

**Numerische Behandlung von  
Differentialgleichungen II  
Finite Elemente**

Vorlesungsskriptum SS 2006

R. Verfürth

Fakultät für Mathematik, Ruhr-Universität Bochum



## Inhaltsverzeichnis

Vorbemerkung	5
Kapitel I. Analytische Grundlagen	7
I.1. Sobolevräume	7
I.2. Abstrakte Variationsprobleme	16
I.3. Schwache Lösungen	20
Kapitel II. Finite Element Räume	27
II.1. Eindimensionale lineare Elemente	27
II.2. Bilineare Rechteckselemente	28
II.3. Lineare Dreieckselemente	37
II.4. Finite Elemente höherer Ordnung	51
Kapitel III. Praktische Aspekte	61
III.1. Randapproximation und numerische Integration	61
III.2. Numerische Lösung der diskreten Probleme	65
III.3. Fehlerschätzer und adaptive Gitterverfeinerung	80
III.4. Implementierung	97
Kapitel IV. Ergänzungen	105
IV.1. Nichtkonforme Methoden	105
IV.2. Gemischte Methoden	112
Zusammenfassung	127
Index	129



## Vorbemerkung

In den ersten drei Kapiteln dieser Vorlesung behandeln wir konforme Finite Element Verfahren für lineare elliptische Differentialgleichungen zweiter Ordnung. Die behandelten Methoden und einige weitere sind für den zweidimensionalen Fall, d.h. Differentialgleichungen in  $\mathbb{R}^2$ , und lineare Dreiecks- und bilineare Viereckselemente als Java-Demonstrationsapplet ALF (Adaptive Linear Finite elements) implementiert. Dieses Applet steht unter der Adresse

<http://www.rub.de/num1/demo/alf/ALF.html>

zur Verfügung. Eine Benutzeranleitung ist unter der Adresse

<http://www.rub.de/num1/demo/alf/ALFUserGuide.htm>

online verfügbar bzw. kann unter der Adresse

<http://www.rub.de/num1/demo/alf/ALFUserGuide.pdf>

als pdf-Datei heruntergeladen werden. Der Benutzer von ALF hat u.a. die Wahl zwischen folgenden Möglichkeiten:

- verschiedene Gebiete einschließlich krummlinig berandeter,
- unterschiedliche gröbste Unterteilungen,
- mehrere Differentialgleichungen u.a.
  - die Poissongleichung mit glatten und singulären Lösungen,
  - Reaktions-Diffusionsgleichungen insbesondere mit Lösungen mit inneren Grenzschichten,
  - Konvektions-Diffusionsgleichungen insbesondere mit Lösungen mit inneren oder Randgrenzschichten,
- verschiedene Optionen zur Berechnung der rechten Seite und der Steifigkeitsmatrix,
- unterschiedliche Löser u.a.
  - CG- und PCG-Verfahren,
  - verschiedene Varianten von Mehrgitterverfahren mit unterschiedlichen Zyklen und verschiedenen Glätttern,
- die Wahl zwischen uniformer Gitterverfeinerung und adaptiver Gitterverfeinerung basierend auf verschiedenen a posteriori Fehlerschätzern.



## KAPITEL I

# Analytische Grundlagen

### I.1. Sobolevräume

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  $|\alpha| = \alpha_1 + \dots + \alpha_d$  und

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

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

$$(I.1.1) \quad \int_{\Omega} \varphi D^\alpha \psi = (-1)^{|\alpha|} \int_{\Omega} \psi D^\alpha \varphi.$$

Gleichung (I.1.1) motiviert die folgende Definition der schwachen Ableitung. Dieser Begriff verallgemeinert denjenigen der klassischen Ableitung.

**DEFINITION I.1.1.** Seien  $\varphi, \psi \in L^1(\Omega)$  und  $\alpha \in \mathbb{N}^d$ . Dann heißt  $\psi$  die  $\alpha$ -TE SCHWACHE ABLEITUNG von  $\varphi$ , kurz  $\psi = D^\alpha \varphi$ , wenn für alle  $\rho \in C_0^\infty(\Omega)$  gilt

$$(I.1.2) \quad \int_{\Omega} \varphi D^\alpha \rho = (-1)^{|\alpha|} \int_{\Omega} \psi \rho.$$

**BEMERKUNG I.1.2.** (1) Die  $\alpha$ -te schwache Ableitung ist, sofern sie existiert, eindeutig (im Sinne von  $L^1$ -Funktionen).

(2) Ist  $\varphi \in C^{|\alpha|}(\Omega)$ , so stimmen die  $\alpha$ -te schwache Ableitung und die klassische  $\alpha$ -te Ableitung überein.

**BEWEIS.** AD (1): Seien  $\varphi, \psi_1, \psi_2 \in L^1(\Omega)$  mit

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

Dann gilt

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

Da  $C_0^\infty(\Omega)$  dicht ist in  $L^1(\Omega)$ , folgt  $\psi_1 = \psi_2$  fast überall.

AD (2): Folgt aus dem Gaußschen Integralsatz (vgl. (I.1.1)).  $\square$

BEISPIEL I.1.3. Sei  $\Omega = (-1, 1)$  und  $\varphi(x) = |x|$ . Dann ist  $\varphi$  im Sinne von Definition I.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.

Mit Hilfe der schwachen Ableitung definieren wir die Sobolevräume. Sie verallgemeinern die klassischen  $C^k(\Omega)$ -Räume und sind ein unverzichtbares Hilfsmittel für die Variationsrechnung (vgl. §§I.2, I.3), auf der die Finite Element Methoden aufbauen.

DEFINITION I.1.4. (1) Für  $k \in \mathbb{N}$  und  $p \in [1, \infty)$  definieren wir den SOBOLEVRAUM  $W^{k,p}(\Omega)$  und seine Norm  $\|\cdot\|_{k,p}$  durch

$$\begin{aligned} W^{k,p}(\Omega) &= \{\varphi \in L^p(\Omega) : D^\alpha \varphi \in L^p(\Omega) \forall |\alpha| \leq k\}, \\ \|\varphi\|_{k,p} &= \left\{ \sum_{|\alpha| \leq k} \|D^\alpha \varphi\|_{L^p(\Omega)}^p \right\}^{1/p}. \end{aligned}$$

(2) Durch

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

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.

