon(v) - j_\varepsilon(u_\varepsilon) \ge \ell(v - u_\varepsilon) \quad \text{für alle } v \in V$$

ist äquivalent zur nicht-linearen Variationsgleichung

$$u_\varepsilon \in V: \quad a(u_\varepsilon, v) + \langle j_\varepsilon'(u_\varepsilon), v\rangle = \ell(v) \quad \text{für alle } v \in V.$$

Lagrange-Multiplikatoren

Der folgende Beweis basiert auf der Äquivalenz von (8.12) und

$$a(u,v) + j(v) \ge \ell(v) \quad \text{für alle } v \in V, \tag{8.15a}$$
$$a(u,u) + j(u) = \ell(u). \tag{8.15b}$$

für Funktionale $j : V \to \mathbb{R}$. Wir betrachten beispielhaft das vereinfachte Gleitreibungsproblem (8.14). Dazu sei

$$\Lambda := \{\mu \in L^\infty(\partial\Omega) : |\mu(x)| \le 1 \text{ auf } \partial\Omega\}.$$

Satz 8.15. *Das Problem (8.14) ist äquivalent mit*

$$(u,\lambda) \in V \times \Lambda: \quad a(u,v) + g\int_{\partial\Omega} \lambda v\,\mathrm{d}s = \ell(v) \quad \textit{für alle } v \in V, \tag{8.16a}$$
$$\lambda u = |u| \quad \textit{auf } \partial\Omega. \tag{8.16b}$$

Beweis. Aus (8.15a) folgt $|L(v)| \leq j(v)$ für alle $v \in V$ mit $L(v) := \ell(v) - a(u,v)$. Daher hängt $L(v)$ nur von der Spur $v|_{\partial\Omega}$ ab, und L ist ein stetiges, lineares Funktional auf $H^{1/2}(\partial\Omega)$. Ferner erhält man die Abschätzung

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

Weil $H^{1/2}(\partial\Omega) \subset L^1(\partial\Omega)$, können wir L nach dem Satz von Hahn-Banach zu $\hat{L} \in [L^1(\partial\Omega)]'$ mit

$$\|\hat{L}\| = \|L\| \leq g$$

fortsetzen. Wegen $[L^1(\partial\Omega)]' = L^\infty(\partial\Omega)$ gibt es nach dem Satz von Riesz $\lambda \in \Lambda$ mit

$$\hat{L}(v) = g \int_{\partial\Omega} \lambda v \,\mathrm{d}s \quad \text{für alle } v \in L^1(\partial\Omega).$$

Daher gilt

$$\ell(v) - a(u,v) = L(v) = \hat{L}(v) = g \int_{\partial\Omega} \lambda v \,\mathrm{d}s \quad \text{für alle } v \in V$$

und somit (8.16a). Für $v = u$ erhält man insbesondere

$$a(u,u) + g \int_{\partial\Omega} \lambda u \,\mathrm{d}s = \ell(u).$$

Zusammen mit (8.15b) folgt

$$\int_{\partial\Omega} |u| - \lambda u \,\mathrm{d}s = 0.$$

Wegen $|\lambda| \leq 1$ auf $\partial\Omega$ folgt auch die zweite Bedingung (8.16b).

Umgekehrt erfülle $(u,\lambda) \in V \times \Lambda$ die Bedingungen (8.16). Dann folgt aus (8.16a) für $v - u$ statt v

$$a(u, v-u) + g \int_{\partial\Omega} \lambda v \,\mathrm{d}s - g \int_{\partial\Omega} \lambda u \,\mathrm{d}s = \ell(v-u).$$

Wegen

$$g \int_{\partial\Omega} \lambda u \,\mathrm{d}s = g \int_{\partial\Omega} |u| \,\mathrm{d}s = j(u), \quad g \int_{\partial\Omega} \lambda v \,\mathrm{d}s \leq g \int_{\partial\Omega} |v| \,\mathrm{d}s = j(v)$$

löst u das Problem (8.14). □

Hieraus kann man eine iterative Technik zur Lösung von (8.14) entwickeln. Sei $\rho > 0$ ein Parameter.

Algorithmus 8.16.
`Input:` $\lambda_0 \in \Lambda$ (z.B. $\lambda_0 = 0$)
`Output:` Folge (u_n, λ_n)

`for` $n = 0, 1, 2, \ldots$
finde $u_n \in V$ als Lösung des Variationsproblems

$$a(u_n, v) = \ell(v) - g \int_{\partial\Omega} \lambda_n v \,\mathrm{d}s \quad \text{für alle } v \in V;$$

bestimme den Lagrange-Multiplikator

$$\lambda_{n+1} = \mathcal{P}_\Lambda(\lambda_n + \rho g u_n),$$

wobei $\mathcal{P}_\Lambda : L^\infty(\partial\Omega) \to \mathbb{R}$ durch $\mathcal{P}_\Lambda \mu := \sup\{-1, \inf\{1, \mu\}\}$ definiert ist.

Man kann zeigen, dass ein $\rho_0 > 0$ existiert, so dass für alle $\rho \in (0, \rho_0)$ gilt

$$(u_n, \lambda_n) \to (u, \lambda) \in V \times \Lambda.$$