Numerische Behandlung von Differentialgleichungen II (Finite Elemente)

Vorlesungsskriptum Sommersemester 1998

R. Verfürth

Fakultät für Mathematik, Ruhr-Universität Bochum, D-44780 Bochum, Germany

Inhalt

1. Sobolev Räume

Im folgenden bezeichnet $\Omega \subset \mathbb{R}^d, d \geq 1$, stets eine offene, beschränkte Menge, $p \in [1, \infty)$ einen Lebesgue-Exponenten mit dualem Exponenten $p' \in (1, \infty]$, $\frac{1}{p} + \frac{1}{p'} = 1$, und $\alpha \in \mathbb{N}^d$ einen Multiindex mit $|\alpha| := \alpha_1 + ... + \alpha_d$ und

$$D^\alpha \varphi := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} ... \partial x_d^{\alpha_d}} \varphi \qquad \forall \varphi \in C^{|\alpha|}(\Omega).$$

Da Ω beschränkt ist, gilt $L^p(\Omega) \subset L^1(\Omega)$, und die kanonische Injektion ist stetig. Aus dem Gauß'schen Integralsatz folgt für alle $\varphi, \psi \in C_0^\infty(\Omega)$ und alle $\alpha \in \mathbb{N}^d$

$$\int_\Omega \varphi D^\alpha \psi = (-1)^{|\alpha|} \int_\Omega \psi D^\alpha \varphi. \tag{1}$$

1.1 Definition: Seien $\varphi, \psi \in L^1(\Omega)$ und $\alpha \in \mathbb{N}^d$. Dann heißt ψ die **α-te schwache Ableitung** von φ, kurz $\psi = D^\alpha \varphi$, wenn für alle $\rho \in C_0^\infty(\Omega)$ gilt

$$\int_\Omega \varphi D^\alpha \rho = (-1)^{|\alpha|} \int_\Omega \psi \rho. \tag{2}$$

□

1.2 Bemerkung: **(1)** Die α-te schwache Ableitung ist, sofern sie existiert, eindeutig (im Sinne von L^1-Funktionen).
(2) Ist $\varphi \in C^{|\alpha|}(\Omega)$, so stimmen die α-te schwache Ableitung und die klassische α-te Ableitung überein. □

Beweis: **ad (1):** Seien $\varphi, \psi_1, \psi_2 \in L^1(\Omega)$ mit

$$(-1)^{|\alpha|} \int_\Omega \psi_1 \rho = \int_\Omega \varphi D^\alpha \rho = (-1)^{|\alpha|} \int_\Omega \psi_2 \rho \qquad \forall \rho \in C_0^\infty(\Omega).$$

Dann gilt

$$\int_\Omega (\psi_1 - \psi_2) \rho = 0 \qquad \forall \rho \in C_0^\infty(\Omega).$$

Da $C_0^\infty(\Omega)$ dicht ist in $L^1(\Omega)$, folgt $\psi_1 = \psi_2$ fast überall.
ad (2): Folgt aus dem Gauß'schen Integralsatz (vgl. (1)). □

1.3 Beispiel: Sei $\Omega = (-1, 1)$ und $\varphi(x) = |x|$. Dann ist φ im Sinne von Def. 1.1 differenzierbar und die Ableitung ist

$$\psi(x) := \begin{cases} -1 & \text{für } -1 < x < 0 \\ 1 & \text{für } 0 < x < 1. \end{cases}$$

Denn für alle $\rho \in C_0^\infty(\Omega)$ gilt

$$\begin{aligned}\int_{-1}^{1} \varphi\rho' &= \int_{-1}^{0} \varphi\rho' + \int_{0}^{1} \varphi\rho' \\ &= \varphi(0)\rho(0) - \varphi(-1)\rho(-1) + \int_{-1}^{0} \rho \\ &\quad - \varphi(0)\rho(0) + \varphi(1)\rho(1) - \int_{0}^{1} \rho \\ &= -\int_{-1}^{1} \psi\rho,\end{aligned}$$

da $\rho(\pm 1) = 0$ ist. □

1.4 Definition: **(1)** Für $k \in \mathbb{N}$ und $p \in [1, \infty)$ definieren wir den Sobolev Raum $W^{k,p}(\Omega)$ und seine Norm $\|.\|_{k,p}$ durch

$$\begin{aligned}W^{k,p}(\Omega) &:= \big\{\varphi \in L^p(\Omega) : D^\alpha\varphi \in L^p(\Omega) \quad \forall|\alpha| \le k\big\}, \\ \|\varphi\|_{k,p} &:= \Big\{\sum_{|\alpha|\le k} \|D^\alpha\varphi\|^p_{L^p(\Omega)}\Big\}^{1/p}.\end{aligned}$$

(2) Durch

$$|\varphi|_{k,p} := \Big\{\sum_{|\alpha|=k} \|D^\alpha\varphi\|^p_{L^p(\Omega)}\Big\}^{1/p}$$

wird für $k \in \mathbb{N}^*$ eine Semi-Norm auf $W^{k,p}(\Omega)$ definiert.
(3) Ist speziell $p = 2$, so schreiben wir $H^k(\Omega)$ statt $W^{k,2}(\Omega)$ und lassen den Index $p = 2$ bei der Norm und Semi-Norm weg. □

1.5 Beispiel: **(1)** Seien φ und Ω wie in Bsp. 1.3. Dann gilt $\varphi \in W^{1,p}(\Omega)$ für alle $p \in [1, \infty)$.
(2) Seien $d \ge 2$, $\Omega = B(0, \frac{1}{2})$ und $\varphi(x) := |x|^s$ mit $s \in \mathbb{R}$, wobei $|.|$ die euklidische Norm in $\mathbb{R}^d$ bezeichnet. Dann gilt

$$D^\alpha\varphi(x) \sim |x|^{s-|\alpha|}$$

und

$$\begin{aligned}&\|D^\alpha\varphi\|^p_{L^p(\Omega)} \sim \omega_{d-1}\int_0^{\frac{1}{2}} r^{(s-|\alpha|)p} r^{d-1} dr < \infty \\ \iff &p(s-|\alpha|) + d - 1 > -1 \\ \iff &s > |\alpha| - \frac{d}{p}.\end{aligned}$$

(3) Sei $d=2$, $\Omega = B(0,\frac{1}{2})$ und $\varphi(x) = \ln\big|\ln(|x|)\big|$. Für $x \neq 0$ und $i \in \{1,2\}$ ist

$$\frac{\partial \varphi}{\partial x_i} = \frac{x_i}{|x|^2\big|\ln(|x|)\big|}.$$

Hieraus folgt

$$\begin{aligned}\int_\Omega \sum_{i=1}^{2} \Big|\frac{\partial \varphi}{\partial x_i}\Big|^2 =& 2\pi \int_0^{1/2} \frac{1}{r^2(\ln r)^2} r dr \\ =& 2\pi \lim_{\varepsilon \to 0}\Big[-\frac{1}{\ln r}\Big]_{r=\varepsilon}^{r=\frac{1}{2}} \\ =& \frac{2\pi}{\ln 2}.\end{aligned}$$

Also ist $\varphi \in H^1(\Omega)$. □

1.6 Satz: **(1)** *$(W^{k,p}(\Omega), \|.\|_{k,p})$ ist ein Banach-Raum.*
(2) *$C^\infty(\Omega)$ ist dicht in $W^{k,p}(\Omega)$.*
(3) *$H^k(\Omega)$ ist ein Hilbertraum mit Skalarprodukt*

$$(\varphi,\psi)_k := \sum_{|\alpha|\leq k} \int_\Omega D^\alpha \varphi D^\alpha \psi.$$

Beweis: **ad (1):** Sei $n_{k,d} := \#\{\alpha \in \mathbb{N}^d : |\alpha| \leq k\}$. Dann können wir $W^{k,p}(\Omega)$ vermittels der Abbildung $i : \varphi \to (D^\alpha \varphi)_{|\alpha|\leq k}$ mit $L^p(\Omega;\mathbb{R}^{n_{k,d}})$ identifizieren. Insbesondere ist dann $\|\varphi\|_{k,p} = \|i(\varphi)\|_{L^p(\Omega:\mathbb{R}^{n_{k,d}})}$. Hieraus folgt sofort die Normeigenschaft von $\|.\|_{k,p}$. Sei nun $(\varphi_n)_{n\in\mathbb{N}} \subset W^{k,p}(\Omega)$ eine Cauchy-Folge. Dann ist $(i(\varphi_n))_{n\in\mathbb{N}} \subset L^p(\Omega;\mathbb{R}^{n_{k,d}})$ ebenfalls eine Cauchy Folge und damit konvergent. Daher gibt es zu jedem $\alpha \in \mathbb{N}^d$ mit $|\alpha| \leq k$ ein $\psi_\alpha \in L^p(\Omega)$, so daß $D^\alpha \varphi_n$ in $L^p(\Omega)$ gegen ψ_α konvergiert. Insbesondere konvergiert $D^\alpha \varphi_n$ punktweise f.ü. gegen ψ_α. Für jedes $\rho \in C_0^\infty(\Omega)$ gilt andererseits

$$\int_\Omega \varphi_n D^\alpha \rho = (-1)^{|\alpha|} \int_\Omega D^\alpha \varphi_n \rho. \tag{3}$$

Wegen des Lebesgueschen Konvergenzsatzes können wir in (3) den Grenzübergang $n \to \infty$ durchführen und erhalten

$$\begin{aligned}\int_\omega \psi_0 D^\alpha \rho &= \lim_{n\to\infty} \int_\Omega \varphi_n D^\alpha \rho \\ &= \lim_{n\to\infty} (-1)^{|\alpha|} \int_\Omega D^\alpha \varphi_n \rho = (-1)^{|\alpha|} \int_\Omega \psi_\alpha \rho.\end{aligned}$$

Also ist ψ_α die α-te schwache Ableitung von ψ_0, und $(\varphi_n)_{n\in\mathbb{N}}$ konvergiert in $W^{k,p}(\Omega)$ gegen ψ_0.
ad (2): Kopiere den Beweis von "$C^\infty(\Omega)$ ist dicht in $L^p(\Omega)$".
ad (3): Offensichtlich ist $(.,.)_k$ bilinear und $\|\varphi\|_k^2 = (\varphi,\varphi)_k$. Damit folgt die Behauptung aus Teil (1). □

Im folgenden werden wir häufig Funktionen begegnen, die stückweise glatt sind. Der folgende Satz gibt uns ein Kriterium, wann solche Funktionen in $W^{k,p}(\Omega)$ sind.

1.7 Satz: *Seien Ω_1, Ω_2 zwei nichtleere, offene, beschränkte und disjunkte Teilmengen von Ω mit stückweise glattem Rand und $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2$. Weiter sei $\varphi \in L^p(\Omega)$ so, daß $\varphi|_{\Omega_i} \in C^k(\Omega_i), i \in \{1,2\}$, $k \geq 1$ ist. Dann ist $\varphi \in W^{k,p}(\Omega)$ genau dann, wenn $\varphi \in C^{k-1}(\Omega)$ ist.*

Beweis: Es reicht den Fall $k = 1$ zu betrachten. Der allgemeine Fall folgt dann durch Induktion. Sei n die äußere Normale an Ω_1 und $[\varphi]$ der Sprung von φ über $\Sigma := \partial\Omega_1 \cap \partial\Omega_2$ in Richtung n, d.h.

$$[\varphi](x) = \lim_{t\to 0+} \varphi(x + tn) - \lim_{t\to 0+} \varphi(t - tn) \quad \forall x \in \Sigma.$$

Seien $\rho \in C_0^\infty(\Omega)$ und $i \in \{1, ...d\}$ beliebig. Dann folgt aus dem Gauß'schen Integralsatz

$$\begin{aligned} -\int_\Omega \varphi \frac{\partial \rho}{\partial x_i} &= -\int_{\Omega_1} \varphi \frac{\partial \rho}{\partial x_i} - \int_{\Omega_2} \varphi \frac{\partial \rho}{\partial x_i} \\ &= \int_{\Omega_1} \frac{\partial \varphi}{\partial x_i}\rho - \int_{\partial\Omega_1} \varphi\rho n_i \\ &\quad + \int_{\Omega_2} \frac{\partial \varphi}{\partial x_i}\rho + \int_{\partial\Omega_2} \varphi\rho n_i \\ &= \int_\Omega \frac{\partial \varphi}{\partial x_i}\rho + \int_\Sigma [\varphi]\rho n_i. \end{aligned}$$

Ist also $\varphi \in W^{1,p}(\Omega)$, so folgt

$$\int_\Sigma [\varphi]\rho n_i = 0 \quad \forall \rho \in C_0^\infty(\Omega), i \in \{1, ..., d\}.$$

Also ist $[\varphi] = 0$ f.ü. auf Σ, d.h. aber $\varphi \in C(\Omega)$.
Ist umgekehrt $\varphi \in C(\Omega)$, so verschwindet $[\varphi]$ auf Σ, und aus obiger Identität folgt $\varphi \in W^{1,p}(\Omega)$. □

1.8 Bemerkung: Gemäß Satz 1.6 (2) ist $C^\infty(\Omega)$ dicht in $W^{k,p}(\Omega)$. Für $C_0^\infty(\Omega)$ gilt dies aber i.a. **nicht**. Betrachte z.B. $d = 1, \Omega = (0,1)$ und $\varphi(x) = x$. Sei $\rho \in C_0^\infty(\Omega)$ beliebig. Wegen $\varphi(0) = \rho(0) = \rho(1) = 0$ folgt aus dem Hauptsatz der Differential- und Integralrechnung und der Cauchy-Schwarz'schen Ungleichung

$$\begin{aligned} 1 &= \varphi(1) - \rho(1) - [\varphi(0) - \rho(0)] \\ &= \int_0^1 [\varphi'(t) - \rho'(t)]dt \\ &\leq \Big\{ \int_0^1 |\varphi' - \rho'|^2 dt \Big\}^{1/2} \\ &\leq \|\varphi - \rho\|_1. \end{aligned}$$

Also kann $C_0^\infty(\Omega)$ nicht dicht in $H^1(\Omega)$ sein. □

1.9 Definition: $W_0^{k,p}(\Omega), k \geq 1$, ist die Vervollständigung von $C_0^\infty(\Omega)$ bzgl. $\|.\|_{k,p}$; $H_0^k(\Omega) := W_0^{k,2}(\Omega)$. □

1.10 Definition: Wir sagen, Ω hat einen **Lipschitz-Rand** bzw. Ω ist ein **Lipschitz-Gebiet**, wenn es ein $N \in \mathbb{N}^*$ und offene Mengen $U_1, ..., U_N \in \mathbb{R}^d$ mit folgenden Eigenschaften gibt:

(1) $\partial\Omega \subset \bigcup_{1 \leq i \leq N} U_i$,

(2) Für jedes $1 \leq i \leq N$ ist $\partial\Omega \cap U_i$ darstellbar als Graph einer Lipschitz-stetigen Funktion.

□

1.11 Bemerkung: **(1)** Ω sei ein Lipschitz-Gebiet. Dann existiert fast überall auf $\partial\Omega$ das äußere Einheitsnormalenfeld n zu Ω.

(2) Ω habe einen stückweise glatten Rand. Zudem gebe es zu jedem $x_0 \in \partial\Omega$ einen nichttrivialen Kegel K_0 mit Basis x_0 und $\Omega \subset \mathbb{R}^d \backslash K_0$ (**Kegelbedingung**). Dann ist Ω ein Lipschitz-Gebiet. □

1.12 Satz: **(Spursatz)** *Seien Ω ein Lipschitz-Gebiet und $k \in \mathbb{N}^*, l \in \{0, ..., k-1\}$. Dann gibt es eine stetige lineare Abbildung $\gamma_l : W^{k,p}(\Omega) \to L^p(\partial\Omega)$ mit der Eigenschaft*

$$\gamma_l(\varphi) = \frac{\partial^l}{\partial n^l}\varphi|_{\partial\Omega} \qquad \forall \varphi \in C^k(\overline{\Omega}).$$

Beweis: R.A. Adams: Sobolev Spaces. Academic Press, 1975.

Idee: Man zeigt zunächst, daß die Restriktionen von $C_0^\infty(\mathbb{R}^d)$-Funktionen auf Ω dicht sind in $W^{k,p}(\Omega)$. Dann führt man eine Überdeckung von $\partial\Omega$ wie in Definition 1.10 ein und rechnet die Eigenschaft von γ_l auf den Karten $\partial\Omega \cap U_i$ für $C_0^\infty(\mathbb{R}^d)$-Funktionen nach. □

1.13 Bemerkung: Die Bezeichnungen und Voraussetzungen seien wie in Satz 1.12. Wegen des Satzes vom abgeschlossenen Graphen ist $\gamma_l(W^{k,p}(\Omega))$ ein abgeschlossener Unterraum von $L^p(\partial\Omega)$. Dieser wird üblicherweise mti $W^{k-l-\frac{1}{p},p}(\partial\Omega)$ bezeichnet. Für unsere Anwendungen sind die Fälle $l = 0$ und $l = 1$ besonders wichtig. Für eine alternative Charakterisierung der Räume $W^{k-l-\frac{1}{p},p}(\partial\Omega)$ analog zu Definition 1.4 (mit Ω ersetzt durch $\partial\Omega$) verweisen wir auf das Buch von R.A. Adams. □

1.14 Satz: $W_0^{k,p}(\Omega) = \{\varphi \in W^{k,p}(\Omega) : \gamma_l(\varphi) = 0 \; \forall 0 \leq l \leq k-1\}$.

Beweis: R.A. Adams; loc. cit.

Idee: Da die γ_l stetig sind, ist $\bigcap_{0 \leq l \leq k-1} \ker\gamma_l$ ein abgeschlossener Unterraum von $W^{k,p}(\Omega)$, der $C_0^\infty(\Omega)$ enthält. Hieraus folgt mit Definition 1.9 die Behauptung. □

1.15 Satz: (Friedrich'sche Ungleichung) *$\|.\|_{k,p}$ und $|.|_{k,p}$ sind äquivalente Normen auf $W_0^{k,p}(\Omega)$.*

Beweis: Offensichtlich gilt $|\varphi|_{k,p} \leq \|\varphi\|_{k,p}$ für alle $\varphi \in W^{k,p}(\Omega)$.
Für die umgekehrte Abschätzung wähle $R \in \mathbb{R}_+^*$, so daß $\Omega \subset B_{|.|_\infty}(0,R)$. Dabei bezeichnet $|.|_\infty$ die Maximum-Norm auf $\mathbb{R}^d$. Sei $\alpha \in \mathbb{N}^d$ mit $|\alpha| = k-1$ und $\varphi \in C_0^\infty(\Omega)$, $\psi := D^\alpha \varphi$. Dann ist $\psi \in C_0^\infty(\Omega)$. Wegen $\Omega \subset B_{|.|_\infty}(0,R)$ folgt für beliebiges $x \in \Omega$ mit der Hölder'schen Ungleichung

$$\begin{aligned}|\psi(x)|^p =& \Big| \int_{-R}^{x} \frac{\partial}{\partial x_1} \psi(y, x_2, ..., x_d) dy \Big|^p \\ \leq & (2R)^{p-1} \int_{-R}^{R} \Big| \frac{\partial \psi}{\partial x_1}(y, x_2, ..., x_d) \Big|^p dy.\end{aligned}$$

Integration über Ω liefert

$$\begin{aligned}\|\psi\|_{L^p(\Omega)}^p =& \int_\Omega |\psi(x)|^p \\ \leq & \int_{B_{|.|_\infty}(0,R)} |\psi(x)|^p \\ \leq & (2R)^{p-1} \int_{-R}^{R} \int_{B_{|.|_\infty}(0,R)} \Big| \frac{\partial \psi}{\partial x_1}(y, x_2, ..., x_d) \Big|^p dy dx \\ = & (2R)^p \int_{B_{|.|_\infty}(0,R)} \Big| \frac{\partial \psi}{\partial x_1} \Big|^p \\ = & (2R)^p \Big\| \frac{\partial \psi}{\partial x_1} \Big\|_{L^p(\Omega)}^p.\end{aligned}$$

Summation über alle Multiindizes $\alpha \in \mathbb{N}^d$ mit $|\alpha| \leq k-1$ ergibt

$$|\varphi|_{k-1,p}^p \leq c_{k-1} |\varphi|_{k,p}^p$$

mit

$$c_{k-1} = (2R)^p \frac{(k+d-2)!}{d!(k-1)!}.$$

Hieraus folgt

$$\begin{aligned}\|\varphi\|_{k,p}^p =& |\varphi|_{k,p}^p + \sum_{l=0}^{k-1} |\varphi|_{l,p}^p \\ \leq & |\varphi|_{k,p}^p + \sum_{l=0}^{k-1} c_l |\varphi|_{l+1,p}^p \\ \leq & \big\{ 1 + c_{k-1} + c_{k-1}c_{k-2} + ... + c_{k-1}...c_1 c_0 \big\} |\varphi|_{k,p}^p.\end{aligned}$$

Hieraus folgt die Behauptung, da $C_0^\infty(\Omega)$ dicht ist in $W_0^{k,p}(\Omega)$. □

1.16 Bemerkung: Aus Satz 1.15 folgt $\|\varphi\|_{k,p} \leq c_k(\Omega)|\varphi|_{k,p}$ für alle $\varphi \in W_0^{k,p}(\Omega)$. Die Konstante $c_k(\Omega)$ hängt nur von k und dem Durchmesser von Ω ab. Eine analoge Abschätzung gilt für alle Funktionen, die auf einem Teil des Randes verschwinden, der positives $(d-1)$-dimensionales Maß hat. □

1.17 Definition: Seien $(X, \|.\|_X)$ und $(Y, \|.\|_Y)$ zwei normierte Vektorräume.
(1) Eine lineare Abbildung $A : X \to Y$ heißt **kompakt**, wenn das Bild $A(\overline{B_X(0;1)})$ der abgeschlossenen X-Einheitskugel in Y kompakt ist.
(2) X ist **stetig eingebettet** in Y, kurz $X \hookrightarrow Y$, wenn $X \subset Y$ und die kanonische Injektion $i : X \to Y$ stetig ist.
(3) X ist **kompakt eingebettet** in Y, kurz $X \overset{c}{\hookrightarrow} Y$, wenn $X \subset Y$ und die kanonische Injektion $i : X \to Y$ kompakt ist.

□

1.18 Bemerkung: **(1)** Gilt $X \hookrightarrow Y$, so gibt es eine Konstante $c > 0$ mit $\|\varphi\|_Y \leq c\|\varphi\|_X$ für alle $\varphi \in X$.
(2) Aus $X \overset{c}{\hookrightarrow} Y$ folgt $X \hookrightarrow Y$.
(3) Ist $X \overset{c}{\hookrightarrow} Y$ und $(\varphi_n)_{n\in\mathbb{N}} \subset X$ eine beschränkte Folge, so besitzt $(\varphi_n)_{n\in\mathbb{N}}$ eine in Y konvergente Teilfolge. □

Beweis: **ad (1):** Folgt aus der Definition der Stetigkeit für lineare Operatoren.
ad (2): Sei $A : X \to Y$ ein kompakter linearer Operator. Dann ist n.V. $A(\overline{B_X(0;1)})$ kompakt und somit insbesondere beschränkt. Also gibt es ein $c > 0$ mit $\|A\varphi\|_Y \leq C$ für alle $\varphi \in X$ mit $\|\varphi\|_X \leq 1$. Also ist A stetig.
ad (3): $(i(\varphi_n))_{n\in\mathbb{N}} \subset Y$ ist in der kompakten Menge $i(\overline{B_{\|.\|_X}(0;R)})$ mit $R := \max\limits_{n\in\mathbb{N}} \|\varphi_n\|_X$ enthalten. □

1.19 Satz: (Sobolev'scher Einbettungssatz) (1) *Sei $p < d$. Dann gilt $W^{k,p}(\Omega) \hookrightarrow W^{k-1,q}(\Omega)$ für alle $q \in [1, \frac{pd}{d-p}]$ und $W^{k,p}(\Omega) \overset{c}{\hookrightarrow} W^{k-1,q}(\Omega)$ für alle $q \in [1, \frac{pd}{d-p})$.*
(2) *Sei $p = d$. Dann gilt $W^{k,p}(\Omega) \overset{c}{\hookrightarrow} W^{k-1,q}(\Omega)$ für alle $q \in [1, \infty)$.*
(3) *Sei $k > \frac{d}{p}$. Dann gilt $W^{k,p}(\Omega) \overset{c}{\hookrightarrow} C^l(\overline{\Omega})$ für alle $l \in \mathbb{N}$ mit $0 \leq l < k - \frac{d}{p}$.*
Beweis: R.A. Adams; loc cit. □

1.20 Bemerkung: **(1)** Sei $d = 2$, $p = 2$ und $\Omega = B(0; \frac{1}{2})$. Dann zeigt Beispiel 1.5 (3), daß die Schranke an q in Satz 1.19 (2) scharf ist.
(2) Sei $d \geq 3$, $p = 2$ und $\Omega = B(0; \frac{1}{2})$. Dann zeigt Beispiel 1.5 (2), daß die Schranke an q in Satz 1.19 (1) scharf ist.
(3) Sei $p = 2$ und $d = 2$. Dann ist $H^1(\Omega) \overset{c}{\hookrightarrow} L^q(\Omega)$ für jedes $q \in [1, \infty)$.
(4) Sei $p = 2$ und $d = 3$. Dann ist $H^1(\Omega) \overset{c}{\hookrightarrow} L^q(\Omega)$ für jedes $q \in [1, 6)$ und $H^1(\Omega) \hookrightarrow L^6(\Omega)$.
(5) Sei $p = 2$ und $d \in \{2, 3\}$. Dann ist $H^2(\Omega) \hookrightarrow C^0(\overline{\Omega})$. Für $H^1(\Omega)$-Funktionen sind Punktwerte dagegen **nicht** definiert. □

1.21 Satz: (Poincaré'sche Ungleichung) $|.|_1$ *und* $\|.\|_1$ *sind äquivalente Normen auf* $V := \{\varphi \in H^1(\Omega) : \int_\Omega \varphi = 0\}$.

Beweis: Wie im Beweis von Satz 1.15 müssen wir nur zeigen, daß es eine Konstante $C > 0$ gibt mit

$$\|\varphi\|_1 \leq C|\varphi|_1 \qquad \forall \varphi \in V. \tag{4}$$

Wir nehmen an, eine solche Konstante existiere nicht. Dann gibt es eine Folge $(\varphi_n)_{n\in\mathbb{N}} \subset V$ mit

$$\|\varphi_n\|_1 = 1 \qquad \forall n \in \mathbb{N} \tag{5}$$

und

$$\lim_{n\to\infty} |\varphi_n|_1 = 0. \tag{6}$$

Wegen Satz 1.19 und Bem. 1.18 (3) gibt es eine Teilfolge $(\varphi_{n_k})_{k\in\mathbb{N}}$ von $(\varphi_n)_{n\in\mathbb{N}}$ und eine Funktion $\varphi \in L^2(\Omega)$ mit

$$\lim_{k\to\infty} \|\varphi_{n_k} - \varphi\|_0 = 0.$$

Wegen (6) konvergiert $(\varphi_{n_k})_{k\in\mathbb{N}}$ sogar in $H^1(\Omega)$. Mithin ist $\varphi \in H^1(\Omega)$ und $|\varphi|_1 = 0$. Daher ist φ konstant. Da V ein abgeschlossener Unterraum von $H^1(\Omega)$ ist, gilt aber $\int_\Omega \varphi = 0$. Also ist $\varphi = 0$ im Widerspruch zu (5). □

1.22 Bemerkung: **(1)** Satz 1.21 kann für $H^1(\Omega)$ nicht gelten, da die rechte Seite von (4) für die konstante Funktion $\varphi \equiv 1$ verschwindet.
(2) Der Beweis von Satz 1.21 ist nicht konstruktiv. Mit anderen Techniken kann man zeigen, daß die Konstante C in (4) proportional zum Durchmesser $d := \sup\limits_{x,y\in\Omega} |x - y|$ von Ω ist. Ist insbesondere Ω konvex, ergibt sich $C \leq \frac{d}{\pi}$.
(3) Analoge Aussagen zu Satz 1.21 gelten für die Räume $\{\varphi \in W^{1,p}(\Omega) : \int_\Omega \varphi = 0\}$ mit $p \in (1, \infty)$. □

2. Abstrakte Variationsprobleme

In diesem Paragraphen erinnern wir an die abstrakten Ergebnisse des § II.5 der Vorlesung "Numerische Behandlung von Differentialgleichungen I". Wegen ihrer Bedeutung für das Folgende führen wir diese Ergebnisse und ihre Beweise nochmals auf.

2.1 Satz: (Lax-Milgram) *Seien* $(X, \|.\|_X)$ *ein Banach Raum,* $l \in \mathcal{L}(X, \mathbb{R})$ *ein stetiges lineares Funktional und* $a \in \mathcal{L}^2(X, \mathbb{R})$ *eine stetige Bilinearform. Zusätzlich sei a symmetrisch, d.h.*

$$a(u, v) = a(v, u) \quad \forall u, v \in X,$$

und koerziv, d.h., es gibt ein $\alpha > 0$ mit

$$a(u,u) \geq \alpha \|u\|_X^2 \quad \forall u \in X.$$

Dann besitzt das Funktional $J \in C^2(X, \mathbb{R})$ mit

$$J(u) := \frac{1}{2} a(u,u) - l(u) \tag{1}$$

ein eindeutiges Minimum u^* in X. Dieses ist die eindeutige Lösung von

$$a(u^*, v) = l(v) \quad \forall v \in X. \tag{2}$$

Beweis: **1. Schritt:** Offensichtlich ist $J \in C^2(X, \mathbb{R})$ mit

$$DJ(u)v = a(u,v) - l(v) \quad \forall u, v \in X.$$

Also ist jeder kritische Punkt von J eine Lösung von (2).
2. Schritt: Seien $u_1, u_2 \in X$ zwei Lösungen von (2). Dann folgt

$$a(u_1 - u_2, v) = 0 \quad \forall v \in X$$

und

$$\alpha \|u_1 - u_2\|_X^2 \leq a(u_1 - u_2, u_1 - u_2) = 0.$$

Also besitzt (2) höchstens eine Lösung.
3. Schritt: Für alle $u \in X$ gilt

$$\begin{aligned} J(u) &\geq \frac{\alpha}{2} \|u\|_X^2 - \|l\|_{\mathcal{L}(X,\mathbb{R})} \|u\|_X \\ &\geq \frac{\alpha}{4} \|u\|_X^2 - \frac{1}{\alpha} \|l\|_{\mathcal{L}(X,\mathbb{R})}^2 \\ &\geq -\frac{1}{\alpha} \|l\|_{\mathcal{L}(X,\mathbb{R})}^2. \end{aligned}$$

Also ist J nach unten beschränkt. Sei

$$\rho := \inf_{u \in X} J(u) \in \mathbb{R}$$

und $(u_n)_{n \in \mathbb{N}}$ eine Minimalfolge, d.h.

$$\rho = \lim_{n \to \infty} J(u_n).$$

Dann folgt für $n, m \in \mathbb{N}$

$$\begin{aligned} \alpha \|u_n - u_m\|_X^2 &\leq a(u_n - u_m, u_n - u_m) \\ &= 8 \Big\{ \frac{1}{2} J(u_n) + \frac{1}{2} J(u_m) - J(\frac{1}{2}(u_n + u_m)) \Big\} \\ &\leq 8 \Big\{ \frac{1}{2} J(u_n) + \frac{1}{2} J(u_m) - \rho \Big\} \\ &\underset{n,m \to \infty}{\longrightarrow} 0. \end{aligned}$$

Also ist $(u_n)_{n \in \mathbb{N}}$ eine Cauchy-Folge und konvergiert gegen ein $u^* \in X$ mit

$$J(u^*) = \rho.$$

Also besitzt J mindestens ein Minimum. Zusammen mit den Schritten 1 und 2 folgt hieraus die Behauptung. □

2.2 Satz: *Die Voraussetzungen und Bezeichnungen seien wie in Satz 2.1. Setze zur Abkürzung*

$$A := \|a\|_{\mathcal{L}^2(X,\mathbb{R})}.$$

Sei $X_h \subset X$ ein endlich dimensionaler Unterraum von X. Bezeichne mit $u \in X$ und $u_h \in X_h$ das eindeutige Minimum von J in X bzw. X_h. Dann gilt

$$\|u - u_h\|_X \leq \frac{A}{\alpha} \inf_{v_h \in X_h} \|u - v_h\|_X.$$

Sei zusätzlich H ein Hilbert Raum mit Skalarprodukt $(.,.)_H$ und Norm $\|.\|_H$ derart, daß $X \hookrightarrow H$ und bzgl. $\|.\|_H$ dicht ist in H. Für jedes $\varphi \in H$ bezeichne $u_\varphi \in X$ die eindeutige Lösung von

$$a(v, u_\varphi) = (\varphi, v)_H \quad \forall v \in X. \tag{3}$$

Dann gilt

$$\|u - u_h\|_H \leq A\|u - u_h\|_X \sup_{\substack{\varphi \in H \\ \|\varphi\|_H = 1}} \inf_{v_h \in X_h} \|u_\varphi - v_h\|_X.$$

Beweis: Wegen Satz 2.1 besitzt J ein eindeutiges Minimum u_h in X_h. Dieses ist eindeutig charakterisiert durch

$$a(u_h, v_h) = l(v_h) \quad \forall v_h \in X_h. \tag{4}$$

Aus (2) und (4) folgt

$$a(u - u_h, v_h) = 0 \quad \forall v_h \in X_h. \tag{5}$$

Hieraus ergibt sich für jedes $v_h \in X_h$

$$\begin{aligned}
\alpha\|u - u_h\|_X^2 &\leq a(u - u_h, u - u_h) \\
&= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \\
&= a(u - u_h, u - v_h) \\
&\leq A\|u - u_h\|_X \|u - v_h\|_X.
\end{aligned}$$

Da v_h beliebig war, folgt hieraus die erste Fehlerabschätzung.
Wegen $X \subset H$ definiert jedes $\varphi \in H$ durch

$$v \to (\varphi, v)_H$$

ein stetiges lineares Funktional auf X. Wegen Satz 2.1 besitzt somit (3) eine eindeutige Lösung $u_\varphi \in X$. Aus (3) und (5) folgt für beliebiges $\varphi \in H$ und beliebiges $v_h \in X_h$

$$\begin{aligned}
(u - u_h, \varphi)_H =& a(u - u_h, u_\varphi) \\
=& a(u - u_h, u_\varphi - v_h) \\
\leq& A\|u - u_h\|_X \|u_\varphi - v_h\|_X.
\end{aligned}$$

Da

$$\|u - u_h\|_H = \sup_{\substack{\varphi \in H \\ \|\varphi\|_H = 1}} (u - u_h, \varphi)_H$$

ist, folgt hieraus die zweite Fehlerabschätzung. □

2.3 Bemerkung: Der erste Teil von Satz 2.2 ist bekannt unter dem Namen *"Cea - Lemma"*; der zweite firmiert unter dem Namen *"Dualitätsargument von Aubin - Nitsche"*. □

Die Voraussetzungen der Sätze 2.1 und 2.2 sind insbesondere für Konvektions-Diffusions Gleichungen, die wir in den nächsten Paragraphen betrachten werden, zu restriktiv. Daher schwächen wir sie in den folgenden beiden Sätzen entsprechend ab.

2.4 Satz: *Seien* $(X, \|.\|_X)$ *und* $(Y, \|.\|_Y)$ *Banach Räume mit* $X \overset{c}{\hookrightarrow} Y$ *und* $a_0, a_1 \in \mathcal{L}^2(X, \mathbb{R})$ *zwei stetige Bilinearformen. Die Bilinearform* a_0 *sei symmetrisch und koerziv. Für die Bilinearform* a_1 *gebe es eine Konstante* $\overline{A} \in \mathbb{R}_+^*$ *mit*

$$a_1(u, v) \leq \overline{A}\|u\|_X\|v\|_Y \quad \forall u, v \in X. \tag{6}$$

Sei $a := a_0 + a_1 \in \mathcal{L}^2(X, \mathbb{R})$*, d.h.*

$$a(u, v) := a_0(u, v) + a_1(u, v) \qquad \forall u, v \in X.$$

Für alle $u \in X \backslash \{0\}$ *gelte schließlich*

$$a(u, u) > 0. \tag{7}$$

Dann besitzen die Probleme

$$a(u, v) = l(v) \quad \forall v \in X \tag{8}$$

und

$$a(v, u) = l(v) \quad \forall v \in X \tag{9}$$

für jedes stetige lineare Funktional $l \in \mathcal{L}(X, \mathbb{R})$ *jeweils eine eindeutige Lösung.*

Beweis: Wir beweisen die Behauptung nur für Problem (8). Der Beweis für das andere Problem ist völlig analog. Sei $l \in \mathcal{L}(X, \mathbb{R})$ beliebig. Wegen Satz 2.1 gibt es genau ein $u_l \in X$ mit

$$a_0(u_l, v) = l(v) \quad \forall v \in X.$$

Dann ist (8) äquivalent zu

$$a_0(u, v) + a_1(u, v) = a_0(u_l, v) \quad \forall v \in X.$$

Wiederum wegen Satz 2.1 gibt es zu jedem $w \in X$ ein eindeutiges $u_w \in X$ mit

$$a_0(u_w, v) = a_1(w, v) \quad \forall v \in X.$$

Die Zuordnung $w \to u_w$ definiert eine lineare Abbildung $K \in \mathcal{L}(X, X)$, und (8) ist damit äquivalent zu

$$(Id + K)u = u_l. \tag{10}$$

Wegen $X \overset{c}{\hookrightarrow} Y$ und (6) ist K kompakt. Daher erfüllt $Id+K$ die Fredholm Alternative: Entweder besitzt (10) für jede rechte Seite eine eindeutige Lösung oder das zugehörige homogene Problem besitzt eine nichttriviale Lösung $u \neq 0$. Wegen (7) besitzt (8) mit $l = 0$ und damit (10) mit $u_l = 0$ aber nur die triviale Lösung. □

2.5 Satz: *Die Bezeichnungen und Voraussetzungen seien wie in Satz 2.4. Zusätzlich sei $X_h \subset X$ ein endlich dimensionaler Unterraum. Dann besitzt das Problem*

$$a(u_h, v_h) = l(v_h) \quad \forall v_h \in X_h \tag{11}$$

für jedes $l \in \mathcal{L}(X, \mathbb{R})$ eine eindeutige Lösung $u_h \in X_h$. Die Bilinearform a sei zusätzlich koerziv, d.h. es gibt ein $\beta > 0$ mit $a(u,u) \geq \beta \|u\|_X^2$ $\forall u \in X$. Dann gilt für die eindeutigen Lösungen u und u_h der Probleme (8) und (11) die Fehlerabschätzung

$$\|u - u_h\|_X \leq \frac{A}{\beta} \inf_{v_h \in X_h} \|u - u_h\|_X. \tag{12}$$

Dabei ist $A := \|a\|_{\mathcal{L}^2(X,\mathbb{R})}$. Seien schließlich H, φ und u_φ wie in Satz 2.2. Dann gilt die Fehlerabschätzung

$$\|u - u_h\|_H \leq A \|u - u_h\|_X \sup_{\substack{\varphi \in H \\ \|\varphi\|_H = 1}} \inf_{u_h \in X_h} \|u_\varphi - u_h\|_X. \tag{13}$$

Beweis: Die eindeutige Lösbarkeit von (11) folgt aus Satz 2.4. Die Fehlerabschätzung (12) folgt wie im Beweis von Satz 2.2. Man beachte, daß wir dort nur die zu (8) und (11) analogen Eigenschaften (2) und (4) ausgenutzt haben. Wegen $X \hookrightarrow H$ definiert jedes $\varphi \in H$ durch $v \to (\varphi, v)_H$ ein stetiges lineares Funktional auf X. Daher folgt die Existenz und Eindeutigkeit von u_φ aus Satz 2.4. Die Fehlerabschätzung (13) folgt dann wie im Beweis von Satz 2.2. □

3. Schwache Lösungen

Im folgenden ist $\Omega \subset \mathbb{R}^d$, $d \geq 2$ eine offene beschränkte Menge mit Lipschitz-Rand $\Gamma := \partial\Omega$ und äußerem Einheitsnormalenfeld n. Wir betrachten skalare, lineare, elliptische Differentialgleichungen 2. Ordnung. Ihre allgemeine Form lautet (vgl. § III.1 der Vorlesung "Numerische Behandlung von Differentialgleichungen I")

$$-\sum_{1 \leq i,j \leq d} \frac{\partial}{\partial x_i}\Big(A_{ij} \frac{\partial u}{\partial x_j}\Big) + \sum_{i=1}^{d} a_i \frac{\partial u}{\partial x_i} + \alpha u = f \text{ in } \Omega. \tag{1}$$

Dabei ist $f \in L^2(\Omega)$, $\alpha \in C(\Omega, \mathbb{R}_+)$, $a = (a_1, ..., a_d) \in C^1(\Omega, \mathbb{R}^d)$ und $A = (A_{ij})_{1 \leq i,j \leq d} \in C^1(\Omega, \mathbb{R}^{d \times d})$ mit $A_{ij}(x) = A_{ji}(x)$ für alle $x \in \Omega$, $1 \leq i, j \leq d$ und

$$\lambda_0 := \inf_{x \in \Omega} \inf_{z \in \mathbb{R}^d} \frac{z^T A(x) z}{z^T z} > 0. \tag{2}$$

Später werden wir die Glattheitsbedingungen an die Koeffizienten α, a und A abschwächen. Zur Vereinfachung der Notation sprechen wir im folgenden von

- einer **Konvektions-Diffusionsgleichung**, wenn α, a und A beliebig sind,
- einer **Reaktions-Diffusionsgleichung**, wenn $a = 0$ ist,
- einer **Membrangleichung**, wenn $\alpha = 0$ und $a = 0$ ist,
- einer **Poisson Gleichung**, wenn $\alpha = 0$, $a = 0$ und $A = I$ ist.