BEISPIEL I.1.5. (1) Seien  $\varphi$  und  $\Omega$  wie in Beispiel I.1.3. Dann gilt  $\varphi \in W^{1,p}(\Omega)$  für alle  $p \in [1, \infty)$ .

(2) Seien  $d \geq 2$ ,  $\Omega = B(0, \frac{1}{2})$  und  $\varphi(x) = |x|^s$  mit  $s \in \mathbb{R}$ , wobei  $|\cdot|$  die euklidische Norm in  $\mathbb{R}^d$  bezeichnet. Dann gilt

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

und

$$\|D^\alpha \varphi\|_{L^p(\Omega)}^p \sim \omega_{d-1} \int_0^{\frac{1}{2}} r^{(s-|\alpha|)p} r^{d-1} dr$$

$$\begin{aligned}
&< \infty \\
&\iff p(s - |\alpha|) + d - 1 > -1 \\
&\iff s > |\alpha| - \frac{d}{p}.
\end{aligned}$$

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

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

Hieraus folgt

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

Also ist  $\varphi \in H^1(\Omega)$ .

SATZ I.1.6. (1)  $(W^{k,p}(\Omega), \|\cdot\|_{k,p})$  ist ein Banachraum.

(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  $W^{k,p}(\Omega)$  mittels der Abbildung  $i : \varphi \mapsto (D^\alpha \varphi)_{|\alpha| \leq k}$  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  $\|\cdot\|_{k,p}$ . Sei nun  $(\varphi_n)_{n \in \mathbb{N}} \subset W^{k,p}(\Omega)$  eine Cauchyfolge. Dann ist  $(i(\varphi_n))_{n \in \mathbb{N}} \subset L^p(\Omega; \mathbb{R}^{n_{k,d}})$  ebenfalls eine Cauchyfolge 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  $\psi_\alpha$  konvergiert. Insbesondere konvergiert  $D^\alpha \varphi_n$  punktweise f.ü. gegen  $\psi_\alpha$ . Für jedes  $\rho \in C_0^\infty(\Omega)$  gilt andererseits

$$(I.1.3) \quad \int_{\Omega} \varphi_n D^\alpha \rho = (-1)^{|\alpha|} \int_{\Omega} D^\alpha \varphi_n \rho.$$

Wegen des Lebesgueschen Konvergenzsatzes können wir in (I.1.3) den Grenzübergang  $n \rightarrow \infty$  durchführen und erhalten

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

$$= (-1)^{|\alpha|} \int_{\Omega} \psi_{\alpha} \rho.$$

Also ist  $\psi_{\alpha}$  die  $\alpha$ -te schwache Ableitung von  $\psi_0$ , und  $(\varphi_n)_{n \in \mathbb{N}}$  konvergiert in  $W^{k,p}(\Omega)$  gegen  $\psi_0$ .

AD (2): Kopiere den Beweis von „ $C^{\infty}(\Omega)$  ist dicht in  $L^p(\Omega)$ “.

AD (3): Offensichtlich ist  $(\cdot, \cdot)_k$  bilinear und  $\|\varphi\|_k^2 = (\varphi, \varphi)_k$ . Damit folgt die Behauptung aus Teil (1).  $\square$

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.

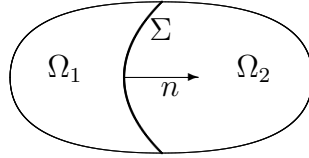


ABBILDUNG I.1.1. Zerlegung von  $\Omega$  in Teilgebiete

**SATZ I.1.7.** Seien  $\Omega_1, \Omega_2$  zwei nicht leere, offene, beschränkte und disjunkte Teilmengen von  $\Omega$  mit stückweise glattem Rand und  $\bar{\Omega} = \bar{\Omega}_1 \cup \bar{\Omega}_2$ . 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  $\Omega_1$  und  $[\varphi]$  der Sprung von  $\varphi$  über  $\Sigma = \partial\Omega_1 \cap \partial\Omega_2$  in Richtung  $n$ , d.h.

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

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

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

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

BEMERKUNG I.1.8. Gemäß Satz I.1.6 (2) ist der Raum  $C^\infty(\Omega)$  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-Schwarzschen Ungleichung

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

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

Bemerkung I.1.8 führt auf folgende Definition.

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

Wie wir aus §III.1 der Vorlesung „Numerische Behandlung von Differentialgleichungen I“ wissen, spielt der Rand von  $\Omega$  eine wesentliche Rolle für die Regularität der Lösungen partieller Differentialgleichungen. Dies gilt auch für das „Randverhalten“ von Funktionen in  $W^{k,p}$ -Räumen. Die folgenden Definition und Bemerkung präzisieren diesen Sachverhalt.

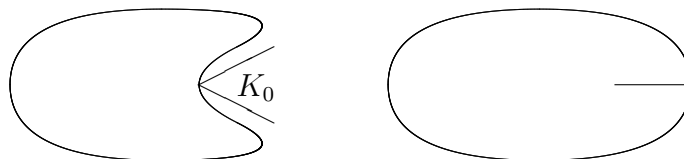


ABBILDUNG I.1.2. links: Lipschitz-Gebiet mit Kegel  $K_0$ ; rechts: nicht Lipschitz-Gebiet

DEFINITION I.1.10. Wir sagen,  $\Omega$  hat einen LIPSCHITZ-RAND bzw.  $\Omega$  ist ein LIPSCHITZ-GEBIET, wenn es ein  $N \in \mathbb{N}^*$  und offene Mengen  $U_1, \dots, U_N \in \mathbb{R}^d$  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.

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

(2)  $\Omega$  habe einen stückweise glatten Rand. Zudem gebe es zu jedem  $x_0 \in \partial\Omega$  einen nicht trivialen Kegel  $K_0$  mit Basis  $x_0$  und  $\Omega \subset \mathbb{R}^d \setminus K_0$  (KEGELBEDINGUNG). Dann ist  $\Omega$  ein Lipschitz-Gebiet.

Der folgende Satz charakterisiert die „Spuren“ von  $W^{k,p}$ -Funktionen, d.h. ihre Einschränkungen auf den Rand von  $\Omega$ .

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

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

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

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

BEMERKUNG I.1.13. Die Bezeichnungen und Voraussetzungen seien wie in Satz I.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 mit  $W^{k-l-\frac{1}{p},p}(\partial\Omega)$  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 I.1.4 mit  $\Omega$  ersetzt durch  $\partial\Omega$  verweisen wir auf das Buch von R.A. Adams.

Der folgende Satz charakterisiert  $W_0^{k,p}$ -Funktionen als die  $W^{k,p}$ -Funktionen, die auf dem Rand von  $\Omega$  „verschwinden“.

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

BEWEIS. R.A. Adams; loc. cit.

BEWEISIDEE: Da die  $\gamma_l$  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 I.1.9 die Behauptung.  $\square$

Der folgende Satz ist ein wichtiges Hilfsmittel.

SATZ I.1.15 (FRIEDRICHSCHES UNGLEICHUNG).  $\|\cdot\|_{k,p}$  und  $|\cdot|_{k,p}$  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  $\Omega$  in dem Würfel  $B_{|\cdot|_\infty}(0, R)$  enthalten ist. Dabei bezeichnet  $|\cdot|_\infty$  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  $\Omega \subset B_{|\cdot|_\infty}(0, R)$  folgt für beliebiges  $x \in \Omega$  mit der Hölderschen Ungleichung

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

Integration über  $\Omega$  liefert

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

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

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

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

**BEMERKUNG I.1.16.** Aus Satz I.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  $\Omega$  ab. Eine analoge Abschätzung gilt für alle Funktionen, die auf einem Teil des Randes verschwinden, der positives  $(d-1)$ -dimensionales Maß hat.

Als nächstes wollen wir Teilmengenbeziehungen zwischen  $W^{k,p}$ - und  $L^q$ -Räumen untersuchen. Dazu benötigen wir die Begriffe der stetigen und kompakten Einbettung.

DEFINITION I.1.17. Seien  $(X, \|\cdot\|_X)$  und  $(Y, \|\cdot\|_Y)$  zwei normierte Vektorräume.

- (1) Eine lineare Abbildung  $A : X \rightarrow Y$  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  $i : X \rightarrow Y$  stetig ist.
- (3)  $X$  ist **KOMPAKT EINGEBETTET** in  $Y$ , kurz  $X \overset{c}{\hookrightarrow} Y$ , wenn  $X \subset Y$  und die kanonische Injektion  $i : X \rightarrow Y$  kompakt ist.

BEMERKUNG I.1.18. (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  $A : X \rightarrow Y$  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  $i(\overline{B_{\|\cdot\|_X}(0;R)})$  mit  $R = \max_{n \in \mathbb{N}} \|\varphi_n\|_X$  enthalten.  $\square$

SATZ I.1.19 (SOBOLEVSCHER EINBETTUNGSSATZ). (1) Sei  $p < d$ . Dann gilt

$$W^{k,p}(\Omega) \hookrightarrow W^{k-1,q}(\Omega) \quad \forall q \in [1, \frac{pd}{d-p}]$$

und

$$W^{k,p}(\Omega) \overset{c}{\hookrightarrow} W^{k-1,q}(\Omega) \quad \forall q \in [1, \frac{pd}{d-p}).$$

(2) Sei  $p = d$ . Dann gilt

$$W^{k,p}(\Omega) \overset{c}{\hookrightarrow} W^{k-1,q}(\Omega) \quad \forall q \in [1, \infty).$$

(3) Sei  $k > \frac{d}{p}$ . Dann gilt

$$W^{k,p}(\Omega) \overset{c}{\hookrightarrow} C^l(\overline{\Omega}) \quad \forall l \in \mathbb{N}, 0 \leq l < k - \frac{d}{p}.$$

BEWEIS. R.A. Adams; loc cit.  $\square$

BEMERKUNG I.1.20. (1) Sei  $d = 2$ ,  $p = 2$  und  $\Omega = B(0; \frac{1}{2})$ . Dann zeigt Beispiel I.1.5 (3), dass die Schranke an  $q$  in Satz I.1.19 (2) scharf ist.

(2) Sei  $d \geq 3$ ,  $p = 2$  und  $\Omega = B(0; \frac{1}{2})$ . Dann zeigt Beispiel I.1.5 (2), dass die Schranke an  $q$  in Satz I.1.19 (1) scharf ist.

(3) Sei  $p = 2$  und  $d = 2$ . Dann ist  $H^1(\Omega) \xrightarrow{c} L^q(\Omega)$  für jedes  $q \in [1, \infty)$ .

(4) Sei  $p = 2$  und  $d = 3$ . Dann ist  $H^1(\Omega) \xrightarrow{c} L^q(\Omega)$  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  $H^2(\Omega) \hookrightarrow C^0(\bar{\Omega})$ . Für  $H^1(\Omega)$ -Funktionen sind Punktwerte dagegen nicht definiert.

Der folgende Satz ist ein Analogon der Friedrichschen Ungleichung.

SATZ I.1.21 (POINCARÉSCHE UNGLEICHUNG).  $|\cdot|_1$  und  $\|\cdot\|_1$  sind äquivalente Normen auf  $V = \{\varphi \in H^1(\Omega) : \int_{\Omega} \varphi = 0\}$ .