Die partielle Differentialgleichung (1) muß mit Randbedingungen versehen werden. Wir betrachten drei Typen von Randbedingungen

- (homogene) **Dirichlet Randbedingungen**: $u = 0$ auf Γ,
- (inhomogene) **Neumann Randbedingungen**: $n^T A\nabla u = g$ auf Γ,
- **gemischte Dirichlet-Neumann Randbedingungen**: $u = 0$ auf Γ_D und $n^T A\nabla u = g$ auf Γ_N.

Dabei ist $g \in L^2(\Gamma)$, $\Gamma_D \cap \Gamma_N = \emptyset$ und $\Gamma = \Gamma_D \cup \Gamma_N$. Wir werden bei gemischten Randbedingungen stets fordern, daß Γ_D ein positives $(d-1)$-dimensionales Maß hat. Die Beschränkung auf homogene Dirichlet Randdaten ist nicht wesentlich, vereinfacht aber die Darstellung.

Sei nun $u \in C^2(\Omega)$ eine Lösung von (1) mit homogenen Dirichlet Randbedingungen und $v \in C_0^\infty(\Omega)$. Multiplikation von (1) mit v , Integration über Ω und Anwenden des Gauß'schen Integralsatzes liefert

$$\begin{aligned}\int_\Omega fv &= -\sum_{1\le i,j\le d}\int_\Omega \frac{\partial}{\partial x_i}\Big(A_{ij}\frac{\partial u}{\partial x_j}\Big)v + \sum_{i=1}^d \int_\Omega \alpha_i \frac{\partial u}{\partial x_i}v + \int_\Omega \alpha uv \\ &= \sum_{1\le i,j\le d}\int_\Omega A_{ij}\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i} + \sum_{i=1}^d \int_\Omega a_i \frac{\partial u}{\partial x_i}v + \int_\Omega \alpha uv \\ &= \int_\Omega \big\{\nabla u^T A\nabla v + a\cdot\nabla uv + \alpha uv\big\}.\end{aligned} \tag{3}$$

Da $C_0^\infty(\Omega)$ dicht ist in $H_0^1(\Omega)$ folgt, daß $u \in H_0^1(\Omega)$ die Gleichung

$$\int_\Omega \big\{\nabla u^T A\nabla v + a\cdot\nabla uv\big\} = \int_\Omega fv \tag{4}$$

für alle $v \in H_0^1(\Omega)$ erfüllt. Umgekehrt folgt aus (3), daß eine Lösung von (4) die Differentialgleichung (1) erfüllt, sofern sie hinreichend glatt, d.h. in $C^2(\Omega)$ ist. In diesem Sinne ist (4) zur Konvektions-Diffusionsgleichung (1) mit homogenen Dirichlet Randbedingungen äquivalent.

Betrachten wir in obigem Argument Funktionen $v \in C^\infty(\overline{\Omega})$, so treten in (3) zusätzlich Randterme $-\int_\Gamma n^T A\nabla uv$ auf. Erfüllt u Neumann Randbedingungen, so gilt für diesen Randterm

$$-\int_\Gamma n^T A\nabla uv = -\int_\Gamma gv.$$

Wir werden daher in diesem Fall (4) durch den zusätzlichen Term $\int_\Gamma gv$ auf der rechten Seite modifizieren. Diese Überlegungen führen auf folgende Definition.

3.1 Definition: **(1)** $u \in H_0^1(\Omega)$ heißt **schwache Lösung** der Konvektions-Diffusions Gleichung mit homogenen Dirichlet Randbedingungen, wenn für alle $v \in H_0^1(\Omega)$ gilt

$$\int_\Omega \{\nabla u^T A \nabla v + a \cdot \nabla u v + \alpha u v\} = \int_\Omega f v.$$

(2) $u \in H_D^1(\Omega) := \{\varphi \in H^1(\Omega) : \varphi = 0 \text{ auf } \Gamma_D\}$ heißt **schwache Lösung** der Konvektions-Diffusions Gleichung mit gemischten Randbedingungen, wenn für alle $v \in H_D^1(\Omega)$ gilt

$$\int_\Omega \{\nabla u^T A \nabla v + a \cdot \nabla u v + \alpha u v\} = \int_\Omega f v + \int_{\Gamma_N} g v.$$

(3) $u \in H^1(\Omega)$ heißt **schwache Lösung** der Konvektions-Diffusions Gleichung mit Neumann Randbedingungen, wenn für alle $v \in H^1(\Omega)$ gilt

$$\int_\Omega \{\nabla u^T A \nabla v + a \cdot \nabla u v + \alpha u v\} = \int_\Omega f v + \int_\Gamma g v.$$

□

3.2 Bemerkung: **(1)** Jede klassische Lösung von (1) ist auch eine schwache Lösung. Jede schwache Lösung, die zweimal stetig differenzierbar ist, ist eine klassische Lösung von (1).
(2) Für schwache Lösungen benötigen wir für die Koeffizienten nur die Regularitätsvoraussetzungen $\alpha \in L^\infty(\Omega), \alpha \geq 0, a \in L^\infty(\Omega, \mathbb{R}^d)$, $A \in L^\infty(\Omega, \mathbb{R}^{d\times d})$.
(3) Bei inhomogenen Dirichlet Randbedingungen $u = u_D$ auf Γ bzw. Γ_D muß in Definition 3.1 die Bedingung $u \in H_0^1(\Omega)$ bzw. $u \in H_D^1(\Omega)$ durch $u \in u_D + H_0^1(\Omega)$ bzw. $u \in u_D + H_D^1(\Omega)$ ersetzt werden. □

3.3 Satz: (Existenz- und Eindeutigkeitssatz für schwache Lösungen)
(1) *Ist* $-\frac{1}{2}\mathrm{div}a+\alpha \geq 0$*, so besitzt die Konvektions-Diffusionsgleichung mit homogenen Dirichlet-Randbedingungen eine eindeutige schwache Lösung.*
(2) *Ist* $-\frac{1}{2}\mathrm{div}a + \alpha \geq 0$ *und* $a \cdot n \geq 0$ *auf* Γ_N*, so besitzt die Konvektions-Diffusions Gleichung mit gemischten Randbedingungen eine eindeutige schwache Lösung.*
(3) *Ist* $\alpha \geq a_0 > 0, -\frac{1}{2}\mathrm{div}a + \alpha \geq 0$ *und* $a \cdot n \geq 0$ *auf* Γ*, so besitzt die Konvektions-Diffusions Gleichung mit Neumann Randbedingungen eine eindeutig schwache Lösung.*
(4) *Ist* $\alpha = 0$*,* $-\frac{1}{2}\mathrm{div}a \geq 0$ *und* $a \cdot n \geq 0$ *auf* Γ *sowie* $\int_\Omega f + \int_\Gamma g = 0$*, so besitzt die Konvektions-Diffusions Gleichung mit Neumann Randbedingungen eine eindeutige schwache Lösung* u *mit* $\int_\Omega u = 0$*.*

Beweis: Wir wenden jeweils Satz 2.4 an.
ad (1): Setze

$$\begin{aligned} X :=& H_0^1(\Omega), \\ Y :=& L^2(\Omega), \\ l(v) :=& \int_\Omega fv, \\ a_0(u,v) :=& \int_\Omega \big\{\nabla u^T A \nabla v + \alpha uv\big\}, \\ a_1(u,v) :=& \int_\Omega a \cdot \nabla uv. \end{aligned}$$

Aus der Cauchy-Schwarz'schen Ungleichung folgt

$$\begin{aligned} |l(v)| \leq& \|f\|_0 \|v\|_0 \leq \|f\|_0 \|v\|_1, \\ |a_0(u,v)| \leq& \|A\|_{L^\infty} |u|_1 |v|_1 + \|\alpha\|_{L^\infty} \|u\|_0 \|v\|_0 \\ \leq& \max\big\{\|A\|_{L^\infty}, \|\alpha\|_{L^\infty}\big\} \|u\|_1 \|v\|_1, \\ |a_1(u,v)| \leq& \|a\|_{L^\infty} |u|_1 \|v\|_0 \leq \|a\|_{L^\infty} \|u\|_1 \|v\|_0. \end{aligned}$$

Wegen $\alpha \geq 0$ und (2) ist

$$\begin{aligned} a_0(u,u) =& \int_\Omega \big\{\nabla u^T A \nabla u + \alpha u^2\big\} \\ \geq& \lambda_0 \int_\Omega \nabla u^T \nabla u \\ =& \lambda_0 |u|_1^2. \end{aligned}$$

Wegen der Friedrich'schen Ungleichung, Satz 1.15, ist also a_0 koerziv. Aus dem Gauß'schen Integralsatz und (2) folgt schließlich

$$\begin{aligned} a(u,u) =& \int_\Omega \big\{\nabla u^T A \nabla u + a \cdot \nabla uu + \alpha u^2\big\} \\ \geq& \lambda_0 |u|_1^2 + \int_\Omega \big\{a \cdot \nabla uu + \alpha u^2\big\} \\ =& \lambda_0 |u|_1^2 + \int_\Omega \big\{\frac{1}{2} a \cdot \nabla(u^2) + \alpha u^2\big\} \\ =& \lambda_0 |u|_1^2 + \int_\Omega \big\{-\frac{1}{2} \mathrm{div} a + \alpha\big\} u^2. \end{aligned}$$

Also ist wegen $-\frac{1}{2}\mathrm{div}a + \alpha \geq 0$ auch a koerziv.
ad (2): In diesem Fall ist

$$\begin{aligned} X :=& H_D^1(\Omega), \\ l(v) :=& \int_\Omega fv + \int_{\Gamma_N} gv. \end{aligned}$$

Die anderen Größen ändern sich nicht. Aus der Cauchy-Schwarzschen Ungleichung und dem Spursatz, Satz 1.12, folgt

$$\begin{aligned} |l(v)| \leq & \|f\|_0\|v\|_0 + \|g\|_{L^2(\Gamma_N)}\|v\|_{L^2(\Gamma_N)} \\ \leq & \big\{\|f\|_0 + c\|g\|_{L^2(\Gamma_N)}\big\}\|v\|_1, \end{aligned}$$

so daß l stetig ist. Die Koerzivität von a_0 bleibt wegen Bem. 1.16 erhalten. Bei der Anwendung des Gauß'schen Integralsatzes in der Abschätzung von $a(u,u)$ tritt der zusätzliche Randterm $\int_{\Gamma_N} n \cdot au^2$ auf. Wegen $n \cdot a \geq 0$ auf Γ_N ist er nicht negativ, und die Koerzivität von a bleibt erhalten.
ad (3): Nun ist $X = H^1(\Omega)$. Die anderen Größen sind wie in (2) mit Γ an Stelle von Γ_N. Wegen $\alpha \geq \alpha_0 > 0$ erhalten wir

$$\begin{aligned} a_0(u,u) \geq & \lambda_0|u|_1^2 + \alpha_0\|u\|_0^2 \\ \geq & \min\{\lambda_0, \alpha_0\}\|u\|_1^2 \end{aligned}$$

und somit die Koerzivität von a_0. Die anderen Abschätzungen ändern sich nicht, insbesondere gilt $a(u,u) \geq 0$ für alle $u \in X$.
ad (4): Alle Größen sind wie in (3). Wegen $\alpha = 0$ erhalten wir

$$a_0(u,u) \geq \lambda_0|u|_1^2.$$

Hieraus und aus der Poincaré'schen Ungleichung, Satz 1.21, folgt die Koerzivität von a_0 auf $V := \{\varphi \in H^1(\Omega) : \int_\Omega \varphi = 0\}$. Die Abschätzung $a(u,u) \geq 0$ für alle $u \in X$ bleibt gültig, ebenso die Stetigkeit von a_0 und a_1. Lediglich bei der Stetigkeit von l ist Sorgfalt geboten. Da $V \cong H^1(\Omega)/\mathbb{R}$ ist, muß für die Stetigkeit von l die Inklusion $\mathbb{R} \subset \ker l$ gelten. Dies ist aber wegen $\int_\Omega f + \int_\Gamma g = 0$ der Fall. □

3.4 Bemerkung: **(1)** Im Fall der Reaktions-Diffusions Gleichung, d.h. $a = 0$, reduzieren sich die Voraussetzung von Satz 3.3 auf $\alpha \geq \alpha_0 > 0$ bei Teil (3) und auf $\alpha = 0$, $\int_\Omega f + \int_\Gamma g = 0$ bei Teil (4).
(2) Gelegentlich treten auch sog. Robin Randbedingungen der Form $\beta u + n^T A\nabla u = g_R$ auf $\Gamma_R \subset \Gamma$ auf. In diesem Fall muß l duch $\int_{\Gamma_R} g_R v$ und a_0 durch $\int_{\Gamma_R} \beta uv$ ergänzt werden. Die Koerzivität von a_0 bleibt erhalten, wenn entweder $\beta \geq 0$ und $\Gamma_D \neq \emptyset$ oder $\beta \geq \beta_0 > 0$ ist. □

Das folgende Beispiel, das wir schon in § III.1 der Vorlesung "Numerische Behandlung von Differentialgleichungen I" kennengelernt haben, zeigt, daß wir eine Regularitätsaussage der Form $u \in H^2$ für schwache Lösungen nur unter zusätzlichen Annahmen an den Rand Γ erwarten können.

3.5 Beispiel: Sei $0 < \alpha < 2\pi$ und Ω_α das Kreissegment

$$\Omega_\alpha := \{x \in \mathbb{R}^2 : x = (r\cos\varphi, r\sin\varphi), 0 < r < 1, 0 < \varphi < \alpha\}.$$

Definiere die Funktion $v : \Omega_\alpha \to \mathbb{R}$ durch

$$v(x) := r^{\pi/\alpha}\sin(\frac{\pi}{\alpha}\varphi) \quad x = (r\cos\varphi, r\sin\varphi).$$

Dann gilt für jedes $x \in \Omega_\alpha$

$$\Delta v(x) = \frac{1}{r}\frac{\partial}{\partial r}(r\frac{\partial v}{\partial r}) + \frac{1}{r^2}\frac{\partial^2 v}{\partial \varphi^2} = 0.$$

Sei $w \in C_0^\infty(\mathbb{R}^2, \mathbb{R})$ mit

$$\text{supp } w \subset B(0, \frac{2}{3}) \quad \text{und} \quad w = 1 \text{ auf } \overline{B(0, \frac{1}{3})}.$$

Definiere

$$u := wv \quad , \quad f := \Delta[(1-w)v].$$

Dann gilt

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega_\alpha \\ u &= 0 \quad \text{auf } \partial\Omega_\alpha. \end{aligned}$$

Offensichtlich ist $(1-w)v \in C^\infty(\mathbb{R}^2, \mathbb{R})$ und somit $f \in C^\infty(\overline{\Omega}_\alpha)$. Ebenso ist $u \in C^\infty(\Omega_\alpha)$. Wegen $u = v$ in $B(0, \frac{1}{3})$ gilt aber

$$u \notin C^\infty(\overline{\Omega}_\alpha).$$

Wie man leicht nachrechnet gilt

$$u \in C^k(\overline{\Omega}_\alpha) \Longleftrightarrow 0 < \alpha \leq \frac{\pi}{k} \quad , k \geq 1$$

und

$$D^k u \in L^2(\Omega_\alpha) \Longleftrightarrow 0 < \alpha < \frac{\pi}{k-1} \quad , k \geq 2.$$

Wir können also bei gegebenem α i.a. keine Abschätzung der Form

$$\|u\|_{C^{k+2}(\overline{\Omega}_\alpha)} \leq c_k \|f\|_{C^k(\overline{\Omega}_\alpha)}$$

oder

$$\|u\_H^{k+2}(\Omega_\alpha) \leq c_k' \|f\|_{H^k(\Omega_\alpha)}$$

erwarten, wie sie für gewöhnliche Differentialgleichungen gelten würde. □

3.6 Satz: (Regularitätssatz) *Sei Γ eine C^1-Mannigfaltigkeit oder Ω konvex und $f \in L^2(\Omega)$. Bei gemischten oder Neumann Randbedingungen gebe es eine Funktion u_g in $H^2(\Omega)$ mit $g = \gamma_0(u_g) := u_g|_{\Gamma_N}$. Dann gilt für die schwache Lösung u der Konvektions-Diffusions Gleichung mit homogenen Dirichlet oder gemischten oder Neumann Randbedingungen die Regularitätsaussage $u \in H^2(\Omega)$ und die a priori Abschätzung*

$$\|u\|_2 \leq c\big\{\|f\|_0 + \|u_g\|_2\big\}.$$

Die Konstante c hängt nur von Ω und den Koeffizienten α, a und A ab.

Beweis: J. Nečas: Les Méthodes directes en théorie des Equations Elliptiques. Masson, 1967 oder D. Gilbarg, N.S. Trudinger: Elliptic Partial Differential Equations of Second Order. Springer, 1983. □

4. Eindimensionale lineare Elemente

Zur Motivation erinnern wir in diesem Paragraphen kurz an die Ergebnisse von § II.5 der Vorlesung "Numerische Behandlung von Differentialgleichungen I". Dort haben wir Finite Element Methoden für die eindimensionale Konvektions-Reaktions Gleichung mit homogenen Dirichlet Randbedingungen auf $\Omega := (0,1)$ kennengelernt. Im Rahmen von § 2 ist

$$\begin{aligned} X =& H_0^1((0,1)), \\ l(v) =& \int_0^1 fv, \\ a(u,v) =& \int_0^1 \big\{Au'v' + \alpha uv\big\}. \end{aligned}$$

Sei $\mathcal{T}_h := \{I_j : 0 \leq j \leq n\}$ mit $I_j := [x_j, x_{j+1}]$ und $0 = x_0 < x_1 < ... < x_{n+1} = 1$ eine Unterteilung von $[0,1]$ in $n+1$ Teilintervalle. Setze

$$h_j := x_{j+1} - x_j \qquad , 0 \leq j \leq n$$

und

$$h := \max_{0 \leq j \leq n} h_j.$$

Im Rahmen von Satz 2.1 setzen wir

$$X_h := S_{h,0}^{1,0} := \{\varphi \in C([0,1]) : \varphi|_{I_j} \in \mathbb{P}_1 \quad \forall 0 \leq j \leq n, \varphi(0) = \varphi(1) = 0\}.$$

Eine Basis von X_h ist gegeben durch die Funktionen $v_1, ..., v_n$ mit

$$v_i(x) = \begin{cases} 0 & \text{für } x \leq x_{i-1} \text{ oder } x \geq x_{i+1}, \\ \frac{x - x_{i-1}}{h_{i-1}} & \text{für } x_{i-1} \leq x \leq x_i, \\ \frac{x_{i+1} - x}{h_i} & \text{für } x_i \leq x \leq x_{i+1}. \end{cases}$$

Das diskrete Problem

$$a(u_h, v_h) = l(v_h) \qquad \forall u_h \in X_h$$

ist dann äquivalent zu einem linearen Gleichungssystem im $\mathbb{R}^n$ mit einer symmetrischen, positiv definiten Tridiagonalmatrix.

Aus Satz 2.1 folgt, daß der Fehler $u - u_h$ zwischen der schwachen Lösung u und der Finite Element Lösung u_h durch den Approximationsfehler

$$\inf_{v_h \in X_h} \|u - v_h\|_1$$

kontrolliert wird. In Satz II.5.14 der Vorlesung "Numerische Behandlung von Differentialgleichungen I" haben wir gezeigt, daß

$$\inf_{v_h \in X_h} \|u - v_h\|_1 \leq \sqrt{2} \inf_{v_h \in X_h} |u - v_h|_1 \leq \sqrt{2} h |u|_2 \tag{1}$$

ist. Mit Satz 2.1 folgt hieraus sofort die Fehlerabschätzung

$$\|u - u_h\|_0 + h|u - u_h|_1 \leq ch^2 |u|_2$$

für die schwache Lösung u und ihre Finite Element Approximation u_h. Der Beweis von (1) beruhte auf der Friedrich'schen Ungleichung angewandt auf den Interpolationsfehler $u - I_h u$. Dabei ist $I_h u \in X_h$ bestimmt durch die Interpolationsbedingungen

$$I_h u(x_i) = u(x_i) \qquad \forall 0 \leq i \leq n+1.$$

Fassen wir noch einmal die wesentlichen Schritte zusammen:

(1) Konstruktion einer Unterteilung $\mathcal{T}_h$ von Ω in einfache Teilgebiete.
(2) Konstruktion eines Finite Elemente Raumes X_h, der aus "einfachen" Funktionen auf den Teilgebieten in $\mathcal{T}_h$ besteht.
(3) Konstruktion einer Basis von X_h, die auf Funktionen mit möglichst kleinem Träger besteht.
(4) Abschätzung des Interpolationsfehlers.

5. Bilineare Rechteckselemente

Im folgenden bezeichnet $\Omega \subset \mathbb{R}^2$ ein zusammenhängendes Polygon mit achsenparallelen Kanten. Wir betrachten zunächst die Reaktion-Diffusions Gleichung mit homogenen Dirichlet Randbedingungen. Im Rahmen von § 2 ist gemäß § 3

$$\begin{aligned} X &= H_0^1(\Omega) \\ l(v) &= \int_\Omega fv \\ a(u,v) &= \int_\Omega \{\nabla u^T A \nabla v + \alpha uv\}. \end{aligned}$$

Sei $\mathcal{T}_h = \{K_i : 1 \le i \le m_h\}$ eine Zerlegung von Ω in achsenparallele Rechtecke, so daß je zwei Rechtecke entweder disjunkt sind oder einen Eckpunkt oder eine Kante gemeinsam haben (**Zulässigkeit**). Für $K \in \mathcal{T}_h$ bezeichnet h_K die längste Kante. Setze

$$h := \max_{K \in \mathcal{T}_h} h_K.$$

Die Eckpunkte der Rechtecke bezeichnen wir mit $\mathcal{N}_h$; $\mathcal{N}_{h,\Omega}$ sind die Eckpunkte im Innern von Ω. Wir numerieren zunächst die Punkte in $\mathcal{N}_{h,\Omega}$ von 1 bis n_h und anschließend die Punkte in $\mathcal{N}_h \backslash \mathcal{N}_{h,\Omega}$, die alle auf dem Rand Γ liegen, von $n_h + 1$ bis $\overline{n}_h$.

Sei $Q_1 := \operatorname{span}\{1, x_1, x_2, x_1 x_2\}$ der Raum aller Polynome in zwei Variablen, die in jeder Variablen höchstens den Grad 1 haben. Im Rahmen von Satz 2.1 setzen wir

$$X_h := \{\varphi \in C(\overline{\Omega}) : \varphi|_K \in Q_1 \quad \forall K \in \mathcal{T}_h, \varphi = 0 \text{ auf } \Gamma\}.$$

Man überlegt sich leicht, daß für jedes achsenparallele Rechteck K und jedes Polynom $p \in Q_1$ das Polynom eindeutig bestimmt wird durch seine Werte in den vier Eckpunkten von K. Da die Einschränkung von p auf eine beliebige Kante von K ein lineares Polynom einer Veränderlichen ist, folgt außerdem, daß für je zwei achsenparallele Rechtecke K_1, K_2, die eine Kante E gemeinsam haben, und je zwei Polynome $p_1, p_2 \in Q_1$ die Funktion

$$\varphi := \begin{cases} p_1 & \text{auf } K_1 \\ p_2 & \text{auf } K_2 \end{cases}$$

genau dann stetig ist, wenn p_1 und p_2 in den Eckpunkten von E übereinstimmen. Daher sind die Funktionen in X_h eindeutig bestimmt durch ihre Werte in den Punkten von $\mathcal{N}_{h,\Omega}$. Insbesondere ist $\dim X_h = n_h$.

Wir wollen nun eine Basis für X_h konstruieren. Sei dazu $z_0 \in \mathcal{N}_{h,\Omega}$ beliebig. Definiere zwei Funktionen einer Veränderlichen wie in § 4 durch

$$\varphi_{z_0}(x_1) := \begin{cases} 0 & \text{falls } x_1 \le z_{0,1} - h_W \text{ oder } x_1 \ge z_{0,1} + h_E, \\ \frac{x_1 - z_{0,1} + h_W}{h_W} & \text{falls } z_{0,1} - h_W \le x_1 \le z_{0,1}, \\ \frac{z_{0,1} + h_E - x_1}{h_E} & \text{falls } z_{0,1} \le x_1 \le z_{0,1} + h_E \end{cases}$$

und

$$\psi_{z_0}(x_2) := \begin{cases} 0 & \text{falls } x_2 \le z_{0,2} - h_S \text{ oder } x_2 \ge z_{0,2} + h_N, \\ \frac{x_2 - z_{0,2} + h_S}{h_S} & \text{falls } z_{0,2} - h_S \le x_2 \le z_{0,2}, \\ \frac{z_{0,2} + h_N - x_2}{h_N} & \text{falls } z_{0,2} \le x_2 \le z_{0,2} + h_N. \end{cases}$$

Dabei bezeichnen h_E, h_W, h_N, h_S die Längen der Kanten, die von z_0 in die vier Himmelsrichtungen ausgehen. Offensichtlich sind die Funktionen φ_{z_0} und ψ_{z_0} stetig und stückweise linear. Daher ist

$$v_{z_0}(x_1, x_2) := \varphi_{z_0}(x_1)\psi_{z_0}(x_2) \in X_h$$

und

$$X_h = \text{span}\{v_z : z \in \mathcal{N}_{h,\Omega}\}.$$

Man beachte, daß der Träger von v_z genau aus den vier Rechtecken besteht, die den Knoten z gemeinsam haben.

Jedes $u_h \in X_h$ läßt sich eindeutig darstellen als

$$u_h = \sum_{z \in \mathcal{N}_{h,\Omega}} \mu_z v_z,$$

und es ist

$$\mu_z = u_h(z) \qquad \forall z \in \mathcal{N}_{h,\Omega}.$$

Daher ist das diskrete Problem

$$a(u_h, v_h) = l(v_h) \qquad \forall v_h \in X_h \tag{1}$$

äquivalent zu einem linearen Gleichungssystem mit n_h Gleichungen und Unbekannten. Die Unbekannten sind genau die Werte von u_h in den Gitterpunkten. Die Matrix des LGS heißt **(System-) Steifigkeitsmatrix**. Wegen der Symmetrie von a ist die Steifigkeitsmatrix symmetrisch. Wegen der Koerzivität von a ist die Steifigkeitsmatrix positiv definit. Da die Träger der Basisfunktionen v_z aus jeweils vier Rechtecken bestehen, hat die Steifigkeitsmatrix höchstens 9 von Null verschiedene Einträge pro Zeile, d.h., sie ist dünn besetzt.

Da die Unbekannten des zu (1) äquivalenten LGS die Werte von u_h in den Gitterpunkten sind, kann man (1) auch als ein Differenzenverfahren interpretieren. Dies gibt dann auch einen Einblick in die Struktur der Steifigkeitsmatrix. Wir betrachten hierzu den einfachsten Spezialfall der Poisson Gleichung auf einem äquidistanten Gitter, d.h., $\alpha = 0$, $A = I$ und alle Kanten haben die Länge h. Für jedes $p \in Q_1$ sind die partiellen Ableitungen $\frac{\partial p}{\partial x_1}$ und $\frac{\partial p}{\partial x_2}$ offensichtlich lineare Funktionen allein der Variablen x_2 bzw. x_1. Daher ist für alle $u_h, v_h \in X_h$ der Ausdruck $\nabla u_h^T \nabla v_h$ stückweise von der Form $\alpha\varphi(x_1) + \beta\psi(x_2)$ mit $\alpha, \beta \in \mathbb{R}$ und $\varphi, \psi \in \mathbb{P}_2$. Da die eindimensionale Simpson Regel die Ordnung 3 hat, folgt aus dem Satz von Fubini, daß ein Ausdruck der Form $\int_K \nabla u_h^T \nabla v_h$ durch die zweidimensionale Simpsonregel

$$\int_K \varphi \sim \frac{h_1 h_2}{36} \sum_{i=0}^{2} \sum_{j=0}^{2} \alpha_{i,j} \varphi(z_1 + \frac{i}{2} h_1, z_2 + \frac{j}{2} h_2)$$

mit

$$\begin{aligned} &\alpha_{0,0} = \alpha_{0,2} = \alpha_{2,0} = \alpha_{2,2} = 1, \\ &\alpha_{0,1} = \alpha_{1,0} = \alpha_{2,1} = \alpha_{1,2} = 4, \\ &\alpha_{1,1} = 16 \end{aligned}$$

exakt integriert wird (vgl. Bsp. II.1.3 (3) der Vorlesung "Einführung in die Numerik"). Hierbei ist K ein achsenparalleles Rechteck mit Kantenlängen h_1 und h_2 und unterer linker Ecke z. Seien $p \in Q_1$ und $p_{\mu,\nu} := p(z_1 + \mu h_1, z_2 + \nu h_2), \mu, \nu \in \{0,1\}$, die Werte von p in den Eckpunkten von K. Dann sind die Werte von ∇p in den Eckpunkten von K gegeben durch

$$\begin{pmatrix} \frac{1}{h_1}[p_{1,1} - p_{0,1}] \\ \frac{1}{h_2}[p_{0,1} - p_{0,0}] \end{pmatrix} \qquad \begin{pmatrix} \frac{1}{h_1}[p_{1,1} - p_{0,1}] \\ \frac{1}{h_2}[p_{1,1} - p_{1,0}] \end{pmatrix}$$

$$\begin{pmatrix} \frac{1}{h_1}[p_{1,0} - p_{0,0}] \\ \frac{1}{h_2}[p_{0,1} - p_{0,0}] \end{pmatrix} \qquad \begin{pmatrix} \frac{1}{h_1}[p_{1,0} - p_{0,0}] \\ \frac{1}{h_2}[p_{1,1} - p_{1,0}] \end{pmatrix}.$$

Die Werte in den restlichen Knoten obiger Quadraturformel ergeben sich durch lineare Interpolation. Wenden wir diese Überlegungen auf $\int_\Omega \nabla u_h^T \nabla v_z$ für alle $z \in \mathcal{N}_{h,\Omega}$ an, so erhalten wir für eine Unterteilung in Quadrate, d.h. $h_1 = h_2 = h$,

$$\int_\Omega \nabla u_h^T \nabla v_z = \sum_{i=-1}^{1} \sum_{j=-1}^{1} \beta_{i,j} u_h(z_1 + ih, z_2 + jh)$$

mit

$$\begin{aligned} \beta_{0,0} &= \frac{8}{3} \\ \beta_{\mu,\nu} &= -\frac{1}{3} \qquad \forall (\mu,\nu) \neq (0,0). \end{aligned}$$

Dies liefert die gewünschte Darstellung der linken Seite von (1) als Differenzenverfahren.

Sei nun $u \in H_0^1(\Omega)$ die eindeutige schwache Lösung der Reaktions-Diffusions Gleichung und $u_h \in X_h$ die eindeutige Lösung von (1). Gemäß Satz 2.1 kann der Fehler $|u - u_h|_1$ durch den Approximationsfehler $\inf\limits_{v_h \in X_h} |u - v_h|_1$ abgeschätzt werden. Um diesen Approximationsfehler zu kontrollieren, betrachten wir wie in § 4 einen geeigneten Interpolationsoperator $I_h : C(\overline{\Omega}) \to X_h$. Dieser ist definiert durch

$$I_h u \in X_h \text{ und } (I_h u)(z) = u(z) \quad \forall z \in \mathcal{N}_h$$

oder äquivalent

$$I_h u = \sum_{z \in \mathcal{N}_{h,\Omega}} u(z) v_z.$$

5.1 Satz: *Für alle $v \in H_0^1(\Omega) \cap H^2(\Omega)$ gilt die Interpolationsfehlerabschätzung*

$$|v - I_h v|_1 \leq \frac{h}{\sqrt{3}} |v|_2.$$

Beweis: Sei zunächst $h > 0$ und $i_h : C([0,h],\mathbb{R}) \to \mathbb{P}_1$ der lineare Interpolationsoperator in den Punkten 0 und h. Mit dem Hauptsatz der Differential- und Integralrechnung und partieller Integration erhalten wir für alle $\varphi \in C^2([0,h],\mathbb{R})$ und alle $t \in \mathbb{R}$

$$\begin{aligned}\varphi(t) =& \varphi(0) + \int_0^t \varphi'(s)ds \\ =& \varphi(0) + t\varphi'(t) - \int_0^t s\varphi''(s)ds\end{aligned}$$

und

$$\begin{aligned}\varphi(t) =& \varphi(h) - \int_t^h \varphi'(s)ds \\ =& \varphi(h) - (h-t)\varphi'(t) - \int_h^h (h-s)\varphi''(s)ds.\end{aligned}$$

Multiplikation der ersten Gleichung mit $\frac{h-t}{h}$ und der zweiten Gleichung mit $\frac{t}{h}$ und anschließende Addition der resultierenden Gleichungen liefert

$$\varphi(t) = i_h\varphi(t) + \frac{h-t}{h}\int_0^t \varphi'(s)ds - \frac{t}{h}\int_t^h \varphi'(s)ds \tag{2}$$

$$= i_h\varphi(t) - \frac{h-t}{h}\int_0^t s\varphi''(s)ds - \frac{t}{h}\int_t^h (h-s)\varphi''(s)ds \tag{3}$$

Sei nun $v \in C_0^\infty(\Omega)$ und K ein achsenparalleles Rechteck mit Kantenlängen h_1 und h_2. Durch Translation des Koordinatensystems können wir erreichen, daß o.E. $K = [0,h_1] \times [0,h_2]$ ist. Anwenden der Gleichung (3) auf die Variable x liefert für alle $(x,y) \in K$

$$v(x,y) = (i_h v(.,y))(x) - \frac{h_1 - x}{h_1}\int_0^x s\frac{\partial^2}{\partial x^2}v(s,y)ds - \frac{x}{h_1}\int_x^{h_1}(h_1-s)\frac{\partial^2}{\partial x^2}v(s,y)ds.$$

Wenden wir Gleichung (2) für festes x auf die Variable y und $\varphi(y) := (i_h v(.,y))(x)$ an, erhalten wir weiter

$$\begin{aligned}(i_h v(.,y))(x) =& \frac{h_1-x}{h_1}\Big\{\frac{h_2-y}{h_2}v(0,0) + \frac{y}{h_2}v(0,h_2) \\ &+ \frac{h_2-y}{h_2}\int_0^y \frac{\partial}{\partial y}v(0,t)dt - \frac{y}{h_2}\int_y^{h_2}\frac{\partial}{\partial y}v(0,t)dt\Big\} \\ &+ \frac{x}{h_1}\Big\{\frac{h_2-y}{h_2}v(h_1,0) + \frac{y}{h_2}v(h_1,h_2) \\ &+ \frac{h_2-y}{h_2}\int_0^y \frac{\partial}{\partial y}v(h_1,t)dt - \frac{y}{h_2}\int_y^{h_2}\frac{\partial}{\partial y}v(h_1,t)dt\Big\}\end{aligned}$$

$$
\begin{aligned}
=&I_h v(x,y) + \frac{h_2-y}{h_2}\int_0^y \frac{\partial}{\partial y}v(0,t)dt - \frac{y}{h_2}\int_y^{h_2}\frac{\partial}{\partial y}v(0,t)dt \\
&+ \frac{x}{h_1}\frac{h_2-y}{h_2}\int_0^y \Big\{\frac{\partial}{\partial y}v(h_1,t) - \frac{\partial}{\partial y}v(0,t)\Big\}dt \\
&- \frac{x}{h_1}\frac{y}{h_2}\int_y^{h_2}\Big\{\frac{\partial}{\partial y}v(h_1,t) - \frac{\partial}{\partial y}v(0,t)\Big\}dt.
\end{aligned}
$$

Da für alle $t \in [0,h_2]$

$$
\frac{\partial}{\partial y}v(h_1,t) - \frac{\partial}{\partial y}v(0,t) = \int_0^{h_1}\frac{\partial^2}{\partial x \partial y}v(s,t)ds
$$

ist, erhalten wir insgesamt die Darstellung

$$
\begin{aligned}
v(x,y) = I_h(x,y) &+ \frac{h_2-y}{h_2}\int_0^y \frac{\partial}{\partial y}v(0,t)dt - \frac{y}{h_2}\int_y^{h_2}\frac{\partial}{\partial y}v(0,t)dt \\
&+ \frac{x}{h_1}\frac{h_2-y}{h_2}\int_0^y\int_0^{h_1}\frac{\partial^2}{\partial x\partial y}v(s,t)dsdt \\
&- \frac{x}{h_1}\frac{y}{h_2}\int_y^{h_2}\int_0^{h_1}\frac{\partial^2}{\partial x\partial y}v(s,t)dsdt \\
&- \frac{h_1-x}{h_1}\int_0^x s\frac{\partial^2}{\partial x^2}v(s,y)ds \\
&- \frac{x}{h_1}\int_x^{h_1}(h_1-s)\frac{\partial^2}{\partial x^2}v(s,y)ds.
\end{aligned}
$$

Differentiation bzgl. x liefert

$$
\begin{aligned}
\frac{\partial}{\partial x}(v - I_h v)(x,y) =&\frac{1}{h_1}\int_0^{h_1}\int_0^y \frac{h_2-y}{h_2}\frac{\partial^2}{\partial x\partial y}v(s,t)dsdt \\
&- \frac{1}{h_1}\int_0^{h_1}\int_y^{h_2}\frac{y}{h_2}\frac{\partial^2}{\partial x\partial y}v(s,t)dsdt \\
&+ \frac{1}{h_1}\int_0^x s\frac{\partial^2}{\partial x^2}v(s,y)ds \\
&- \frac{1}{h_1}\int_x^{h_1}(h_1-s)\frac{\partial^2}{\partial x^2}v(s,y)ds \\
=&\frac{1}{h_1}\int_0^{h_1}\int_0^{h_2}K_1(t,y)\frac{\partial^2 v}{\partial x\partial y}(s,t)dsdt \\
&+ \frac{1}{h_1}\int_0^{h_1}K_2(s,x)\frac{\partial^2 v}{\partial x^2}(s,y)ds
\end{aligned}
$$

mit

$$
K_1(t,y) := \begin{cases} \frac{h_2-y}{h_2} & \text{für } 0 \le t < y \\ -\frac{y}{h_2} & \text{für } y < t \le h_2 \end{cases}
$$

$$
K_2(s,x) := \begin{cases} s & \text{für } 0 \le s < x \\ -(h_1-s) & \text{für } x < s \le h_1 \end{cases}.
$$

Quadrieren dieser Identität und Anwenden der Cauchy-Schwarz'schen Ungleichung ergibt

$$\begin{aligned}
&|\frac{\partial}{\partial x}(v - I_h v)(x,y)|^2 \\
\leq &\frac{2}{h_1^2}\int_0^{h_1}\int_0^{h_2}|K_1(t,y)|^2 dsdt \int_0^{h_1}\int_0^{h_2}|\frac{\partial^2 v}{\partial x \partial y}(s,t)|^2 dsdt \\
&+ \frac{2}{h_1^2}\int_0^{h_1}|K_2(s,x)|^2 ds \int_0^{h_1}|\frac{\partial^2 v}{\partial x^2}(s,y)|^2 ds.
\end{aligned}$$

Integration über K liefert

$$\begin{aligned}
&\int_K |\frac{\partial}{\partial x}(v - I_h v)|^2 \\
\leq &\frac{2}{h_1^2}\int_0^{h_1}\int_0^{h_2}\int_0^{h_1}\int_0^{h_2}|K_1(t,y)|^2 dsdtdxdy \|\frac{\partial^2 v}{\partial x \partial y}\|^2_{L^2(K)} \\
&+ \frac{2}{h_1^2}\int_0^{h_1}\int_0^{h_1}|K_2(s,x)|^2 dsdx \|\frac{\partial^2 v}{\partial x^2}\|^2_{L^2(K)}.
\end{aligned}$$

Offensichtlich ist

$$\begin{aligned}
&\frac{2}{h_1^2}\int_0^{h_1}\int_0^{h_2}\int_0^{h_1}\int_0^{h_2}|K_1(t,y)|^2 dsdtdxdy \\
=&2\int_0^{h_2}\int_0^{h_2}|K_1(t,y)|^2 dtdy \\
=&2\int_0^{h_2}\frac{1}{h_2^2}y(h_2-y)^2 + \frac{1}{h_2^2}(h_2-y)y^2 dy \\
=&\frac{2}{h_2}\int_0^{h_2} y(h_2-y)dy \\
=&\frac{1}{3}h_2^2
\end{aligned}$$

und

$$\begin{aligned}
&\frac{2}{h_1^2}\int_0^{h_1}\int_0^{h_1}|K_2(s,x)|^2 dsdx \\
=&\frac{2}{h_1^2}\int_0^{h_1}\frac{1}{3}x^3 + \frac{1}{3}(h_1-x)^3 dx \\
=&\frac{2}{3h_1^2}\Big\{\frac{1}{4}h_1^4 + \frac{1}{4}h_1^4\Big\} \\
=&\frac{1}{3}h_1^2.
\end{aligned}$$

Insgesamt haben wir also die Abschätzung

$$\begin{aligned}
\|\frac{\partial}{\partial x}(v - I_h v)\|^2_{L^2(K)} \leq &\frac{h_2^2}{3}\|\frac{\partial^2 v}{\partial x \partial y}\|^2_{L^2(K)} + \frac{h_1^2}{3}\|\frac{\partial^2 v}{\partial x^2}\|^2_{L^2(K)} \\
\leq &\frac{h^2}{3}\Big\{\|\frac{\partial^2 v}{\partial x \partial y}\|^2_{L^2(K)} + \|\frac{\partial^2 v}{\partial x^2}\|^2_{L^2(K)}\Big\}.
\end{aligned}$$

Vertauschen wir die Rollen von x und y, so erhalten wir mit der gleichen Rechnung

$$\|\frac{\partial}{\partial y}(v - I_h v)\|_{L^2(K)} \leq \frac{h^2}{3}\Big\{\|\frac{\partial^2 v}{\partial x \partial y}\|^2_{L^2(K)} + \|\frac{\partial^2 v}{\partial y^2}\|^2_{L^2(K)}\Big\}$$

Addition dieser beiden Abschätzungen und Summation über alle Rechtecke beweist wegen

$$|v|_2^2 = \|\frac{\partial^2 v}{\partial x^2}\|_0^2 + 2\|\frac{\partial^2 v}{\partial x \partial y}\|_0^2 + \|\frac{\partial^2 v}{\partial y^2}\|_0^2$$

und Definition 1.9 die Behauptung. □

Aus Satz 2.2, Satz 3.6 und Satz 5.1 folgt unmittelbar die folgende Fehlerabschätzung.