BEWEIS. Wie im Beweis von Satz I.1.15 müssen wir nur zeigen, dass es eine Konstante  $C > 0$  gibt mit

$$(I.1.4) \quad \|\varphi\|_1 \leq C|\varphi|_1 \quad \forall \varphi \in V.$$

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

$$(I.1.5) \quad \|\varphi_n\|_1 = 1 \quad \forall n \in \mathbb{N}$$

und

$$(I.1.6) \quad \lim_{n \rightarrow \infty} |\varphi_n|_1 = 0.$$

Wegen Satz I.1.19 und Bemerkung I.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

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

Wegen (I.1.6) 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  $\varphi$  konstant. Da  $V$  ein abgeschlossener Unterraum von  $H^1(\Omega)$  ist, gilt aber  $\int_{\Omega} \varphi = 0$ . Also ist  $\varphi = 0$  im Widerspruch zu (I.1.5).  $\square$

BEMERKUNG I.1.22. (1) Satz I.1.21 kann für  $H^1(\Omega)$  nicht gelten, da die rechte Seite von (I.1.4) für die konstante Funktion  $\varphi = 1$  verschwindet.

(2) Der Beweis von Satz I.1.21 ist nicht konstruktiv. Mit anderen Techniken kann man zeigen, dass die Konstante  $C$  in (I.1.4) proportional zum Durchmesser  $d = \sup_{x, y \in \Omega} |x - y|$  von  $\Omega$  ist. Ist insbesondere  $\Omega$  konvex,

ergibt sich  $C \leq \frac{d}{\pi}$ .

(3) Analoge Aussagen zu Satz I.1.21 gelten für die Räume

$$\{\varphi \in W^{1,p}(\Omega) : \int_{\Omega} \varphi = 0\}$$

mit  $p \in (1, \infty)$ .

## I.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.

Der folgende Satz sichert die eindeutige Lösbarkeit von Variationsproblemen und charakterisiert deren Lösungen.

**SATZ I.2.1 (SATZ VON LAX-MILGRAM).** *Seien  $(X, \|\cdot\|_X)$  ein Banachraum,  $l \in \mathcal{L}(X, \mathbb{R})$  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

$$(I.2.1) \quad J(u) = \frac{1}{2}a(u, u) - l(u)$$

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

$$(I.2.2) \quad a(u^*, v) = l(v) \quad \forall v \in X.$$

**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 (I.2.2).

2. SCHRITT: Seien  $u_1, u_2 \in X$  zwei Lösungen von (I.2.2). Dann folgt

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

und

$$\begin{aligned} \alpha \|u_1 - u_2\|_X^2 &\leq a(u_1 - u_2, u_1 - u_2) \\ &= 0. \end{aligned}$$

Also besitzt (I.2.2) 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.

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

Dann folgt für  $n, m \in \mathbb{N}$

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

Also ist  $(u_n)_{n \in \mathbb{N}}$  eine Cauchyfolge 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.  $\square$

Alle Finite Element Methoden beruhen auf dem Ansatz, die unendlich dimensionalen Räume  $X$  durch geeignete endlich dimensionale Räume  $X_h$  zu ersetzen. Die Methoden unterscheiden sich dann u.a. in der Wahl der diskreten Räume  $X_h$ . Satz I.2.1 angewandt auf einen solchen Raum  $X_h$  liefert sofort die eindeutige Lösbarkeit des diskreten Problems und führt es auf ein äquivalentes endlich dimensionales System linearer Gleichungen zurück. Der folgende Satz zeigt, dass der Fehler zwischen den Lösungen der Variationsprobleme in  $X$  und in  $X_h$  bestimmt ist durch die Approximationsgüte des Raumes  $X_h \subset X$ . Diese Größe ist unabhängig von den Formen  $a$  und  $l$  und damit von der speziellen Differentialgleichung. Daher erhalten wir mit dem folgenden Satz Fehlerabschätzungen für eine ganze Klasse von Differentialgleichungen und Diskretisierungen.

**SATZ I.2.2 (CÉA-LEMMA).** *Die Voraussetzungen und Bezeichnungen seien wie in Satz I.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*

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

BEWEIS. Wegen Satz I.2.1 besitzt  $J$  ein eindeutiges Minimum  $u_h$  in  $X_h$ . Dieses ist eindeutig charakterisiert durch

$$(I.2.3) \quad a(u_h, v_h) = l(v_h) \quad \forall v_h \in X_h.$$

Aus (I.2.2) und (I.2.3) folgt

$$(I.2.4) \quad a(u - u_h, v_h) = 0 \quad \forall v_h \in X_h.$$

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.  $\square$

Bei den Anwendungen der folgenden Abschnitte wird  $\|\cdot\|_X$  in der Regel die  $H^1$ -Norm oder eine ähnliche Norm sein. Häufig wollen wir aber auch Fehlerabschätzungen in der  $L^2$ -Norm oder einer vergleichbaren Norm herleiten. Der folgende Satz zeigt, dass wir dann unter bestimmten Bedingungen eine verbesserte Abschätzung erhalten.

SATZ I.2.3 (SATZ VON AUBIN-NITSCHKE). *Sei zusätzlich  $H$  ein Hilbertraum mit Skalarprodukt  $(\cdot, \cdot)_H$  und Norm  $\|\cdot\|_H$  derart, dass  $X \hookrightarrow H$  und bzgl.  $\|\cdot\|_H$  dicht ist in  $H$ . Für jedes  $\varphi \in H$  bezeichne  $u_\varphi \in X$  die eindeutige Lösung von*

$$(I.2.5) \quad a(v, u_\varphi) = (\varphi, v)_H \quad \forall v \in X.$$

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  $X \subset H$  definiert jedes  $\varphi \in H$  durch

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

ein stetiges lineares Funktional auf  $X$ . Wegen Satz I.2.1 besitzt somit (I.2.5) eine eindeutige Lösung  $u_\varphi \in X$ . Aus (I.2.5) und (I.2.4) 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 Fehlerabschätzung.  $\square$

Die Voraussetzungen der Sätze I.2.1 – I.2.3 sind insbesondere für Konvektions-Diffusionsgleichungen, die wir in den nächsten Paragraphen betrachten werden, zu restriktiv. Daher schwächen wir sie in den folgenden beiden Sätzen entsprechend ab.