5.2 Satz: *Seien $u \in H_0^1(\Omega)$ die schwache Lösung der Reaktions-Diffusions Gleichung mit homogenen Dirichlet Randbedingungen und $u_h \in X_h$ die Lösung des diskreten Problems (1). Es sei $u \in H^2(\Omega)$. Dann gilt*

$$|u - u_h|_1 \leq c_1 h |u|_2.$$

Ist Ω zusätzlich konvex, so ist

$$\|u - u_h\|_0 \leq c_2 h^2 |u|_2.$$

Die Konstanten c_1 und c_2 hängen nur von Ω und den Koeffizienten α und A ab.

5.3 Bemerkung: **(1)** Bei inhomogenen Dirichlet Randbedingungen $u = u_D$ auf Γ sucht man die Lösung u_h von (1) statt in X_h in $I_h u_D + X_h$, d.h. man setzt u_h in der Form

$$u_h = \sum_{z \in \mathcal{N}_{h,\Omega}} \mu_z v_z + \sum_{z \in \mathcal{N}_h \setminus \mathcal{N}_{h,\Omega}} u_D(z) v_z$$

mit unbekannten Koeffizienten μ_z an. Satz 5.2 bleibt gültig.
(2) Bei gemischten oder Neumann Randbedingungen muß man analog zu Definition 3.1 die rechte Seite von (1) durch $\int_{\Gamma_N} gv$ ergänzen. Satz 5.2 bleibt gültig.
(3) Unter den Voraussetzungen von Satz 3.3 gilt Satz 5.2 auch für Konvektions-Diffusionsgleichungen (die Bilinearform a muß natürlich wie in § 3 beschrieben angepaßt werden). Die Konstanten c_1 und c_2 verhalten sich dann im wesentlichen wie $\lambda_0^{-1} \max\{\|A\|_{L^\infty}, \|a\|_{L^\infty}, \|\alpha\|_{L^\infty}\}$, wobei λ_0 wie in Gleichung (2) von § 3 ist. Für Diffusions-dominante Probleme, d.h. $\|a\|_{L^\infty} \sim \|A\|_{L^\infty} \sim \lambda_0$, ist diese Abschätzung gut. Für Konvektions-dominante Probleme, d.h. $\|A\|_{L^\infty} \sim \lambda_0 << \|a\|_{L^\infty}$, ist sie dagegen unbrauchbar. Außerdem entspricht die Diskretisierung in (1) einer zentralen Differenzendiskretisierung des Konvektionstermes $a \cdot \nabla u$. Daher treten bei der numerischen Lösung u_h unphysikalsiche Oszillationen auf, wenn die **Péclet-Zahl**

$\lambda_0^{-1}\|a\|_{L^\infty}h$ groß ist. Im nächsten Paragraphen beschreiben wir für diesen Fall eine Modifikation von (1), die die genannten Schwierigkeiten vermeidet und die auch auf Rechteckselemente angewendet werden kann (vgl. (5) in § 6 und Satz 6.5 ff.). □

6. Lineare Dreieckselemente

Im folgenden bezeichnet $|.|$ die euklidische Norm auf dem $\mathbb{R}^d$ und $|||.|||$ die zugehörige Matrixnorm. $\Omega \subset \mathbb{R}^2$ ist ein offenes, beschränktes, zusammenhängendes Gebiet mit polygonalem Rand Γ. Wie in § 5 betrachten wir zunächst die Reaktions-Diffusions Gleichung mit homogenen Dirichlet Randbedingungen und setzen im Rahmen von §2

$$\begin{aligned} X :=& H_0^1(\Omega), \\ l(v) :=& \int_\Omega fv, \\ a(u,v) :=& \int_\Omega \{\nabla u^T A \nabla v + \alpha uv\}. \end{aligned}$$

Sei $\mathcal{T}_h = \{K_i : 1 \le i \le m_h\}$ eine Unterteilung von Ω in Dreiecke, derart daß je zwei Dreiecke entweder disjunkt sind oder einen Eckpunkt oder eine Kante gemeinsam haben **(Zulässigkeit)**. Für jedes $K \in \mathcal{T}_h$ bezeichnen wir mit h_K den Durchmesser von K und mit ρ_K den Durchmesser des größten in K eingeschriebenen Kreises. Die Dreiecke sollen vergleichbare Größe haben, d.h. h_K/ρ_K soll unabhängig von K und h durch eine Konstante $c_{\mathcal{T}}$ nach oben beschränkt sein **(Regularität)**. Dies ist äquivalent zu der Bedingung, daß der kleinste Winkel aller Dreiecke gleichmäßig von Null weg beschränkt sein soll **(Winkelbedingung)**. Setze

$$h := \max_{K \in \mathcal{T}_h} h_K.$$

Für $K \in \mathcal{T}_h$ bezeichnen wir die Eckpunkte mit $z_{K,1}, ..., z_{K,3}$. Dabei soll aus praktischen Gründen die Numerierung so sein, daß der Rand von K im mathematisch positiven Sinn durchlaufen wird, wenn die $z_{K,i}$ in aufsteigender Reihenfolge durchlaufen werden. Die Menge aller Eckpunkte der Dreiecke bezeichnen wir wieder mit $\mathcal{N}_h$; $\mathcal{N}_{h,\Omega}$ sind wieder die Eckpunkte im Innern von Ω. Wir numerieren die Punkte in $\mathcal{N}_{h,\Omega}$ zunächst von 1 bis n_h und anschließend die Punkte in $\mathcal{N}_h \backslash \mathcal{N}_{h,\Omega}$, die auf dem Rand Γ liegen, von $n_h + 1$ bis $\overline{n}_h$.

Im Rahmen von Satz 2.1 setzen wir

$$X_h := \{\varphi \in C(\overline{\Omega}) : \varphi|_K \in \mathbb{P}_1 \quad \forall K \in \mathcal{T}_h, \varphi = 0 \text{ auf } \Gamma\}.$$

Da eine Ebene im $\mathbb{R}^3$ durch drei nicht kolineare Punkte eindeutig bestimmt ist, ist für jedes $K \in \mathcal{T}_h$ und jedes $p \in \mathbb{P}_1$ das Polynom p eindeutig bestimmt durch seine

Werte in den drei Eckpunkten von K. Da die Einschränkung von p auf eine beliebige Kante eine lineare Funktion einer Veränderlichen ist, folgt außerdem, daß für je zwei Dreiecke K_1, K_2, die eine Kante E gemeinsam haben, und für je zwei Polynome $p_1, p_2 \in \mathbb{P}_1$ die Funktion

$$\varphi := \begin{cases} p_1 & \text{auf } K_1 \\ p_2 & \text{auf } K_2 \end{cases}$$

genau dann stetig ist, wenn p_1 und p_2 in den Eckpunkten von E übereinstimmen. Daher sind die Funktionen in X_h eindeutig bestimmt durch ihre Werte in den Punkten von $\mathcal{N}_{h,\Omega}$. Insbesondere ist

$$\dim X_h = n_h.$$

Wir wollen nun eine Basis für X_h konstruieren. Sei dazu $K \in \mathcal{T}_h$ beliebig. Für $i \in \{1,2,3\}$ definieren wir die Funktion $\lambda_{K,i}$ durch

$$\lambda_{K,i}(x) := \frac{\det(z_{K,i+1} - x, z_{K,i+2} - z_{K,i+1})}{\det(z_{K,i+1} - z_{K,i}, z_{K,i+2} - z_{K,i})}. \tag{1}$$

Dabei sind die Indizes $i+1$ und $i+2$ modulo 3 zu interpretieren, d.h. $4 \leftrightarrow 1$, $5 \leftrightarrow 2$ usw.. Die Funktionen $\lambda_{K,1}, ..., \lambda_{K,3}$ heißen die **Schwerpunktskoordinaten** von K . Sie haben offensichtlich folgende Eigenschaften

$$\begin{aligned} &\lambda_{K,i} \in \mathbb{P}_1, \\ &\lambda_{K,i}(z_{K,j}) = \delta_{ij}, \\ &\lambda_{K,1} + \lambda_{K,2} + \lambda_{K,3} = 1. \end{aligned} \tag{2}$$

Man beachte, daß $\frac{1}{2}\det(z_{K,i+1} - z_{K,i}, z_{K,i+2} - z_{K,i})$ für alle $i \in \{1,2,3\}$ die Fläche von K angibt. Für beliebiges $z \in \mathcal{N}_h$ definieren wir

$$\omega_z := \bigcup_{\substack{K \in \mathcal{T}_h \\ z \in K}} K$$

und

$$v_z(x) := \begin{cases} 0 & \text{falls } x \notin \omega_z \\ \lambda_{K,i}(x) & \text{falls } x \in K \subset \omega_z \text{ und } z = z_{K,i}. \end{cases}$$

Offensichtlich sind die Funktionen v_z stetig, stückweise linear, haben den Wert 1 im Punkt z und verschwinden in allen anderen Punkten von $\mathcal{N}_h$. Daher ist

$$X_h = \operatorname{span}\{v_z : z \in \mathcal{N}_h\}.$$

Man beachte, daß ω_z genau aus den Dreiecken besteht, die den Punkt z als Eckpunkt haben. Die Funktionen v_z heißen **nodale Basis** von X_h.

Jedes $u_h \in X_h$ läßt sich eindeutig darstellen als

$$u_h = \sum_{z \in \mathcal{N}_h} \mu_z v_z$$

und es ist

$$\mu_z = u_h(z).$$

Daher ist das diskrete Problem

$$a(u_h, v_h) = l(u_h) \quad \forall v_h \in X_h \tag{3}$$

äquivalent zu einem linearen Gleichungssystem mit n_h Gleichungen und Unbekannten. Wie in § 5 ist die Steifigkeitsmatrix symmetrisch, positiv definit und dünn besetzt.

Da die Unbekannten des zu (3) äquivalenten LGS die Werte von u_h in den Gitterpunkten $\mathcal{N}_h$ sind, kann (3) wieder als ein Differenzenverfahren interpretiert werden. Um einen Eindruck über die Struktur dieses Differenzenverfahrens zu erhalten, betrachten wir den einfachsten Spezialfall der Poisson Gleichung auf einer sog. **Courant Triangulierung**. Diese besteht aus gleichschenklig rechtwinkligen Dreiecken mit Katheten der Länge h parallel zu den Koordinatenachsen und Hypotenusen parallel zu einer festen Richtung (dies ist entweder die Winkelhalbierende des 1. und 3. Quadranten oder diejenige des 2. und 4. Quadranten). Seien K ein Dreieck der Triangulierung und $u_h, v_h \in X_h$. Man überlegt sich leicht, daß der Ausdruck $\int_K \nabla u_h^T \nabla v_h$ translations- und rotationsinvariant ist. Daher ist K o.E. das Dreieck mit den Eckpunkten $z_1 := (0,0), z_2 := (h,0)$ und $z_3 := (0,h)$. Setze $u_i := u_h(z_i), v_i := v_h(z_i)$. Da u_h und v_h linear sind, sind ihre Gradienten konstant auf K und haben den Wert

$$\nabla u_h = \begin{pmatrix} \frac{1}{h}(u_2 - u_1) \\ \\ \frac{1}{h}(u_3 - u_1) \end{pmatrix}, \nabla v_h = \begin{pmatrix} \frac{1}{h}(v_2 - v_1) \\ \\ \frac{1}{h}(v_3 - v_1) \end{pmatrix}.$$

Daher ist

$$\int_K \nabla u_h^T \nabla v_h = \frac{1}{2}\big\{(u_2 - u_1)(v_2 - v_1) + (u_3 - u_1)(v_3 - v_1)\big\}.$$

Hieraus folgt mit leichter Rechnung, daß die linke Seite von (3) dem Differenzenschema

$$4u_h(z) - u_h(z + he_1) - u_h(z - he_1) - u_h(z + he_2) - u_h(z - he_2)$$

entspricht. Dabei ist $z \in \mathcal{N}_{h,\Omega}$ beliebig, und e_1, e_2 sind die kanonischen Einheitsvektoren des $\mathbb{R}^2$.

Zur Abschätzung des Fehlers der Finite Element Diskretisierung (3) definieren wir wie in § 5 einen Interpolationsoperator $I_h : C(\overline{\Omega}) \to X_h$ durch

$$I_h u \in X_h \text{ und } (I_h u)(z) = u(z) \quad \forall z \in \mathcal{N}_{h,\Omega}$$

oder äquivalent

$$I_h u = \sum_{z \in \mathcal{N}_{h,\Omega}} u(z) v_z.$$

Bei der Analyse des Interpolationsfehlers spielt das **Referenzelement** $\hat{K}$ mit den Eckpunkten $(0,0),(1,0),(0,1)$ eine ausgezeichnete Rolle. Jedes $K \in \mathcal{T}_h$ ist **affin äquivalent** zu $\hat{K}$. D.h., es gibt eine reguläre Matrix $B_K \in \mathbb{R}^{2\times 2}$ und einen Vektor $b_K \in \mathbb{R}^2$, so daß die affine Transformation

$$\begin{aligned} F_K : & I\!R^2 \to \mathbb{R}^2 \\ & \hat{x} \to B_K \hat{x} + b_k \end{aligned} \tag{4}$$

das Dreieck $\hat{K}$ bijektiv auf das Dreieck K abbildet. O.E. können wir dabei annehmen, daß F_K orientierungstreu ist.

Wir schätzen zunächst den Interpolationsfehler auf dem Referenzelement ab.

6.1 Satz: *Bezeichne mit $\hat{\pi} : C(\hat{K}) \to \mathbb{P}_1$ den linearen Interpolationsoperator zu den Eckpunkten von $\hat{K}$. Dann gilt für alle $v \in H^2(\hat{K})$ die Fehlerabschätzung*

$$|v - \hat{\pi} v|_{H^1(\hat{K})} \le 2.5 |v|_{H^2(\hat{K})}.$$

Beweis: Gemäß Bem. 1.20 (5) gilt $H^2(\hat{K}) \hookrightarrow C(\hat{K})$. Daher ist $\hat{\pi} v$ für $v \in H^2(\hat{K})$ wohldefiniert. Da gemäß Satz 1.6 (2) $C^\infty(\hat{K})$ dicht ist in $H^2(\hat{K})$, reicht es, ein beliebiges $v \in C^\infty(\hat{K})$ zu betrachten. Da $\hat{\pi} v$ linear ist, ist $\nabla(\hat{\pi} v)$ konstant und

$$\nabla(\hat{\pi} v) = \begin{pmatrix} v(1,0) - v(0,0) \\ v(0,1) - v(0,0) \end{pmatrix}.$$

Für beliebiges $(x,y) \in \hat{K}$ gilt daher

$$\begin{aligned} & \frac{\partial v}{\partial x}(x,y) - \frac{\partial}{\partial x}(\hat{\pi} v)(x,y) \\ =& \frac{\partial v}{\partial x}(x,y) - \{v(1,0) - v(0,0)\} \\ =& \frac{\partial v}{\partial x}(x,y) - \int_0^1 \frac{\partial v}{\partial x}(s,0) ds \\ =& \int_0^1 \Big\{ \frac{\partial v}{\partial x}(x,y) - \frac{\partial v}{\partial x}(s,0) \Big\} ds \end{aligned}$$

$$
\begin{aligned}
&= \int_0^x \Big\{ \frac{\partial v}{\partial x}(x,y) - \frac{\partial v}{\partial x}(s,y) \Big\} ds + \int_0^x \Big\{ \frac{\partial v}{\partial x}(s,y) - \frac{\partial v}{\partial x}(s,0) \Big\} ds \\
&\quad + \int_x^1 \Big\{ \frac{\partial v}{\partial x}(x,y) - \frac{\partial v}{\partial x}(x,1-s) \Big\} ds + \int_x^1 \Big\{ \frac{\partial v}{\partial x}(x,1-s) - \frac{\partial v}{\partial x}(s,1-s) \Big\} ds \\
&\quad + \int_x^1 \Big\{ \frac{\partial v}{\partial x}(s,1-s) - \frac{\partial v}{\partial x}(s,0) \Big\} ds \\
&= \int_0^x \int_s^x \frac{\partial^2 v}{\partial x^2}(\sigma,y) d\sigma ds + \int_0^x \int_0^y \frac{\partial^2 v}{\partial x \partial y}(s,t) dt ds \\
&\quad + \int_x^1 \int_{1-s}^y \frac{\partial^2 v}{\partial x \partial y}(x,t) dt ds - \int_x^1 \int_x^s \frac{\partial^2 v}{\partial x^2}(\sigma,1-s) d\sigma ds \\
&\quad + \int_x^1 \int_0^{1-s} \frac{\partial^2 v}{\partial x \partial y}(s,t) dt ds \\
&=: \sum_{i=1}^5 I_i(x,y).
\end{aligned}
$$

Quadrieren und integrieren dieser Gleichung liefert wegen der Cauchy-Schwarz'schen Ungleichung für endliche Summen die Abschätzung

$$
\Big\| \frac{\partial}{\partial x}(v - \hat{\pi} v) \Big\|^2_{L^2(\hat{K})} = \int_{\hat{K}} \Big| \frac{\partial v}{\partial x} - \frac{\partial}{\partial x}(\hat{\pi} v) \Big|^2 \leq 5 \sum_{i=1}^5 \int_{\hat{K}} |I_i|^2.
$$

Mit Hilfe der Cauchy-Schwarz'schen Ungleichung für Integrale können die Ausdrücke $\int_{\hat{K}} |I_i|^2$ wie folgt abgeschätzt werden:

$$
\begin{aligned}
\int_{\hat{K}} |I_1|^2 &\leq \int_{\hat{K}} \Big\{ \int_0^x \int_s^x d\sigma ds \Big\} \Big\{ \int_0^x \int_s^x \Big| \frac{\partial^2 v}{\partial x^2}(\sigma,y) \Big|^2 d\sigma ds \Big\} \\
&\leq \int_{\hat{K}} \Big\{ \frac{1}{2} x^2 \cdot x \int_0^x \Big| \frac{\partial^2 v}{\partial x^2}(\sigma,y) \Big|^2 d\sigma \Big\} \\
&\leq \frac{1}{2} \int_0^1 \int_0^{1-y} \int_0^x \Big| \frac{\partial^2 v}{\partial x^2}(\sigma,y) \Big|^2 d\sigma dx dy \\
&\leq \frac{1}{2} \int_0^1 \int_0^{1-y} \int_0^{1-y} \Big| \frac{\partial^2 v}{\partial x^2}(\sigma,y) \Big|^2 d\sigma dx dy \\
&\leq \frac{1}{2} \Big\| \frac{\partial^2 v}{\partial x^2} \Big\|^2_{L^2(\hat{K})} \\
\int_{\hat{K}} |I_2|^2 &\leq \int_{\hat{K}} \Big\{ \int_0^x \int_0^y dt ds \Big\} \Big\{ \int_0^x \int_0^y \Big| \frac{\partial^2 v}{\partial x \partial y}(s,t) \Big|^2 dt ds \Big\} \\
&= \int_0^1 \int_0^{1-x} \Big\{ xy \int_0^x \int_0^y \Big| \frac{\partial^2 v}{\partial x \partial y}(s,t) \Big|^2 dt ds \Big\} dy dx \\
&\leq \int_0^1 \int_0^{1-x} \int_0^1 \int_0^{1-x} \Big| \frac{\partial^2 v}{\partial x \partial y}(s,t) \Big|^2 dt ds dy dx \\
&\leq \frac{1}{2} \Big\| \frac{\partial^2 v}{\partial x \partial y} \Big\|^2_{L^2(\hat{K})}
\end{aligned}
$$

$$\begin{aligned}
\int_{\hat K}|I_3|^2 \le& \int_{\hat K}\Big\{\int_x^1\Big|\int_{1-s}^y dt\Big|ds\Big\}\Big\{\int_x^1\Big|\int_{1-s}^y\Big|\frac{\partial^2 v}{\partial x\partial y}(x,t)\Big|^2 dt\Big|ds\Big\}\\
\le& \int_{\hat K}\Big\{\int_x^1\int_0^{1-x}\Big|\frac{\partial^2 v}{\partial x\partial y}(x,t)\Big|^2 dtds\Big\}\\
=& \int_0^1\int_0^{1-x}\int_0^1\int_0^{1-x}\Big|\frac{\partial^2 v}{\partial x\partial y}(x,t)\Big|^2 dtdsdydx\\
\le& \frac{1}{2}\Big\|\frac{\partial^2 v}{\partial x\partial y}\Big\|^2_{L^2(\hat K)}
\end{aligned}$$

$$\begin{aligned}
\int_{\hat K}|I_4|^2 \le& \int_{\hat K}\Big\{\int_x^1\int_x^s d\sigma ds\Big\}\Big\{\int_x^1\int_x^s\Big|\frac{\partial^2 v}{\partial x^2}(\sigma,1-s)\Big|^2 d\sigma ds\Big\}\\
=& \int_{\hat K}\Big\{\frac{1}{2}(1-x)^2\int_0^{1-x}\int_x^{1-t}\Big|\frac{\partial^2 v}{\partial x^2}(\sigma,t)\Big|^2 d\sigma dt\Big\}\\
\le& \frac{1}{2}\int_{\hat K}\int_0^1\int_0^{1-t}\Big|\frac{\partial^2 v}{\partial x^2}(\sigma,t)\Big|^2 d\sigma dt\\
\le& \frac{1}{4}\Big\|\frac{\partial^2 v}{\partial x^2}\Big\|^2_{L^2(\hat K)}
\end{aligned}$$

$$\begin{aligned}
\int_{\hat K}|I_5|^2 \le& \int_{\hat K}\Big\{\int_x^1\int_0^{1-s} dtds\Big\}\Big\{\int_x^1\int_0^{1-s}\Big|\frac{\partial^2 v}{\partial x\partial y}(s,t)\Big|^2 dtds\Big\}\\
=& \int_{\hat K}\Big\{\frac{1}{2}(1-x)^2\int_x^1\int_0^{1-s}\Big|\frac{\partial^2 v}{\partial x\partial y}(s,t)\Big|^2 dtds\\
\le& \frac{1}{2}\int_{\hat K}\int_0^1\int_0^{1-s}\Big|\frac{\partial^2 v}{\partial x\partial y}(s,t)\Big|^2 dtds\\
=& \frac{1}{4}\Big\|\frac{\partial^2 v}{\partial x\partial y}\Big\|^2_{L^2(\hat K)}.
\end{aligned}$$

Insgesamt ergibt sich

$$\begin{aligned}
\Big\|\frac{\partial}{\partial x}(v-\hat\pi v)\Big\|^2_{L^2(\hat K)} \le& 5\Big\{\frac{3}{4}\Big\|\frac{\partial^2 v}{\partial x^2}\Big\|^2_{L^2(\hat K)}+\frac{5}{4}\Big\|\frac{\partial^2 v}{\partial x\partial y}\Big\|^2_{L^2(\hat K)}\Big\}\\
\le& \frac{25}{4}\Big\{\Big\|\frac{\partial^2 v}{\partial x^2}\Big\|^2_{L^2(\hat K)}+\Big\|\frac{\partial^2 v}{\partial x\partial y}\Big\|^2_{L^2(\hat K)}\Big\}.
\end{aligned}$$

Vertauschen wir die Rollen von x und y erhalten wir mit den gleichen Argumenten die Abschätzung

$$\Big\|\frac{\partial}{\partial y}(v-\hat\pi v)\Big\|^2_{L^2(\hat K)} \le \frac{25}{4}\Big\{\Big\|\frac{\partial^2 v}{\partial y^2}\Big\|^2_{L^2(\hat K)}+\Big\|\frac{\partial^2 v}{\partial x\partial y}\Big\|^2_{L^2(\hat K)}\Big\}.$$

Hieraus folgt die Behauptung. □

Als nächstes schätzen wir die Spektralnormen $|||B_K|||$ und $|||B_K^{-1}|||$ der Transformationsmatrizen in (4) ab.

6.2 Satz: *Für jedes $K \in \mathcal{T}_h$ gilt für die Matrix B_K aus (4)*

$$|||B_K||| \leq \frac{2+\sqrt{2}}{2} h_K \quad , \quad |||B_K^{-1}||| \leq \frac{\sqrt{2}}{\rho_K}.$$

Beweis: Bezeichne mit $h_{\hat{K}}$ und $\rho_{\hat{K}}$ den Durchmesser von $\hat{K}$ und des größten, in $\hat{K}$ eingeschriebenen Kreises. Eine leichte Rechnung liefert

$$h_{\hat{K}} = \sqrt{2} \quad , \quad \rho_{\hat{K}} = 2 - \sqrt{2}.$$

Sei nun $\hat{z} \in \mathbb{R}^2$ mit $|\hat{z}| = \rho_{\hat{K}}$ beliebig. Dann gibt es Punkte $\hat{x}, \hat{y} \in \hat{K}$ mit $\hat{x} - \hat{y} = \hat{z}$. Da $F_K : \hat{K} \to K$ bijektiv ist, folgt

$$|B_K \hat{z}| = |F_K(\hat{x}) - F_K(\hat{y})| \leq h_K.$$

Also ist

$$|||B_K||| = \frac{1}{\rho_{\hat{K}}} \sup_{\substack{\hat{z} \in \mathbf{R}^2 \\ |\hat{z}| = \rho_{\hat{K}}}} |B_K \hat{z}| \leq \frac{h_K}{\rho_{\hat{K}}} = \frac{2+\sqrt{2}}{2} h_K.$$

Vertauschen der Rollen von K und $\hat{K}$ liefert

$$|||B_K^{-1}||| \leq \frac{h_{\hat{K}}}{\rho_K} = \frac{\sqrt{2}}{\rho_K}.$$

□

Durch Transformation auf das Referenzelement können wir jetzt den Interpolationsfehler auf einem beliebigen Dreieck $K \in \mathcal{T}_h$ und damit auf ganz Ω abschätzen.

6.3 Satz: **(1)** *Für jedes $K \in \mathcal{T}_h$ und jedes $v \in H^2(K)$ gilt*

$$|v - I_h v|_{H^1(K)} \leq 10.5 \frac{h_K^2}{\rho_K} |v|_{H^2(K)}.$$

(2) *Für jedes $v \in H_0^1(\Omega) \cap H^2(\Omega)$ gilt*

$$|v - I_h v|_1 \leq 10.5 c_{\mathcal{T}} h |v|_2.$$

Beweis: **ad (1)**: Seien $K \in \mathcal{T}_h$ und $v \in H^2(K)$ beliebig. Mit der affinen Transformation $F_K : \hat{K} \to K$ aus (4) definieren wir

$$\hat{v} := v \circ F_K.$$

Dann folgt

$$I_h v = (\hat{\pi} \hat{v}) \circ F_K^{-1}.$$

Der Transformationssatz liefert daher

$$\begin{aligned}
|v - I_h v|^2_{H^1(K)} &= \int_K \|D(v - I_h v)\|^2_{\mathcal{L}(\mathbb{R}^2,\mathbb{R}^2)} \\
&= \int_K \|D([\hat{v} - \hat{\pi}\hat{v}] \circ F_K^{-1})\|^2_{\mathcal{L}(\mathbb{R}^2,\mathbb{R}^2)} \\
&= \int_K \|B_K^{-1}[D(\hat{v} - \hat{\pi}\hat{v})] \circ F_K^{-1}\|^2_{\mathcal{L}(\mathbb{R}^2,\mathbb{R}^2)} \\
&\leq \int_K |||B_K^{-1}|||^2 \|[D(\hat{v} - \hat{\pi}\hat{v})] \circ F_K^{-1}\|^2_{\mathcal{L}(\mathbb{R}^2,\mathbb{R}^2)} \\
&= |\det B_K| |||B_K^{-1}|||^2 \int_{\hat{K}} \|D(\hat{v} - \hat{\pi}\hat{v})\|^2_{\mathcal{L}(\mathbb{R}^2,\mathbb{R}^2)} \\
&= |\det B_K| |||B_K^{-1}|||^2 |\hat{v} - \hat{\pi}\hat{v}|^2_{H^1(\hat{K})}.
\end{aligned}$$

Ebenso ergibt sich

$$\begin{aligned}
|\hat{v}|^2_{H^2(\hat{K})} &= \int_{\hat{K}} \|D^2\hat{v}\|^2_{\mathcal{L}^2(\mathbb{R}^2,\mathbb{R}^2)} \\
&= \int_{\hat{K}} \|D^2(v \circ F_K)\|^2_{\mathcal{L}^2(\mathbb{R}^2,\mathbb{R}^2)} \\
&= \int_{\hat{K}} \|[D^2 v] \circ F_K(B_K ., B_K .)\|^2_{\mathcal{L}^2(\mathbb{R}^2,\mathbb{R}^2)} \\
&\leq \int_{\hat{K}} |||B_K|||^4 \|[D^2 v] \circ F_K\|^2_{\mathcal{L}^2(\mathbb{R}^2,\mathbb{R}^2)} \\
&= |||B_K|||^4 |\det B_K|^{-1} \int_K \|D^2 v\|^2_{\mathcal{L}^2(\mathbb{R}^2,\mathbb{R}^2)} \\
&= |||B_K|||^4 |\det B_K|^{-1} |v|^2_{H^2(K)}.
\end{aligned}$$

Aus diesen Abschätzungen und den Sätzen 6.1 und 6.2 folgt

$$\begin{aligned}
|v - I_h v|_{H^1(K)} &\leq |\det B_K|^{1/2} |||B_K^{-1}||| |\hat{v} - \hat{\pi}\hat{v}|_{H^1(\hat{K})} \\
&\leq 2.5 |\det B_K|^{1/2} |||B_K^{-1}||| |\hat{v}|_{H^2(\hat{K})} \\
&\leq 2.5 |||B_K^{-1}||| |||B_K|||^2 |v|_{H^2(K)} \\
&\leq 2.5 \frac{(2+\sqrt{2})^2}{4} \sqrt{2} \frac{h_K^2}{\rho_K} |v|_{H^2(K)} \\
&\leq \frac{5}{4}(4 + 3\sqrt{2}) \frac{h_K^2}{\rho_K} |v|_{H^2(K)}.
\end{aligned}$$

Wegen $\frac{5}{4}(4 + 3\sqrt{2}) \leq 10.5$ folgt hieraus die Behauptung.
ad (2): Folgt wegen $h = \max_{K \in \mathcal{T}_h} h_K$ und $c_{\mathcal{T}} = \max_{K \in \mathcal{T}_h} h_K/\rho_K$ aus (1) durch Quadrieren und Aufsummieren. □

Aus Satz 2.2, Satz 3.6 und Satz 6.3 folgt unmittelbar die folgende Fehlerabschätzung.

6.4 Satz: *Seien $u \in H_0^1(\Omega)$ die schwache Lösung der Reaktions-Diffusions Gleichung mit homogenen Dirichlet Randbedingungen und $u_h \in X_h$ die Lösung von (4). Es sei $u \in H^2(\Omega)$. Dann gilt die Fehlerabschätzung*

$$|u - u_h|_1 \leq c_1 h |u|_2.$$

Ist zusätzlich Ω konvex, so ist

$$\|u - u_h\|_0 \leq c_2 h^2 |u|_2.$$

Die Konstanten c_1 und c_2 hängen nur von $c_{\mathcal{T}}$, Ω und den Koeffizienten α und A ab.

Bemerkung 5.3 zur Behandlung von inhomogenen Dirichlet Randbedingungen, von gemischten und Neumann Randbedingungen und von Konvektions-Diffusions Gleichungen überträgt sich direkt auf die vorliegende Diskretisierung. Insbesondere treten im Konvektions-dominanten Fall Schwierigkeiten durch die zentrale Differenzendiskretisierung des Konvektionstermes auf.

Im folgenden wollen wir ein Verfahren beschreiben, das diese Schwierigkeiten vermeidet und im Konvektions-dominanten wie im Diffusions-dominanten Fall gute Ergebnisse liefert. Dieses Verfahren firmiert in der Literatur unter den Bezeichnungen **Stromlinien-Diffusions Methode** (englisch **Streamline-diffusion finite element method**, kurz **SDFEM**) bzw. **streamline upwind Petrov-Galerkin** Verfahren, kurz **SUPG**. Für die Darstellung des zugrunde liegenden Prinzips beschränken wir uns auf den einfachsten Spezialfall konstanter Koeffizienten $A = \varepsilon I$, $a \in \mathbb{R}^2 \backslash \{0\}$, $\alpha = 0$ und homogener Dirichlet Randbedingungen. Dabei ist die Skalierung so gewählt, daß

$$|a| = 1 \quad \text{und} \quad 0 < \varepsilon << 1$$

ist. Andere Randbedingungen und variable Koeffizienten werden im Prinzip genauso behandelt, erfordern aber größeren technischen Aufwand.

Im Rahmen von Satz 2.1 ist jetzt

$$\begin{aligned} X :=& H_0^1(\Omega), \\ l(v) :=& \int_\Omega f v, \\ a(u,v) :=& \int_\Omega \{\varepsilon \nabla u^T \nabla v + a \cdot \nabla u v\}. \end{aligned}$$

Der diskrete Raum X_h ist wie in (3). Wir definieren auf X_h eine gitterabhängige

Bilinearform a_h, eine Linearform l_h und eine Norm $|.|_{1,h}$ durch

$$\begin{aligned} a_h(u_h, v_h) :=& a(u_h, v_h) + \sum_{K \in \mathcal{T}_h} \delta_K \int_K a \cdot \nabla u_h a \cdot \nabla v_h, \\ l_h(u_h) :=& l(v_h) + \sum_{K \in \mathcal{T}_h} \delta_K \int_K f a \cdot \nabla v_h, \\ |u_h|_{1,h} :=& \Big\{ \varepsilon |u_h|_1^2 + \sum_{K \in \mathcal{T}_h} \delta_K \|a \cdot \nabla u_h\|_{L^2(K)}^2 \Big\}^{1/2}. \end{aligned}$$

Das neue diskrete Problem lautet dann

$$a_h(u_h, v_h) = l_h(v_h) \quad \forall v_h \in X_h. \tag{5}$$

Dabei sind die δ_K nicht negative Parameter, die wir später bestimmen werden. Die δ_K-Terme werden eine zusätzliche Stabilisierung ergeben. Der zusätzliche Term in a_h entspricht einer Diskretisierung von $-\frac{\partial^2}{\partial a^2} u$. Daher wird eine zusätzliche Diffusion in Stromrichtung eingeführt. Insbesondere kann man erhoffen, daß senkrecht zu der Stromrichtung keine künstliche Diffusion, d.h. kein Verschmieren auftritt. Formal kann man sich vorstellen, daß die Konvektions-Diffusions Gleichung statt mit v_h mit $v_h + \sum_{K \in \mathcal{T}_h} \delta_K a \cdot \nabla v_h$ getestet wird. Daher sind im allgemeinen Fall noch zusätzliche Terme der Form

$$\sum_{K \in \mathcal{T}_h} \delta_K \int_K \{ -\operatorname{div}(A \nabla u_h) + \alpha u_h \} a \cdot \nabla v_h$$

in a_h aufzunehmen.

6.5 Satz: *Die Bilinearform a_h ist koerziv bzgl. $|.|_{1,h}$, d.h.*

$$a_h(u_h, u_h) \geq |u_h|_{1,h}^2 \quad \forall u_h \in X_h.$$

Beweis: Aus dem Beweis von Satz 3.6 (1) ergibt sich für den aktuellen Spezialfall die Abschätzung

$$a(v, v) \geq \varepsilon |v|_1^2 \quad \forall v \in H_0^1(\Omega).$$

Hieraus und aus der Definition von a_h und $|.|_{1,h}$ folgt sofort die Behauptung. □

6.6 Bemerkung: **(1)** Aus Satz 6.5 folgt, daß das diskrete Problem (5) eindeutig lösbar ist.
(2) Satz 6.5 benötigt außer der Annahme $\delta_K \geq 0$ keine weiteren Voraussetzungen an die δ_K.
(3) Im allgemeinen Fall bleibt Satz 6.5 gültig, wenn wie in Satz 3.6 $-\frac{1}{2} \operatorname{div} a + \alpha \geq 0$ und $h_K^{-1} \|A\|_{L^\infty(K)} \delta_K^{1/2} \leq 1$ ist. □

Für die Fehlerabschätzung der Diskretisierung (5) benötigen wir eine Abschätzung der L^2-Norm des Interpolationsfehlers.

6.7 Satz: **(1)** *Für alle $v \in H^2(\hat{K})$ gilt*

$$\|v - \hat{\pi}v\|_{L^2(\hat{K})} \leq \sqrt{10}|v|_{H^2(\hat{K})}.$$

(2) *Für alle $K \in \mathcal{T}_h$ und alle $v \in H^2(K)$ gilt*

$$\|v - I_h v\|_{L^2(K)} \leq 10h_K^2|v|_{H^2(K)}.$$

Beweis: **ad (1):** Wie im Beweis von Satz 6.1 genügt es, Funktionen $v \in C^\infty(\hat{K})$ zu betrachten. Sei $(x, y) \in \hat{K}$ beliebig. Wegen $v(0,0) = (\hat{\pi}v)(0,0)$ gilt

$$\begin{aligned}
(v - \hat{\pi}v)(x,y) =& (v - \hat{\pi}v)(x,y) - (v - \hat{\pi}v)(x,0) \\
&+ (v - \hat{\pi}v)(x,0) - (v - \hat{\pi}v)(0,0) \\
=& \int_0^y \frac{\partial}{\partial y}(v - \hat{\pi}v)(x,t)dt \\
&+ \int_0^x \frac{\partial}{\partial x}(v - \hat{\pi}v)(s,0)ds \\
=& \int_0^y \frac{\partial}{\partial y}(v - \hat{\pi}v)(x,t)dt \\
&+ \int_0^x \frac{\partial}{\partial x}(v - \hat{\pi}v)(s,y)ds \\
&- \int_0^x \int_0^y \frac{\partial^2}{\partial x \partial y}(v - \hat{\pi}v)(s,t)dtds \\
=:& \sum_{i=1}^3 I_i(x,y).
\end{aligned}$$

Quadrieren und Integrieren über $\hat{K}$ dieser Gleichung liefert mit der Cauchy-Schwarz'schen Ungleichung für endliche Summen

$$\|v - \hat{\pi}v\|^2_{L^2(\hat{K})} \leq 3\sum_{i=1}^3 \int_{\hat{K}} I_i^2.$$

Mit der Cauchy-Schwarz'schen Ungleichung für Integrale erhalten wir für die ersten beiden Summanden

$$\begin{aligned}
\int_{\hat{K}} I_1^2 \leq& \int_{\hat{K}} y \int_0^y \left|\frac{\partial}{\partial y}(v - \hat{\pi}v)(x,t)\right|^2 dt \\
\leq& \int_0^1 \int_0^{1-x} y \int_0^{1-x} \left|\frac{\partial}{\partial y}(v - \hat{\pi}v)(x,t)\right|^2 dtdydx \\
\leq& \frac{1}{2}\left\|\frac{\partial}{\partial y}(v - \hat{\pi}v)\right\|^2_{L^2(\hat{K})} \\
\int_{\hat{K}} I_2^2 \leq& \int_{\hat{K}} x \int_0^x \left|\frac{\partial}{\partial x}(v - \hat{\pi}v)(s,y)\right|^2 ds \\
\leq& \frac{1}{2}\left\|\frac{\partial}{\partial x}(v - \hat{\pi}v)\right\|^2_{L^2(\hat{K})}.
\end{aligned}$$

Bei der Abschätzung des dritten Summanden beachten wir, daß $\frac{\partial^2}{\partial x \partial y}(\hat{\pi} v) = 0$ ist, da $\hat{\pi} v \in \mathbb{P}_1$ ist. Das liefert

$$\begin{aligned}
\int_{\hat{K}} I_3^2 \leq & \int_{\hat{K}} xy \int_0^x \int_0^y \big| \frac{\partial^2 v}{\partial x \partial y}(s,t) \big|^2 dt ds \\
\leq & \int_{\hat{K}} xy \int_0^1 \int_0^{1-x} \big| \frac{\partial^2 v}{\partial x \partial y}(s,t) \big|^2 dt ds \\
= & \big\| \frac{\partial^2 v}{\partial x \partial y} \big\|^2_{L^2(\hat{K})} \int_0^1 \int_0^{1-x} xy dy dx \\
= & \big\| \frac{\partial^2 v}{\partial x \partial y} \big\|^2_{L^2(\hat{K})} \int_0^1 \frac{1}{2} x(1-x)^2 dx \\
= & \frac{1}{24} \big\| \frac{\partial^2 v}{\partial x \partial y} \big\|^2_{L^2(\hat{K})}.
\end{aligned}$$

Aus diesen Abschätzungen und dem Beweis von Satz 6.1 folgt insgesamt

$$\begin{aligned}
\big\| v - \hat{\pi} v \big\|^2_{L^2(\hat{K})} \leq & \frac{3}{2} \big\| \frac{\partial}{\partial x}(v - \hat{\pi} v) \big\|^2_{L^2(\hat{K})} + \frac{3}{2} \big\| \frac{\partial}{\partial y}(v - \hat{\pi} v) \big\|^2_{L^2(\hat{K})} + \frac{1}{8} \big\| \frac{\partial^2 v}{\partial x \partial y} \big\|^2_{L^2(\hat{K})} \\
\leq & \frac{45}{8} \big\| \frac{\partial^2 v}{\partial x^2} \big\|_{L^2(\hat{K})} + \frac{45}{8} \big\| \frac{\partial^2 v}{\partial y^2} \big\|_{L^2(\hat{K})} + \Big(\frac{75}{4} + \frac{1}{8} \Big) \big\| \frac{\partial^2 v}{\partial x \partial y} \big\|^2_{L^2(\hat{K})} \\
\leq & 10 |v|^2_{H^2(\hat{K})}.
\end{aligned}$$

ad (2): Wir gehen wir im Beweis von Satz 6.3 vor und erhalten mit Teil (1) und $\hat{v} := v \circ F_K$

$$\begin{aligned}
& \|v - I_h v\|_{L^2(K)} \\
= & |\det B_K|^{1/2} \|\hat{v} - \hat{\pi} \hat{v}\|_{L^2(\hat{K})} \\
\leq & |\det B_K|^{1/2} \sqrt{10} |\hat{v}|_{H^2(\hat{K})} \\
\leq & \sqrt{10} \|\|B_K\|\|^2 |v|_{H^2(K)} \\
\leq & \sqrt{10} \frac{(2+\sqrt{2})^2}{4} h_K^2 |v|_{H^2(K)} \\
\leq & 10 h_K^2 |v|_{H^2(K)}.
\end{aligned}$$

□

6.8 Satz: *Sei $u \in H_0^1(\Omega)$ die Lösung der Konvektions-Diffusions Gleichung mit $A = \varepsilon I, a \in \mathbb{R}^2, |a| = 1$, $\alpha = 0$ und homogenen Dirichlet-Randbedingungen und $u_h \in X_h$ die Lösung der SDFEM Diskretisierung (5). Es sei $u \in H^2(\Omega)$. Dann gilt die Fehlerabschätzung*

$$|u - u_h|_{1,h} \leq 32.5 c_{\mathcal{T}} \Big\{ \sum_{K \in \mathcal{T}_h} \big[\varepsilon^2 \delta_K + \varepsilon h_K^2 + \delta_K h_K^2 + \delta_K^{-1} h_K^4 \big] |u|^2_{H^2(K)} \Big\}^{1/2}$$

Dabei ist $c_{\mathcal{T}} = \max_{K\in\mathcal{T}_h} h_K/\rho_K$. *Die Fehlerabschätzung ist optimal für die Wahl*

$$\delta_K = \frac{h_K^2}{\sqrt{\varepsilon^2 + h_K^2}}.$$

Beweis: Wegen der Dreiecksungleichung gilt

$$|u - u_h|_{1,h} \le |u - I_h u|_{1,h} + |I_h u - u_h|_{1,h}.$$

Aus Satz 6.3 folgt

$$|u - I_h u|_{1,h} \le 10.5 c_{\mathcal{T}} \Big\{ \sum_{K\in\mathcal{T}_h} (\varepsilon + \delta_K) h_K^2 |u|^2_{H^2(K)} \Big\}^{1/2}.$$

Setze zur Abkürzung $w_h := u_h - I_h u$. Aus Satz 6.5 folgt

$$|w_h|_{1,h} \le a_h(w_h, w_h) = a_h(u - I_h u, w_h) + a_h(u_h - u, w_h).$$

Da $u \in H^2(\Omega)$ ist, folgt aus der Definition von a_h, l_h und (5)

$$\begin{aligned}
&a_h(u_h - u, w_h)\\
=&l_h(w_h) - a_h(u, w_h)\\
=&l(w_h) + \sum_{K\in\mathcal{T}_h} \delta_K \int_K f a_h \cdot \nabla w_h - a(u, w_h) - \sum_{K\in\mathcal{T}_h} \delta_K \int_K a\cdot\nabla u a\cdot\nabla w_h\\
=&\sum_{K\in\mathcal{T}_h} \delta_K \int_K \big\{f - a\cdot\nabla u\big\} a\cdot\nabla w_h\\
=&\sum_{K\in\mathcal{T}_h} \delta_K \int_K \big\{-\varepsilon\Delta u\big\} a\cdot\nabla w_h\\
\le&\sum_{K\in\mathcal{T}_h} \varepsilon\delta_K |u|_{H^2(K)} \|a\cdot\nabla w_h\|_{L^2(K)}\\
\le&\Big\{ \sum_{K\in\mathcal{T}_h} \varepsilon^2\delta_K |u|^2_{H^2(K)} \Big\}^{1/2} |w_h|_{1,h}.
\end{aligned}$$

Hierbei haben wir die Cauchy-Schwarz'sche Ungleichung zunächst für Integrale und danach für endliche Summen ausgenutzt. Mittels partieller Integration für den Konvektionsterm erhalten wir mit Satz 6.3 und Satz 6.7 weiterhin

$$\begin{aligned}
&a_h(u - I_h u, w_h)\\
=&\varepsilon \int_\Omega \nabla(u - I_h u)^T \nabla w_h + \int_\Omega a\cdot\nabla(u - I_h u) w_h\\
&+ \sum_{K\in\mathcal{T}_h} \delta_K \int_K a\cdot\nabla(u - I_h u) a\cdot\nabla w_h\\
=&\varepsilon \int_\Omega \nabla(u - I_h u)^T \nabla w_h - \int_\Omega a\cdot\nabla w_h (u - I_h u)\\
&+ \sum_{K\in\mathcal{T}_h} \delta_K \int_K a\cdot\nabla(u - I_h u) a\cdot\nabla w_h
\end{aligned}$$

$$
\begin{aligned}
\leq & \varepsilon |u - I_h u|_1 |w_h|_1 + \sum_{K \in \mathcal{T}_h} \|u - I_h u\|_{L^2(K)} \|a \cdot \nabla w_h\|_{L^2(K)} \\
& + \sum_{K \in \mathcal{T}_h} \delta_K \|a \cdot \nabla (u - I_h u)\|_{L^2(K)} \|a \cdot \nabla w_h\|_{L^2(K)} \\
\leq & 2 |w_h|_{1,h} \Big\{ \sum_{K \in \mathcal{T}_h} [\varepsilon + \delta_K] |u - I_h u|^2_{H^1(K)} \\
& \qquad + \sum_{K \in \mathcal{T}_h} \delta_K^{-1} \|u - I_h u\|^2_{L^2(K)} \Big\}^{1/2} \\
\leq & 2 |w_h|_{1,h} \Big\{ \sum_{K \in \mathcal{T}_h} [\varepsilon + \delta_K] 10.5^2 c_{\mathcal{T}}^2 h_K^2 |u|^2_{H^2(K)} \\
& \qquad + \sum_{K \in \mathcal{T}_h} \delta_K^{-1} 100 h_K^4 |u|^2_{H^2(K)} \Big\}^{1/2} \\
\leq & 2 \cdot 10.5 c_{\mathcal{T}} |w_h|_{1,h} \Big\{ \sum_{K \in \mathcal{T}_h} h_K^2 [\varepsilon + \delta_K + \delta_K^{-1} h_K^2] |u|^2_{H^2(K)} \Big\}^{1/2}.
\end{aligned}
$$

Aus den letzten beiden Abschätzungen folgt wegen $c_{\mathcal{T}} \geq 1$

$$
\begin{aligned}
|u_h - I_h u|_{1,h} = & |w_h|_{1,h} \\
& \leq 22 c_{\mathcal{T}} \Big\{ \sum_{K \in \mathcal{T}_h} [\varepsilon h_K^2 + h_K^2 \delta_K + \delta_K^{-1} h_K^4 + \varepsilon^2 \delta_K] |u|^2_{H^2(K)} \Big\}^{1/2}.
\end{aligned}
$$

Zusammen mit der bereits bewiesenen Abschätzung für $|u - I_h u|_{1,h}$ folgt hieraus die Fehlerabschätzung für $|u - u_h|_{1,h}$. Offensichtlich ist sie optimal, wenn die δ-Terme und die δ^{-1}-Terme gleich sind. Hieraus folgt die Behauptung über die optimale Wahl von δ. □

6.9 Bemerkung: Es gilt $\delta_K \sim h_K$, wenn $\varepsilon << h_K$ ist. In diesem Fall liefert Satz 6.8 eine optimale Fehlerabschätzung der Form

$$\|a \cdot \nabla (u - u_h)\|_0 \leq ch|u|_2$$

in Stromrichtung. Im Fall $h_K \leq \varepsilon$, in dem auch die Diskretisierung (3) gut funktioniert, ist $\delta_K \sim h_K^2 \varepsilon^{-1}$, und Satz 6.8 liefert eine Fehlerabschätzung der Form

$$|u - u_h|_1 \leq ch|u|_2,$$

die vergleichbar ist zu derjenigen von Satz 6.4. □

7. Finite Elemente höherer Ordnung

In diesem Paragraphen verallgemeinern wir die Vorgehensweise der §§ 5, 6. Dabei ist $d \in \{2, 3\}$ und $\Omega \subset \mathbb{R}^d$ ein offenes, beschränktes, zusammenhängendes Polyedergebiet, d.h. der Rand Γ von Ω besteht stückweise aus Hyperebenen.

Die Grundidee ist folgende:

(1) Unterteile Ω in Teilgebiete $K_1, ..., K_{m_h}$ mit einfacher geometrischer Struktur. In § 5 waren dies Rechtecke, in § 6 Dreiecke.

(2) Approximiere die Sobolev Räume $W^{k,p}(\Omega)$ durch endlich dimensionale Räume X_h, so daß jedes $v \in X_h$ eingeschränkt auf ein beliebiges Element K_i eine einfache Struktur hat. In § 5 waren dies die bilinearen Polynome in Q_1, in § 6 die linearen Polynome in $\mathbb{P}_1$.

(3) Konstruiere eine Basis von X_h, so daß jede Basisfunktion eine einfache Struktur und einen kleinen Träger hat. In § 5, 6 waren die Basisfunktionen die stetigen, stückweise bilinearen bzw. linearen Funktionen, die in einem Elementeckpunkt den Wert 1 haben und in allen anderen Elementeckpunkten verschwinden.

(4) Um den Fehler der Finite Element Diskretisierung abzuschätzen, konstruiere einen einfachen Interpolationsoperator $I_h : W^{k,p} \to X_h$ und schätze den Interpolationsfehler ab. Letzteres geschah in § 6 zunächst auf einem Referenzelement.

Die in (1) geforderte einfache geometrische Struktur der Elemente K_i läßt sich wie folgt konkretisieren: Es gibt ein Referenzelement $\hat{K} \subset \mathbb{R}^d$, so daß jedes K_i zu $\hat{K}$ diffeomorph ist und daß der entsprechende Diffeomorphismus $F_{K_i} : \hat{K} \to K_i$ eine einfache Gestalt hat. Wie in der Praxis üblich, werden wir zwei Typen von Referenzelementen betrachten:

(a) den **Referenz d-Simplex** $\hat{K} := \{\hat{x} \in \mathbb{R}^d : x_1 + ... + x_d \leq 1, x_i \geq 0, 1 \leq i \leq d\}$ und

(b) den **Referenz d-Würfel** $\hat{K} := [0, 1]^d$.

Wir beschränken uns im folgenden auf **affin äquivalente Finite Elemente**, d.h. jedes Element K_i ist das Bild des Referenz-Simplex oder -Würfels unter einer *affinen* Transformation F_{K_i}. Ist $\hat{K}$ der Referenz-Simplex, so ist K_i ein allgemeiner d-Simplex. Ist dagegen $\hat{K}$ der Referenz-Würfel, so ist K_i ein d-Epiped, d.h. ein Parallelogramm, falls $d = 2$ ist, bzw. ein Parallelepiped, falls $d = 3$ ist. Wenn $\hat{K}$ der Referenz-Würfel ist, werden in der Praxis häufig allgemeinere Elemente, sog. **isoparametrische Elemente**, betrachtet, bei denen die Diffeomorphismen F_{K_i} Polynome höheren Grades sind. Wir beschränken uns auf affin äquivalente Elemente, weil dies den technischen Aufwand bei Fehlerabschätzungen erheblich reduziert.

Sei also $\mathcal{T}_h = \{K_i : 1 \leq i \leq m_h\}$ eine Unterteilung von Ω, die folgende Bedingungen erfüllt:

(1) **affine Äquivalenz:** zu jedem $K \in \mathcal{T}_h$ gibt es einen affinen Diffeomorphismus F_K des Referenz-Simplexes oder -Würfels $\hat{K}$ auf K.

(2) **Zulässigkeit:** Je zwei Elemente $K_1, K_2 \in \mathcal{T}_h$ sind entweder disjunkt oder haben einen Eckpunkt, oder eine Kante oder, falls $d = 3$ ist, eine Seitenfläche gemeinsam.

(1) **Regularität:** Der Quotient h_K/ρ_K ist durch eine Konstante $c_{\mathcal{T}}$, die nicht von K oder h abhängt, nach oben beschränkt.

Dabei ist ähnlich wie in § 6 h_K der Durchmesser von K und ρ_K der Durchmesser des größten, in K eingeschriebenen d-Balles.
Wie in den §§ 5,6 bezeichnen wir mit $\mathcal{N}_h$ und $\mathcal{N}_{h,\Omega}$ die Menge aller Eckpunkte aller $K \in \mathcal{T}_h$ bzw. der Eckpunkte im Innern von Ω.

7.1 Definition: **(1)** Bezeichne die Eckpunkte des Referenz-Simplexes mit $\hat{z}_1 := e_1, ..., \hat{z}_d := e_d$ und $\hat{z}_{d+1} := 0$. Dann ist für $k \in \mathbb{N}^*$

$$\hat{\Sigma}_k := \Big\{ x = \sum_{i=1}^{d+1} \mu_i \hat{z}_i : \mu_i \in \{0, \frac{1}{k}, ..., \frac{k-1}{k}, 1\}, \sum_{i=1}^{d+1} \mu_i = 1 \Big\},$$

falls $\hat{K}$ der Referenz-Simplex ist, und

$$\hat{\Sigma}_k := \Big\{ (\frac{\mu_1}{k}, ..., \frac{\mu_d}{k}) : \mu_i \in \{0, 1, ..., k\} \Big\},$$

falls $\hat{K}$ der Referenz-Würfel ist.
(2) Für $K = F_K(\hat{K}) \in \mathcal{T}_h$ und $k \in \mathbb{N}^*$ ist

$$\Sigma_k := \Sigma_k(K) := F_K(\hat{\Sigma}_k).$$

(3) $\mathcal{G}_h := \bigcup_{K \in \mathcal{T}_h} \Sigma_k(K), \mathcal{G}_{h,\Omega} := \mathcal{G}_h \cap \Omega$.
(4) Setze $Q_0 := \mathbb{P}_0 := \mathbb{R}$ und definiere für $k \in \mathbb{N}^*$

$$Q_k := \text{span}\{x^\alpha : \alpha \in \mathbb{N}^d, \max_{1 \le i \le d} \alpha_i \le k\}$$
$$I\!P_k := \text{span}\{x^\alpha : \alpha \in \mathbb{N}^d, \alpha_1 + ... + \alpha_d \le k\}.$$

Dabei ist $x^\alpha := x_1^{\alpha_1} ... x_d^{\alpha_d}$. Setze

$$R_k := R_k(\hat{K}) := \begin{cases} Q_k & \text{falls } \hat{K} \text{ der Referenz-Würfel,} \\ I\!P_k & \text{falls } \hat{K} \text{ der Referenz-Simplex.} \end{cases}$$

□

7.2 Bemerkung: **(1)** Sei K ein allgemeiner Simplex. Bezeichne die Eckpunkte von K mit $z_1, ..., z_{d+1}$. Dann ist

$$\Sigma_k(K) = \Big\{ \sum_{i=1}^{d+1} \mu_i z_i : \mu_i \in \{0, \frac{1}{k}, ..., \frac{k-1}{k}, 1\}, \sum_{i=1}^{d+1} \mu_i = 1 \Big\}.$$

Analog kann man für ein allgemeines Epiped K die Punkte von $\Sigma_k(K)$ bestimmen, indem man die Kanten von K äquidistant unterteilt und die so entstandenen Punkte miteinander verbindet.
(2) Die Menge $\mathcal{G}_h$ heißt auch **Gitter**. Die Punkte von $\mathcal{G}_h$ werden manchmal auch **Knoten** genannt.
(3) Da die F_K affin sind, beschreiben die Transformationen $\varphi \to \varphi \circ F_K$ und $\psi \to \psi \circ F_K^{-1}$ Isomorphismen von $C(K)$ auf $C(\hat{K})$ bzw. von $C(\hat{K})$ auf $C(K)$, die für jedes $k \in \mathbb{N}$ die Polynomräume $\mathbb{P}_k$ und Q_k invariant lassen. □

7.3 Satz: *Sei $k \in \mathbb{N}^*$. Dann ist jedes $p \in R_k$ eindeutig bestimmt durch seine Werte auf $\hat{\Sigma}_k$ bzw. auf $\Sigma_k(K), K \in \mathcal{T}_h$.*

Beweis: Wegen der Definition von $\Sigma_k(K)$ und Bem. 7.2 (3) reicht es, die Behauptung für $\hat{\Sigma}_k$ zu zeigen. Eine leichte Rechnung zeigt, daß

$$\dim R_k = \#\hat{\Sigma}_k$$

ist. Daher reicht es, eine der folgenden Aussagen (a) oder (b) zu zeigen:
(a) Zu jedem Vektor $(b_z)_{z\in\hat{\Sigma}_k}$ existiert ein $\varphi \in R_k$ mit $\varphi(z) = b_z$ für alle $z \in \hat{\Sigma}_k$.
(b) Ist $\varphi \in R_k$ und $\varphi(z) = 0$ für alle $z \in \hat{\Sigma}_k$, so ist $\varphi = 0$.
Fall 1: $\hat{K}$ ist der Referenz-Würfel: Definiere für $i \in \mathbb{N}_k^d$ die Funktion μ_i durch

$$\mu_i(x) := \prod_{j=1}^{d} \prod_{\substack{l_j=0 \\ l_j \neq i_j}}^{k} \frac{kx_j - l_j}{i_j - l_j}.$$

Dann ist $\mu_i \in Q_k$ mit $\mu_i(\frac{1}{k}i) = 1$ und $\mu_i(\frac{1}{k}i') = 0$ für alle $i' \in \mathbb{N}_k^d \backslash \{i\}$. Sei nun $(b_i)_{i\in\mathbb{N}_k^d}$ beliebig. Dann leistet die Funktion

$$\varphi(x) := \sum_{i\in\mathbb{N}_k^d} b_i \mu_i(x)$$

das in Eigenschaft (a) Geforderte.
Fall 2: $\hat{K}$ ist der Referenz-Simplex: Definiere die Funktion $\hat{\lambda}_1, ..., \hat{\lambda}_{d+1}$ durch

$$\hat{\lambda}_i(x) := x_i \quad 1 \le i \le d$$
$$\hat{\lambda}_{d+1}(x) := 1 - \sum_{i=1}^{d} x_i.$$

Die Funktionen $\hat{\lambda}_1, ..., \hat{\lambda}_{d+1}$ heißen die **Schwerpunktskoordinaten** von $\hat{K}$ (vgl. § 6) und haben offensichtlich die Eigenschaft $\hat{\lambda}_i \in \mathbb{P}_1$, $1 \le i \le d+1$, $\hat{\lambda}_i(\hat{z}_j) = \delta_{ij}$, $1 \le i, j \le d+1$.

$\underline{k=1:}$ Zu gegebenem $(b_i)_{1\le i\le d+1}$ leistet offensichtlich

$$\varphi(x) := \sum_{i=1}^{d+1} b_i \hat{\lambda}_i(x)$$

das in Eigenschaft (a) Geforderte.

$\underline{k=2:}$ Bezeichne mit $\hat{z}_{ij} := \frac{1}{2}(\hat{z}_i + \hat{z}_j), 1 \le i < j \le d+1$, die Kantenmittelpunkte von $\hat{K}$. Definiere

$$\begin{aligned} \mu_i &:= \hat{\lambda}_i[2\hat{\lambda}_i - 1] \quad , 1 \le i \le d+1, \\ \mu_{ij} &:= 4\hat{\lambda}_i\hat{\lambda}_j \qquad , 1 \le i < j \le d+1. \end{aligned}$$

Dann gilt offensichtlich $\mu_k, \mu_{ij} \in \mathbb{P}_2$ und

$$\begin{aligned} &\mu_i(\hat{z}_j) = \delta_{ij} \qquad , \quad \mu_i(\hat{z}_{kl}) = 0 \quad \forall j,k,l \\ &\mu_{ij}(\hat{z}_{kl}) = \delta_{ik}\delta_{jl} \quad , \quad \mu_{ij}(\hat{z}_m) = 0 \quad \forall i,j,k,l,m. \end{aligned}$$

Daher leistet zu gegebenem $(b_z)_{z\in\hat{\Sigma}_2} = (b_i, b_{kl})$ die Funktion

$$\varphi(x) := \sum_{i=1}^{d+1} b_i\mu_i(x) + \sum_{1\le k<l\le d+1} b_{kl}\mu_{kl}(x)$$

das in Eigenschaft (a) Geforderte.

$\underline{k=3:}$ Definiere

$$\begin{aligned} \hat{z}_{iij} &:= \frac{1}{3}(2\hat{z}_i + \hat{z}_j) \qquad , 1 \le i,j \le d+1, j \ne i \\ \hat{z}_{ijk} &:= \frac{1}{3}(\hat{z}_i + \hat{z}_j + \hat{z}_k) \quad , 1 \le i < j < k \le d+1 \end{aligned}$$

und

$$\begin{aligned} \mu_i &:= \frac{1}{2}\hat{\lambda}_i[3\hat{\lambda}_i - 1][3\hat{\lambda}_i - 2] \quad , 1 \le i \le d+1, \\ \mu_{iij} &:= \frac{9}{2}\hat{\lambda}_i\hat{\lambda}_j[3\hat{\lambda}_i - 1] \qquad , 1 \le i,j \le d+1, j \ne i, \\ \mu_{ijk} &:= 27\hat{\lambda}_i\hat{\lambda}_j\hat{\lambda}_k \qquad , 1 \le i < l < k \le d+1. \end{aligned}$$

Offensichtlich gilt $\mu_i, \mu_{iij}, \mu_{ijk} \in \mathbb{P}_3$. Eine leichte Rechnung liefert für die relevanten Indizes

$$\begin{aligned} \mu_i(\hat{z}_j) &= \delta_{ij} \quad , \quad \mu_i(\hat{z}_{jjk}) = 0 \quad , \quad \mu_i(\hat{z}_{jkl}) = 0, \\ \mu_{iij}(\hat{z}_{kkl}) &= \delta_{ik}\delta_{jl} \quad , \quad \mu_{iij}(\hat{z}_k) = 0 \quad , \quad \mu_{iij}(\hat{z}_{klm}) = 0, \\ \mu_{ijk}(\hat{z}_{lmn}) &= \delta_{il}\delta_{jm}\delta_{kn} \quad , \quad \mu_{ijk}(\hat{z}_l) = 0 \quad , \quad \mu_{ijk}(\hat{z}_{llm}) = 0. \end{aligned}$$

Daher leistet für gegebenes $(b_z)_{z\in\hat{\Sigma}_3} = (b_i, b_{jjk}, b_{lmn})$ die Funktion

$$\varphi(x) := \sum_{i=1}^{d+1} b_i\mu_i(x) + \sum_{\substack{1\le i,j\le d+1 \\ i\ne j}} b_{ij}\mu_{ij}(x) + \sum_{1\le i<j<k\le d+1} b_{ijk}\mu_{ijk}(x)$$

das in Eigenschaft (a) Geforderte.

$\underline{k \geq 4:}$ Sei $\varphi \in \mathbb{P}_k$ mit $\varphi(z) = 0$ für alle $z \in \hat{\Sigma}_k$. Dann verschwindet φ auf allen Kanten bzw., falls $d = 3$ ist, auf allen Seitenflächen von $\hat{K}$. Daher gibt es ein $\psi \in \mathbb{P}_{k-d-1}$ mit

$$\varphi = \hat{\lambda}_1 ... \hat{\lambda}_{d+1} \psi \quad \text{und} \quad \psi(z) = 0 \quad \forall z \in \hat{\Sigma}_k \cap \overset{\circ}{\hat{K}}.$$

Damit folgt die Eigenschaft (b) durch Induktion über k. □

7.4 Satz: *Seien $k \in \mathbb{N}^*$, $K_1, K_2 \in \mathcal{T}_h$ mit $K_1 \cap K_2 \neq \emptyset$ und $p_1, p_2 \in R_k$. Definiere die Funktion φ auf $K_1 \cup K_2$ durch*

$$\varphi|_{K_i} = p_i \quad , i = 1, 2.$$

Dann ist $\varphi \in C(K_1 \cup K_2)$ genau dann, wenn für alle $x \in F_{K_1}(\hat{\Sigma}_k) \cap F_{K_2}(\hat{\Sigma}_k)$ gilt

$$p_1(x) = p_2(x).$$

Beweis: $\underline{\text{"} \Longrightarrow \text{"}:}$ offensichtlich.

$\underline{\text{"} \Longleftarrow \text{"}:}$ Offensichtlich ist nur etwas zu zeigen, wenn K_1 und K_2 eine Kante oder, falls $d = 3$ ist, eine Seitenfläche gemeinsam haben.

Wir betrachten zunächst den Fall, daß $d = 2$ ist und K_1 und K_2 eine Kante E gemeinsam haben. Setze $\hat{E} := \hat{K} \cap \{x_2 = 0\}$. Dann ist $\hat{E}$ das Einheitsintervall, d.h. der Standard 1-Simplex. Die Transformationen F_{K_1} und F_{K_2} können so gewählt werden, daß $E = F_{K_1}(\hat{E}) = F_{K_2}(\hat{E})$ ist. Definiere $q := p_1 \circ F_{K_1|\hat{E}} - p_2 \circ F_{K_2|\hat{E}}$. Wegen Bem. 7.2 (3) ist $q \in R_k$ ein Polynom einer Veränderlichen. Nach Voraussetzung verschwindet q in allen Punkten von $\hat{\Sigma}_k \cap \hat{E}$. Also ist $q = 0$ und damit φ stetig.

Betrachte nun den Fall, daß $d = 3$ ist und K_1 und K_2 eine Seitenfläche gemeinsam haben. Setze $\hat{E} := \hat{K} \cap \{x_3 = 0\}$ und definiere q wie oben. Dann ist $q \in R_k$ ein Polynom in zwei Veränderlichen, das in den Punkten von $\hat{\Sigma}_k \cap \hat{E}$ verschwindet. Da $\hat{E}$ der Standard 2-Simplex bzw. Standard 2-Würfel ist, folgt aus Satz 7.3, daß $q = 0$ und damit φ stetig ist.

Ist schließlich $d = 3$ und haben K_1 und K_2 eine Kante gemeinsam, setzen wir $\hat{E} := \hat{K} \cap \{x_3 = x_2 = 0\}$ und gehen ansonsten wie im Fall $d = 2$ vor. □

Nach diesen Vorbereitungen können wir nun die Finite Element Räume höherer Ordnung zu $\mathcal{T}_h$ wie folgt definieren:

$$\begin{aligned} S_h^{k,-1} &:= \{\varphi \in L^1(\Omega) : \varphi|_K \in R_k \quad \forall K \in \mathcal{T}_h\}, \\ S_h^{k,0} &:= S_h^{k,-1} \cap C(\overline{\Omega}), \\ S_{h,0}^{k,0} &:= \{\varphi \in S_h^{k,0} : \varphi = 0 \text{ auf } \Gamma\}. \end{aligned}$$

Wegen Satz 1.7 ist $S_h^{k,0} \subset W^{1,p}(\Omega)$ und $S_{h,0}^{k,0} \subset W_0^{k,p}(\Omega)$ für jedes $1 \leq p < \infty$. Wir können daher im Rahmen von Satz 2.2 $X_h = S_{h,0}^{k,0}$ mit beliebigem $k \in \mathbb{N}^*$ wählen

und dadurch die diskreten Probleme der §§ 5, 6 verallgemeinern. Aus den Sätzen 7.3 und 7.4 folgt, daß die Funktionen in $S_h^{k,0}$ und $S_{h,0}^{k,0}$ eindeutig bestimmt sind durch ihre Werte in den Gitterpunkten $\mathcal{G}_h$ bzw. $\mathcal{G}_{h,\Omega}$. Insbesondere gibt es zu jedem $z \in \mathcal{G}_h$ eine eindeutige Funktion $v_z \in S_h^{k,0}$ mit $v_z(z) = 1$ und $v_z(z') = 0$ für alle $z' \in \mathcal{G}_h \backslash \{z\}$. Die Funktionen $v_z, z \in \mathcal{G}_h$, heißen **nodale Basis** von $S_h^{k,0}$. Jedes $u_h \in S_h^{k,0}$ läßt sich eindeutig darstellen als

$$u_h = \sum_{z \in \mathcal{G}_h} \mu_z v_z,$$

und die μ_z sind genau die Werte $u_h(z)$. Daher läßt sich das entsprechende diskrete Problem mit $X_h = S_{h,0}^{k,0}$ wieder als Differenzenverfahren interpretieren.

Ähnlich wie in den §§ 5, 6 definieren wir einen Interpolationsoperator $I_h : H^2(\Omega) \to S_h^{k,0}$ durch

$$(I_h u)(z) = u(z) \quad \forall z \in \mathcal{G}_h$$

oder äquivalent

$$I_h u = \sum_{z \in \mathcal{G}_h} u(z) v_z.$$

Man beachte, daß $I_h(H_0^1(\Omega) \cap H^2(\Omega)) \subset S_{h,0}^{k,0}$ ist. Wie in § 6 geschieht die Abschätzung des Interpolationsfehlers zunächst auf dem Referenzelement. Dazu definieren wir $\hat{\Pi}_k : H^2(\hat{K}) \to R_k$ durch

$$(\hat{\Pi}_k u)(z) = u(z) \quad \forall z \in \hat{\Sigma}_k.$$

Wegen Satz 7.3 und Bem. 1.20 ist diese Definition sinnvoll. Aus Bem. 7.2 (3) folgt außerdem für jedes $u \in H^2(\Omega)$ und jedes $K \in \mathcal{T}_h$ die Identität

$$(I_h u)|_K = [\hat{\Pi}_k(u|_K \circ F_K)] \circ F_K^{-1}.$$

Im folgenden versehen wir Größen wie $\hat{K}, \hat{\Sigma}_k$ oder $\hat{\Pi}_k$ mit einem zusätzlichen Index S oder W, wenn wir hervorheben wollen, daß sie sich auf den Referenz-Simplex (S) bzw. auf den Referenz-Würfel (W) beziehen.

7.5 Satz: *Sei $k \in \mathbb{N}^*$. Dann wird durch*

$$[u]_{k+1} := |u|_{k+1} + \sum_{z \in \hat{\Sigma}_{k,S}} |u(z)|$$

eine Norm auf $H^{k+1}(\hat{K})$ definiert, die zu $\|.\|_{k+1}$ äquivalent ist.

Beweis: Da gemäß Satz 1.19 $H^{k+1}(\hat{K}) \hookrightarrow C(\hat{K})$ ist, ist $[.]_{k+1}$ wohldefiniert, und es gibt eine Konstante $c_1 \in \mathbb{R}_+^*$ mit

$$[u]_{k+1} \leq c_1 \|u\|_{k+1} \quad \forall u \in H^{k+1}(\hat{K}).$$

Wir müssen also noch zeigen, daß es eine Konstante $c_2 \in \mathbb{R}_+^*$ gibt mit

$$\|u\|_{k+1} \leq c_2[u]_{k+1} \quad \forall u \in H^{k+1}(\hat{K}).$$

Angenommen, eine solche Konstante existiere nicht. Dann gibt es eine Folge $(u_n)_{n\in\mathbb{N}}$ $\subset H^{k+1}(\hat{K})$ mit

$$\|u_n\|_{k+1} = 1 \quad \forall n \in \mathbb{N} \tag{1}$$

und

$$\lim_{n\to\infty} [u_n]_{k+1} = 0. \tag{2}$$

Wegen Satz 1.19 und Bem. 1.18 gibt es eine Teilfolge $(u_{n_m})_{m\in\mathbb{N}}$ von $(u_n)_{n\in\mathbb{N}}$ und ein $u \in H^k(\hat{K})$ mit

$$\lim_{m\to\infty} \|u_{n_m} - u\|_k = 0.$$

Wegen (2) ist insbesondere

$$\lim_{m\to\infty} |u_{n_m} - u|_{k+1} = 0.$$

Daher ist sogar $u \in H^{k+1}(\hat{K})$ mit $|u|_{k+1} = 0$, und es gilt

$$\lim_{m\to\infty} \|u_{n_m} - u\|_{k+1} = 0.$$

Wegen $|u|_{k+1} = 0$ ist $u \in \mathbb{P}_k$. Wegen Satz 1.19 gilt

$$u(z) = \lim_{m\to\infty} u_{n_m}(z) \quad \forall z \in \hat{\Sigma}_{k,S}.$$

Hieraus und aus (2) folgt aber

$$u(z) = 0 \quad \forall z \in \hat{\Sigma}_{k,S}.$$

Wegen Satz 7.3 ist also $u = 0$ im Widerspruch zu (1). □

7.6 Satz: *Sei $k \in \mathbb{N}^*$. Dann gibt es eine Konstante c, die von k abhängt, mit*

$$\|u - \hat{\Pi}_k u\|_{k+1} \leq c|u|_{k+1} \quad \forall u \in H^{k+1}(\hat{K}).$$

Beweis: Fall 1: $\hat{K} = \hat{K}_S$: Aus Satz 7.5 folgt für beliebiges $u \in H^{k+1}(\hat{K})$

$$\|u - \hat{\Pi}_k u\|_{k+1} \leq c_2[u - \hat{\Pi}_k u]_{k+1} = c_2|u|_{k+1},$$

da $\hat{\Pi}_k u \in \mathbb{P}_k$ und $u(z) - \hat{\Pi}_k u(z) = 0$ ist für alle $z \in \hat{\Sigma}_{k,S}$.

Fall 2: $\hat{K} = \hat{K}_W$: Sei $u \in H^{k+1}(\hat{K})$ beliebig. Da $\hat{K}_S \subset \hat{K}_W$ ist, folgt aus der Dreiecksungleichung

$$\|u - \hat{\Pi}_{k,W} u\|_{k+1} \leq \|u - \hat{\Pi}_{k,S} u\|_{k+1} + \|\hat{\Pi}_{k,S} u - \hat{\Pi}_{k,W} u\|_{k+1}.$$

Wegen $\mathbb{P}_k \subset Q_k$ und Satz 7.3 gilt

$$\hat{\Pi}_{k,W}(\hat{\Pi}_{k,S} u) = \hat{\Pi}_{k,S} u$$

und somit

$$\hat{\Pi}_{k,S} u - \hat{\Pi}_{k,W} u = \hat{\Pi}_{k,W}(\hat{\Pi}_{k,S} u - u).$$

Bezeichne mit $\hat{v}_z$, $z \in \hat{\Sigma}_{k,W}$, die nodale Basis zu $\hat{K}$ und Q_k, d.h.

$$\begin{aligned} &\hat{v}_z \in Q_k, \\ &\hat{v}_z(z) = 1, \\ &\hat{v}_z(z') = 0 \quad \forall z' \in \hat{\Sigma}_{k,W} \backslash \{z\}. \end{aligned}$$

Dann folgt für beliebiges $\varphi \in H^{k+1}(\hat{K})$ mit Satz 1.19

$$\begin{aligned} \|\hat{\Pi}_{k,W} \varphi\|_{k+1} =& \| \sum_{z \in \hat{\Sigma}_{k,W}} \varphi(z) \hat{v}_z \|_{k+1} \\ \leq& \sum_{z \in \hat{\Sigma}_{k,W}} |\varphi(z)| \|\hat{v}_z\|_{k+1} \\ \leq& \|\varphi\|_{C(\hat{K})} \sum_{z \in \hat{\Sigma}_{k,W}} \|\hat{v}_z\|_{k+1} \\ \leq& c \|\varphi\|_{k+1} \sum_{z \in \hat{\Sigma}_{k,W}} \|\hat{v}_z\|_{k+1} \\ =& c' \|\varphi\|_{k+1}. \end{aligned}$$

Aus dem soeben Gezeigten und dem Fall 1 folgt

$$\begin{aligned} \|u - \hat{\Pi}_{k,W} u\|_{k+1} \leq& \|u - \hat{\Pi}_{k,S} u\|_{k+1} + \|\hat{\Pi}_{k,W}(\hat{\Pi}_{k,S} u - u)\|_{k+1} \\ \leq& (1 + c') \|u - \hat{\Pi}_{k,S} u\|_{k+1} \\ \leq& (1 + c') c_2 |u|_{k+1}. \end{aligned}$$

□

7.7 Satz: *Sei $k \in \mathbb{N}^*$. Dann gelten für alle $u \in H^{k+1}(\Omega)$ die Interpolationsfehlerabschätzungen*

$$\begin{aligned} &\|u - I_h u\|_0 \leq c_1 h^{k+1} |u|_{k+1} \\ &|u - I_h u|_1 \leq c_2 h^k |u|_{k+1}. \end{aligned}$$

Dabei ist $h := \max_{K\in\mathcal{T}_h} h_K$. *Die Konstanten* c_1 *und* c_2 *hängen von* Ω, k *und der Konstanten* $c_{\mathcal{T}}$ *in der Regularitätsbedingung an* $\mathcal{T}_h$ *ab.*

Beweis: Sei $K \in \mathcal{T}_h$ und $F_K : x \to b_K + B_K x$ eine affine Transformation des Referenzelements $\hat{K}$ auf K. Wie im Beweis von Satz 6.3 erhalten wir

$$\begin{aligned}
\|u - I_h u\|_{L^2(K)} =& |\det B_K|^{1/2} \|(u - I_h u) \circ F_K\|_{L^2(\hat{K})},\\
|u - I_h u|_{H^1(K)} \leq& |\det B_K|^{1/2} \|\|B_K^{-1}\|\| (u - I_h u) \circ F_K|_{H^1(\hat{K})},\\
|u \circ F_K|_{H^{k+1}(\hat{K})} \leq& |\det B_K|^{-1/2} \|\|B_K\|\|^{k+1} |u|_{H^{k+1}(\hat{K})}.
\end{aligned}$$

Wie im Beweis von Satz 6.2 folgt

$$\|\|B_K\|\| \leq h_K/\rho_{\hat{K}} \quad , \quad \|\|B_K^{-1}\|\| \leq h_{\hat{K}}/\rho_K.$$

Aus diesen Abschätzungen und Satz 7.6 ergibt sich wegen $I_h u \circ F_K = \hat{\Pi}_k(u \circ F_K)$

$$\begin{aligned}
\|u - I_h u\|_{L^2(K)} \leq& c(h_K/\rho_{\hat{K}})^{k+1} |u|_{H^{k+1}(K)},\\
|u - I_h u|_{H^1(K)} \leq& c h_{\hat{K}}/\rho_K (h_K/\rho_{\hat{K}})^{k+1} |u|_{H^{k+1}(K)},\\
=& c h_{\hat{K}} \rho_{\hat{K}}^{-k-1} \Big(\frac{h_K}{\rho_K}\Big) h_K^k |u|_{H^{k+1}(K)}.
\end{aligned}$$

Hieraus folgt die Behauptung durch Quadrieren und Summieren über alle Elemente $K \in \mathcal{T}_h$. □

7.8 Bemerkung: Die Ergebnisse der Sätze 7.5, 7.7 gelten mit den offensichtlichen Modifikationen für alle $W^{k+1,p}$-Räume mit $1 \leq p < \infty$. □

8. Randapproximation und numerische Integration

In den §§ 5 – 7 haben wir stets vorausgesetzt, daß Ω ein Polyedergebiet ist, d.h. der Rand Γ besteht stückweise aus Hyperebenen. I.a. ist jedoch der Rand von Ω gekrümmt. Daher ersetzt man in der Praxis Ω durch ein approximierendes Polyeder Ω_h, derart daß die Eckpunkte von Ω_h auf Γ liegen und die Kanten von Ω_h die Länge $O(h)$ haben. Sei Γ_h der Rand von Ω_h. Wenn Γ stückweise C^2 ist, folgt dann

$$\begin{aligned}
\operatorname{dist}(\Gamma, \Gamma_h) =& \sup_{x\in\Gamma} \inf_{y\in\Gamma_h} |x - y|\\
=& \sup_{y\in\Gamma_h} \inf_{x\in\Gamma} |x - y| = O(h^2).
\end{aligned}$$

Daher kann der Fehler der Finite Element Approximation bestenfalls von der Ordnung $O(h^2)$ sein.

$\mathcal{T}_h$ ist dann eine Unterteilung von Ω_h, die den Bedingungen der §§ 5 – 7 genügt. Zusätzlich müssen die Eckpunkte von Ω_h in der Menge $\mathcal{N}_h$ der Elementeckpunkte enthalten sein. Für die Konstruktion von Ω_h und $\mathcal{T}_h$ muß der Benutzer den Rand Γ stückweise als Graph oder implizit als $\{F(x) = 0\}$ zusammen mit F und $\operatorname{grad} F$ angeben.