SATZ I.2.4. Seien  $(X, \|\cdot\|_X)$  und  $(Y, \|\cdot\|_Y)$  Banachräume 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  $\bar{A} \in \mathbb{R}_+^*$  mit

$$(I.2.6) \quad a_1(u, v) \leq \bar{A} \|u\|_X \|v\|_Y \quad \forall u, v \in X.$$

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

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

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

$$(I.2.7) \quad a(u, u) > 0.$$

Dann besitzen die Probleme

$$(I.2.8) \quad a(u, v) = l(v) \quad \forall v \in X$$

und

$$(I.2.9) \quad a(v, u) = l(v) \quad \forall v \in X$$

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

BEWEIS. Wir beweisen die Behauptung nur für Problem (I.2.8). Der Beweis für das andere Problem ist völlig analog. Sei  $l \in \mathcal{L}(X, \mathbb{R})$  beliebig. Wegen Satz I.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 (I.2.8) äquivalent zu

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

Wiederum wegen Satz I.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  $w \mapsto u_w$  definiert eine lineare Abbildung  $K \in \mathcal{L}(X, X)$ , und (I.2.8) ist damit äquivalent zu

$$(I.2.10) \quad (Id + K)u = u_l.$$

Wegen  $X \overset{c}{\hookrightarrow} Y$  und (I.2.6) ist  $K$  kompakt. Daher erfüllt  $Id + K$  die Fredholm Alternative:

Entweder besitzt (I.2.10) für jede rechte Seite eine eindeutige Lösung oder das zugehörige homogene Problem besitzt eine nicht triviale Lösung  $u \neq 0$ .

Wegen (I.2.7) besitzt (I.2.8) mit  $l = 0$  und damit (I.2.10) mit  $u_l = 0$  aber nur die triviale Lösung.  $\square$

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

$$(I.2.11) \quad a(u_h, v_h) = l(v_h) \quad \forall v_h \in X_h$$

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  $a(u, u) \geq \beta \|u\|_X^2$  für alle  $u \in X$ . Dann gilt für die eindeutigen Lösungen  $u$  und  $u_h$  der Probleme (I.2.8) und (I.2.11) die Fehlerabschätzung

$$(I.2.12) \quad \|u - u_h\|_X \leq \frac{A}{\beta} \inf_{v_h \in X_h} \|u - v_h\|_X.$$

Dabei ist  $A = \|a\|_{\mathcal{L}^2(X, \mathbb{R})}$ . Seien schließlich  $H, \varphi$  und  $u_\varphi$  wie in Satz I.2.3. Dann gilt die Fehlerabschätzung

$$(I.2.13) \quad \|u - u_h\|_H \leq A \|u - u_h\|_X \sup_{\substack{\varphi \in H \\ \|\varphi\|_H=1}} \inf_{u_h \in X_h} \|u_\varphi - u_h\|_X.$$

BEWEIS. Die eindeutige Lösbarkeit von (I.2.11) folgt aus Satz I.2.4. Die Fehlerabschätzung (I.2.12) folgt wie im Beweis von Satz I.2.2. Man beachte, dass wir dort nur die zu (I.2.8) und (I.2.11) analogen Eigenschaften (I.2.2) und (I.2.3) ausgenutzt haben. Wegen  $X \hookrightarrow H$  definiert jedes  $\varphi \in H$  durch  $v \rightarrow (\varphi, v)_H$  ein stetiges lineares Funktional auf  $X$ . Daher folgt die Existenz und Eindeutigkeit von  $u_\varphi$  aus Satz I.2.4. Die Fehlerabschätzung (I.2.13) folgt dann wie im Beweis von Satz I.2.3.  $\square$

### I.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“)

$$(I.3.1) \quad - \sum_{1 \leq i, j \leq d} \frac{\partial}{\partial x_i} \left( A_{ij} \frac{\partial u}{\partial x_j} \right) + \sum_{i=1}^d a_i \frac{\partial u}{\partial x_i} + \alpha u = f \quad \text{in } \Omega.$$

Dabei ist  $f \in L^2(\Omega)$ ,  $\alpha \in C(\Omega, \mathbb{R}_+)$ ,  $a = (a_1, \dots, a_d) \in C^1(\Omega, \mathbb{R}^d)$  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

$$(I.3.2) \quad \lambda_0 = \inf_{x \in \Omega} \inf_{z \in \mathbb{R}^d} \frac{z^T A(x) z}{z^T z} > 0.$$

Später werden wir die Glattheitsbedingungen an die Koeffizienten  $\alpha, a$  und  $A$  abschwächen. Zur Vereinfachung der Notation sprechen wir im Folgenden von

- einer KONVEKTIONS-DIFFUSIONSGLEICHUNG, wenn  $\alpha, a$  und  $A$  beliebig sind,
- einer REAKTIONS-DIFFUSIONSGLEICHUNG, wenn  $a = 0$  ist,
- einer MEMBRANGLEICHUNG, wenn  $\alpha = 0$  und  $a = 0$  ist,
- einer POISSONGLEICHUNG, wenn  $\alpha = 0, a = 0$  und  $A = I$  ist.

Die partielle Differentialgleichung (I.3.1) muss mit Randbedingungen versehen werden. Wir betrachten drei Typen von Randbedingungen

- (homogene) DIRICHLET-RANDBEDINGUNGEN:  $u = 0$  auf  $\Gamma$ ,
- (inhomogene) NEUMANN-RANDBEDINGUNGEN:  $n^T A \nabla u = g$  auf  $\Gamma$ ,
- GEMISCHTE DIRICHLET-NEUMANN-RANDBEDINGUNGEN:  $u = 0$  auf  $\Gamma_D$  und  $n^T A \nabla u = g$  auf  $\Gamma_N$ .