Falls Ω konvex ist, gilt $\Omega_h \subset \Omega$ und die Fehlerabschätzungen der §§ 5 – 7 können übertragen werden, wobei die Randapproximation einen zusätzlichen Konsistenzfehler von $O(h^2)$ beiträgt. Ist Ω nicht konvex, gilt $\Omega_h \backslash \Omega \neq \emptyset$. Dann müssen die Daten A, a, α und f auf Ω_h fortgesetzt werden. Dies geschieht am einfachsten durch Reflexion: Ist $x \in \Omega_h \backslash \Omega$, setze $f(x) := f(\tilde{x})$ und analog für A, a und α. Dabei wird $\tilde{x}$ wie folgt bestimmt: Berechne die orthogonale Projektion $\overline{x}$ von x auf Γ und wähle $\tilde{x}$ auf der Geraden durch x und $\overline{x}$, so daß $\tilde{x}$ in Ω liegt und den gleichen Abstand zu $\overline{x}$ hat wie x. Mit diesen Modifikationen bleiben die Fehlerabschätzungen der §§ 5 – 7 erhalten. Dabei müssen alle Normen auf $\Omega \cap \Omega_h$ statt auf Ω bezogen werden. Die Randapproximation trägt wieder einen Konsistenzfehler der Ordnung $O(h^2)$ bei.

Um den Konsistenzfehler durch die Randapproximation zu verringern, kann man auch krummlinige oder isoparametrische Elemente betrachten. Dabei ist mit der Notation von § 7 die Transformation $F_K : \hat{K} \to K$ nicht mehr affin, sondern ein allgemeiner Diffeomorphismus (krummlinige Elemente) oder komponentenweise eine Funktion aus R_k (isoparametrische Elemente). In diesem Sinn sind die linearen Dreieckselemente aus § 6 isoparametrische Elemente der Ordnung 1. Bei Verwendung isoparametrischer Elemente der Ordnung $k \geq 2$ kann man eine Randapproximation der Ordnung $O(h^{k+1})$ erreichen. Dementsprechend ist der Konsistenzfehler durch die Randapproximation auch von der Ordnung $O(h^{k+1})$. Die größere Genauigkeit wird natürlich durch einen erhöhten Aufwand erkauft.

Die Finite Element Diskretisierungen der §§ 5 – 7 führen auf lineare Gleichungssysteme. Zur Berechnung der rechten Seite und der Steifigkeitsmatrix müssen Integrale der Form $\int_K \varphi$ mit $\varphi = fv, \varphi = \nabla u^T A \nabla v$ o.ä. berechnet werden. Wir haben bisher stets angenommen, daß dies exakt geschieht. Außer in einfachen Spezialfällen wird man aber in der Praxis diese Integrale nur näherungsweise mit einem Quadraturverfahren berechnen. Passende Quadraturformeln werden wegen der Ergebnisse von § 7 am besten aus solchen für das Referenzelement hergeleitet. Sei also

$$Q_{\hat{K}}(\hat{\varphi}) := \sum_{l=1}^{L} \hat{\omega}_l \hat{\varphi}(\hat{b}_l)$$

eine Quadraturformel für $\int_{\hat{K}} \hat{\varphi}$. Sie hat die Ordnung k, wenn für alle $\hat{p} \in \mathbb{P}_k$ gilt

$$Q_{\hat{K}}(\hat{p}) = \int_{\hat{K}} \hat{p}.$$

Sei nun $K \in \mathcal{T}_h$ und $F_K : \hat{K} \to K$ eine affine Transformation mit $B_K := DF_K$. Da F_K affin ist, folgt aus Satz II.1.4 der Vorlesung "Einführung in die Numerik", daß

$$Q_K(\varphi) := \sum_{l=1}^{L} \omega_l \varphi(b_l)$$

mit

$$\omega_l = |\det B_K| \hat{\omega}_l \quad , \quad b_l = F_K(\hat{b}_l)$$

eine Quadraturformel der Ordnung k für $\int_K \varphi$ ist, d.h.

$$Q_K(\varphi) = \int_K \varphi \qquad \forall \varphi \in R_k.$$

Man kann zeigen, daß die Fehlerabschätzungen der §§ 5 – 7 für Finite Element Diskretisierungen der Ordnung m, d.h. $X_h = S^{m,0}_{h,0}$, gültig bleiben, wenn die Quadraturformel $Q_{\hat{K}}$ mindestens die Ordnung $2m-2$ hat (s. § 4.1 in Ph. G. Ciarlet: The Finite Element Method for Elliptic Problems). Insbesondere reicht also für lineare simpliziale Elemente ($m = 1, \hat{K} = \hat{K}_S$) und d-lineare Würfelelemente ($m = 1, \hat{K} = \hat{K}_W$) eine Quadraturformel der Ordnung 0 aus, d.h. nur die konstanten Funktionen müssen exakt integriert werden.

8.1 Beispiel: **(1)** $\hat{K} = \hat{K}_S, d \in \{2,3\}, L = 1, \hat{\omega}_1 = \frac{1}{d!}, \hat{b}_1 = \frac{1}{d+1} \sum_{i=1}^{d+1} \hat{z}_i$ (Schwerpunkt) liefert eine Quadraturformel der Ordnung 1 für den Referenz d-Simplex.
(2) $\hat{K} = \hat{K}_S, d = 2, L = 3, \hat{\omega}_l = \frac{1}{6}, \hat{b}_l := \frac{1}{2}(\hat{z}_l + \hat{z}_{l+1})$ (Kantenmittelpunkte) liefert eine Quadraturformel der Ordnung 2 für das Referenzdreieck.
(3) $\hat{K} = \hat{K}_S, d = 2, L = 7$

$$Q_{\hat{K}}(\varphi) = \frac{1}{120}\Big\{3\sum_{i=1}^{3} \varphi(\hat{z}_i) + 8 \sum_{1 \le i < j \le 3} \varphi(\hat{z}_{ij}) + 27\varphi(\hat{z}_{123})\Big\}$$

liefert eine Formel der Ordnung 3 für das Referenzdreieck.
(4) Sei

$$\tilde{Q}(\psi) = \sum_{l=1}^{L} \tilde{\omega}_l \psi(\tilde{x}_l)$$

eine Quadraturformel der Ordnung k für $\int_0^1 \psi(x)dx$. Dann ist gemäß Beispiel II.1.3 (3) der Vorlesung "Einführung in die Numerik"

$$Q_{\hat{K}}(\hat{\varphi}) := \sum_{l_1=1}^{L} ... \sum_{l_d=1}^{L} \tilde{\omega}_{l_1} ... \tilde{\omega}_{l_d} \hat{\varphi}(\tilde{x}_{l_1}, ..., \tilde{x}_{l_d})$$

eine Quadraturformel der Ordnung k für den Referenz d-Würfel.

(5) Die Mittelpunktsregel liefert die Formel

$$Q_{\hat{K}}(\hat{\varphi}) = \hat{\varphi}\Big(\frac{1}{2}, \frac{1}{2}\Big)$$

der Ordnung 1 für das Referenz-Quadrat.
(6) Die Trapezregel liefert die Formel

$$Q_{\hat{K}}(\hat{\varphi}) = \frac{1}{4}\sum_{i=0}^{1}\sum_{j=0}^{1}\hat{\varphi}(i,j)$$

der Ordnung 1 für das Referenz-Quadrat.
(7) Die Simpsonregel liefert mit

$$(a_{ij})_{0\le i,j\le 2} := \begin{pmatrix} 1 & 4 & 1 \\ 4 & 16 & 4 \\ 1 & 4 & 1 \end{pmatrix}$$

die Formel

$$Q_{\hat{K}}(\hat{\varphi}) = \frac{1}{36}\sum_{i=0}^{2}\sum_{j=0}^{2} a_{ij}\varphi\Big(\frac{i}{2}, \frac{j}{2}\Big)$$

der Ordnung 3 für das Referenz-Quadrat. □

9. Numerische Lösung der diskreten Probleme

Um die wesentlichen Punkte besser herausarbeiten zu können und um unnötige technische Schwierigkeiten zu vermeiden, betrachten wir im folgenden die Reaktions-Diffusions Gleichung mit homogenen Dirichlet Randbedingungen und Dreieckselemente der Ordnung $k \ge 1$, d.h. $\mathcal{T}_h$ besteht aus Dreiecken und $X_h = S^{k,0}_{h,0}$.

Da die Träger der nodalen Basisfunktionen nur aus wenigen Dreiecken bestehen, ist die Steifigkeitsmatrix der Finite Element Diskretisierung dünn besetzt. Wegen des Speicherplatzbedarfes und des Rechenaufwandes sind direkte Gleichungslöser wie die Cholesky Zerlegung nur für grobe Gitter, d.h. wenige Dreiecke effizient. Für feinere Diskretisierungen sind iterative Lösungsverfahren vorzuziehen. Da die Effizienz dieser Verfahren wesentlich von der Kondition der Steifigkeitsmatrix abhängt, wollen wir diese zunächst abschätzen. Dazu definieren wir ein gitterabhängiges Skalarprodukt $(.,.)_h$ und eine zugehörige Norm $\|.\|_h$ auf X_h durch

$$\begin{aligned}(\varphi,\psi)_h :=& \sum_{K\in\mathcal{T}_h}\sum_{z\in\Sigma_k(K)} h_K^2\varphi(z)\psi(z),\\ \|\varphi\|_h :=& (\varphi,\varphi)_h^{1/2}.\end{aligned}$$

$\|.\|_h$ ist eine skalierte euklidische Norm auf $\mathbb{R}^{n_h}$, $n_h := \#\mathcal{G}_{h,\Omega}$. Die Matrix, die zu $(.,.)_h$ und der nodalen Basis gehört, ist diagonal. Insbesondere können Gleichungssysteme der Form

$$(u,v)_h = l(v) \qquad \forall v \in X_h$$

leicht gelöst werden.

9.1 Satz: **(1)** *$\|.\|_h$ und $\|.\|_0$ sind äquivalente Normen auf X_h. Die entsprechenden Konstanten hängen nur von k and der Konstanten $c_{\mathcal{T}}$ in der Regularitätsbedingung an $\mathcal{T}_h$ ab.*
(2) *Für alle $v \in X_h$ gilt die* **inverse Abschätzung**

$$|v|_1 \le c(\min_{K\in\mathcal{T}_h} h_K)^{-1}\|v\|_0.$$

Die Konstante c hängt von k und $c_{\mathcal{T}}$ ab.
Beweis: **ad (1):** Wegen Satz 7.3 ist $\left\{\sum_{z\in\hat{\Sigma}_k}|\varphi(z)|^2\right\}^{1/2}$ eine Norm auf $\mathbb{P}_k$. Da $\mathbb{P}_k$ endlich dimensional ist, gibt es zwei Konstanten $\hat{c}_1, \hat{c}_2$, die nur von k abhängen, mit

$$\hat{c}_1\|\varphi\|_{L^2(\hat{K})} \le \Big\{\sum_{z\in\hat{\Sigma}_k}\varphi(z)^2\Big\}^{1/2} \le \hat{c}_2\|\varphi\|_{L^2(\hat{K})} \quad \forall\varphi\in\mathbb{P}_k.$$

Sei nun $K \in \mathcal{T}_h$ beliebig und $F_K : \hat{K} \to K$ eine affine Transformation. Wegen

$$c_1 h_K^2 \le |\det DF_K| \le c_2 h_K^2$$

folgt aus obiger Abschätzung mit dem Transformationssatz

$$\begin{aligned}
\hat{c}_1\|\varphi\|_{L^2(K)} =& \hat{c}_1|\det DF_K|^{1/2}\|\varphi\circ F_K\|_{L^2(\hat{K})}\\
\le& c_2^{1/2}h_K\Big\{\sum_{z\in\hat{\Sigma}_k}|\varphi\circ F_K(z)|^2\Big\}^{1/2}\\
=& c_2^{1/2}h_K\Big\{\sum_{z\in\Sigma_k(K)}|\varphi(z)|^2\Big\}^{1/2}\\
\le& c_2^{1/2}h_K\hat{c}_2\|\varphi\circ F_K\|_{L^2(\hat{K})}\\
=& c_2^{1/2}\hat{c}_2 h_K|\det DF_K|^{-1/2}\|\varphi\|_{L^2(K)}\\
\le& c_2^{1/2}\hat{c}_2 c_1^{-1/2}\|\varphi\|_{L^2(K)}.
\end{aligned}$$

Hieraus folgt die Behauptung (1) durch Quadrieren und Summieren über alle Dreiecke. Man beachte, daß die Konstanten c_1, c_2 von $c_{\mathcal{T}}$ abhängen.
ad (2): Da $|.|_{H^1(\hat{K})}$ eine Norm auf $\mathbb{P}_k/\mathbb{R}$ und $\mathbb{P}_k/\mathbb{R}$ endlich dimensional ist, gibt es eine Konstante $\hat{c}$ mit

$$|\varphi|_{H^1(\hat{K})} \le \hat{c}\|\varphi\|_{L^2(\hat{K})} \quad \forall\varphi\in\mathbb{P}_k/\mathbb{R}.$$

Da die linke Seite dieser Ungleichung für konstante Funktionen verschwindet, gilt die Ungleichung sogar auf ganz $\mathbb{P}_k$. Seien nun K und F_K wie in Teil (1). Dann folgt durch Transformation auf das Referenzelement

$$\begin{aligned}
|\varphi|_{H^1(K)} \leq & |\det DF_K|^{1/2} \|\|DF_K^{-1}\|\| |\varphi \circ F_K|_{H^1(\hat{K})} \\
\leq & \hat{c} |\det DF_K|^{1/2} \|\|DF_K^{-1}\|\| \|\varphi \circ F_K\|_{L^2(\hat{K})} \\
= & \hat{c} \|\|DF_K^{-1}\|\| \|\varphi\|_{L^2(K)} \\
\leq & \hat{c} h_{\hat{K}} / \rho_K \|\varphi\|_{L^2(K)} \\
\leq & c' h_K^{-1} \|\varphi\|_{L^2(K)}.
\end{aligned}$$

Dabei hängt c' von $c_{\mathcal{T}}$ ab. Quadrieren und Summieren dieser Ungleichung beweist Teil (2). □

Aus Satz 9.1 (1), dem Beweis des Existenzsatzes 3.3 und der Friedrich'schen Ungleichung 1.15 folgt für alle $v \in X_h$

$$a(v,v) \geq \beta |v|_1^2 \geq \beta' \|v\|_0^2 \geq \beta'' \|v\|_h^2.$$

Ebenso folgt aus dem Beweis von Satz 3.3 und Satz 9.1 für alle $v, w \in X_h$

$$\begin{aligned}
a(v,w) \leq & B |v|_1 |w|_1 \\
\leq & c^2 B (\min_{K \in \mathcal{T}_h} h_K)^{-2} \|v\|_0 \|w\|_0 \\
\leq & c' (\min_{K \in \mathcal{T}_h} h_K)^{-2} \|v\|_h \|w\|_h.
\end{aligned}$$

Also hat die Steifigkeitsmatrix die Kondition $0((\min_{K \in \mathcal{T}_h} h_K)^{-2})$. Daher scheiden das Jacobi und Gauß-Seidel Verfahren als Löser aus. Da die Bilinearform a symmetrisch ist, ist die Steifigkeitsmatrix symmetrisch, und ein CG-Verfahren oder besser ein PCG-Verfahren können als Löser benutzt werden. Als Vorkonditionierer eignet sich z.B. der SSOR-Vorkonditionierer aus § III.6 der Vorlesung "Numerische Behandlung von Differentialgleichungen I".

In der Vorlesung "Numerische Behandlung von Differentialgleichungen I" haben wir gesehen, daß Mehrgitter (MG-)Verfahren sehr effiziente Löser für Differenzendiskretisierungen elliptischer Differentialgleichungen liefern. Wir wollen daher nun ein MG-Verfahren für Finite Element Diskretisierung analysieren. Dazu nehmen wir an, daß wir eine Hierarchie geschachtelter Triangulierungen $\mathcal{T}_{h_0} \subset \mathcal{T}_{h_1} \subset ... \subset \mathcal{T}_{h_R}$ haben. Es gilt, die Finite Element Diskretisierung zur feinsten Triangulierung numerisch zu lösen. Dazu machen wir folgende Annahmen:

(1) Ω ist konvex.

(2) Jede Triangulierung $\mathcal{T}_{h_m}$ ist **uniform**, d.h.

$$h_m := \max_{K \in \mathcal{T}_{h_m}} h_K \leq c \min_{K \in \mathcal{T}_{h_m}} h_K$$

mit einer von m unabhängigen Konstanten c.

(3) $h_{m-1} \leq c h_m$ für alle m mit einer von m unabhängigen Konstanten c.

Zur Vereinfachung der Notation ersetzen wir einen Index h_m in der Regel durch m.

9.2 Algorithmus: (MG-Verfahren mit V-Zyklus und Jacobi Glättung)

0. *Gegeben eine Näherung $u_m^0 \in X_m$ für die Lösung des diskreten Problems.*
1. **(Vorglättung)** *Für $i = 1, ..., \nu_1$ berechne u_m^i aus u_m^{i-1} als Lösung von*

$$(u_m^i - u_m^{i-1}, v)_m = \omega_m^{-1}\{l_m(v) - a(u_m^{i-1}, v)\} \quad \forall v \in X_m.$$

2. **(Grobgitterkorrektur)** *Berechne*

$$l_{m-1}(v) := l_m(v) - a(u_m^{\nu_1}, v) \quad \forall v \in X_{m-1}.$$

Falls $m \geq 1$, wende das MG-Verfahren mit Startwert $u_{m-1}^0 = 0$ auf das Problem

$$a(u_{m-1}^*, v) = l_{m-1}(v) \quad \forall v \in X_{m-1}$$

an. Das Ergebnis sei $\tilde{u}_{m-1}$. Falls $m = 1$ ist, berechne $\tilde{u}_{m-1} := u_{m-1}^*$. Setze

$$u_m^{\nu_1+1} := u_m^{\nu_1} + \tilde{u}_{m-1}.$$

3. **(Nachglättung)** *Für $i = \nu_1 + 2, ..., \nu_1 + \nu_2 + 1$ berechne u_m^i aus u_m^{i-1} als Lösung von*

$$(u_m^i - u_m^{i-1}, v)_m = \omega_m^{-1}\{l_m(v) - a(u_m^{i-1}, v)\} \quad \forall v \in X_m.$$

9.3 Bemerkung: **(1)** Es ist $l_R(v) = l(v) = \int_\Omega fv$. Die rechten Seiten l_m zu den gröberen Triangulierungen werden rekursiv berechnet.
(2) Die Dämpfungsparameter ω_m werden später bestimmt.
(3) Die Gleichungssysteme in den Glättungsschritten haben eine diagonale Koeffizientenmatrix.
(4) Die Grobgitterkorrektur nutzt aus, daß die Finite Element Räume geschachtelt sind, d.h. $X_{m-1} \subset X_m$. In der Praxis stellt man u_m und u_{m-1} als Vektoren dar, deren Komponenten die Werte in den entsprechenden Gitterpunkten sind. Insbesondere muß u_{m-1} vom Gitter $\mathcal{G}_{m-1}$ auf das Gitter $\mathcal{G}_m$ interpoliert werden. □

Da die Bilinearform a symmetrisch ist, besitzt sie auf jedem X_m einen vollständigen Satz $\lambda_{m,1}, ..., \lambda_{m,n_m}$, $n_m := \dim X_m$, von Eigenwerten und zugehörigen, bzgl. $(.,.)_m$ orthonormierten Eigenfunktionen $\psi_{m,1}, ..., \psi_{m,n_m}$:

$$a(\psi_{m,\mu}, v) = \lambda_{m,\mu}(\psi_{m,\mu}, v)_m \quad \forall v \in X_m$$
$$(\psi_{m,\mu}, \psi_{m,\nu})_m = \delta_{\mu\nu}.$$

O.E. können wir die Eigenwerte der Größe nach ordnen

$$0 < \lambda_{m,1} < ... < \lambda_{m,n_m} =: \Lambda_m.$$

Da die Triangulierungen uniform sind, folgt aus Satz 9.1 $\Lambda_m \sim h_m^{-2}$. Wir nehmen im folgenden an, daß

$$\omega_m = \Lambda_m$$

ist. Es reicht jedoch, wenn $\omega_m \geq \Lambda_m$ und $\omega_m \sim h_m^{-2}$ ist.

Jedes $v \in X_m$ läßt sich eindeutig darstellen als

$$v = \sum_{\mu=1}^{n_m} c_\mu \psi_{m,\mu}.$$

Für $s \in \mathbb{R}$ können wir daher durch

$$|||v|||_{s,m} := \Big\{ \sum_{\mu=1}^{n_m} \lambda_{m,\mu}^s c_\mu^2 \Big\}^{1/2}$$

eine Norm auf X_m definieren. Aus den Voraussetzungen folgt für alle $v \in X_m$

$$\begin{aligned} |||v|||_{0,m} &= \|v\|_{h_m} \simeq \|v\|_{L^2(\Omega)}, \\ |||v|||_{1,m} &= a(v,v)^{1/2} \cong \|v\|_{H^1(\Omega)}. \end{aligned}$$

Dabei bedeutet "$\simeq$" Äquivalenz von Normen mit von m unabhängigen Konstanten. Sind $v, w \in X_m$ mit $v = \Sigma c_\mu \psi_{m,\mu} \quad , \quad w = \Sigma d_\mu \psi_{m,\mu}$, so folgt aus der Cauchy-Schwarz'schen Ungleichung für Summen und der Orthogonalität der Eigenfunktionen

$$\begin{aligned} a(v,w) =& \sum_{\mu=1}^{n_m} \lambda_{m,\mu} c_\mu d_\mu \\ \leq& \Big\{ \sum_{\mu=1}^{n_m} c_\mu^2 \Big\}^{1/2} \Big\{ \sum_{\mu=1}^{n_m} \lambda_{m,\mu}^2 d_\mu^2 \Big\}^{1/2} \\ =& |||v|||_{0,m} |||w|||_{2,m}. \end{aligned} \tag{1}$$

9.4 Satz: *Bezeichne mit $Q_m : X_m \to X_{m-1}$ die* **Ritz-Projektion**, *d.h. $Q_m v \in X_{m-1}$ und*

$$a(Q_m v, w) = a(v,w) \quad \forall w \in X_{m-1}.$$

Dann gilt für alle $v \in X_m$

$$|||v - Q_m v|||_{1,m} \leq c h_m |||v|||_{2,m}.$$

Die Konstante c hängt nicht von m ab.

Beweis: Aus der Definition der Ritz-Projektion und (1) folgt

$$\begin{aligned} |||v - Q_m v|||_{1,m}^2 =& a(v - Q_m v, v - Q_m v) \\ =& a(v - Q_m v, v) \\ \leq& |||v - Q_m v|||_{0,m} |||v|||_{2,m} \\ \leq& c \|v - Q_m v\|_{0,m} |||v|||_{2,m}. \end{aligned}$$

Da Ω konvex ist, folgt aus Satz 2.2

$$\begin{aligned}\|v - Q_m v\|_0 \leq & c h_{m-1} \|v - Q_m v\|_1 \\ \leq & c' h_{m-1} |||v - Q_m v|||_{1,m}.\end{aligned}$$

Wegen $h_{m-1} \leq c h_m$ folgt hieraus die Behauptung. □

Als nächstes definieren wir einen Operator $J : X_m \to X_m$ durch

$$(Jv, w)_m = (v, w)_m - \Lambda_m^{-1} a(v, w) \quad \forall w \in X_m.$$

J beschreibt die Fehlerfortpflanzung in den Glättungsschritten von Algorithmus 9.2. Ist $v \in \Sigma c_\mu \psi_{m,\mu}$, so folgt

$$Jv = \sum_{\mu=1}^{n_m} c_\mu (1 - \frac{\lambda_{m,\mu}}{\Lambda_m}) \psi_{m,\mu}.$$

Insbesondere ist J symmetrisch positiv semi-definit bzgl. des Skalarproduktes $a(.,.)$. Definiere für $v \in X_m$

$$\begin{aligned}|v| := & a(v, Jv)^{1/2} \\ \rho(v) := & \begin{cases} \frac{|v|^2}{|||v|||_{1,m}^2} & \text{, falls } v \neq 0, \\ 0 & \text{, falls } v = 0. \end{cases}\end{aligned}$$

Dann gilt offensichtlich für $v \in X_m$

$$\begin{aligned}|v| &= |||J^{1/2} v|||_{1,m}, \\ 0 \leq \rho(v) &\leq 1.\end{aligned}$$

9.5 Satz: *Sei $v \in X_m$ und $\rho := \rho(J^\nu v)$. Dann gilt*

$$|||J^\nu v|||_{1,m} \leq \rho^\nu |||v|||_{1,m}.$$

Beweis: Schreibe $v = \Sigma c_\mu \psi_{m,\mu}$ und setze zur Abkürzung $\sigma_\mu := 1 - \lambda_{m,\mu}/\Lambda_m$. Dann folgt mit der Hölder'schen Ungleichung

$$\begin{aligned}|||J^\nu v|||_{1,m}^2 = & \sum_{\mu=1}^{n_m} \lambda_{m,\mu} \sigma_\mu^{2\nu} c_\mu^2 \\ \leq & \Big\{ \sum_{\mu=1}^{n_m} \lambda_{m,\mu} \sigma_\mu^{2\nu+1} c_\mu^2 \Big\}^{2\nu/2\nu+1} \Big\{ \sum_{\mu=1}^{n_m} \lambda_{m,\mu} c_\mu^2 \Big\}^{1/2\nu+1} \\ = & |||J^{\nu+1/2} v|||_{1,m}^{4\nu/2\nu+1} |||v|||_{1,m}^{2/2\nu+1}.\end{aligned}$$

Bilden wir die $(\nu + \frac{1}{2})$-te Potenz dieser Ungleichung, so erhalten wir

$$\begin{aligned}|||J^\nu v|||_{1,m}^{2\nu+1} \leq & |||J^{\nu+1/2} v|||_{1,m}^{2\nu} |||v|||_{1,m} \\ = & |J^\nu v|^{2\nu} |||v|||_{1,m}.\end{aligned}$$

Hieraus folgt

$$|||J^\nu v|||_{1,m} \leq \Big\{ \frac{|J^\nu v|}{|||J^\nu v|||_{1,m}} \Big\}^{2\nu} |||v|||_{1,m} = \rho^\nu |||v|||_{1,m}.$$

□

9.6 Satz: *Sei $v \in X_m$ und $\rho := \rho(v)$. Dann gilt*

$$|||[I - Q_m]v|||_{1,m} \leq \min\{1, c\sqrt{1-\rho}\}|||v|||_{1,m}.$$

Dabei ist c die Konstante aus Satz 9.4.

Beweis: Da $I - Q_m$ eine Projektion ist, ist

$$|||[I - Q_m]v|||_{1,m} \leq |||v|||_{1,m}.$$

Weiter ist mit $v = \Sigma c_\mu \psi_{m,\mu}$

$$\begin{aligned}
|||v|||_{1,m}^2 - |v|^2 =& \sum_{\mu=1}^{n_m} c_\mu^2 \lambda_{m,\mu} - \sum_{\mu=1}^{n_m} c_\mu^2 \lambda_{m,\mu}(1 - \frac{\lambda_{m,\mu}}{\Lambda_m}) \\
=& \sum_{\mu=1}^{n_m} \Lambda_m^{-1} \lambda_{m,\mu}^2 c_\mu^2 \\
=& \Lambda_m^{-1} |||v|||_{2,m}^2.
\end{aligned}$$

Hieraus und aus Satz 9.4 folgt

$$\begin{aligned}
|||[I - Q_m]v|||_{1,m}^2 \leq& c^2 h_m^2 |||v|||_{2,m}^2 \\
=& c^2 h_m^2 \Lambda_m (1-\rho) |||v|||_{1,m}^2.
\end{aligned}$$

Da $\Lambda_m \sim h_m^{-2}$ ist, folgt hieraus die Behauptung. □

Nach diesen Vorbereitungen können wir nun die Konvergenz von Algorithmus 9.2 beweisen.

9.7 Satz: *Bezeichne mit δ_m die Konvergenzrate von Algorithmus 9.2 mit $\nu_1 = \nu_2 = \nu$ auf dem m-ten Gitter gemessen in der $|||.|||_{1,m}$-Norm. Dann gilt mit der Konstanten c aus Satz 9.4*

$$\delta_m \leq \frac{c}{c + 2\nu}.$$

Beweis: Bezeichne mit u_m^* die Lösung der Finite Element Probleme auf dem m-ten Gitter und mit e_m^i den Fehler im i-ten Schritt von Algorithmus 9.2. Dann gilt

$$\begin{aligned}
e_m^{\nu+1} =& e_m^\nu - u_{m-1}^* + u_{m-1}^* - \tilde{u}_{m-1} \\
=& e_m^\nu - u_{m-1}^* + \delta_{m-1} \frac{1}{\delta_{m-1}} (u_{m-1}^* - \tilde{u}_{m-1}) \\
=& [I - Q_m] e_m^\nu + \delta_{m-1} w_{m-1}
\end{aligned}$$

mit $w_{m-1} := \delta_{m-1}^{-1}(u_{m-1}^* - \tilde{u}_{m-1}) \in X_{m-1}$ und

$$|||w_{m-1}|||_{1,m-1} \leq |||u_{m-1}^*|||_{1,m-1}.$$

Wegen der Galerkin Orthogonalität

$$a((I-Q_m)v,w)=0 \quad \forall v\in X_m, w\in X_{m-1}$$

folgt

$$\begin{aligned}
&|||[I-Q_m]e_m^\nu + w_{m-1}|||_{1,m}^2\\
=&|||[I-Q_m]e_m^\nu|||_{1,m}^2 + |||w_{m-1}|||_{1,m}^2\\
=&|||[I-Q_m]e_m^\nu|||_{1,m}^2 + |||w_{m-1}|||_{1,m-1}^2\\
\leq&|||[I-Q_m]e_m^\nu|||_{1,m}^2 + |||u_{m-1}^*|||_{1,m-1}^2\\
=&|||[I-Q_m]e_m^\nu|||_{1,m}^2 + |||u_{m-1}^*|||_{1,m}^2\\
=&|||[I-Q_m]e_m^\nu + \underbrace{u_{m-1}^*}_{=Q_m e_m^\nu}|||_{1,m}^2\\
=&|||e_m^\nu|||_{1,m}^2.
\end{aligned}$$

Sei nun $w\in X_m$ mit $|||w|||_{1,m}=1$ beliebig. Dann gilt

$$\begin{aligned}
&a(e_m^{2\nu+1},w)\\
=&a(J^\nu e_m^{\nu+1},w)\\
=&a(e_m^{\nu+1},J^\nu w)\\
=&a([I-Q_m]e_m^\nu+\delta_{m-1}w_{m-1},J^\nu w)\\
=&(1-\delta_{m-1})a([I-Q_m]e_m^\nu,J^\nu w)\\
&+\delta_{m-1}a([I-Q_m]e_m^\nu+w_{m-1},J^\nu w).
\end{aligned}$$

Da $I-Q_m$ ein Projektor bzgl. des Skalarproduktes $a(.,.)$ ist, folgt für den ersten Summanden

$$\begin{aligned}
&a([I-Q_m]e_m^\nu,J^\nu w)\\
=&a([I-Q_m]e_m^\nu,[I-Q_m]J^\nu w)\\
\leq&|||[I-Q_m]e_m^\nu|||_{1,m}|||[I-Q_m]J^\nu w|||_{1,m}.
\end{aligned}$$

Für den zweiten Summanden gilt

$$\begin{aligned}
&a([I-Q_m]e_m^\nu+w_{m-1},J^\nu w)\\
\leq&|||[I-Q_m]e_m^\nu+w_{m-1}|||_{1,m}|||J^\nu w|||_{1,m}\\
\leq&|||e_m^\nu|||_{1,m}|||J^\nu w|||_{1,m}.
\end{aligned}$$

Aus diesen beiden Abschätzungen folgt mit der Cauchy-Schwarz'schen Ungleichung

$$\begin{aligned}
&a(e_m^{2\nu+1},w)\\
\leq&(1-\delta_{m-1})|||[I-Q_m]e_m^\nu|||_{1,m}|||[I-Q_m]J^\nu w|||_{1,m}\\
&+\delta_{m-1}|||e_m^\nu|||_{1,m}|||J^\nu w|||_{1,m}\\
\leq&\{(1-\delta_{m-1})|||[I-Q_m]e_m^\nu|||_{1,m}^2+\delta_{m-1}|||e_m^\nu|||_{1,m}^2\}^{1/2}\\
&\{(1-\delta_{m-1})|||[I-Q_m]J^\nu w|||_{1,m}^2+\delta_{m-1}|||J^\nu w|||_{1,m}^2\}^{1/2}.
\end{aligned}$$

Wegen der Sätze 9.5 und 9.6 ist

$$\begin{aligned}
&(1-\delta_{m-1})|||[I-Q_m]e_m^\nu|||_{1,m}^2+\delta_{m-1}|||e_m^\nu|||_{1,m}^2\\
=&(1-\delta_{m-1})|||[I-Q_m]J^\nu e_m^0|||_{1,m}^2+\delta_{m-1}|||J^\nu e_m^0|||_{1,m}^2\\
\le&\Big\{(1-\delta_{m-1})\min\{1,c\sqrt{1-\rho(J^\nu e_m^0)}\}^2+\delta_{m-1}\Big\}\rho(J^\nu e_m^0)^{2\nu}|||e_m^0|||_{1,m}^2
\end{aligned}$$

und

$$\begin{aligned}
&(1-\delta_{m-1})|||[I-Q_m]J^\nu w|||_{1,m}^2+\delta_{m-1}|||J^\nu w|||_{1,m}^2\\
\le&\Big\{(1-\delta_{m-1})\min\{1,c\sqrt{1-\rho(J^\nu w)}\}^2+\delta_{m-1}\Big\}\rho(J^\nu w)^{2\nu}|||w|||_{1,m}^2.
\end{aligned}$$

Da

$$|||e^{2\nu+1}|||_{1,m}=\sup_{\substack{w\in X_m\\ |||w|||_{1,m}=1}} a(e^{2\nu+1},w)$$

ist, folgt hieraus

$$|||e^{2\nu+1}|||_{1,m}\le|||e^0|||_{1,m}\max_{0\le\rho\le1}\Big\{\rho^{2\nu}\big[\delta_{m-1}+(1-\delta_{m-1})\min\{1,c(1-\rho)\}\big]\Big\}.$$

Also ist

$$\delta_m\le\max_{0\le\rho\le1}\Big\{\rho^{2\nu}\big[\delta_{m-1}+(1-\delta_{m-1})\min\{1,c(1-\rho)\}\big]\Big\}.$$

Da auf dem gröbsten Gitter exakt gelöst wird, gilt

$$\delta_0=0\le\frac{c}{c+2\nu}.$$

Wir nehmen nun an, die Behauptung sei für $m-1$ gezeigt. Man überlegt sich leicht, daß die Funktion

$$\delta\to\max_{0\le\rho\le1}\Big\{\rho^{2\nu}\big[\delta+(1-\delta)\min\{1,c(1-\rho)\}\big]\Big\}$$

auf $[0,1]$ monoton wachsend ist. Daher folgt aus der Induktionsannahme

$$\begin{aligned}
\delta_m\le&\max_{0\le\rho\le1}\Big\{\rho^{2\nu}\big[\frac{c}{c+2\nu}+\frac{2\nu}{c+2\nu}\min\{1,c(1-\rho)\}\big]\Big\}\\
\le&\max_{0\le\rho\le1}\Big\{\rho^{2\nu}\big[\frac{c}{c+2\nu}+\frac{c2\nu}{c+2\nu}(1-\rho)\big]\Big\}\\
=&\frac{c}{c+2\nu}\max_{0\le\rho\le1}\Big\{\rho^{2\nu}[2\nu+1-2\nu\rho]\Big\}\\
=&\frac{c}{c+2\nu}.
\end{aligned}$$

Denn die Funktion $\rho\to\rho^{2\nu}[2\nu+1-2\nu\rho]$ ist monoton wachsend auf $[0,1]$ und nimmt den Wert 1 im Punkt 1 an. □

9.8 Satz: *Sei $\tilde{\delta}_m$ die Konvergenzrate von Algorithmus 9.2 mit $\nu_1 = \nu, \nu_2 = 0$ auf dem m-ten Gitter gemessen in der $|||.|||_{1,m}$-Norm. Dann gilt mit der Konstanten c aus Satz 9.4*

$$\tilde{\delta}_m \leq \Big[\frac{c}{c+2\nu}\Big]^{1/2}.$$

Beweis: Sei M der Operator $e_m^0 \to e_m^{\nu+1}$ des Algorithmus 9.2 mit $\nu_1 = \nu, \nu_2 = 0$. Dann beschreibt der bzgl. des Skalarproduktes $a(.,.)$ adjungierte Operator M^* die Fehlerfortpflanzung von Algorithmus 9.2 mit $\nu_1 = 0, \nu_2 = \nu$. Insbesondere beschreibt M^*M die Fehlerfortpflanzung von Algorithmus 9.2 mit $\nu_1 = \nu_2 = \nu$. Damit folgt aus Satz 9.7

$$\begin{aligned}
&\|M\|^2_{\mathcal{L}((X_m,|||.|||_{1,m}),(X_m,|||.|||_{1,m}))} \\
=&\|M^*M\|_{\mathcal{L}((X_m,|||.|||_{1,m}),(X_m,|||.|||_{1,m}))} \\
\leq&\frac{c}{c+2\nu}.
\end{aligned}$$

□

Der Beweis von Satz 9.7 beruht wesentlich auf den Annahmen (1) – (3), d.h. daß Ω konvex ist und die Gitter uniform sind. Die Konvexität von Ω wird in Satz 9.4 benötigt, da dort das Dualitätsargument von Aubin-Nitsche benutzt wird, das die H^2-Regularität der Differentialgleichung voraussetzt. Die Uniformität der Gitter wird für die inverse Abschätzung zur Kontrolle des größten Eigenwertes der Steifigkeitsmatrix benötigt. Diese Einschränkungen sind für die Praxis zu restriktiv. Ebenso hat sich gezeigt, daß man mit anderen Glättern als der simplen Jacobi Iteration in Algorithmus 9.2 wesentlich bessere Konvergenzraten erzielen kann.

Die genannten Nachteile können mit einem allgemeineren, abstrakten Zugang, der in den letzten Jahren entwickelt wurde und als **Teilraum-Korrektur-Methode (TRK)** bekannt ist, vermieden werden. Wir wollen diesen Zugang und die entsprechende Konvergenzanalyse im folgenden kurz darstellen. Dazu betrachten wir folgende abstrakte Situation:

- V ist ein endlich dimensionaler Hilbertraum mit Skalarprodukt $(.,.)$.
- $V_1, ..., V_N$ sind Untervektorräume von V mit $V = \sum_{i=1}^N V_i$ (diese Zerlegung ist in der Regel weder direkt noch gar orthogonal).
- $Q_i : V \to V_i$ sind die orthogonalen Projektionen bzgl. $(.,.)$.
- $A : V \to V$ ist ein symmetrischer, positiv definiter Operator.
- $A_i : V_i \to V_i$ ist die Einschränkung von A auf V_i, d.h. $(A_i u, v) = (Au, v) \quad \forall u, v \in V_i$.
- $R_i : V_i \to V_i$ ist eine leicht berechenbare, symmetrisch positiv definite Approximation an A_i^{-1}.

Im Rahmen dieser Vorlesung ist V ein Finite Element Raum, $(.,.)$ ist das L^2-Skalarprodukt oder ein dazu äquivalentes Skalarprodukt wie z.B. $(.,.)_h$ und A wird durch eine symmetrische, koerzive Bilinearform erzeugt.
Zu lösen ist das Problem

$$Au = f.$$

Dies geschieht mit folgendem Algorithmus:

9.9 Algorithmus: (Teilraum-Korrektur-Methode)

0. *Gegeben: Startnäherung* $u_0 \in V$.
1. *Für* $n = 0, 1, ...$ *und* $j = 1, ..., N$ *berechne*

$$u_{n+j/N} := u_{n+(j-1)/N} + R_j Q_j (f - A u_{n+(j-1)/N}).$$

9.10 Beispiel: **(1)** Sei $V = \mathbb{R}^N$ und $V_i = \text{span}\{e_i\}$. Dann ist Algorithmus 9.9 das Gauß-Seidel Verfahren.
(2) Sei $V = X_{h_N}$, $V_i = X_{h_i}$, $R_0 = A_0^{-1}$ und $R_i = \omega_i^{-1} I, i > 0$. Dann ist Algorithmus 9.9 der Mehrgitteralgorithmus 9.2 mit $\nu_1 = 1, \nu_2 = 0$. Durch passende Wahl der R_i kann man auch den Fall $\nu_1 > 1$ und andere Glätter berücksichtigen. □

Für die Konvergenzanalyse von Algorithmus 9.9 setzen wir

$$\begin{aligned} \lambda &:= \min_{1 \le i \le N} \lambda_{\min}(R_i A_i), \\ \Lambda &:= \max_{1 \le i \le N} \lambda_{\max}(R_i A_i), \\ T_i &:= R_i Q_i A. \end{aligned}$$

Aus den Voraussetzungen folgt $0 < \lambda \le \Lambda$. Durch entsprechende Skalierung kann man o.E. erreichen, daß $\Lambda < 2$ ist. Wir bezeichnen mit $|||.|||$ die zu A gehörige Energienorm, d.h.

$$|||u||| := (Au, u)^{1/2} \quad \forall u \in V.$$

Für Finite Element Diskretisierungen der Reaktions-Diffusions Gleichung ist insbesondere $|||.|||$ zur H^1-Norm äquivalent.