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  $\Gamma_D$  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 (I.3.1) mit homogenen Dirichlet Randbedingungen und  $v \in C_0^\infty(\Omega)$ . Multiplikation von (I.3.1) mit  $v$ , Integration über  $\Omega$  und Anwenden des Gaußschen Integralsatzes liefert

$$\begin{aligned}
 \int_{\Omega} f v &= - \sum_{1 \leq i, j \leq d} \int_{\Omega} \frac{\partial}{\partial x_i} \left( A_{ij} \frac{\partial u}{\partial x_j} \right) v + \sum_{i=1}^d \int_{\Omega} \alpha_i \frac{\partial u}{\partial x_i} v + \int_{\Omega} \alpha u v \\
 \text{(I.3.3)} \quad &= \sum_{1 \leq i, j \leq d} \int_{\Omega} A_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} + \sum_{i=1}^d \int_{\Omega} a_i \frac{\partial u}{\partial x_i} v + \int_{\Omega} \alpha u v \\
 &= \int_{\Omega} \{ \nabla u^T A \nabla v + a \cdot \nabla u v + \alpha u v \}.
 \end{aligned}$$

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

$$\text{(I.3.4)} \quad \int_{\Omega} \{ \nabla u^T A \nabla v + a \cdot \nabla u v \} = \int_{\Omega} f v$$

für alle  $v \in H_0^1(\Omega)$  erfüllt. Umgekehrt folgt aus (I.3.3), dass eine Lösung von (I.3.4) die Differentialgleichung (I.3.1) erfüllt, sofern sie hinreichend glatt, d.h. in  $C^2(\Omega)$  ist. In diesem Sinne ist (I.3.4) zur Konvektions-Diffusionsgleichung (I.3.1) mit homogenen Dirichlet Randbedingungen äquivalent.

Betrachten wir in obigem Argument Funktionen  $v \in C^\infty(\overline{\Omega})$ , so treten in (I.3.3) zusätzlich Randterme  $-\int_{\Gamma} n^T A \nabla u v$  auf. Erfüllt  $u$  Neumann Randbedingungen, so gilt für diesen Randterm

$$- \int_{\Gamma} n^T A \nabla u v = - \int_{\Gamma} g v.$$

Wir werden daher in diesem Fall (I.3.4) durch den zusätzlichen Term  $\int_{\Gamma} gv$  auf der rechten Seite modifizieren. Diese Überlegungen führen auf folgende Definition.

DEFINITION I.3.1. (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

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

(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

$$\int_{\Omega} \{ \nabla u^T A \nabla v + a \cdot \nabla uv + \alpha uv \} = \int_{\Omega} fv + \int_{\Gamma_N} gv.$$

(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

$$\int_{\Omega} \{ \nabla u^T A \nabla v + a \cdot \nabla uv + \alpha uv \} = \int_{\Omega} fv + \int_{\Gamma} gv.$$

BEMERKUNG I.3.2. (1) Jede klassische Lösung von (I.3.1) ist auch eine schwache Lösung. Jede schwache Lösung, die zweimal stetig differenzierbar ist, ist eine klassische Lösung von (I.3.1).

(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  $\Gamma$  bzw.  $\Gamma_D$  muss in Definition I.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.

SATZ I.3.3 (EXISTENZ- UND EINDEUTIGKEITSSATZ FÜR SCHWACHE LÖSUNGEN). (1) Ist  $-\frac{1}{2} \operatorname{div} a + \alpha \geq 0$ , so besitzt die Konvektions-Diffusionsgleichung mit homogenen Dirichlet-Randbedingungen eine eindeutige schwache Lösung.

(2) Ist  $-\frac{1}{2} \operatorname{div} a + \alpha \geq 0$  und  $a \cdot n \geq 0$  auf  $\Gamma_N$ , so besitzt die Konvektions-Diffusionsgleichung mit gemischten Randbedingungen eine eindeutige schwache Lösung.

(3) Ist  $\alpha \geq a_0 > 0$ ,  $-\frac{1}{2} \operatorname{div} a + \alpha \geq 0$  und  $a \cdot n \geq 0$  auf  $\Gamma$ , so besitzt die Konvektions-Diffusionsgleichung mit Neumann-Randbedingungen eine eindeutig schwache Lösung.

(4) Ist  $\alpha = 0$ ,  $\operatorname{div} a = 0$  und  $a \cdot n = 0$  auf  $\Gamma$  sowie  $\int_{\Omega} f + \int_{\Gamma} g = 0$ , so besitzt die Konvektions-Diffusionsgleichung mit Neumann-Randbedingungen eine eindeutige schwache Lösung  $u$  mit  $\int_{\Omega} u = 0$ .

BEWEIS. Wir wenden jeweils Satz I.2.4 (S. 19) an.  
AD (1): Setze

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

Aus der Cauchy-Schwarzschen Ungleichung folgt

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

Wegen  $\alpha \geq 0$  und (I.3.2) ist

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

Wegen der Friedrichschen Ungleichung, Satz I.1.15 (S. 12), ist also  $a_0$  koerziv. Aus dem Gaußschen Integralsatz und (I.3.2) folgt schließlich

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

Also ist wegen  $-\frac{1}{2} \operatorname{div} a + \alpha \geq 0$  auch  $a$  koerziv.

AD (2): In diesem Fall ist

$$X = H_D^1(\Omega),$$

$$l(v) = \int_{\Omega} f v + \int_{\Gamma_N} g v.$$

Die anderen Größen ändern sich nicht. Aus der Cauchy-Schwarzschen Ungleichung und dem Spursatz, Satz [I.1.12](#) (S. [12](#)), folgt

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

so dass  $l$  stetig ist. Die Koerzivität von  $a_0$  bleibt wegen Bemerkung [I.1.16](#) (S. [13](#)) erhalten. Bei der Anwendung des Gaußschen Integralsatzes in der Abschätzung von  $a(u, u)$  tritt der zusätzliche Randterm  $\int_{\Gamma_N} n \cdot a u^2$  auf. Wegen  $n \cdot a \geq 0$  auf  $\Gamma_N$  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  $\Gamma$  an Stelle von  $\Gamma_N$ . 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  $a_0$ . Aus den Voraussetzungen und dem Gaußschen Integralsatz folgt weiter

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

Also gilt

$$a(u, u) = 0 \quad \iff \quad u = \text{konstant.}$$

Für konstantes  $u$  folgt aber

$$\begin{aligned} a(u, u) &= \int_{\Omega} \alpha u^2 \\ &\geq \alpha_0 \|u\|_0^2. \end{aligned}$$

Also ist  $a(u, u) > 0$  für alle  $u \in X \setminus \{0\}$ .