9.11 Satz: *Es gebe zwei Konstanten* K_0 *und* K_1 *mit*

$$\Big\{ \sum_{i=1}^N |||v_i|||^2 \Big\}^{1/2} \le K_0 |||v||| \qquad \forall v = \sum_{i=1}^N v_i \in V$$

und

$$\sum_{1 \le i,j \le N} (Av_i, w_j) \le K_1 \Big\{ \sum_{i=1}^N |||v_i|||^2 \Big\}^{1/2} \Big\{ \sum_{j=1}^N |||w_j|||^2 \Big\}^{1/2} \quad \forall v_i \in V_i, w_j \in V_j.$$

Weiter sei $\Lambda < 2$. Dann ist die Konvergenzrate von Algorithmus 9.9 gemessen in der Norm $|||.|||$ kleiner oder gleich

$$\Big[1 - \big(\frac{2}{\Lambda} - 1\big)\big(\frac{\lambda}{\Lambda K_0 K_1}\big)^2\Big]^{1/2}.$$

Beweis: Sei u die exakte Lösung von $Au = f$. Aus Algorithmus 9.9 folgt

$$u - u_{n+j/N} = (I - T_j)(u - u_{n+(j-1)/N})$$

und somit

$$u - u_{n+1} = (I - T_N)...(I - T_1)(u - u_n).$$

Setze zur Abkürzung

$$\begin{aligned} E_0 &:= I \\ E_j &:= (I - T_j)...(I - T_1) \quad , 1 \le j \le N. \end{aligned}$$

Dann müssen wir zeigen, daß für alle $v \in V$ gilt

$$|||E_N v||| \le \Big[1 - \big(\frac{2}{\Lambda} - 1\big)\big(\frac{\lambda}{\Lambda K_0 K_1}\big)^2\Big]^{1/2} |||v|||. \tag{2}$$

Aus der Definition der E_j folgt sofort

$$-E_j + E_{j-1} = T_j E_{j-1} \qquad \forall 1 \le j \le N.$$

Bezeichne mit $P_i : V_i \to V_i$ die orthogonale Projektion bzgl. des Skalarproduktes $(A.,.)$. Dann folgt

$$Q_i A = A_i P_i$$

und damit

$$\begin{aligned} (R_i A_i)^{-1} T_i =& (R_i A_i)^{-1} R_i Q_i A \\ =& (R_i A_i)^{-1} R_i A_i P_i = P_i, \end{aligned}$$

d.h. $(R_i A_i)^{-1} T_i$ ist die orthogonale Projektion von V auf V_i bzgl. $(A.,.)$. Damit folgt für beliebiges $v \in V$

$$\begin{aligned} & |||E_{j-1} v|||^2 - |||E_j v|||^2 \\ =& |||T_j E_{j-1} v + E_j v|||^2 - |||E_j v|||^2 \\ =& (AT_j E_{j-1} v, T_j E_{j-1} v) + 2(AT_j E_{j-1} v, E_j v) \\ =& (AT_j E_{j-1} v, T_j E_{j-1} v) + 2(AT_j E_{j-1} v, (I - T_j) E_{j-1} v) \\ =& (AT_j E_{j-1} v, (2I - T_j) E_{j-1} v) \\ \ge& (2 - \Lambda)(AT_j E_{j-1} v, E_{j-1} v). \end{aligned}$$

Summieren wir diese Ungleichung von $j = 1$ bis N, so erhalten wir

$$\begin{aligned} &|||v|||^2 - |||E_N v|||^2 \\ =&\sum_{j=1}^{N} \Big\{ |||E_{j-1}v|||^2 - |||E_j v|||^2 \Big\} \\ \geq&(2-\Lambda) \sum_{j=1}^{N} (AT_j E_{j-1}v, E_{j-1}v). \end{aligned} \tag{3}$$

Schreibe $v = \sum_{i=1}^{N} v_i$. Da $(R_i A_i)^{-1} T_i$ die orthogonale Projektion bzgl. $(A.,.)$ ist, folgt aus der Chauchy-Schwarz'schen Ungleichung und der ersten Voraussetzung des Satzes 9.11

$$\begin{aligned} |||v|||^2 =&(Av, v) \\ =&\sum_{i=1}^{N} (Av, v_i) \\ =&\sum_{i=1}^{N} (A(R_i A_i)^{-1} T_i v_i v_i) \\ \leq&\lambda^{-1} \sum_{i=1}^{N} (AT_i v, v_i) \\ \leq&\lambda^{-1} \sum_{i=1}^{N} |||T_i v||| |||v_i||| \\ \leq&\lambda^{-1} \Big\{ \sum_{i=1}^{N} |||T_i v|||^2 \Big\}^{1/2} \Big\{ \sum_{i=1}^{N} |||v_i|||^2 \Big\}^{1/2} \\ \leq&\lambda^{-1} K_0 |||v||| \Big\{ \sum_{i=1}^{N} |||T_i v|||^2 \Big\}^{1/2}. \end{aligned}$$

Also ist

$$|||v||| \leq \lambda^{-1} K_0 \Big\{ \sum_{i=1}^{N} |||T_i v|||^2 \Big\}^{1/2}. \tag{4}$$

Aus der zweiten Voraussetzung von Satz 9.11 folgt

$$\begin{aligned} \sum_{i=1}^{N} |||T_i v|||^2 =&\sum_{i=1}^{N} (AT_i v, T_i v) \\ =&\sum_{i=1}^{N} \Big\{ \sum_{j=1}^{i-1} (AT_i v, T_i (E_{j-1} - E_j)v) + (AT_i v, T_i E_{i-1} v) \Big\} \\ =&\sum_{i=1}^{N} \Big\{ \sum_{j=1}^{i-1} (AT_i v, T_i T_j E_{j-1} v) + (AT_i v, T_i E_{i-1} v) \Big\} \end{aligned}$$

$$
\begin{aligned}
&\leq \Lambda \sum_{i=1}^{N} \sum_{j=1}^{i-1} (AT_i v, T_j E_{j-1} v) \\
&\leq \Lambda K_1 \Big\{ \sum_{i=1}^{N} |||T_i v|||^2 \Big\}^{1/2} \Big\{ \sum_{j=1}^{N} |||T_j E_{j-1} v|||^2 \Big\}^{1/2}
\end{aligned}
$$

und somit

$$
\begin{aligned}
\sum_{i=1}^{N} |||T_i v|||^2 &\leq \Lambda^2 K_1^2 \sum_{j=1}^{N} |||T_j E_{j-1} v|||^2 \\
&= \Lambda^2 K_1^2 \sum_{j=1}^{N} (AT_j E_{j-1} v, T_j E_{j-1} v) \qquad (5) \\
&\leq \Lambda^3 K_1^2 \sum_{j=1}^{N} (AT_j E_{j-1} v, E_{j-1} v).
\end{aligned}
$$

Aus den Abschätzungen (3) – (5) erhalten wir

$$
\begin{aligned}
&|||v|||^2 - |||E_N v|||^2 \\
\geq &(2-\Lambda) \sum_{i=1}^{N} (AT_j E_{j-1} v, E_{j-1} v) \\
\geq &(2-\Lambda) \Lambda^{-3} K_1^{-2} \sum_{i=1}^{N} |||T_i v|||^2 \\
\geq &(2-\Lambda) \Lambda^{-3} \lambda^2 K_0^{-2} K_1^{-2} |||v|||^2.
\end{aligned}
$$

Hieraus folgt offensichtlich (2) und damit die Behauptung des Satzes. □

9.12 Bemerkung: **(1)** Wegen der Cauchy-Schwarz'schen Ungleichung ist die zweite Voraussetzung von Satz 9.11 stets mit $K_1 \leq N$ erfüllt. Für Mehrgitterverfahren bedeutet die Abschätzung $K_1 \leq N$, daß die Konvergenzrate von Algorithmus 9.9 sich schlimmstenfalls wie $1/\ln|h_N|$ verhält. Häufig kann man jedoch eine **verschärfte Cauchy-Schwarz'sche Ungleichung**

$$
(Av_i, w_j) \leq \gamma^{|i-j|} |||v_i||| |||w_j|||
$$

mit $\gamma \in (0,1)$ beweisen. Dann ist $K_1 \leq \frac{1}{1-\gamma}$.

(2) Da nach Voraussetzung $V = \sum_{i=1}^{N} V_i$ ist, ist die Abbildung

$$
V_1 \times ... \times V_N \ni (v_1, ..., v_N) \to \sum_{i=1}^{N} v_i \in V
$$

linear, stetig und surjektiv. Aus dem Satz von der offenen Abbildung (§ II.5 Yosida: Functional Analysis) folgt daher, daß die erste Voraussetzung von Satz 9.11 ebenfalls stets erfüllt. Die Schwierigkeit besteht darin, eine explizite Abschätzung von K_0, die möglichst nicht von N abhängt, zu finden. Mit Hilfe allgemeiner Sätze der Approximationstheorie und tiefliegender Charakterisierungen der Sobolev Räume ist dies im Rahmen von Mehrgitterverfahren für Finite Element Diskretisierungen in der Tat möglich. □

10. A posteriori Fehlerschätzung und adaptive Gitterverfeinerung

In den §§ 6 – 7 haben wir sog. **a priori Fehlerabschätzungen** gezeigt, d.h. wir haben den Fehler der Finite Element Diskretisierung abgeschätzt, ohne dazu das entsprechende diskrete Problem zu lösen. Die Fehlerabschätzungen sind alle **asymptotisch**, d.h. sie sagen etwas über die Konvergenzgeschwindigkeit des Fehlers aus, wenn die Elementgrößen gegen Null streben. Für ein gegebenes Problem und eine gegebene Unterteilung sind sie aber unbrauchbar, da sie u.a. von Normen der unbekannten Lösung der Differentialgleichung abhängen. Für die Praxis stellt sich aber natürlich die Frage nach dem tatsächlichen Fehler der berechneten Finite Element Lösung. Zudem will man i.a. eine bestimmte Genauigkeit mit minimalem Aufwand, d.h. möglichst wenigen Elementen erreichen. Diese Fragen werden durch **a posteriori Fehlerabschätzungen** und **adaptive Gitterverfeinerungstechniken** gelöst. Hiermit wollen wir uns in diesem Abschnitt beschäftigen.

Um die wesentlichen Prinzipien herauszuarbeiten und um technische Schwierigkeiten zu vermeiden, beschränken wir uns auf die Reaktions-Diffusions Gleichung in $\Omega \subset \mathbb{R}^2$ mit homogenen Dirichlet Randbedingungen und Dreieckselemente der Ordnung $k \geq 1$, d.h. $\mathcal{T}_h$ besteht aus Dreiecken und $X_h = S^{k,0}_{h,0}$.

Zunächst müssen wir einige zusätzliche Notationen einführen. Die Menge aller Dreieckskanten, die im Innern von Ω liegen, bezeichnen wir mit $\mathcal{E}_h$. Jedes E ist dann die gemeinsame Kante von genau zwei Dreiecken, die wir mit K_{E1} und K_{E2} bezeichnen. Jedem $E \in \mathcal{E}_h$ ordnen wir einen Einheitsvektor n_E, der senkrecht auf E steht, zu und bezeichnen für stückweise stetige Funktionen φ mit $[\varphi]_E$ den Sprung von φ über E in Richtung n_E, d.h.

$$[\varphi]_E(x) = \lim_{t \to 0+} \varphi(x + t n_E) - \lim_{t \to 0+} \varphi(x - t n_E) \quad \forall x \in E.$$

Der Sprung $[\varphi]_E$ hängt von der Orientierung von n_E ab. Größen der Form $[n_E \nabla \varphi]_E$ sind aber von der Orientierung von n_E unabhängig. Für $E \in \mathcal{E}_h$ bezeichnet h_E die Länge von E. Wegen der Regularitätsbedingung der §§ 6 – 7 können Größen der Form $h_K/h_{K'}$ und h_K/h_E mit $K \cap K' \neq \emptyset$ und $E \cap K \neq \emptyset$ nach oben und unten

durch die Konstante $c_{\mathcal{T}}$ abgeschätzt werden. Schließlich benötigen wir verschiedene Umgebungen von Punkten $z \in \mathcal{N}_h$, Kanten $E \in \mathcal{E}_h$ und Dreiecken $K \in \mathcal{T}_h$:

$$\begin{aligned}
\omega_z &:= \mathrm{supp} v_z = \bigcup_{z \in K'} K', \\
\omega_K &:= \bigcup_{K \cap K' \in \mathcal{E}_h} K', \\
\tilde{\omega}_K &:= \bigcup_{K \cap K' \neq \emptyset} K', \\
\omega_E &:= \bigcup_{E \subset \partial K'} K', \\
\tilde{\omega}_E &:= \bigcup_{E \cap K' \neq \emptyset} K'.
\end{aligned}$$

Dabei ist $v_z \in S^{1,0}_{h,0}$ die nodale Basisfunktion zu z, und $K \cap K' \in \mathcal{E}_h$ bedeutet, daß K und K' eine Kante gemeinsam haben.

Im folgenden bezeichnen $u \in H^1_0(\Omega)$ und $u_h \in X_h$ die schwache Lösung der Reaktions-Diffusions Gleichung und ihre Finite Element Approximation. Wegen des Existenzsatzes 3.3 und der Friedrich'schen Ungleichung 1.15 gibt es eine Konstante $\beta > 0$ mit

$$\beta \|u - u_h\|_1^2 \leq a(u - u_h, u - u_h).$$

β hängt von Ω und den Koeffizienten A und α des Differentialoperators ab. Aus dieser Abschätzung folgt insbesondere **Stabilität**, d.h.

$$\|u - u_h\|_1 \leq \frac{1}{\beta} \sup_{\substack{v \in H^1_0(\Omega) \\ \|v\|_1 = 1}} a(u - u_h, v). \tag{1}$$

Da $X_h \subset H^1_0(\Omega)$ ist, haben wir **Galerkin-Orthogonalität** des Fehlers, d.h.

$$a(u - u_h, v_h) = 0 \quad \forall v_h \in X_h. \tag{2}$$

Die Abbildung $v \to a(u - u_h, v) = l(v) - a(u_h, v)$ definiert ein stetiges lineares Funktional auf $H^1_0(\Omega)$. Dies ist das **Residuum**. Als nächstes wollen wir eine Darstellung des Residuums angeben, die praktisch handhabbar ist. Sei dazu $v \in H^1_0(\Omega)$ mit $\|v\|_1 = 1$ beliebig, aber fest. Anwenden des Gauß'schen Integralsatzes auf jedem Element $K \in \mathcal{T}_h$ liefert

$$\begin{aligned}
& a(u - u_h, v) \\
=& l(v) - a(u_h, v) \\
=& \int_\Omega fv - \int_\Omega \left\{ \nabla u_h^T A \nabla v + \alpha u_h v \right\} \\
=& \sum_{K \in \mathcal{T}_h} \left\{ \int_K fv - \int_K \nabla u_h^T A \nabla v - \int_K \alpha u_h v \right\}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{K\in\mathcal{T}_h} \Big\{ \int_K fv + \int_K \nabla\cdot(A\nabla u_h)v - \int_{\partial K} n_K\cdot A\nabla u_h v - \int_K \alpha u_h v \Big\} \\
&= \sum_{K\in\mathcal{T}_h} \int_K \big\{ f + \nabla\cdot(A\nabla u_h) - \alpha u_h \big\} v - \sum_{E\in\mathcal{E}_h} \int_E [n_E\cdot A\nabla u_h]_E v.
\end{aligned} \tag{3}
$$

Dabei bezeichnet n_K die äußere Normale zu K. Wegen der Galerkin Orthogonalität (2) können wir auf der rechten Seite von (3) eine beliebiges Element $v_h \in X_h$ von v subtrahieren. Aus der Cauchy-Schwarz'schen Ungleichung für Integrale folgt dann

$$
\begin{aligned}
a(u-u_h,v) \le & \sum_{K\in\mathcal{T}_h} \|f + \nabla\cdot(A\nabla u_h) - \alpha u_h\|_{L^2(K)} \|v-v_h\|_{L^2(K)} \\
& + \sum_{E\in\mathcal{E}_h} \|[n_E\cdot A\nabla u_h]_E\|_{L^2(E)} \|v-v_h\|_{L^2(E)}.
\end{aligned} \tag{4}
$$

Als nächstes müssen wir $v_h \in X_h$ geschickt wählen, so daß die Normen $\|v-v_h\|_{L^2(K)}$ und $\|v-v_h\|_{L^2(E)}$ möglichst klein werden. Wegen der Ergebnisse der §§ 6,7 sind wir versucht, $v_h = I_h v$ zu wählen. Dies ist aber nicht möglich, da v nur aus $H_0^1(\Omega)$ ist und damit $I_h v$ gar nicht definiert ist. Statt dessen konstruieren wir einen sog. **Quasi-Interpolationsoperator** $R_h : H_0^1(\Omega) \to S_{h,0}^{1,0}$ wie folgt. Für jedes $z \in \mathcal{N}_h$ bezeichnen wir mit $\pi_z : L^2(\omega_z) \to \mathbb{R}$ die orthogonale Projektion, d.h.

$$
\pi_z\varphi := \frac{1}{|\omega_z|} \int_{\omega_z} \varphi.
$$

Dabei ist $|\omega_z|$ die Fläche von ω_z. Dann ist

$$
R_h\varphi := \sum_{z\in\mathcal{N}_{h,\Omega}} (\pi_z\varphi) v_z. \tag{5}
$$

Man beachte, daß $R_h\varphi$ für alle $\varphi \in L^1(\Omega)$ definiert ist und daß nur über die Dreieckseckpunkte im Innern von Ω summiert wird.

10.1 Satz: *Es gibt zwei Konstanten c_1, c_2 die nur von der Konstanten $c_{\mathcal{T}}$ in der Regularitätsbedingung an $\mathcal{T}_h$ abhängen, so daß für alle $v \in H_0^1(\Omega)$, alle Dreiecke $K \in \mathcal{T}_h$ und alle Kanten $E \in \mathcal{E}_h$ gilt*

$$
\begin{aligned}
\|v - R_h v\|_{L^2(K)} &\le c_1 h_K |v|_{H^1(\tilde\omega_K)}, \\
\|v - R_h v\|_{L^2(E)} &\le c_2 h_E^{1/2} |v|_{H^1(\tilde\omega_E)}.
\end{aligned}
$$

Beweis: **1. Schritt:** Aus der Poincaré'schen Ungleichung 1.21 folgt für jedes $z \in \mathcal{N}_h$ und jedes $\varphi \in H^1(\omega_z)$

$$
\|\varphi - \pi_z\varphi\|_{L^2(\omega_z)} \le \operatorname{diam}(\omega_z) |\varphi|_{H^1(\omega_z)}.
$$

Wegen der Regularitätsannahme an $\mathcal{T}_h$ gibt es eine Konstante c' mit

$$\operatorname{diam}(\omega_z) \leq c' h_K$$

für jedes Dreieck K, das z als Eckpunkt hat.

2. Schritt: Bezeichne mit $\hat{E}$ die horizontale Kante des Referenz-Dreiecks $\hat{K}$. Wegen des Spursatzes 1.12 gibt es eine Konstante $\hat{c}$, so daß für alle $\varphi \in H^1(\hat{K})$ gilt

$$\|\varphi\|_{L^2(\hat{E})} \leq \hat{c} \|\varphi\|_{H^1(\hat{K})}.$$

Seien nun $K \in \mathcal{T}_h$ ein beliebiges Dreieck, E eine Kante von K und $\varphi \in H^1(K)$. Wähle die affine Transformation $F_K : \hat{K} \to K$ so, daß $\hat{E}$ auf E abgebildet wird. Setze $\hat{\varphi} := \varphi \circ F_K \in H^1(\hat{K})$. Wegen der Regularitätsannahme an $\mathcal{T}_h$ gilt

$$h_E^{1/2} |\det DF_K|^{-1/2} \leq c h_K^{-1/2} \quad , \quad h_E^{1/2} |\det DF_K|^{-1/2} \|\!| DF_K^{-1} |\!\| \leq c h_K^{1/2}.$$

Damit folgt durch Transformation auf $\hat{K}$

$$\begin{aligned}
\|\varphi\|_{L^2(E)} =& h_E^{1/2} \|\hat{\varphi}\|_{L^2(\hat{E})} \\
\leq& \hat{c} h_E^{1/2} \|\hat{\varphi}\|_{H^1(\hat{K})} \\
=& \hat{c} h_E^{1/2} \Big\{ \|\hat{\varphi}\|^2_{L^2(\hat{K})} + |\hat{\varphi}|^2_{H^1(\hat{K})} \Big\}^{1/2} \\
\leq& \hat{c} h_E^{1/2} \Big\{ |\det DF_K|^{-1} \|\varphi\|^2_{L^2(\hat{K})} \\
& \qquad + |\det DF_K|^{-1} \|\!| DF_K^{-1} |\!\|^2 |\varphi|^2_{H^1(K)} \Big\}^{1/2} \\
\leq& \hat{c} c \Big\{ h_K^{-1/2} \|\varphi\|_{L^2(K)} + h_K^{1/2} |\varphi|_{H^1(K)} \Big\}.
\end{aligned}$$

3. Schritt: Sei $K \in \mathcal{T}_h$ ein Dreieck, das keinen Eckpunkt auf dem Rand Γ hat. Bezeichne die Menge der Eckpunkte von K mit $\mathcal{N}(K)$. Dann ist

$$\sum_{z \in \mathcal{N}(K)} v_z = 1 \qquad \text{auf} \quad K.$$

Damit folgt aus Schritt 1

$$\begin{aligned}
\|v - R_h v\|_{L^2(K)} =& \| \sum_{z \in \mathcal{N}(K)} v_z (v - \pi_z v) \|_{L^2(K)} \\
\leq& \sum_{z \in \mathcal{N}(K)} \|v_z (v - \pi_z v)\|_{L^2(K)} \\
\leq& \sum_{z \in \mathcal{N}(K)} \|v - \pi_z v\|_{L^2(K)} \\
\leq& \sum_{z \in \mathcal{N}(K)} \|v - \pi_z v\|_{L^2(\omega_z)} \\
\leq& \sum_{z \in \mathcal{N}(K)} c h_K |v|_{H^1(\omega_z)} \\
\leq& c' h_K |v|_{H^1(\tilde{\omega}_K)}.
\end{aligned}$$

4. Schritt: Betrachte nun ein Dreieck K, das mindestens einen Eckpunkt auf dem Rand Γ hat. Dann gilt auf K

$$\begin{aligned} v - R_h v &= \sum_{z \in \mathcal{N}(K)} v_z v - \sum_{z \in \mathcal{N}(K) \cap \mathcal{N}_{h,\Omega}} v_z \pi_z v \\ &= \sum_{z \in \mathcal{N}(K)} v_z (v - \pi_z v) + \sum_{z \in \mathcal{N}(K) \setminus \mathcal{N}_{h,\Omega}} v_z \pi_z v \end{aligned}$$

und somit

$$\begin{aligned} &\|v - R_h v\|_{L^2(K)} \\ \leq &\sum_{z \in \mathcal{N}(K)} \|v_z (v - \pi_z v)\|_{L^2(K)} + \sum_{z \in \mathcal{N}(K) \setminus \mathcal{N}_{h,\Omega}} \|v_z \pi_z v\|_{L^2(K)}. \end{aligned}$$

Der erste Summand wurde bereits in Schritt 3 abgeschätzt. Sei also $z \in \mathcal{N}(K) \setminus \mathcal{N}_{h,\Omega}$ ein Eckpunkt von K, der auf Γ liegt. Dann ist

$$\|v_z \pi_z v\|_{L^2(K)} \leq c h_K |\pi_z v|.$$

Da $z \in \Gamma$ ist, gibt es ein Dreieck $K' \in \mathcal{T}_h$ und eine Kante E' von K', so daß z ein Endpunkt von E' und $E' \subset \Gamma$ ist. Da u auf E' verschwindet, folgt mit Schritt 2

$$\begin{aligned} |\pi_z v| =& h_{E'}^{-1/2} \|\pi_z v\|_{L^2(E')} \\ =& h_{E'}^{-1/2} \|v - \pi_z v\|_{L^2(E')} \\ \leq& \hat{c} c \Big\{ h_{E'}^{-1/2} h_{K'}^{-1/2} \|v - \pi_z v\|_{L^2(K')} + h_{E'}^{-1/2} h_{K'}^{1/2} |v - \pi_z v|_{H^1(K')} \Big\} \\ \leq& \hat{c} c' \Big\{ h_K^{-1} \|v - \pi_z v\|_{L^2(K')} + |v|_{H^1(K')} \Big\}. \end{aligned}$$

Dabei haben wir ausgenutzt, daß $|v - \pi_z v|_{H^1(K')} = |v|_{H^1(K')}$ ist, da $\pi_z v$ konstant ist. Aus diesen Abschätzungen folgt die Behauptung für K.

5. Schritt: Sei $E \in \mathcal{E}_h$ eine Kante, die keinen Eckpunkt auf dem Rand Γ hat. Bezeichne mit $\mathcal{N}(E)$ die Menge der Eckpunkte von E. Dann ist

$$\sum_{z \in \mathcal{N}(E)} v_z = 1 \quad \text{auf } E.$$

Hieraus und aus Schritt 1 und 2 folgt

$$\begin{aligned} \|v - R_h v\|_{L^2(E)} =& \| \sum_{z \in \mathcal{N}(E)} v_z (v - \pi_z v) \|_{L^2(E)} \\ \leq& \sum_{z \in \mathcal{N}(E)} \|v_z (v - \pi_z v)\|_{L^2(E)} \\ \leq& \sum_{z \in \mathcal{N}(E)} \|v - \pi_z v\|_{L^2(E)} \end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{z\in\mathcal{N}(E)} \hat{c}c\Big\{h_{K_E}^{-1/2}\|v-\pi_z v\|_{L^2(K_E)} \\
&\qquad\qquad + h_{K_E}^{1/2}|v-\pi_z v|_{H^1(K_E)}\Big\} \\
&\leq \sum_{z\in\mathcal{N}(E)} \hat{c}ch_{K_E}^{1/2}|v|_{H^1(\omega_z)} \\
&\leq \hat{c}c' h_E^{1/2}|v|_{H^1(\tilde{\omega}_E)}.
\end{aligned}
$$

Dabei bezeichnet K_E ein Dreieck, das E als Kante hat.

6. Schritt: Betrachte nun eine Kante E, die einen Endpunkt auf dem Rand Γ hat. Dann ist auf E

$$
v - R_h v = \sum_{z\in\mathcal{N}(E)} v_z(v-\pi_z v) + \sum_{z\in\mathcal{N}(E)\setminus\mathcal{N}_{h,\Omega}} v_z \pi_z v.
$$

Der erste Summand wird wie in Schritt 5 abgeschätzt. Der zweite Summand wird mit den gleichen Methoden wie in Schritt 4 behandelt. Hieraus folgt dann die Behauptung für E. □

Wir greifen nun die Abschätzung (4) wieder auf. Wir setzen $v_h = R_h v$, benutzen Satz 10.1 und wenden die Cauchy-Schwarz'sche Ungleichung für endliche Summen an:

$$
\begin{aligned}
&a(u-u_h, v) \\
\leq& c_1 \sum_{K\in\mathcal{T}_h} h_K \|f + \nabla\cdot(A\nabla u_h) - \alpha u_h\|_{L^2(K)} |v|_{H^1(\tilde{\omega}_K)} \\
&+ c_2 \sum_{E\in\mathcal{E}_h} h_E^{1/2} \|[n_E \cdot A\nabla u_h]_E\|_{L^2(E)} |v|_{H^1(\tilde{\omega}_E)} \\
\leq& c_1 \Big\{\sum_{K\in\mathcal{T}_h} h_K^2 \|f + \nabla\cdot(A\nabla u_h) - \alpha u_h\|_{L^2(K)}^2\Big\}^{1/2} \Big\{\sum_{K\in\mathcal{T}_h} |v|_{H^1(\tilde{\omega}_K)}^2\Big\}^{1/2} \\
&+ c_2 \Big\{\sum_{E\in\mathcal{E}_h} h_E \|[n_E\cdot A\nabla u_h]_E\|_{L^2(E)}^2\Big\}^{1/2} \Big\{\sum_{E\in\mathcal{E}_h} |v|_{H^1(\tilde{\omega}_E)}^2\Big\}^{1/2} \\
\leq& c'|v|_1 \Big\{\sum_{K\in\mathcal{T}_h} h_K^2 \|f + \nabla\cdot(A\nabla u_h) - \alpha u_h\|_{L^2(K)}^2 \\
&\qquad + \sum_{E\in\mathcal{E}_h} h_E \|[n_E\cdot A\nabla u_h]_E\|_{L^2(E)}^2\Big\}^{1/2}.
\end{aligned}
$$

Dabei haben wir im letzten Schritt die Regularitätsbedingung an $\mathcal{T}_h$ ausgenutzt. Hieraus und aus (1) folgt insgesamt

$$
\|u - u_h\|_1 \leq c\eta \tag{6}
$$

mit

$$\begin{aligned}\eta :=& \Big\{ \sum_{K\in\mathcal{T}_h} \eta_K^2 \Big\}^{1/2}, \\ \eta_K :=& \Big\{ h_K^2 \|f + \nabla\cdot(A\nabla u_h) - \alpha u_h\|^2_{L^2(K)} \\ &+ \frac{1}{2} \sum_{E\subset\partial K\setminus\Gamma} h_E \|[n_E\cdot A\nabla u_h]_E\|^2_{L^2(E)} \Big\}^{1/2}. \end{aligned} \tag{7}$$

Der Faktor $\frac{1}{2}$ vor der zweiten Summe in η_K berücksichtigt, daß bei Summation über alle Dreiecke jede innere Kante doppelt gezählt wird.

Ungleichung (6) ist eine **a posteriori Fehlerabschätzung**. Die Größe η kann aus den gegebenen Daten f, A, α und der berechneten numerischen Lösung u_h a posteriori berechnet werden. Sie heißt daher auch **a posteriori Fehlerschätzer**. Ungleichung (6) zeigt, daß der Fehlerschätzer **zuverlässig** ist, d.h. ist $\eta \le \varepsilon$, so ist der Fehler ebenfalls (bis auf einen Faktor) nicht größer als ε. Die Kontrolle von η erlaubt also, eine vorgegebene Toleranz zu erreichen. Um dies mit einem minimalen Aufwand zu erreichen, reicht die obere Schranke (6) nicht aus. Wir müssen zusätzlich garantieren, daß η den Fehler nicht überschätzt und die räumliche Verteilung des Fehlers auch richtig widerspiegelt. Dies nennt man **Effizienz**. Sie ist gegeben, wenn es gelingt, den Fehler auch nach unten durch η abzuschätzen.

Um dies zu erreichen, benötigen wir einige zusätzliche Notationen. Bezeichne mit f_h irgendeine, im folgenden feste Finite Element Approximation an f, z.B. die L^2-Projektion auf die stückweise konstanten Funktionen $S_h^{0,-1}$. Für ein gegebenes Dreieck K bezeichne mit z_1, z_2, z_3 seine Eckpunkte und setze

$$\psi_K := 27 v_{z_1} v_{z_2} v_{z_3} \quad \text{auf } K.$$

Wie man leicht nachprüft, hat ψ_K folgende Eigenschaften

$$\begin{aligned} &\psi_K \in \mathbb{P}_3, \\ &\psi_K \ge 0 \quad \text{auf } K, \\ &\psi_K = 0 \quad \text{auf } \partial K, \\ &\max_{x\in K} \psi_K(x) = 1. \end{aligned}$$

Insbesondere kann also ψ_K durch Null zu einer Funktion aus $H_0^1(\Omega)$ forgesetzt werden. Für eine Kante $E \in \mathcal{E}_h$ numerieren wir die Eckpunkte der angrenzenden Dreiecke K_{E1} und K_{E2} so, daß die Endpunkte von E zuerst numeriert werden. Bezeichnen $v_{z_1,i}, v_{z_2,i}, v_{z_3,i}$ die nodalen Basisfunktionen zu $K_{Ei}, i = 1, 2$, so definieren wir

$$\psi_E := 4 v_{z_1,i} v_{z_2,i} \quad \text{auf } K_{Ei}, i = 1, 2.$$

Offensichtlich hat ψ_E folgende Eigenschaften

$$\begin{aligned}
&\psi_E \in C(\omega_E)\\
&\psi_{E|K_{Ei}} \in \mathbb{P}_2, i = 1, 2,\\
&\psi_E \geq 0 \quad \text{auf } \omega_E,\\
&\psi_E = 0 \quad \text{auf } \partial\omega_E,\\
&\max_{x\in E} \psi_E(x) = 1.
\end{aligned}$$

Für eine Dreieckskante $E \in \mathcal{E}_h$ bezeichnen wir mit $\mathbb{P}_k(E)$ die Polynome von Grad $\leq k$ in einer Variablen auf E. Jedes $\varphi \in \mathbb{P}_k(E)$ kann in kanonischer Weise zu einem Polynom vom Grad $\leq k$ in zwei Variablen auf $\mathbb{R}^2$ fortgesetzt werden. Diese Fortsetzung bezeichnen wir wieder mit φ.

10.2 Satz: *Für jedes Dreieck $K \in \mathcal{T}_h$, jede Kante $E \in \mathcal{E}_h$, jedes $v \in \mathbb{P}_k$ und jedes $\sigma \in \mathbb{P}_k(E)$ gilt*

$$\begin{aligned}
&c_1\|v\|_{L^2(K)} \leq \Big\{\int_K \psi_K v^2\Big\}^{1/2} \leq \|v\|_{L^2(K)},\\
&c_2 h_K^{-1}\|\psi_K v\|_{L^2(K)} \leq |\psi_K v|_{H^1(K)} \leq c_3 h_K^{-1}\|\psi_K v\|_{L^2(K)},\\
&c_4\|\sigma\|_{L^2(E)} \leq \Big\{\int_E \psi_E \sigma^2\Big\}^{1/2} \leq \|\sigma\|_{L^2(E)},\\
&c_5 h_E^{-1}\|\psi_E\sigma\|_{L^2(\omega_E)} \leq |\psi_E\sigma|_{H^1(\omega_E)} \leq c_6 h_E^{-1}\|\psi_E\sigma\|_{L^2(\omega_E)},\\
&\|\psi_E\sigma\|_{L^2(\omega_E)} \leq c_7 h_E^{1/2}\|\sigma\|_{L^2(E)}.
\end{aligned}$$

Die Konstanten $c_1, ..., c_7$ hängen nur von k und der Größe $c_{\mathcal{T}}$ in der Regularitätsannahme an $\mathcal{T}_h$ ab.

Beweis: Die obere Schranke in der ersten Abschätzung folgt aus der Cauchy-Schwarz'schen Ungleichung. Wie man leicht nachrechnet, definiert $\int_{\hat{K}} \psi_K \circ F_K w^2$ eine Norm auf $\mathbb{P}_k$. Da auf endlich dimensionalen Räumen alle Normen äquivalent sind, gibt es eine Konstante $\hat{c}$ mit

$$\hat{c}\|w\|_{L^2(\hat{K})} \leq \Big\{\int_{\hat{K}} \psi_K \circ F_K w^2\Big\}^{1/2} \quad \forall w \in \mathbb{P}_k.$$

Anwenden des Transformationssatzes liefert

$$\begin{aligned}
\hat{c}\|v\|_{L^2(K)} =& \hat{c}|\det DF_K|^{1/2}\|v \circ F_K\|_{L^2(\hat{K})}\\
\leq& |\det DF_K|^{1/2}\Big\{\int_{\hat{K}} \psi_K \circ F_K (v \circ F_K)^2\Big\}^{1/2}\\
=& \Big\{\int_K \psi_K v^2\Big\}^{1/2}
\end{aligned}$$

und beweist die untere Schranke der ersten Abschätzung.
Da $\psi_K \circ F_K$ in den Eckpunkten von $\hat{K}$ verschwindet, definiert $|\psi_K \circ F_K w|_{H^1(\hat{K})}$ eine Norm auf $\mathbb{P}_k$. Mithin gibt es zwei Konstanten $\hat{c}_1$ und $\hat{c}_2$ mit

$$\begin{aligned}\hat{c}_1\|\psi_K \circ F_K w\|_{L^2(\hat{K})} \leq & |\psi_K \circ F_K w|_{H^1(\hat{K})} \\ \leq & \hat{c}_2|\psi_K \circ F_K w|_{L^2(\hat{K})} \quad \forall w \in \mathbb{P}_k.\end{aligned}$$

Hieraus und aus dem Transformationssatz folgt wieder die Behauptung.
Die Abschätzungen für σ folgen mit den gleichen Argumenten wie diejenigen für v. □

Sei nun $K \in \mathcal{T}_h$ beliebig. Setze

$$w_K := \psi_K(f_h + \nabla \cdot (A\nabla u_h) - \alpha u_h).$$

Wegen Satz 10.2 ist

$$c_1^2\|f_h + \nabla \cdot (A\nabla u_h) - \alpha u_h\|^2_{L^2(K)} \leq \int_K w_K(f_h + \nabla \cdot (A\nabla u_h) - \alpha u_h).$$

Setzen wir w_K als Testfunktion v in (3) ein und berücksichtigen, daß w_K auf ∂K und außerhalb von K verschwindet, so erhalten wir

$$\begin{aligned}&\int_K w_K(f_h + \nabla \cdot (A\nabla u_h) - \alpha u_h) \\ =&\int_K w_K(f + \nabla \cdot (A\nabla u_h) - \alpha u_h) + \int_K w_K(f_h - f) \\ =&a(u - u_h, w_K) + \int_K w_K(f_h - f) \\ \leq&\|a\|_{\mathcal{L}^2}\|u - u_h\|_{H^1(K)}\|w_K\|_{H^1(K)} + \|f_h - f\|_{L^2(K)}\|w_K\|_{L^2(K)}.\end{aligned}$$

Offensichtlich ist

$$\|w_K\|_{L^2(K)} \leq \|f_h + \nabla \cdot (A\nabla u_h) - \alpha u_h\|_{L^2(K)}.$$

Aus Satz 10.2 folgt weiter

$$\begin{aligned}\|w_K\|_{H^1(K)} =&\Big\{\|w_K\|^2_{L^2(K)} + |w_K|^2_{H^1(K)}\Big\}^{1/2} \\ \leq&\Big\{1 + c_3^2 h_K^{-2}\Big\}^{1/2}\|w_K\|_{L^2(K)} \\ \leq&ch_K^{-1}\|f_h + \nabla \cdot (A\nabla u_h) - \alpha u_h\|_{L^2(K)}.\end{aligned}$$

Aus diesen Abschätzungen und der Dreiecksungleichung ergibt sich

$$h_K\|f + \nabla \cdot (A\nabla u_h) - \alpha u_h\|_{L^2(K)} \leq c\Big\{\|u - u_h\|_{H^1(K)} + h_K\|f - f_h\|_{L^2(K)}\Big\}. \quad (8)$$

Sei nun $E \in \mathcal{E}_h$ eine Kante. Setze

$$w_E := \psi_E [n_E \cdot A\nabla u_h]_E.$$

Wegen Satz 10.2 ist

$$c_4^2 \|[n_E \cdot A\nabla u_h]_E\|^2_{L^2(E)} \leq \int_E w_E [n_E \cdot A\nabla u_h]_E.$$

Setzen wir w_E als Testfunktion v in (3) ein und berücksichtigen, daß w_E auf $\partial\omega_E$ und außerhalb ω_E verschwindet, so folgt

$$\begin{aligned}
&\int_E w_E [n_E \cdot A\nabla u_h]_E \\
=&\int_{\omega_E} (f + \nabla \cdot (A\nabla u_h) - \alpha u_h) w_E - a(u - u_h, w_E) \\
\leq&\|f + \nabla \cdot (A\nabla u_h) - \alpha u_h\|_{L^2(\omega_E)} \|w_E\|_{L^2(\omega_E)} + \|a\|_{\mathcal{L}^2} \|u - u_h\|_{H^1(\omega_E)} \|w_E\|_{H^1(\omega_E)}
\end{aligned}$$

Wegen Satz 10.2 ist

$$\begin{aligned}
\|w_E\|_{L^2(\omega_E)} \leq& c_7 h_E^{1/2} \|w_E\|_{L^2(E)} \\
\leq& c_7 h_E^{1/2} \|[n_E \cdot A\nabla u_h]_E\|_{L^2(E)}
\end{aligned}$$

und

$$\begin{aligned}
\|w_E\|_{H^1(\omega_E)} \leq& c h_E^{-1} \|w_E\|_{L^2(\omega_E)} \\
\leq& c' h_E^{-1/2} \|[n_E \cdot A\nabla u_h]_E\|_{L^2(E)}.
\end{aligned}$$

Aus diesen Abschätzungen und (8) ergibt sich insgesamt

$$h_E^{1/2} \|[n_E \cdot A\nabla u_h]_E\|_{L^2(E)} \leq c\big\{\|u - u_h\|_{H^1(\omega_E)} + h_E \|f - f_h\|_{L^2(\omega_E)}\big\}. \tag{9}$$

Aus den Abschätzungen (8), (9) und der Definition (7) erhalten wir folgende **lokale untere Fehlerschranke**

$$\eta_K \leq c\big\{\|u - u_h\|_{H^1(\omega_K)} + h_K \|f - f_h\|_{L^2(\omega_K)}\big\}. \tag{10}$$

Summation über alle Dreiecke liefert zudem die **globale untere Fehlerschranke**

$$\eta \leq c\Big\{\|u - u_h\|_1 + \Big\{\sum_{K \in \mathcal{T}_h} h_K^2 \|f - f_h\|^2_{L^2(K)}\Big\}^{1/2}\Big\}. \tag{11}$$

In beiden Abschätzungen sind die $f - f_h$-Terme Störterme höherer Ordnung, die zudem a priori allein aus der Kenntnis der Daten kontrolliert werden können, ohne eine Differentialgleichung oder Diskretisierung zu lösen.