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 Poincaréschen Ungleichung, Satz [I.1.21](#) (S. [15](#)), folgt die Koerzivität von  $a_0$  auf

$$V = \{\varphi \in H^1(\Omega) : \int_{\Omega} \varphi = 0\}.$$

Wie in (3) folgt  $a(u, u) > 0$  für alle  $u \in V \setminus \{0\}$ . Wegen  $V \subset X$  sind  $a_0$ ,  $a_1$ ,  $a$  und  $l$  stetig auf  $V$ . Damit liefert Satz 1.2.4 (S. 19) die Existenz einer eindeutigen Lösung  $u^* \in V$  von

$$a(u^*, v) = l(v) \quad \forall v \in V.$$

Wir müssen noch zeigen, dass diese Gleichung sogar für alle  $v \in X$  gilt. Sei dazu  $v \in X$  beliebig. Setze

$$(I.3.5) \quad m_v = \frac{1}{|\Omega|} \int v, \quad \bar{v} = v - m_v.$$

Dann ist  $\bar{v} \in V$  und daher

$$\begin{aligned} a(u^*, v) - l(v) &= a(u^*, \bar{v}) - l(\bar{v}) + a(u^*, m_v) - l(m_v) \\ &= a(u^*, m_v) - l(m_v). \end{aligned}$$

Wegen der Kompatibilitätsbedingung an  $l$  ist

$$l(m_v) = 0.$$

Mit dem Gaußschen Integralsatz folgt

$$\begin{aligned} a(u^*, m_v) &= \int_{\Omega} (a \cdot \nabla u^*) m_v \\ &= - \int_{\Omega} (\operatorname{div} a) u^* m_v + \int_{\Gamma} a \cdot n u^* m_v \\ &= 0. \end{aligned}$$

Also ist  $u^*$  eine schwache Lösung der Konvektions-Diffusionsgleichung.  $\square$

**BEMERKUNG I.3.4.** (1) Im Fall der Reaktions-Diffusionsgleichung, d.h.  $a = 0$ , reduzieren sich die Voraussetzung von Satz 1.3.3 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  $\beta u + n^T A \nabla u = g_R$  auf  $\Gamma_R \subset \Gamma$  auf. In diesem Fall muss  $l$  durch  $\int_{\Gamma_R} g_R v$  und  $a_0$  durch  $\int_{\Gamma_R} \beta u v$  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“ kennen gelernt haben, zeigt, dass wir eine Regularitätsaussage der Form  $u \in H^2$  für schwache Lösungen nur unter zusätzlichen Annahmen an den Rand  $\Gamma$  erwarten können.

**BEISPIEL I.3.5.** Sei  $0 < \alpha < 2\pi$  und  $\Omega_{\alpha}$  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  $v : \Omega_{\alpha} \rightarrow \mathbb{R}$  durch

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

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

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

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

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

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

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

und

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

Wir können also bei gegebenem  $\alpha$  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.

**SATZ I.3.6 (REGULARITÄTSSATZ).** *Sei  $\Gamma$  eine  $C^1$ -Mannigfaltigkeit oder  $\Omega$  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-Diffusionsgleichung mit homogenen Dirichlet- oder gemischten oder Neumann-Randbedingungen die Regularitätsaussage  $u \in H^2(\Omega)$  und die a priori Abschätzung*

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

Die Konstante  $c$  hängt nur von  $\Omega$  und den Koeffizienten  $\alpha, a$  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.  $\square$

## KAPITEL II

### Finite Element Räume

#### II.1. 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-Reaktionsgleichung mit homogenen Dirichlet-Randbedingungen auf  $\Omega = (0, 1)$  kennen gelernt. Im Rahmen von §I.2 ist

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

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

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

und

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

Im Rahmen von Satz I.2.1 (S. 16) setzen wir

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

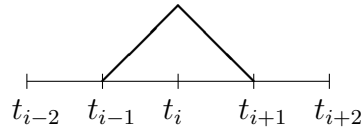
Eine Basis von  $X_h$  ist gegeben durch die Funktionen  $v_1, \dots, v_n$  mit

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

(vgl. Abbildung II.1.1). Das diskrete Problem

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

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

ABBILDUNG II.1.1. Basisfunktion  $v_i$ 

Aus Satz I.2.2 (S. 17) 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

$$(II.1.1) \quad \inf_{v_h \in X_h} \|u - v_h\|_1 \leq \sqrt{2} \inf_{v_h \in X_h} |u - v_h|_1 \leq \sqrt{2}h|u|_2$$

ist. Mit Satz I.2.3 (S. 18) 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 (II.1.1) beruhte auf der Friedrichschen Ungleichung angewandt auf den Interpolationsfehler  $u - I_h u$ . Dabei ist  $I_h u \in X_h$  bestimmt durch die Interpolationsbedingungen

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

Fassen wir noch einmal die wesentlichen Schritte zusammen:

- Konstruktion einer Unterteilung  $\mathcal{T}_h$  von  $\Omega$  in einfache Teilgebiete.
- Konstruktion eines Finite Elemente Raumes  $X_h$ , der aus „einfachen“ Funktionen auf den Teilgebieten in  $\mathcal{T}_h$  besteht.
- Konstruktion einer Basis von  $X_h$ , die aus Funktionen mit möglichst kleinem Träger besteht.
- Abschätzung des Interpolationsfehlers.