10.3 Bemerkung: **(1)** Die Abschätzungen (6), (10) und (11) zeigen, daß der Fehlerschätzer η zuverlässig und effizient ist. Es ist nicht verwunderlich, daß wir nur in einer Richtung lokale Schranken erhalten. Denn die Abschätzung (6) benötigt den inversen Differentialoperator, der ein globaler Operator ist. Die Abschätzung (10) dagegen benutzt nur den Differentialoperator, der ein lokaler Operator ist.
(2) Die Größe $f + \nabla \cdot (A\nabla u_h) - \alpha u_h$ ist das Residuum der Finite Element Approximation u_h bzgl. der starken Form der Differentialgleichung. Die Größe $[n_E \cdot \nabla A u_h]_E$ ist der Sprung des Spuroperators, der auf kanonische Weise die starke und schwache Form der Differentialgleichung verknüpft.
(3) Ähnliche Ergebnisse gelten für die Konvektions-Diffusions Gleichung. A priori hängen die Konstanten in den Fehlerabschätzungen von der Peclet-Zahl ab und wachsen für abnehmende Diffusion an. Dieser Effekt kann durch eine verbesserte Analyse weitgehend vermieden werden.
(4) Es gibt auch andere Fehlerschätzer, die z.B. auf der Lösung lokaler, diskreter Dirichlet- oder Neumann-Probleme beruhen. Mit technischem Mehraufwand lassen sich für diese Schätzer zu (6), (10) und (11) analoge Abschätzungen beweisen. □

Wir wenden uns nun dem Problem der adaptiven Gitterverfeinerung zu. Zunächst könnte man versuchen, den geschätzten Fehler η über alle Unterteilungen $\mathcal{T}_h$ mit einer gegebenen Elementzahl zu minimieren. Dies ist jedoch ein hochgradig nichtlineares, extrem aufwendiges Optimierungsproblem. Einfache heuristische Argumente zeigen andererseits, daß bei einer optimalen Triangulierung alle Elemente etwa den gleichen Beitrag zu η liefern. Dies legt es nahe, Elemente, die einen zu großen Beitrag η_K liefern, zu unterteilen, und führt auf folgenden Algorithmus.

10.4 Algorithmus: (Adaptive Gitterverfeinerung)

0. *Bestimme eine grobe Triangulierung $\mathcal{T}_{h_0}$ von Ω. Setze $k := 0$.*
1. *Löse das diskrete Problem zur Triangulierung $\mathcal{T}_{h_k}$.*
2. *Berechne $\eta_K, K \in \mathcal{T}_{h_k}$, und $\eta_k := \max_{K \in \mathcal{T}_{h_k}} \eta_K$.*
3. *Falls $\eta_k \leq \varepsilon$ ist,* STOP. *Sonst gehe zu* 4.
4. *Verfeinere alle Dreiecke $K \in \mathcal{T}_{h_k}$ mit $\eta_K \geq \gamma \eta_k$. Verfeinere evtl. weitere Dreiecke, um eine zulässige Triangulierung $\mathcal{T}_{h_{k+1}}$ zu erhalten. Erhöhe k um 1 und gehe nach 1 zurück.*

In Algorithmus 10.4 ist γ ein zu wählender Parameter mit $0 < \gamma < 1$. Ist $\gamma \sim 0$, werden viele Elemente verfeinert; ist $\gamma \sim 1$, werden nur sehr wenige Elemente unterteilt. In der Praxis wählt man häufig $\gamma = 0.5$. Algorithmus 10.4 kann auch ggf. durch eine Vergröberungsstrategie ergänzt werden.

Für die praktische Realisierung von Algorithmus 10.4 müssen wir noch beschreiben, wie Dreiecke verfeinert werden und wie die Zulässigkeit der verfeinerten Triangulierung gesichert werden kann. Dabei müssen wir beachten, daß alle Triangulierun-

gen die Regularitätsbedingung erfüllen sollen, d.h. die Dreieckswinkel sollen nicht zu klein oder zu groß werden. Dazu führen wir folgende Sprechweise ein:

- Ein Dreieck wird **rot** unterteilt, wenn seine Kantelmittelpunkte miteinander verbunden werden.
- Ein Dreieck wird **blau** unterteilt, wenn der Mittelpunkt der längsten Kante mit dem gegenüberliegenden Eckpunkt und dem Mittelpunkt einer weiteren Kante verbunden wird.
- Ein Dreieck wird **grün** unterteilt, wenn der Mittelpunkt der längsten Kante mit dem gegenüberliegenden Eckpunkt verbunden wird.
- Ein Dreieck hat einen (oder mehrere) **hängenden Knoten**, wenn K nicht unterteilt wurde, aber eines (oder mehrere) der angrenzenden Dreiecke unterteilt wurde.

Dabei bedeutet "angrenzend", daß die betreffenden Dreiecke eine Kante gemeinsam haben.
Eine rote Unterteilung erzeugt offensichtlich ähnliche Dreiecke und verändert somit die Winkel nicht. Die Vorgabe, primär die längste Kante zu unterteilen, sichert bei der grünen und blauen Unterteilung, daß der kleinste Winkel nicht verkleinert wird. Offensichtlich ist eine Triangulierung genau dann zulässig, wenn kein Dreieck hängende Knoten hat.
Schritt 4 von Algorithmus 10.4 wird nun gemäß folgender Regeln durchgeführt:

1. Ist $\eta_K \geq \gamma\eta_k$, unterteile K rot.
2. Hat K drei hängende Knoten, unterteile K rot.
3. Hat K zwei hängende Knoten, von denen keiner auf der längsten Kante liegt, unterteile K rot.
4. Hat K zwei hängende Knoten, von denen einer auf der längsten Kante liegt, unterteile K blau.
5. Hat K einen hängenden Knoten, unterteile K blau, wenn der hängende Knoten nicht auf der längsten Kante liegt, sonst unterteile K grün.

Man kann zeigen, daß dieser Algorithmus in endlich vielen Schritten eine verfeinerte Triangulierung erzeugt, die den oben diskutierten Vorgaben genügt.

11. Implementierung

In diesem Abschnitt gehen wir kurz auf die Implementierung der Methoden der letzten Paragraphen und die dafür benötigten Datenstrukturen ein. Um die Notationen und den technischen Aufwand möglichst gering zu halten, beschränken wir uns dabei in der Regel auf lineare Dreieckselemente.

Sei also $\mathcal{T}_h$ eine Triangulierung des polygonalen Gebietes $\Omega \subset \mathbb{R}^2$. Bezeichne mit NT die Zahl aller Dreiecke und mit NV die Zahl aller Dreieckseckpunkte. Die Dreieckseckpunkte, d.h. die Punkte in $\mathcal{N}_h$ (alllgemeiner in $\mathcal{G}_h$) werden beliebig numeriert. Die Koordinaten werden in zwei Feldern $VX(NV)$ und $VY(NV)$ gespeichert. In jedem Dreieck K werden die Eckpunkte lokal im mathematisch positiven Sinne durchnumeriert. Der Zusammenhang zwischen lokaler und globaler Numerierung wird durch ein Feld $ITNODE(3, NT)$ hergestellt. Dabei gilt

$$ITNODE(i, K) = l \iff \text{Knoten } i \text{ von Dreieck } K \text{ hat die globale Nummer } l.$$

Jedem Knoten wird eine Eigenschaft durch das Feld $IVERT(NV)$ zugeordnet. Dabei ist

$$IVERT(i) = \begin{cases} 0 & \text{Knoten } i \text{ liegt in } \Omega, \\ j > 0 & \text{Knoten } i \text{ liegt auf dem } j\text{-ten Dirichlet Randstück}, \\ j < 0 & \text{Knoten } i \text{ liegt auf dem } j\text{-ten Neumann Randstück}. \end{cases}$$

Weiter müssen wir die Nachbarschaftsbeziehungen zwischen den Dreiecken kontrollieren. Dazu vereinbaren wir, daß die j-te Kante stets dem j-ten Eckpunkt gegenüberliegt, und definieren das Feld $ITEDGE(3, NT)$ durch

$$ITEDGE(i, K) = j > 0 \iff \text{Kante } i \text{ von Dreieck } K \text{ grenzt an Dreieck } j.$$

Für Kanten auf dem Rand von Γ kann $ITEDGE$ genutzt werden, um den Typ des Randes, d.h. gerades oder j-tes gekrümmtes Randstück, zu beschreiben.
Wenn wir ein Mehrgitterverfahren wie in § 9 zur numerischen Lösung benutzen oder eine adaptive Gitterverfeinerung wie in § 10 durchführen, müssen wir auch mehrere Gitterhierarchien verwalten. Dies geschieht durch das Feld $ITMARK(3, NTT)$. Dabei ist NTT die Summe der Zahl aller Dreiecke auf allen Niveaus. Weiter ist

$$ITMARK(1, K) = j \iff \text{Dreieck } K \text{ ist durch Unterteilung des Dreiecks } j \text{ entstanden (eine Hierarchiestufe zurückgehen),}$$

$$ITMARK(2, K) = \begin{cases} 4 & \text{Dreieck } K \text{ wird rot unterteilt,} \\ 0 & \text{Dreieck } K \text{ wird nicht unterteilt,} \\ j & \text{Dreieck } K \text{ wird grün unterteilt,} \\ & \text{indem die } j\text{-te Kante halbiert wird,} \end{cases}$$

(eine Hierarchiestufe vorrücken)

$$ITMARK(3, K) = j \iff \text{Dreieck } j \text{ ist ein Nachkomme von Dreieck } K.$$

Dabei gilt für die Numerierung der Nachkommen folgende Konvention:

- bei roter Unterteilung erhält das Zentrumsdreieck, dessen Eckpunkte die Kantenmittelpunkte von K sind, die Nummer $ITMARK(3,K)$ und das Dreieck, das den i-ten Eckpunkt von K erbt, die Nummer $ITMARK(3,K)+i$,
- bei grüner Unterteilung erhält das Dreieck, das die Knoten $ITMARK(2,K)$ und $ITMARK(2,K)+1$ von K erbt, die Nummer $ITMARK(3,K)$ und das andere Dreieck die Nummer $ITMARK(3,K)+1$.

Hierbei sind Ausdrücke der Form $i+j$ stets modulo 3 zu verstehen, d.h. $4 \leftrightarrow 1$, $5 \leftrightarrow 2$ usw. Ganz analog kann man auch blaue Unterteilungen codieren.

Die gröbste Triangulierung kann man auf zwei Weisen erzeugen. Bei der ersten Möglichkeit gibt man die Größen $ITNODE$, $IVERT$, VX und VY für die Dreiecke und Eckpunkte des gröbsten Gitters von Hand ein. Zusätzlich gibt man die Werte von $ITEDGE$ für die Randkanten ein. Die Nachbarschaftsbeziehungen für die inneren Kanten lassen sich aus $ITNODE$ errechnen. Nach Eingabe der gröbsten Triangulierung erzeugt man weitere Triangulierungen durch gleichmäßige oder adaptive Verfeinerungen.
Bei der zweiten Variante erzeugt man die gröbste Triangulierung automatisch aus einer orientierten Liste von Eckpunkten auf dem Rand Γ. Für diese Punkte muß man VX, VY und $IVERT$ vorgeben. Weitere Gitterpunkte und die Triangulierung werden aufgrund einer der folgenden beiden Strategien, die auch kombiniert werden können, erzeugt:

(a) Schlage eine Ecke ab, d.h. wähle ein i, verbinde die Punkte $i-1$ und $i+1$ der Liste und erzeuge so ein Dreieck, dessen Eckpunkte die Punkte $i-1$, i und $i+1$ der Liste sind.
(b) Verbinde alle Punkte der Liste mit ihrem Schwerpunkt.

Beide Varianten sind natürlich nur zulässig, wenn die so erzeugten Dreiecke in Ω liegen. Zusätzlich sollte man die Qualität der erzeugten Dreiecke kontrollieren, um zu spitze oder zu stumpfe Dreiecke zu vermeiden (wegen der Regularitätsbedingung der §§ 6, 7). Ein Maß für die Qualität eines Dreieckes ist das Verhältnis der Summe der Quadrate seiner Kantenlängen zu seiner Fläche.

Für das Aufstellen des linearen Gleichungssystems muß man die Größen

$$b_i := l(v_i) \quad \text{und} \quad a_{ij} := a(v_i, v_j)$$

berechnen. Dabei bezeichnet v_i die nodale Basisfunktion zum i-ten Knoten. Bezeichne dazu für $K \in \mathcal{T}_h$ mit l_K bzw. a_K die Einschränkung von l bzw. a auf das Element K, d.h. die Integrale werden nur bzgl. K genommen. Dann erfolgt die Berechnung der rechten Seite nach folgendem Schema:

11.1 Algorithmus: (Berechnung der rechten Seite b)
$b := 0$
for all triangles K do
for all vertices I of K do

$$i := ITNODE(I, K)$$
$$b_i := b_i + l_K(v_i)$$

enddo
enddo.

Dabei wird $l_K(v_i)$ mit einer Quadraturformel berechnet. Für die Formel aus Beispiel 8.1 (2) ergibt sich z.B.

$$l_K(v_i) = \frac{1}{12} D \Big\{ f\big(\frac{1}{2}(z_I + z_{I+1})\big) + f\big(\frac{1}{2}(z_I + z_{I+2})\big) \Big\}.$$

Dabei sind z_1, z_2, z_3 die Eckpunkte von K und

$$D = \det(z_2 - z_1, z_3 - z_1).$$

Außerdem sind die Indizes $I + 1$ und $I + 2$ wie zuvor beschrieben modulo 3 zu verstehen.

Ganz analog berechnet man die Steifigkeitsmatrix gemäß folgendem Schema:

11.1 Algorithmus: (Aufstellen der Steifigkeitsmatrix a)
$a := 0$
for all triangles K do
for all vertices I of K do

$$i := ITNODE(I, K)$$
$$j := ITNODE(I + 1, K)$$
$$a_{ii} := a_{ii} + a_K(v_i, v_i)$$

if $j > i$ then

$$a_{ij} := a_{ij} + a_K(v_i, v_j)$$

endif
enddo
enddo.

Dabei werden die Größen $a_K(v_i, v_i)$ und $a_K(v_i, v_j)$ wieder mittels numerischer Integration berechnet. Zusätzlich sind jetzt aber zwei Punkte zu beachten.

Erstens sind die Ableitungen der Basisfunktion v_i stückweise konstant. Bezeichne mit z_1, z_2, z_3 die Eckpunkte von K und wähle die Transformation $F_K : \hat{K} \to K$ so, daß der Nullpunkt auf z_I abgebildet wird. Dann ist

$$B_K := DF_K = (z_{I+1} - z_I, z_{I+2} - z_I)$$

und

$$\begin{pmatrix} -1 \\ -1 \end{pmatrix} = \nabla_{\hat{K}}(v_i \circ F_K) = B_K^T \nabla_K v_i.$$

Hieraus folgt sofort

$$\nabla_K v_i = \frac{1}{D} \begin{pmatrix} z_{I+2,1} - z_{I,1} - (z_{I+2,2} - z_{I,2}) \\ z_{I+1,2} - z_{I,2} - (z_{I+1,1} - z_{I,1}) \end{pmatrix}$$

mit

$$D := \det B_K = \det(z_2 - z_1, z_3 - z_1).$$

Eine analoge Formel gilt natürlich für $\nabla_K v_j$.
Zweitens ist die Steifigkeitsmatrix dünn besetzt. Das Element i, j ist höchstens dann von Null verschieden, wenn die entsprechenden Eckpunkte auf einer gemeinsamen Dreieckskante liegen. Diese dünne Besetzung muß in Algorithmus 11.2 unbedingt beachtet werden. Wenn man das diskrete Problem mit einem iterativen Verfahren wie dem CG- oder MG-Verfahren löst, benötigt man nur die von Null verschiedenen Einträge rechts von der Diagonalen und die Diagonalelemente. In diesem Fall speichert man diese Elemente hintereinander in einem Feld $STIFF(NN)$ ab. Zusätzlich benötigt man zwei Felder $IC(NN)$ und $ID(NV)$. Dabei gibt $IC(k)$ den Spaltenindex des k-ten Eintrages von $STIFF$ an. $ID(i)$ gibt die Position des i-ten Diagonalelementes der Steifigkeitsmatrix in $STIFF$ an. Insbesondere stehen die Elemente der i-ten Zeile in den Positionen $ID(i)$ bis $ID(i+1) - 1$ mit $ID(NV+1) := NN + 1$. Die Felder ID und IC müssen zusätzlich in Algorithmus 11.2 belegt werden. Den benötigten Speicherplatz, d.h. NN, kann man leicht a priori abschätzen. Bezeichnet NE die Zahl aller Dreieckskanten, so ist $NN \leq NV + NE$. Wegen der Euler'schen Polyederformel ist $NE = NT + NV - 1$.

12. Nichtkonforme Methoden

Bisher haben wir stets konforme Finite Element Diskretisierungen betrachtet, bei denen der diskrete Raum X_h im unendlich dimensionalen Raum X des Variationsproblems enthalten ist. Nun wollen wir sog. **nicht-konforme Methoden** betrachten, bei denen diese Inklusion nicht mehr gilt.

Zunächst betrachten wir wie in § 2 eine abstrakte Situation. Sei dazu $(X, \|.\|_X)$ ein Banach Raum, $l \in \mathcal{L}(X, \mathbb{R})$ und $a \in \mathcal{L}^2(X, \mathbb{R})$ symmetrisch und koerziv. Zu lösen ist das Variationsproblem

$$a(u, v) = l(v) \quad \forall v \in X. \tag{1}$$

Weiter seien $(X_h, \|.\|_{X_h})$ ein endlich dimensionaler Banach Raum, $l_h \in \mathcal{L}(X_h, \mathbb{R})$ und $a_h \in \mathcal{L}^2(X_h, \mathbb{R})$ symmetrisch und koerziv. Die Normen von l_h und a_h sowie die Koerzivitätskonstante α_h von a_h seien gleichmäßig in h nach oben bzw. – für α_h – nach unten weg von 0 beschränkt. Die diskrete Approximation von (1) lautet

$$a_h(u_h, v_h) = l_h(v_h) \quad \forall v_h \in X_h. \tag{2}$$

Wegen Satz 2.1 haben (1) und (2) jeweils eine eindeutige Lösung u bzw. u_h. (Man beachte, daß hierfür die gleichmäßige Beschränktheit von $\|l_h\|_{\mathcal{L}}, \|a_h\|_{\mathcal{L}^2}$ und α_h nicht benötigt wird.) Da wir an der nicht-konformen Situation interessiert sind, ist $X_h \not\subset X$. Allerdings setzen wir voraus, daß a_h, l_h und $\|.\|_{X_h}$ auch für Elemente von X einen Sinn machen und für solche Elemente mit a, l bzw. $\|.\|_X$ übereinstimmen. Der folgende Satz ist unter diesen Annahmen das nicht-konforme Analogon zum Céa-Lemma (erster Teil von Satz 2.2)

12.1 Satz: (Zweites Strang Lemma) *Unter den obigen Voraussetzungen gilt die Fehlerabschätzung*

$$\|u - u_h\|_{X_h} \leq \left(1 + \frac{A_h}{\alpha_h}\right) \inf_{v_h \in X_h} \|u - v_h\|_{X_h} + \frac{1}{\alpha_h} \sup_{\substack{w_h \in X_h \\ \|w_h\|_{X_h}=1}} |a_h(u, w_h) - l_h(w_h)|.$$

Dabei ist $A_h := \|a_h\|_{\mathcal{L}^2}$ und α_h die Koerzivitätskonstante von a_h.

Beweis: Sei $v_h \in X_h$ beliebig und $w_h := u_h - v_h$. Aus der Koerzivität von a_h folgt dann

$$\begin{aligned}
\alpha_h \|w_h\|_{X_h}^2 \leq & a_h(w_h, w_h) \\
= & a_h(u_h - v_h, w_h) \\
= & a_h(u - v_h, w_h) + a_h(u_h, w_h) - a_h(u, w_h) \\
= & a_h(u - v_h, w_h) + l_h(w_h) - a_h(u, w_h) \\
\leq & A_h \|u - v_h\|_{X_h} \|w_h\|_{X_h} \\
& + |l_h(w_h) - a_h(u, w_h)|.
\end{aligned}$$

Wegen

$$\|u - u_h\|_{X_h} \leq \|u - v_h\|_{X_h} + \|w_h\|_{X_h}$$

folgt hieraus die Behauptung. □

Der folgende Satz ist die nicht-konforme Variante des Dualitätsargumentes von Aubin-Nitsche (zweiter Teil von Satz 2.2).

12.2 Satz: *Die Bezeichnungen und Voraussetzungen seien wie in Satz 12.1. Zusätzlich sei H ein Hilbertraum mit Skalarprodukt $(.,.)_H$ und Norm $\|.\|_H$, so daß $X \hookrightarrow H$ dicht und $X_h \subset H$ ist. Für jedes $\varphi \in H$ seien $u_\varphi \in X$ die eindeutige Lösung von*

$$a(v, u_\varphi) = (\varphi, v)_H \quad \forall v \in X$$

und $u_{\varphi,h} \in X_h$ die eindeutige Lösung von

$$a_h(v_h, u_{\varphi,h}) = (\varphi, v_h)_H \quad \forall v_h \in X_h.$$

Dann gilt die Fehlerabschätzung

$$\begin{aligned}\|u - u_h\|_H \leq \sup_{\substack{\varphi \in H \\ \|\varphi\|_H = 1}} \Big\{ & A_h \|u - u_h\|_{X_h} \|u_\varphi - u_{\varphi,h}\|_{X_h} \\ & + |a_h(u - u_h, u_\varphi) - (\varphi, u - u_h)_H| \\ & + |a_h(u, u_\varphi - u_{\varphi,h}) - l_h(u_\varphi - u_{\varphi,h})| \Big\}.\end{aligned}$$

Beweis: Da $X \hookrightarrow H$ dicht und $X_h \subset H$ ist, ist

$$\|u - u_h\|_H = \sup_{\substack{\varphi \in H \\ \|\varphi\|_H = 1}} (\varphi, u - u_h)_H. \tag{3}$$

Sei nun $\varphi \in H$ mit $\|\varphi\|_H = 1$ beliebig. Dann folgt

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

Für den zweiten und dritten Summanden dieser Gleichung folgt

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

und

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

Also ist

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

Hieraus und aus (3) folgt die Behauptung. □

Wir wenden diese abstrakten Ergebnisse auf ein einfaches, aber typisches Modellproblem an: der **Crouzeix-Raviart Diskretisierung** der ebenen Poisson Gleichung mit homogenen Dirichlet Randbedingungen. Dabei ist $\Omega \subset \mathbb{R}^2$ ein beschränktes, zusammenhängendes Gebiet mit stückweise geradem Rand Γ, $X = H_0^1(\Omega)$, $H = L^2(\Omega)$, $l(v) = \int_\Omega fv$ und $a(u,v) = \int_\Omega \nabla u^T \nabla v$.
Wie in § 6 bezeichnet $\mathcal{T}_h = \{K_i : 1 \le i \le m_h\}$ eine zulässige und reguläre Unterteilung von Ω in Dreiecke. Wir bezeichnen die Dreieckskanten in $\mathcal{T}_h$ mit $\mathcal{E}_h$ und die inneren Kanten mit $\mathcal{E}_{h,\Omega}$. Für $E \in \mathcal{E}_h$ ist m_E der Kantenmittelpunkt. Da jede Ebene durch drei nicht kolineare Punkte eindeutig bestimmt ist, ist für jedes $K \in \mathcal{T}_h$ jedes lineare Polynom $p \in \mathbb{P}_1$ eindeutig bestimmt durch seine Werte in den Kantenmittelpunkten von K. Daher ist folgende Definition sinnvoll

$$\begin{aligned}X_h := \Big\{\varphi \in L^2(\Omega) : \varphi|_K \in \mathbb{P}_1 \quad &\forall K \in \mathcal{T}_h, \varphi \text{ ist stetig in} \\ &m_E, E \in \mathcal{E}_{h,\Omega}, \varphi(m_E) = 0 \quad \forall E \in \mathcal{E}_h \backslash \mathcal{E}_{h,\Omega}\Big\}.\end{aligned}$$

X_h heißt der Finite Element Raum von **Crouzeix-Raviart**.
Jedes $u_h \in X_h$ ist eindeutig bestimmt durch seine Werte in den Kantenmittelpunkten $m_E, E \in \mathcal{E}_{h,\Omega}$. Bezeichnen z_1, z_2, z_3 die Eckpunkte von K und $v_{z_1}, v_{z_2}, v_{z_3}$ die zugehörigen nodalen Basisfunktionen aus § 6, so folgt mit einer leichten Rechnung

$$u_{h|K} = \sum_{i=1}^{3} u_h(m_i) w_i,$$

wobei m_i der Mittelpunkt der i-ten Kante von K und

$$w_i = v_{z_{i+1}} + v_{z_{i+2}} - v_{z_i}$$

ist. (Dabei liegt wie in § 11 die i-te Kante vereinbarungsgemäß dem i-ten Eckpunkt gegenüber.) Die Funktionen in X_h sind nicht stetig und verschwinden nicht identisch auf Γ. Daher ist $X_h \not\subset X$. Es ist aber $X_h \subset H = L^2(\Omega)$. Außerdem ist offensichtlich $S_{h,0}^{1,0} \subset X_h$. Die Norm $\|.\|_{X_h}$, die Linearform l_h und die Bilinearform a_h werden definiert durch

$$\|u_h\|_{X_h} := \Big\{ \sum_{K \in \mathcal{T}_h} |u_h|^2_{H^1(K)} \Big\}^{1/2},$$

$$l_h(u_h) := \sum_{K \in \mathcal{T}_h} \int_K f u_h,$$

$$a_h(u_h, v_h) := \sum_{K \in \mathcal{T}_h} \int_K \nabla u_h^T \nabla v_h.$$

Offensichtlich sind die Voraussetzungen der Sätze 12.1 und 12.2 erfüllt. Insbesondere ist $A_h = \alpha_h = 1$.

12.3 Satz: *Sei $u \in H_0^1(\Omega) \cap H^2(\Omega)$ die Lösung von (1) und*

$$L_u(w) := a_h(u,w) - l_h(w) \quad \forall w \in X \oplus X_h.$$

Dann gilt für alle $w \in X \oplus X_h$

$$|L_u(w)| \le ch|u|_2 \|w\|_{X_h}.$$

Die Konstante c hängt nur von der Konstanten $c_{\mathcal{T}}$ der Regularitätsbedingung an $\mathcal{T}_h$ ab.

Beweis: Sei $w \in X \oplus X_h$ beliebig. Durch elementweise partielle Integration folgt mit den gleichen Notationen wie in § 10

$$\begin{aligned}
L_u(w) &= \sum_{K\in\mathcal{T}_h} \int_K \nabla u^T \nabla w - \sum_{K\in\mathcal{T}_h} \underbrace{\int_K fw}_{=\int_K -\Delta uw} \\
&= \sum_{K\in\mathcal{T}_h} \int_K \nabla u^T \nabla w + \sum_{K\in\mathcal{T}_h} \underbrace{\int_K \Delta uw}_{=-\int_K \nabla u^T\nabla w + \int_{\partial K} \partial_{n_K} uw} \\
&= \sum_{K\in\mathcal{T}_h} \int_{\partial K} \partial_{n_K} uw \\
&= \sum_{E\in\mathcal{E}_{h,\Omega}} \int_E [\partial_{n_E} uw]_E + \sum_{E\in\mathcal{E}_h\setminus\mathcal{E}_{h,\Omega}} \int_E \partial_n uw.
\end{aligned}$$

Da $u \in H^2(\Omega)$ ist, gilt für jede innere Kante $E \in \mathcal{E}_{h,\Omega}$

$$\int_E [\partial_{n_E} u]_E = 0.$$

Ist dagegen $E \in \mathcal{E}_h \setminus \mathcal{E}_{h,\Omega}$ eine Randkante, so ist entweder $w = 0$ auf E, falls $w \in X$ ist, oder

$$\int_E w = h_E w(m_E) = 0,$$

falls $w \in X_h$ ist. Definieren wir daher für jede Kante $E \in \mathcal{E}_h$ den Mittelwert von w auf E durch

$$\overline{w}_E := h_E^{-1} \int_E w,$$

so folgt

$$L_u(w) = \sum_{E\in\mathcal{E}_{h,\Omega}} \int_E [\partial_{n_E} u(w - \overline{w}_E)]_E + \sum_{E\in\mathcal{E}_h\setminus\mathcal{E}_{h,\Omega}} \int_E \partial_n u(w - \overline{w}_E).$$

Bezeichne mit $I_h : X \cap H^2(\Omega) \to S^{1,0}_{h,0} \subset X_h$ den nodalen Interpolationsoperator aus § 6. Da für jede Kante E

$$\int_E (w - \overline{w}_E) = 0$$

und $\partial_{n_E}(I_h u)_{|E}$ konstant ist, folgt

$$\begin{aligned} L_u(w) = &\sum_{E\in\mathcal{E}_{h,\Omega}} \int_E [\partial_{n_E}(u - I_h u)(w - \overline{w}_E)]_E \\ &+ \sum_{E\in\mathcal{E}_h\setminus\mathcal{E}_{h,\Omega}} \int_E \partial_n(u - I_h u)(w - \overline{w}_E)]_E. \end{aligned} \tag{4}$$

Wegen $|n| = |n_E| = 1$ folgt hieraus mit der Cauchy-Schwarz'schen Ungleichung

$$|L_u(w)| \le \sum_{E\in\mathcal{E}_h} \|\nabla(u - I_h u)\|_{L^2(E)} \|w - \overline{w}_E\|_{L^2(E)}.$$

Betrachte nun eine beliebige Kante $E \in \mathcal{E}_h$ und ein Dreieck $K \in \mathcal{T}_h$, das E als Kante hat. Bezeichne mit $\hat{E}$ die horizontale Kathete des Referenz-Dreieckes $\hat{K}$ und mit $F_K : \hat{K} \to K$ eine affine Transformation von $\hat{K}$ auf K, die $\hat{E}$ auf E abbildet. Setze $\hat{w} := w \circ F_K$. Aus dem Transformationssatz folgt

$$\overline{w}_E = h_E^{-1} \int_E w = \int_{\hat{E}} \hat{w}.$$

Also ist

$$\int_{\hat{E}} (\hat{w} - \overline{w}_E) = 0.$$

Wie im Beweis der Poincaré'schen Ungleichung 1.21 folgt, daß es eine Konstante $\hat{c}$ gibt mit

$$\|\varphi\|_{L^2(\hat{K})} \le \hat{c} |\varphi|_{H^1(\hat{K})}$$

für alle $\varphi \in H^1(\hat{K})$ mit $\int_{\hat{E}} \varphi = 0$. Hieraus folgt mit dem Transformationssatz und dem Spursatz 1.12

$$\begin{aligned} \|w - \overline{w}_E\|_{L^2(E)} &= h_E^{1/2} \|\hat{w} - \overline{w}_E\|_{L^2(\hat{E})} \\ &\le h_E^{1/2} c \Big\{ \|\hat{w} - \overline{w}_E\|^2_{L^2(\hat{K})} + |\hat{w} - \overline{w}_E|^2_{H^1(\hat{K})} \Big\}^{1/2} \\ &\le h_E^{1/2} c (1 + \hat{c}^2)^{1/2} |\hat{w} - \overline{w}_E|_{H^1(\hat{K})} \\ &= h_E^{1/2} c (1 + \hat{c}^2)^{1/2} |\hat{w}|_{H^1(\hat{K})} \\ &\le h_E^{1/2} c (1 + \hat{c}^2)^{1/2} |\det DF_K|^{-1/2} \|\!|DF_K|\!\| |w|_{H^1(K)} \\ &\le c' h_E^{1/2} |w|_{H^1(K)}. \end{aligned} \tag{5}$$

Hierbei haben wir im letzten Schritt ausgenutzt, daß wegen Satz 6.2 und der Regularität von $\mathcal{T}_h$ der Ausdruck $|\det DF_K|^{-1/2}|||DF_K|||$ durch eine von $c_{\mathcal{T}}$ abhängige Konstante beschränkt ist.
Mit den gleichen Argumenten folgt aus dem Spursatz 1.12, Satz 6.1 und Satz 6.7 durch Transformation auf das Referenzelement mit $\hat{u} := u \circ F_K$

$$\begin{aligned}
&\|\nabla(u - I_h u)\|_{L^2(E)} \\
=&h_E^{-1/2}\|\nabla(\hat{u} - \hat{\pi}\hat{u})\|_{L^2(\hat{E})} \\
\leq&ch_E^{-1/2}\|\hat{u} - \hat{\pi}\hat{u}\|_{H^2(\hat{K})} \\
=&ch_E^{-1/2}\left\{\|\hat{u} - \hat{\pi}\hat{u}\|^2_{L^2(\hat{K})} + |\hat{u} - \hat{\pi}\hat{u}|^2_{H^1(\hat{K})} + |\hat{u}|^2_{H^2(\hat{K})}\right\}^{1/2} \qquad (6) \\
\leq&c'h_E^{-1/2}|\hat{u}|_{H^2(\hat{K})} \\
\leq&c'h_E^{-1/2}|\det DF_K|^{-1/2}|||DF_K|||^2|u|_{H^2(\hat{K})} \\
\leq&c''h_E^{1/2}|u|_{H^2(K)}.
\end{aligned}$$

Hierbei haben wir im letzten Schritt ausgenutzt, daß wegen Satz 6.2 und der Regularität von $\mathcal{T}_h$ die Größe $h_E^{-1}|\det DF_K|^{-1/2}|||DF_K|||^2$ durch eine von $c_{\mathcal{T}}$ abhängige Konstante beschränkt ist.
Aus den Abschätzungen (4) – (6) und der Cauchy-Schwarz'schen Ungleichung für Summen folgt

$$|L_u(u)| \leq c \sum_{E\in\mathcal{E}_h} h_E|w|_{H^1(K)}|u|_{H^2(K)} \leq c'h|w|_{X_h}|u|_2.$$

□

12.4 Satz: *Seien $u \in H_0^1(\Omega) \cap H^2(\Omega)$ die schwache Lösung der Poisson Gleichung und $u_h \in X_h$ die Lösung der Crouzeix-Raviart Diskretisierung. Ω sei konvex. Dann gelten die Fehlerabschätzungen*

$$\begin{aligned}
\|u - u_h\|_{X_h} &\leq c_1 h|u|_2, \\
\|u - u_h\|_0 &\leq c_2 h^2|u|_2.
\end{aligned}$$

Die Konstanten c_1, c_2 hängen von Ω und der Konstanten $c_{\mathcal{T}}$ aus der Regularitätsbedingung an $\mathcal{T}_h$ ab.

Beweis: Da $S^{1,0}_{h,0} \subset X_h$ ist, folgt aus Satz 6.3

$$\begin{aligned}
\inf_{v_h\in X_h}\|u - v_h\|_{X_h} \leq& \inf_{w_h\in S^{1,0}_{h,0}}\|u - w_h\|_{X_h} \\
=& \inf_{w_h\in S^{1,0}_{h,0}}|u - w_h|_1 \\
\leq&|u - I_h u|_1 \\
\leq&ch|u|_2.
\end{aligned}$$

Damit folgt die erste Fehlerabschätzung aus Satz 12.1 und Satz 12.3.
Satz 12.3 mit $w = u - u_h$ liefert andererseits

$$\begin{aligned}&|a_h(u-u_h,u_\varphi)-(\varphi,u-u_h)_0|\\ \leq& ch|u_\varphi|_2\|u-u_h\|_{X_h}\\ \leq& c'h\|\varphi\|_0\|u-u_h\|_{X_h}.\end{aligned}$$

Satz 12.3 mit $w = u_\varphi - u_{\varphi,h}$ und der erste Teil des Beweises ergeben weiter

$$\begin{aligned}&|a_h(u,u_\varphi-u_{\varphi,h})-l_h(u_\varphi-u_{\varphi,h})|\\ \leq& ch\|u_\varphi-u_{\varphi,h}\|_{X_h}|u|_2\\ \leq& c'h^2|u_\varphi|_2|u|_2\\ \leq& c''h^2\|\varphi\|_0|u|_2.\end{aligned}$$

Damit folgt die zweite Fehlerabschätzung des Satzes aus der ersten Abschätzung und Satz 12.2. □

12.5 Bemerkung: **(1)** Sei $J(u) := \frac{1}{2}a_h(u,u) - l_h(u)$. Wegen $S^{1,0}_{h,0} \subset X$ ist

$$\inf_{u\in X} J(u) \leq \inf_{u_h\in S^{1,0}_{h,0}} J(u_h).$$

Wegen $S^{1,0}_{h,0} \subset X_h$ ist ebenso

$$\inf_{u_h\in X_h} J(u_h) \leq \inf_{u_h\in S^{1,0}_{h,0}} J(u_h).$$

Daher ist das Minimum von J bei nicht-konformen Methoden kleiner als bei konformen Methoden und häufig dichter am kontinuierlichen Minimum.
(2) Satz 12.4 benötigt im Gegensatz zu Satz 6.4 nicht nur die H^2-Regularität der aktuellen Lösung u sondern die H^2-Regularität für jede rechte Seite in $L^2(\Omega)$. Daher muß Ω als konvex vorausgesetzt werden. Diese zusätzliche Einschränkung ist nicht Beweistechnik. Nicht-konforme Methoden reagieren häufig empfindlicher als konforme Methoden auf mangelnde H^2-Regularität z.B. aufgrund einspringender Ecken.
(3) Konstruktionsgemäß gilt

$$L_u(w) = 0 \quad \forall w \in H^1_0(\Omega).$$

Dies gilt aber nicht für L^2-Funktionen, da L_u auf $L^2(\Omega)$ nicht stetig fortsetzbar ist. □

13. Gemischte Methoden

Um technische Schwierigkeiten zu vermeiden und die wesentlichen Aspekte besser herauszuarbeiten, beschränken wir uns im folgenden auf ein einfaches Modellbeispiel: die Poisson Gleichung in einem zweidimensionalen, beschränkten, *konvexen* Polygon Ω mit homogenen Dirichlet Randbedingungen

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \Gamma. \end{aligned} \tag{1}$$

Dies ist ein extrem einfaches Modell für die vertikale Auslenkung u einer ebenen, dünnen, eingespannten Membran unter Einfluß einer vertikalen Last f. Die bisher betrachteten Methoden liefern Approximationen der Auslenkung. Unter mechanischen Gesichtspunkten sind aber häufig die aus der Belastung resultierenden inneren Spannungskräfte viel wichtiger. Im Rahmen dieses einfachen Modells ist dies der Gradient ∇u der Auslenkung. Diese Größe muß bei den bisher betrachteten Methoden durch Differentiation aus der Auslenkung u berechnet werden und wird i.a. um eine h-Potenz schlechter approximiert als die Auslenkung. Wegen der Bedeutung dieser Größe sucht man nach Verfahren, die sie direkt und genauer approximieren.

Dies leisten sog. **gemischte Finite Element Methoden**. Dazu führen wir $\sigma := \nabla u$ als zusätzliche Variable ein. Damit geht (1) über in das folgende Differentialgleichungssystem 1. Ordnung

$$\begin{aligned} \sigma - \nabla u &= 0 \quad \text{in } \Omega \\ -\operatorname{div} \sigma &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \Gamma. \end{aligned} \tag{2}$$

Multiplizieren wir die erste Gleichung von (2) mit einem hinreichend glatten Vektorfeld τ, integrieren über Ω, wenden den Gauß'schen Integralsatz an und nutzen die Randbedingungen für u aus, erhalten wir

$$\begin{aligned} 0 &= \int_\Omega \sigma \cdot \tau - \int_\Omega \tau \cdot \nabla u \\ &= \int_\Omega \sigma \cdot \tau + \int_\Omega u \operatorname{div} \tau - \int_{\partial\Omega} m \underbrace{u}_{=0} \tau \cdot n \\ &= \int_\Omega \sigma \cdot \tau + \int_\Omega u \operatorname{div} \tau. \end{aligned}$$

Offensichtlich ist die zweite Gleichung von (2) äquivalent zu

$$-\int_\Omega v \operatorname{div} \sigma = \int_\Omega f v \qquad \forall v \in L^2(\Omega).$$

Diese Beobachtung führt auf folgende schwache Formulierung von (1):

$$\begin{aligned} &\text{Finde } [\sigma, u] \in X := M \times Q, \text{ so daß} \\ &\int_\Omega \sigma \cdot \tau + \int_\Omega u \operatorname{div} \tau = 0 \qquad \forall \tau \in M \\ &\qquad - \int_\Omega v \operatorname{div} \sigma = \int_\Omega f v \quad \forall v \in Q. \end{aligned} \tag{3}$$

Dabei ist

$$\begin{aligned} M :=& H(\operatorname{div}, \Omega) := \{\sigma \in L^2(\Omega; \mathbb{R}^2) : \operatorname{div} \sigma \in L^2(\Omega; \mathbb{R})\}, \\ Q :=& L^2(\Omega). \end{aligned}$$

M wird versehen mit der Norm

$$\|\sigma\|_{H(\operatorname{div})} := \{\|\sigma\|_0^2 + \|\operatorname{div} \sigma\|_0^2\}^{1/2}.$$

Wie in Satz 1.6 kann man zeigen, daß M mit dieser Norm ein Banach Raum ist.

Die Herleitung von (3) zeigt, daß dieses Problem im üblichen Sinne zu (2) äquivalent ist: Jede klassische Lösung von (2) löst auch (3) und jede hinreichend glatte Lösung von (3) ist auch eine klassische Lösung von (2). Problem (3) ist ein sog. **gemischtes** oder **Sattelpunktsproblem**. Es stellt die Euler-Lagrange Gleichungen des folgenden Optimierungsproblems mit Nebenbedingungen dar:

$$J(\sigma) := \frac{1}{2} \int_\Omega \sigma \cdot \sigma \longrightarrow \min \text{ in } M_f := \{\sigma \in M : -\operatorname{div} \sigma = f\}.$$

Für die Analyse von (3) führen wir folgende Abkürzungen ein:

$$\begin{aligned} \|[\sigma, u]\|_X :=& \{\|\sigma\|_0^2 + \|\operatorname{div} \sigma\|_0^2 + \|u\|_0^2\}^{1/2}, \\ a([\sigma, u], [\tau, v]) :=& \int_\Omega \sigma \cdot \tau + \int_\Omega u \operatorname{div} \tau - \int_\Omega v \operatorname{div} \sigma, \\ l([\tau, v]) :=& \int_\Omega f v. \end{aligned}$$

Dann ist (3) offensichtlich äquivalent zu

$$a([\sigma, u], [\tau, v]) = l([\tau, v]) \quad \forall [\tau, v] \in X. \tag{4}$$

Auf den ersten Blick scheint Problem (4) in den abstrakten Rahmen von § 2 zu passen. Aber die Bilinearform a ist nicht symmetrisch und nicht koerziv. Dies spiegelt die Sattelpunktsstruktur von (3) wider. Statt der Koerzivität erfüllt a allerdings eine sog. **inf-sup Bedingung**:

13.1 Satz: (inf-sup Bedingung) *Es gibt eine Konstante $\alpha > 0$, die nur von Ω abhängt, mit*

$$\inf_{[\sigma,u]\in X\setminus\{0\}} \sup_{[\tau,v]\in X\setminus\{0\}} \frac{a([\sigma,u],[\tau,v])}{\|[\sigma,u]\|_X \|[\tau,v]\|_X} \geq \alpha.$$

Beweis: Sei $[\sigma,u] \in X\setminus\{0\}$ beliebig, aber fest. Dann gilt

$$a([\sigma,u],[\sigma,u]) = \|\sigma\|_0^2$$

und wegen $\operatorname{div}\sigma \in L^2(\Omega)$

$$a([\sigma,u],[0,-\operatorname{div}\sigma]) = \|\operatorname{div}\sigma\|_0^2.$$

Da Ω beschränkt ist, gibt es ein $R > 0$ mit $\Omega \subset (-R,R)^2$. Setze u durch 0 auf ganz $\mathbb{R}^2$ fort und definiere

$$\tau_u(x) := e_1 \int_{-R}^{x_1} u(s,x_2)ds \quad \forall x = (x_1,x_2) \in \Omega.$$

Dabei ist e_1 der erste Einheitsvektor in $\mathbb{R}^2$. Offensichtlich ist $\tau_u \in M$ und erfüllt

$$\operatorname{div}\tau_u = u\,,\ \|\tau_u\|_0 \leq c_0\|u\|_0.$$

Dabei hängt die Konstante c_0 nur vom Durchmesser von Ω ab. O.E. ist $c_0 \geq 1$. Hieraus folgt

$$\begin{aligned} a([\sigma,u],[\tau_u,0]) &= \int_\Omega \sigma\cdot\tau_u + \|u\|_0^2 \\ &\geq -\|\sigma\|_0\|\tau_u\|_0 + \|u\|_0^2 \\ &\geq -\|\sigma\|_0 c_0\|u\|_0 + \|u\|_0^2 \\ &\geq \frac{1}{2}\|u\|_0^2 - \frac{1}{2}c_0^2\|\sigma\|_0^2. \end{aligned}$$

Setze

$$\rho := c_0^2\sigma + \tau_u\,,\ w := c_0^2 u - \frac{1}{2}\operatorname{div}\sigma.$$

Dann folgt

$$\begin{aligned} a([\sigma,u],[\rho,w]) =& c_0^2 a([\sigma,u],[\sigma,u]) + a([\sigma,u],[\tau_u,0]) + \frac{1}{2}a([\sigma,u],[0,-\operatorname{div}\sigma]) \\ \geq& c_0^2\|\sigma\|_0^2 + \frac{1}{2}\|u\|_0^2 - \frac{1}{2}c_0^2\|\sigma\|_0^2 + \frac{1}{2}\|\operatorname{div}\sigma\|_0^2 \\ \geq& \frac{1}{2}\|[\sigma,u]\|_X^2 \end{aligned}$$

und

$$\begin{aligned}\|[\rho,w]\|_X \leq & c_0^2\|[\sigma,u]\|_X + \|[\tau_u,0]\|_X + \frac{1}{2}\|[0,-\operatorname{div}\sigma]\|_X \\ = & c_0^2\|[\sigma,u]\|_X + \{\|u\|_0^2 + \underbrace{\|\tau_u\|_0^2}_{\leq c_0^2\|u\|_0^2}\}^{1/2} + \frac{1}{2}\|\operatorname{div}\sigma\|_0 \\ \leq & \left\{c_0^4 + c_0^2 + 1 + \frac{1}{4}\right\}^{1/2} \|[\sigma,u]\|_X \\ \leq & 2c_0^2\|[\sigma,u]\|_X.\end{aligned}$$

Aus diesen Abschätzungen folgt die Behauptung mit $\alpha = 1/(4c_0^2)$. □

13.2 Satz: *Problem (3) besitzt eine eindeutige Lösung.*

Beweis: Sei $u \in H_0^1(\Omega)$ die schwache Lösung von (1) im Sinne von Definition 3.1. Da Ω konvex ist, folgt aus Satz 3.6 $u \in H^2(\Omega)$. Also ist $\sigma := \nabla u \in H(\operatorname{div},\Omega)$. Aus der Herleitung von (3) folgt, daß $[\sigma,u]$ eine Lösung von (3) ist. Wir müssen also noch die Eindeutigkeit zeigen. Dazu reicht es, zu zeigen, daß das homogene Problem, d.h. (3) mit $f = 0$ bzw. (4) mit $l = 0$, nur die triviale Lösung hat. Ist aber $[\sigma,u]$ eine Lösung des homogenen Problems (4), so folgt aus Satz 13.1

$$\alpha\|[\sigma,u]\|_X \leq \sup_{[\tau,v]\in X\setminus\{0\}} \frac{a([\sigma,u],[\tau,v])}{\|[\tau,v]\|_X} = 0.$$

Also ist $\sigma = u = 0$. □

Für die Diskretisierung von (3) betrachten wir nur das einfachste Beispiel, das sog. **Raviart-Thomas Element** niedrigster Ordnung. Dazu bezeichnet $\mathcal{T}_h$ eine Familie zulässiger und regulärer Triangulierungen von Ω und $\mathcal{E}_h$ die Menge der Dreieckskanten in $\mathcal{T}_h$. Für ein Dreieck K sei

$$RT(K) := \left\{ \begin{pmatrix}\alpha\\ \beta\end{pmatrix} + \gamma \begin{pmatrix}x_1\\ x_2\end{pmatrix} : \alpha,\beta,\gamma \in \mathbb{R} \right\}.$$

13.3 Satz: *Sei K ein Dreieck und n_K das äußere Einheitsnormalenfeld zu ∂K. Dann gilt für jedes $\sigma \in RT(K)$:*
(1) *$\sigma \cdot n_K$ ist konstant auf den Kanten von K.*
(2) *σ ist eindeutig bestimmt durch die Werte von $\sigma \cdot n_K$ auf den Kanten von K.*

Beweis: **ad (1):** Die Funktion $x \longrightarrow x \cdot n_K$ ist konstant auf den Kanten von K.
ad (2): Wegen $\dim RT(K) = 3$ müssen wir nur zeigen, daß 0 die einzige Funktion $\sigma \in RT(K)$ ist, für die $\sigma \cdot n_K$ auf allen Kanten von K verschwindet. Sei $\sigma = \begin{pmatrix}\alpha\\ \beta\end{pmatrix} + \gamma \begin{pmatrix}x_1\\ x_2\end{pmatrix}$ eine solche Funktion. Dann folgt aus dem Gauß'schen Integralsatz

$$\begin{aligned} 0 = \int_{\partial K} \sigma \cdot n_K = \int_K \operatorname{div}\sigma = \int_K 2\gamma \\ \Longrightarrow \gamma = 0.\end{aligned}$$

Also steht der Vektor $\begin{pmatrix}\alpha\\ \beta\end{pmatrix}$ senkrecht auf dem zweidimensionalen Raum, der von den Richtungsvektoren der drei Kanten von K aufgespannt wird. Also ist auch $\alpha = \beta = 0$. □

Definiere

$$\begin{aligned} RT_h^{-1} :=& \left\{\sigma : \Omega \longrightarrow \mathbb{R}^2 : \sigma_{|K} \in RT(K)\, \forall K \in \mathcal{T}_h\right\} \\ RT_h :=& RT_h^{-1} \cap H(\mathrm{div}, \Omega). \end{aligned}$$

Wie im Beweis von Satz 1.7 folgt, daß $\sigma \in RT_h^{-1}$ genau dann in RT_h liegt, wenn $\sigma \cdot n_K$ stetig ist über alle Dreieckskanten, die in Ω liegen. Wegen Satz 13.3. ist daher $RT_h \neq \{0\}$. Die Freiheitsgrade der Funktionen $\sigma_h \in RT_h$ sind genau die Werte von $\sigma_h \cdot n_K$ auf den Kanten in $\mathcal{E}_h$. Insbesondere ist $\dim RT_h = \sharp\mathcal{E}_h$. Wir setzen nun

$$\begin{aligned} M_h :=& RT_h, \\ Q_h :=& S_h^{0,-1}, \\ X_h :=& M_h \times Q_h \end{aligned}$$

und approximieren Problem (4) durch

$$\begin{aligned} &\text{Finde } [\sigma_h, u_h] \in X_h, \text{ so daß} \\ &a([\sigma_h, u_h], [\tau_h, v_h]) = l([\tau_h, v_h]) \quad \forall [\tau_h, v_h] \in X_h \end{aligned} \tag{5}$$

oder in anderer Schreibweise

$$\begin{aligned} &\text{Finde } [\sigma_h, u_h] \in X_h, \text{ so daß} \\ &\int_\Omega \sigma_h \cdot \tau_h + \int_\Omega u_h \,\mathrm{div}\, \tau_h = 0 \qquad \forall \tau_h \in M_h \\ &\qquad - \int_\Omega v_h \,\mathrm{div}\, \sigma_h = \int_\Omega f v_h \quad \forall v_h \in Q_h. \end{aligned} \tag{6}$$

13.4 Satz: **(1)** *Es ist* $\mathrm{div}\, M_h = Q_h$. *Zu jedem* $u_h \in Q_h$ *gibt es ein* $\tau_{h,u_h} \in M_h$ *mit*

$$\mathrm{div}\, \tau_{h,u_h} = u_h\,,\ \|\tau_{h,u_h}\|_0 \leq c_1 \|u_h\|_0.$$

Die Konstante c_1 *hängt nur vom Durchmesser von* Ω *und der Konstanten* $c_{\mathcal{T}}$ *aus der Regularitätsbedingung an* $\mathcal{T}_h$ *ab.*
(2) (inf-sup Bedingung) *Es gibt eine Konstante* $\beta > 0$, *die nur von der Konstanten* c_1 *aus Teil (1) abhängt, mit*

$$\inf_{[\sigma_h,u_h]\in X_h\setminus\{0\}} \sup_{[\tau_h,v_h]\in X_h\setminus\{0\}} \frac{a([\sigma_h, u_h], [\tau_h, v_h])}{\|[\sigma_h, u_h]\|_X \|[\tau_h, v_h]\|_X} \geq \beta.$$

(3) *Problem (5) bzw. (6) besitzt eine eindeutige Lösung.*

Beweis: **ad (1):** Aus der Definition von RT_h folgt sofort $\operatorname{div} M_h \subset Q_h$. Sei nun $u_h \in Q_h$ beliebig und

$$\tau_{u_h} := e_1 \int_{-R}^{x_1} u_h(s, x_2) ds \quad \forall x = (x_1, x_2) \in \Omega$$

wie im Beweis von Satz 13.1. Wir definieren einen Operator $I_h : H(\operatorname{div}, \Omega) \longrightarrow M_h$ durch

$$(I_h \tau) \cdot n_K = h_E^{-1} \int_E \tau \cdot n_K \quad \forall K \in \mathcal{T}_h, E \in \mathcal{E}_h, E \subset \partial K.$$

Dabei ist h_E die Länge von E. Wegen Satz 13.3 ist diese Definition sinnvoll. Für jedes $K \in \mathcal{T}_h$ folgt mit dem Gauß'schen Integralsatz

$$\begin{aligned}
\int_K \operatorname{div}(I_h \tau_{u_h}) &= \sum_{\substack{E \in \mathcal{E}_h \\ E \subset \partial K}} \int_E (I_h \tau_{u_h}) \cdot n_K \\
&= \sum_{\substack{E \in \mathcal{E}_h \\ E \subset \partial K}} \int_E \tau_{u_h} \cdot n_K \\
&= \int_K \operatorname{div} \tau_{u_h} \\
&= \int_K u_h.
\end{aligned}$$

Da u_h und $\operatorname{div}(I_h \tau_{u_h})$ auf K konstant sind, bedeutet dies

$$\operatorname{div}(I_h \tau_{u_h}) = u_h.$$

Mit dem üblichen Skalierungsargument, d.h. Transformation auf das Referenzelement und Äquivalenz von Normen auf endlich dimensionalen Räumen, zeigt man, daß

$$\|I_h \tau_{u_h}\|_0 \leq c_2 \|\tau_{u_h}\|_{H(\operatorname{div})}$$

ist mit einer Konstanten c_2, die nur von $c_{\mathcal{T}}$ abhängt. Also leistet $\tau_{h,u_h} := I_h \tau_{u_h}$ das Gewünschte mit $c_1 = c_2(1 + c_0^2)^{1/2}$ und c_0 wie im Beweis von Satz 13.1.
ad (2): Wegen Teil (1) können wir den Beweis von Satz 13.1 kopieren. Dabei übernimmt τ_{h,u_h} aus Teil (1) die Rolle von τ_u aus dem Beweis von Satz 13.1.
ad (3): Wie im Beweis von Satz 13.2 folgt aus Teil (2), daß das homogene Problem (5), d.h. (5) mit $l = 0$, nur die triviale Lösung besitzt. Da (5) ein lineares Gleichungssystem mit der gleichen Anzahl von Gleichungen und Unbekannten ist, folgt hieraus die Behauptung. □

13.5 Satz: *Seien $[\sigma, u] \in X$ und $[\sigma_h, u_h] \in X_h$ die eindeutigen Lösungen der Probleme (4) und (5). Es sei $\sigma \in H^1(\Omega)^2$, $\operatorname{div}\sigma \in H^1(\Omega)$ und $u \in H^1(\Omega)$. Dann gilt die Fehlerabschätzung*

$$\|\sigma - \sigma_h\|_0 + \|\operatorname{div}(\sigma - \sigma_h)\|_0 + \|u - u_h\|_0 \leq ch\left\{|\sigma|_1 + |\operatorname{div}\sigma|_1 + |u|_1\right\}.$$

Beweis: Sei $[\tau_h, v_h] \in X_h$ beliebig. Mit der Dreiecksungleichung folgt

$$\|[\sigma - \sigma_h, u - u_h]\|_X \leq \|[\sigma - \tau_h, u - v_h]\|_X + \|[\tau_h - \sigma_h, v_h - u_h]\|_X.$$

Wegen $X_h \subset X$ folgt aus (4) und (5) die Galerkin Orthogonalität

$$a([\sigma - \sigma_h, u - u_h], [\rho_h, w_h]) = 0 \quad \forall [\rho_h, w_h] \in X_h.$$

Hieraus und aus Satz 13.4 (3) folgt

$$\begin{aligned}
&\beta\|[\tau_h - \sigma_h, v_h - u_h]\|_X \\
\leq &\sup_{[\rho_h, w_h] \in X_h \setminus \{0\}} \frac{a([\tau_h - \sigma_h, v_h - u_h], [\rho_h, w_h])}{\|[\rho_h, w_h]\|_X} \\
= &\sup_{[\rho_h, w_h] \in X_h \setminus \{0\}} \frac{a([\tau_h - \sigma, v_h - u], [\rho_h, w_h])}{\|[\rho_h, w_h]\|_X} \\
\leq &\|[\tau_h - \sigma, v_h - u]\|_X.
\end{aligned}$$

Hierbei haben wir im letzten Schritt die Cauchy-Schwarz'sche Ungleichung für Integrale und Summen und die Definition von $\|.\|_X$ ausgenutzt. Da $[\tau_h, v_h] \in X_h$ beliebig war, beweisen diese Abschätzungen das folgende Analogon zum Céa-Lemma, Satz 2.2,

$$\|[\sigma - \sigma_h, u - u_h]\|_X \leq \left(1 + \frac{1}{\beta}\right) \inf_{[\tau_h, v_h] \in X_h} \|[\sigma - \tau_h, u - v_h]\|_X.$$

Bezeichne mit $\pi_h : L^2(\Omega) \longrightarrow Q_h$ die L^2-Projektion. Dann folgt aus obiger Abschätzung

$$\begin{aligned}
&\left\{\|\sigma - \sigma_h\|_0^2 + \|\operatorname{div}(\sigma - \sigma_h)\|_0^2 + \|u - u_h\|_0^2\right\}^{1/2} \\
\leq &\left(1 + \frac{1}{\beta}\right) \|[\sigma - \ I_h\sigma, u - \pi_h u]\|_X \\
= &\left(1 + \frac{1}{\beta}\right) \left\{\|\sigma - I_h\sigma\|_0^2 + \|\operatorname{div}(\sigma - I_h\sigma)\|_0^2 + \|u - \pi_h u\|_0^2\right\}^{1/2}.
\end{aligned}$$

Dabei ist I_h der Interpolationsoperator aus (7). Aus der Poincaré'schen Ungleichung, Satz 1.21, folgt

$$\begin{aligned}\|u-\pi_h u\|_0 =& \left\{\sum_{K\in\mathcal{T}_h}\|u-\pi_h u\|^2_{L^2(K)}\right\}^{1/2}\\ =& \left\{\sum_{K\in\mathcal{T}_h}\|u-\frac{1}{|K|}\int_K u\|^2_{L^2(K)}\right\}^{1/2}\\ \leq& \left\{\sum_{K\in\mathcal{T}_h} c^2h_K^2|u|^2_{H^1(K)}\right\}^{1/2}\\ \leq& ch|u|_1.\end{aligned}$$

Durch Transformation auf das Referenzdreieck folgt wie im Beweis von Satz 6.3 für jedes $K\in\mathcal{T}_h$

$$\begin{aligned}\|\sigma-I_h\sigma\|_{L^2(K)} \leq& c\inf_{\tau_K\in RT(K)}\|\sigma-\tau_K\|_{L^2(K)}\\ \leq& c\inf_{\rho_K\in\mathbb{R}^2}\|\sigma-\rho_K\|_{L^2(K)}\\ \leq& c'h_K|\sigma|_{H^1(K)}.\end{aligned}$$

Dabei haben wir wieder im letzten Schritt die Poincaré'sche Ungleichung, Satz 1.21, ausgenutzt. Quadrieren dieser Abschätzung und Summieren über alle Dreiecke liefert

$$\|\sigma-I_h\sigma\|_0 \leq ch|\sigma|_1.$$

Sei nun $v\in L^2(\Omega)$ beliebig. Dann ist

$$\begin{aligned}&\int_\Omega \operatorname{div}(\sigma-I_h\sigma)v\\ =&\int_\Omega \operatorname{div}(\sigma-I_h\sigma)(v-\pi_h v)+\int_\Omega \operatorname{div}(\sigma-I_h\sigma)\pi_h v.\end{aligned}$$

Wegen (7) erhalten wir für den zweiten Summanden

$$\begin{aligned}&\int_\Omega \operatorname{div}(\sigma-I_h\sigma)\pi_h v\\ =&\sum_{K\in\mathcal{T}_h}\int_K \operatorname{div}(\sigma-I_h\sigma)\pi_h v\\ =&\sum_{K\in\mathcal{T}_h}\int_{\partial K} n_K\cdot(\sigma-I_h\sigma)\pi_h v\\ =&\sum_{K\in\mathcal{T}_h}\sum_{\substack{E\in\mathcal{E}_h\\ E\subset\partial K}}\left\{\underbrace{\int_E n_K\cdot\sigma-\int_E n_K\cdot(I_h\sigma)}_{=0}\right\}\pi_h v\\ =&0.\end{aligned}$$

Für den ersten Summanden folgt aus der Cauchy-Schwarz'schen und der Poincaré'-schen Ungleichung

$$\begin{aligned}
&\int_\Omega \operatorname{div}(\sigma - I_h\sigma)(v - \pi_h v) \\
\leq &\|\operatorname{div}(\sigma - \sigma_h)\|_0 \left\{ \sum_{K\in\mathcal{T}_h} \|v - \pi_h v\|^2_{L^2(K)} \right\}^{1/2} \\
\leq &c\|\operatorname{div}(\sigma - \sigma_h)\|_0 \left\{ \sum_{K\in\mathcal{T}_h} h_K^2 |v|^2_{H^1(K)} \right\}^{1/2}.
\end{aligned}$$

Wählen wir spezeill $v = \operatorname{div}(\sigma - I_h\sigma)$ und beachten, daß $\operatorname{div} I_h\sigma$ elementweise konstant ist, erhalten wir aus obiger Abschätzung

$$\begin{aligned}
&\|\operatorname{div}(\sigma - I_h\sigma)\|_0^2 \\
\leq &c\|\operatorname{div}(\sigma - I_h\sigma)\|_0 \left\{ \sum_{K\in\mathcal{T}_h} h_K^2 |\operatorname{div}\sigma|^2_{H^1(K)} \right\}^{1/2} \\
\leq &ch\|\operatorname{div}(\sigma - I_h\sigma)\|_0 |\operatorname{div}\sigma|_1
\end{aligned}$$

und damit

$$\|\operatorname{div}(\sigma - I_h\sigma)\|_0 \leq ch|\operatorname{div}\sigma|_1.$$

Hieraus folgt die Fehlerabschätzung des Satzes. □

13.6 Bemerkung: Da Ω konvex ist, ist $u \in H^2(\Omega)$ und $\sigma = \nabla u \in H^1(\Omega)$. Wegen $\operatorname{div}\sigma = -f$ sind daher die Regularitätsannahmen von Satz 13.5 erfüllt, wenn $f \in H^1(\Omega)$ ist. □

Das lineare Gleichungssystem (6) hat folgende Struktur

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \sigma_h \\ u_h \end{pmatrix} = \begin{pmatrix} 0 \\ -b_h \end{pmatrix}.$$

Die Koeffizientenmatrix dieses LGS ist symmetrisch, aber indefinit. Dies spiegelt die Sattelpunktsstruktur von (3) wider. Die Matrix A ist symmetrisch, positiv definit, und die Matrix B hat wegen Satz 13.4 (1) maximalen Rang. Die Größe des LGS (6) ist $\sharp\mathcal{E}_h + \sharp\mathcal{T}_h$. Wegen der Indefinitheit kann es nicht mit einem CG-Verfahren gelöst werden. Wir wollen nun ein äquivalentes diskretes Problem herleiten, das auf ein kleineres, symmetrisches, positiv definites LGS führt. Bezeichne dazu wie in § 12 mit $\mathcal{E}_{h,\Omega}$ die Menge aller Kanten im Innern von Ω. Jedem $E \in \mathcal{E}_{h,\Omega}$ ordnen wir wieder einen dazu orthogonalen Einheitsvektor n_E zu und bezeichnen mit $[\varphi]_E$ denSprung von φ über E in Richtung n_E. Setze

$$\Sigma := \bigcup_{E\in\mathcal{E}_{h,\Omega}} E$$

und bezeichne mit

$$S_{h,\Sigma}^{0,-1} := \{\lambda : \Sigma \longrightarrow \mathbb{R} : \lambda_{|E} \in \mathbb{R} \quad \forall E \in \mathcal{E}_{h,\Omega}\}$$

die stückweise konstanten Funktionen auf Σ.

13.7 Satz: **(1)** $RT_h = \left\{\sigma \in RT_h^{-1} : \sum_{E\in\mathcal{E}_{h,\Omega}} \int_E [\sigma \cdot n_E]_E \lambda = 0 \quad \forall \lambda \in S_{h,\Sigma}^{0,-1}\right\}$.
(2) *Sei* $\varphi \in \mathcal{L}(RT_h^{-1}, \mathbb{R})$ *mit* $\varphi(\sigma) = 0 \, \forall \sigma \in RT_h$. *Dann gibt es genau ein* $\lambda_\varphi \in S_{h,\Sigma}^{0,-1}$ *mit*

$$\varphi(\sigma) = \sum_{E\in\mathcal{E}_{h,\Omega}} \int_E [\sigma \cdot n_E]_E \lambda_\varphi \quad \forall \sigma \in RT_h^{-1}.$$

Beweis: **ad (1):** Wie im Beweis von Satz 1.7 folgt, daß $\sigma \in RT_h^{-1}$ genau dann in RT_h liegt, wenn $\sigma \cdot n_K$ stetig ist über alle inneren Kanten, d.h. wenn $[\sigma \cdot n_E]_E$ für alle $E \in \mathcal{E}_{h,\Omega}$ verschwindet. Wegen Satz 13.3 (1) ist dies genau dann der Fall, wenn gilt

$$\sum_{E\in\mathcal{E}_{h,\Omega}} \int_E [\sigma \cdot n_E]_E \lambda = 0 \quad \forall \lambda \in S_{h,\Sigma}^{0,-1}.$$

ad (2): Sei $\varphi \in \mathcal{L}(RT_h^{-1}, \mathbb{R})$ eine lineare Abbildung, die auf RT_h verschwindet. Wegen des Rangsatzes gibt es ein $\lambda_\varphi \in S_{h,\Sigma}^{0,-1}$ mit der gewünschten Eigenschaft. Wir müssen also nur noch die Eindeutigkeit von λ_φ zeigen. Sei dazu $\mu \in S_{h,\Sigma}^{0,-1}$ mit

$$\sum_{E\in\mathcal{E}_{h,\Omega}} \int_E [\sigma \cdot n_E]_E \mu = 0 \quad \forall \sigma \in RT_h^{-1}.$$

Dann müssen wir $\mu = 0$ zeigen. Sei dazu $E^* \in \mathcal{E}_{h,\Omega}$ beliebig und $K^* \in \mathcal{T}_h$ ein Dreieck, das E^* als Kante hat. Wegen Satz 13.3 gibt es ein $\sigma^* \in RT_h^{-1}$ mit

$$\begin{aligned} \sigma^*_{|K} &= 0 \quad \text{für alle } K \in \mathcal{T}_h \backslash \{K^*\} \\ \sigma^* \cdot n_E &= 0 \quad \text{für alle Kanten } E \text{ von } K^* \text{ mit } E \neq E^* \\ \sigma^* \cdot n_{E^*} &= 1. \end{aligned}$$

Damit folgt

$$\begin{aligned} 0 &= \sum_{E\in\mathcal{E}_{h,\Omega}} \int_E [\sigma^* \cdot n_E]_E \mu \\ &= \int_{E^*} [\sigma^* \cdot n_{E^*}]_E \mu \\ &= \pm \, |E^*| \mu_{|E^*}. \end{aligned}$$

Da E^* beliebig war, folgt $\mu = 0$. □

Wir betrachten nun das folgende Problem:

$$\begin{aligned}
&\text{Finde } \tilde{\sigma}_h \in RT_h^{-1}, \tilde{u}_h \in S_h^{0,-1}, \tilde{\mu}_h \in S_{h,\Sigma}^{0,-1}, \text{ so daß} \\
&\int_\Omega \tilde{\sigma}_h \cdot \tau_h + \int_\Omega \tilde{u}_h \operatorname{div} \tau_h + \sum_{E \in \mathcal{E}_{h,\Omega}} \int_E [\tau_h \cdot n_E]_E \tilde{\mu}_h = 0 \qquad \forall \tau_h \in RT_h^{-1} \\
&\qquad - \int_\Omega v_h \operatorname{div} \tilde{\sigma}_h \qquad\qquad = \int_\Omega f v_h \quad \forall v_h \in S_h^{0,-1} \qquad (8) \\
&\qquad\qquad \sum_{E \in \mathcal{E}_{h,\Omega}} \int_E [\tilde{\sigma}_h \cdot n_E]_E \lambda_h = 0 \qquad \forall \lambda_h \in S_{h,\Sigma}^{0,-1}.
\end{aligned}$$

13.8 Satz: *Die Probleme (6) und (8) sind äquivalent.*

Beweis: Sei $\tilde{\sigma}_h$, $\tilde{u}_h$, $\tilde{\mu}_h$ eine beliebige Lösung von (8). Aus der dritten Gelichung von (8) und Satz 13.7 (1) folgt $\tilde{\sigma}_h \in RT_h = M_h$. Indem wir in der ersten Gleichung von (8) nur Vektorfelder $\tau_h \in RT_h$ als Testfunktionen betrachten, sehen wir, daß $\tilde{\sigma}_h$, $\tilde{u}_h$ eine Lösung von (6) ist.
Betrachte nun umgekehrt die Lösung σ_h, u_h von (6). Diese erfüllt wegen (6) und Satz 13.7 (1) die zweite und dritte Gleichung von (8). Wegen (6) verschwindet zudem die lineare Abbildung

$$\tau \longrightarrow \int_\Omega \sigma_h \cdot \tau + \int_\Omega u_h \operatorname{div} \tau$$

auf RT_h. Wegen Satz 13.7 (2) gibt es daher genau ein $\mu_h \in S_{h,\Sigma}^{0,-1}$ mit

$$\int_\Omega \sigma_h \cdot \tau_h + \int_\Omega u_h \operatorname{div} \tau_h + \sum_{E \in \mathcal{E}_{h,\Omega}} \int_E [\tau_h \cdot n_E]_E \mu_h = 0 \quad \forall \tau_h \in RT_h^{-1}.$$

Also ist σ_h, u_h, μ_h eine Lösung von (8). □

Problem (8) ist ein LGS der Form

$$\begin{pmatrix} \tilde{A} & B^T & C^T \\ B & 0 & 0 \\ C & 0 & 0 \end{pmatrix} \begin{pmatrix} \sigma_h \\ u_h \\ \mu_h \end{pmatrix} = \begin{pmatrix} 0 \\ -b_h \\ 0 \end{pmatrix}.$$

Da es die Größe $\sharp\mathcal{E}_h + \sharp\mathcal{T}_h + \sharp\mathcal{E}_{h,\Omega}$ hat, haben wir auf den ersten Blick im Vergleich zu (6) nichts gewonnen. Da die Funktionen in RT_h^{-1} keine globalen Stetigkeitsbedingungen erfüllen müssen, gibt es eine Basis von RT_h^{-1} aus Funktionen, deren Träger jeweils auf ein einziges Dreieck konzentriert ist. Daher ist $\tilde{A}$ ein blockdiagonale Matrix; die Zahl der Blöcke ist $\sharp\mathcal{T}_h$. Wegen Satz 13.3 (2) ist jeder Block eine 3×3 Matrix. Wir können daher die Unbekannte σ_h elementweise durch u_h und μ_h ausdrücken und erhalten

$$\sigma_h = -\tilde{A}^{-1}\{B^T u_h + C^T \mu_h\}.$$

Einsetzen in die Gleichung für u_h liefert

$$-b_h = B\sigma_h = -B\tilde{A}^{-1}B^T u_h - B\tilde{A}^{-1}C^T\mu_h.$$

Da die Funktionen u_h elementweise konstant sind, ist die Matrix $B\tilde{A}^{-1}B^T$ diagonal. Wir können daher u_h durch μ_h ausdrücken und erhalten

$$u_h = \{B\tilde{A}^{-1}B^T\}^{-1}\{b_h - B\tilde{A}^{-1}C^T\mu_h\}.$$

Setzen wir dies in die Gleichung für μ_h ein, erhalten wir

$$\begin{aligned}
0 =& C\sigma_h \\
=& - C\tilde{A}^{-1}B^T u_h - C\tilde{A}^{-1}C^T\mu_h \\
=& - C\tilde{A}^{-1}B^T\{B\tilde{A}^{-1}B^T\}^{-1}\{b_h - B\tilde{A}^{-1}C^T\mu_h\} - C\tilde{A}^{-1}C^T\mu_h \\
=& \{C\tilde{A}^{-1}B^T\{B\tilde{A}^{-1}B^T\}^{-1}B\tilde{A}^{-1}C^T - C\tilde{A}^{-1}C^T\}\mu_h - C\tilde{A}^{-1}B^T\{B\tilde{A}^{-1}B^T\}^{-1}b_h.
\end{aligned}$$

Wir können also die Unbekannten σ_h und u_h elementweise eliminieren und erhalten ein LGS der Form

$$H\mu_h = g_h \tag{9}$$

für μ_h. Es hat nur noch die Größe $\sharp\mathcal{E}_{h,\Omega}$. Die Matrix

$$H = C\tilde{A}^{-1}B^T\{B\tilde{A}^{-1}B^T\}^{-1}B\tilde{A}^{-1}C^T - C\tilde{A}^{-1}C^T$$

ist offensichtlich symmetrisch. Da die Matrix des LGS (8) wegen Satz 13.4 (3) und Satz 13.8 regulär ist, ist H auch positiv definit. Daher kann Problem (9) z.B. mit einem CG- oder PCG-Verfahren gelöst werden.

Die Freiheitsgrade von μ_h sind die Werte in den Mittelpunkten der inneren Kanten. Gleiches gilt für die nicht-konforme Crouzeix-Raviart Diskretisierung aus § 12. In der Tat sind das Problem (9) und die Crouzeix-Raviart Diskretisierung eng verwandt. Aufgrund dieser Verwandschaft kann man zudem die Unbekannte μ_h benutzen, um eine Approximation der Ordnung $O(h^2)$ für die Verschiebung zu erhalten.

13.9 Satz: *Seien $[\sigma, u] \in X$ und $[\sigma_h, u_h, \mu_h] \in RT_h^{-1} \times S_h^{0,-1} \times S_{h,\Sigma}^{0,-1}$ die eindeutigen Lösungen der Probleme (3) und (8). Bezeichne mit CR_h den Raum der Crouzeix-Raviart Elemente aus § 12 und definiere $\hat{u}_h \in CR_h$ durch*

$$\hat{u}_h(m_E) = -\mu_{h|E} \quad \forall E \in \mathcal{E}_{h,\Omega}.$$

Dabei ist m_E der Mittelpunkt der Kante E. Dann gilt

$$\|u - \hat{u}_h\|_0 \leq ch^2\{|f|_1 + \|u\|_2\}.$$

Beweis: Da die Mittelpunktsregel für lineare Funktionen exakt ist, folgt aus der Definition von $\hat{u}_h$

$$\int_E (\hat{u}_h + \mu_h) = 0 \quad \forall E \in \mathcal{E}_{h,\Omega}.$$

Definiere analog $u_h^* \in CR_h$ durch

$$\int_E (u_h^* - u) = 0 \quad \forall E \in \mathcal{E}_{h,\Omega}.$$

Schließlich sei $\overline{u}_h := \pi_h u$ die L^2-Projektion von u auf $S_h^{0,-1}$. Dann folgt mit elementweiser partieller Integration für alle $\tau_h \in RT_h^{-1}$

$$\begin{aligned}
&\int_\Omega \sigma \cdot \tau_h + \int_\Omega \overline{u}_h \operatorname{div} \tau_h \\
&= \int_\Omega \sigma \cdot \tau_h + \int_\Omega u \operatorname{div} \tau_h \\
&= \int_\Omega \sigma \cdot \tau_h + \sum_{K \in \mathcal{T}_h} \int_K u \operatorname{div} \tau_h \\
&= \int_\Omega \sigma \cdot \tau_h + \sum_{K \in \mathcal{T}_h} \left\{ - \int_K \underbrace{\nabla u}_{=\sigma} \cdot \tau_h + \int_{\partial K} u n_K \cdot \tau_h \right\} \\
&= \sum_{K \in \mathcal{T}_h} \int_{\partial K} u n_K \cdot \tau_h \\
&= \sum_{E \in \mathcal{E}_{h,\Omega}} \int_{\partial E} u [n_E \cdot \tau_h]_E \\
&= \sum_{E \in \mathcal{E}_{h,\Omega}} \int_{\partial E} u_h^* [n_E \cdot \tau_h]_E.
\end{aligned}$$

Hieraus und aus der ersten Gleichung von (8) erhalten wir

$$\begin{aligned}
&\sum_{E \in \mathcal{E}_{h,\Omega}} \int_{\partial E} (\hat{u}_h - u_h^*) [n_E \cdot \tau_h]_E \\
&= \sum_{E \in \mathcal{E}_{h,\Omega}} \left\{ - \int_{\partial E} \mu_h [n_E \cdot \tau_h]_E - \int_{\partial E} u_h^* [n_E \cdot \tau_h]_E \right\} \\
&= \int_\Omega (\sigma_h - \sigma) \cdot \tau_h + \int_\Omega (u_h - \overline{u}_h) \operatorname{div} \tau_h.
\end{aligned}$$

Mit dem üblichen Skalierungsargument, d.h. Transformation auf das Referenzelement und Äquivalenz von Normen auf endlich dimensionalen Räumen, folgt andererseits, daß es eine Konstante c gibt, die nur von der Konstanten $c_{\mathcal{T}}$ in der Regu-

laritätsbedingung an $\mathcal{T}_h$ abhängt, mit

$$\begin{aligned}
&\|\hat{u}_h - u_h^*\|_0 \\
&\leq \sup_{\tau_h \in RT_h^{-1}} \frac{\left\{ \sum_{E \in \mathcal{E}_{h,\Omega}} \int_{\partial E} (\hat{u}_h - u_h^*)[n_E \cdot \tau_h]_E \right\}}{\left\{ \sum_{K \in \mathcal{T}_h} \left[h_K^{-2} \|\tau_h\|^2_{L^2(K)} + \|\operatorname{div} \tau_h\|^2_{L^2(K)} \right] \right\}^{1/2}}.
\end{aligned}$$

Damit folgt aus obiger Identität

$$\begin{aligned}
\|\hat{u}_h - u_h^*\|_0 \leq & c \left\{ \sum_{K \in \mathcal{T}_h} \left[h_K^2 \|\sigma - \sigma_h\|^2_{L^2(K)} + \|u_h - \overline{u}_h\|^2_{L^2(K)} \right] \right\}^{1/2} \\
\leq & c' \Big[h \|\sigma - \sigma_h\|_0 + \|u_h - \overline{u}_h\|_0 \Big].
\end{aligned}$$

Gemäß Satz 3.5 und Bem. 3.6 ist

$$h \|\sigma - \sigma_h\|_0 \leq c h^2 \Big[|f|_1 + \|u\|_2 \Big].$$

Wir wollen zeigen, daß gleiches für $\|u_h - \overline{u}_h\|_0$ gilt. Dazu benutzen wir ein Dualitätsargument und bezeichnen mit z die schwache Lösung von

$$\begin{aligned}
\Delta z = & \overline{u}_h - u_h \quad \text{in } \Omega \\
z = & 0 \qquad\qquad \text{auf } \Gamma
\end{aligned}$$

und setzen $\varphi := \nabla z$. Mit dem Interpolationsoperator I_h aus (7) folgt dann

$$\begin{aligned}
\|\overline{u}_h - u_h\|_0^2 &= \int_\Omega (\overline{u}_h - u_h) \operatorname{div} \varphi \\
&= \int_\Omega (\overline{u}_h - u_h) \operatorname{div}(I_h \varphi) \\
&= \int_\Omega (u - u_h) \operatorname{div}(I_h \varphi) \\
&= \int_\Omega (\sigma_h - \sigma) I_h \varphi \\
&= \int_\Omega (\sigma_h - \sigma)(I_h \varphi - \varphi) + \int_\Omega (\sigma_h - \sigma) \varphi \\
&= \int_\Omega (\sigma_h - \sigma)(I_h \varphi - \varphi) + \int_\Omega (\sigma_h - \sigma) \cdot \nabla z \\
&= \int_\Omega (\sigma_h - \sigma)(I_h \varphi - \varphi) - \int_\Omega \operatorname{div}(\sigma_h - \sigma) z \\
&= \int_\Omega (\sigma_h - \sigma)(I_h \varphi - \varphi) - \int_\Omega \operatorname{div}(\sigma_h - \sigma)(z - \pi_h z) \\
&\leq \|\sigma - \sigma_h\|_0 \|\varphi - I_h \varphi\|_0 + \|\operatorname{div}(\sigma - \sigma_h)\|_0 \|z - \pi_h z\|_0 \\
&\leq c h \Big\{ \|\sigma - \sigma_h\|_0 |\varphi|_1 + \|\operatorname{div}(\sigma - \sigma_h)\|_0 |z|_1 \Big\}.
\end{aligned}$$

Wegen

$$|z|_1 + |\varphi|_1 \leq c\|\overline{u}_h - u_h\|_0$$

und Satz 3.5 liefert dies

$$\begin{aligned}\|\overline{u}_h - u_h\|_0 \leq & ch\Big\{\|\sigma - \sigma_h\|_0 + \|\operatorname{div}(\sigma - \sigma_h)\|_0\Big\} \\ \leq & c'h^2\Big\{|f|_1 + \|u\|_2\Big\}.\end{aligned}$$

Dies beweist die Fehlerabschätzung des Satzes. □