## II.2. Bilineare Rechteckselemente

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

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

Sei  $\mathcal{T}_h = \{K_i : 1 \leq i \leq m_h\}$  eine Zerlegung von  $\Omega$  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  $\Omega$ . Wir nummerieren zunächst die Punkte in  $\mathcal{N}_{h,\Omega}$  von 1 bis  $n_h$  und anschließend die Punkte in  $\mathcal{N}_h \setminus \mathcal{N}_{h,\Omega}$ , die alle auf dem Rand  $\Gamma$  liegen, von  $n_h + 1$  bis  $\bar{n}_h$  (vgl. Abbildung II.2.1).

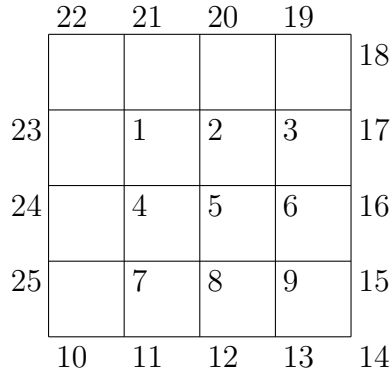


ABBILDUNG II.2.1. Unterteilung  $\mathcal{T}_h$  und Nummerierung der Punkte in  $\mathcal{N}_h$

Sei  $Q_1 = \text{span}\{1, x_1, x_2, x_1x_2\}$  der Raum aller Polynome in zwei Variablen, die in jeder Variablen höchstens den Grad 1 haben. Im Rahmen von Satz I.2.2 (S. 17) setzen wir

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

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 §I.1 durch

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

und

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

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  $\varphi_{z_0}$  und  $\psi_{z_0}$  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

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

Daher ist das diskrete Problem

$$(II.2.1) \quad a(u_h, v_h) = l(v_h) \quad \forall v_h \in X_h$$

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 (II.2.1) äquivalenten LGS die Werte von  $u_h$  in den Gitterpunkten sind, kann man (II.2.1) auch als ein Differenzenverfahren interpretieren. Dies gibt dann auch einen Einblick in die Struktur der Steifigkeitsmatrix.

Wir betrachten hierzu den einfachsten Spezialfall der Poissongleichung

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 Simpsonregel 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. Beispiel 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  $\nabla p$  in den Eckpunkten von  $K$  gegeben durch

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

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

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

$$\beta_{0,0} = \frac{8}{3}$$

$$\beta_{\mu,\nu} = -\frac{1}{3} \quad \forall (\mu, \nu) \neq (0, 0).$$

Dies liefert die gewünschte Darstellung der linken Seite von (II.2.1) 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 (II.2.1). Gemäß Satz I.2.2 (S. 17) kann der Fehler  $|u - u_h|_1$  durch den Approximationsfehler  $\inf_{v_h \in X_h} |u - v_h|_1$  abgeschätzt werden. Um diesen Approximationsfehler zu kontrollieren, betrachten wir wie in §II.1 einen geeigneten Interpolationsoperator  $I_h : C(\bar{\Omega}) \rightarrow X_h$ . Dieser ist definiert durch

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

oder äquivalent

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

**SATZ II.2.1.** *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  $i_h : C([0, h], \mathbb{R}) \rightarrow \mathbb{P}_1$  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

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

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

$$(II.2.2) \quad \varphi(t) = i_h \varphi(t) + \frac{h-t}{h} \int_0^t \varphi'(s) ds - \frac{t}{h} \int_t^h \varphi'(s) ds$$

$$(II.2.3) \quad = i_h \varphi(t) - \frac{h-t}{h} \int_0^t s \varphi''(s) ds - \frac{t}{h} \int_t^h (h-s) \varphi''(s) ds$$

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 (II.2.3) auf die Variable  $x$  liefert für alle  $(x, y) \in K$

$$\begin{aligned} v(x, y) &= (i_h v(\cdot, y))(x) \\ &\quad - \frac{h_1 - x}{h_1} \int_0^x s \frac{\partial^2}{\partial x^2} v(s, y) ds \\ &\quad - \frac{x}{h_1} \int_x^{h_1} (h_1 - s) \frac{\partial^2}{\partial x^2} v(s, y) ds. \end{aligned}$$

Wenden wir Gleichung (II.2.2) für festes  $x$  auf die Variable  $y$  und  $\varphi(y) = (i_h v(\cdot, y))(x)$  an, erhalten wir weiter

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

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

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

Differentiation bzgl.  $x$  liefert

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

mit

$$\begin{aligned}
K_1(t, y) &= \begin{cases} \frac{h_2 - y}{h_2} & \text{für } 0 \leq t < y \\ -\frac{y}{h_2} & \text{für } y < t \leq h_2 \end{cases} \\
K_2(s, x) &= \begin{cases} s & \text{für } 0 \leq s < x \\ -(h_1 - s) & \text{für } x < s \leq h_1. \end{cases}
\end{aligned}$$

Quadrieren dieser Identität und Anwenden der Cauchy-Schwarzschen Ungleichung ergibt

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

$$+ \frac{2}{h_1^2} \int_0^{h_1} |K_2(s, x)|^2 ds \int_0^{h_1} \left| \frac{\partial^2 v}{\partial x^2}(s, y) \right|^2 ds.$$

Integration über  $K$  liefert

$$\begin{aligned} & \int_K \left| \frac{\partial}{\partial x}(v - I_h v) \right|^2 \\ & \leq \frac{2}{h_1^2} \int_0^{h_1} \int_0^{h_2} \int_0^{h_1} \int_0^{h_2} |K_1(t, y)|^2 ds dt dx dy \left\| \frac{\partial^2 v}{\partial x \partial y} \right\|_{L^2(K)}^2 \\ & \quad + \frac{2}{h_1^2} \int_0^{h_1} \int_0^{h_1} |K_2(s, x)|^2 ds dx \left\| \frac{\partial^2 v}{\partial x^2} \right\|_{L^2(K)}^2. \end{aligned}$$

Offensichtlich ist

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

und

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

Insgesamt haben wir also die Abschätzung

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

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

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

Addition dieser beiden Abschätzungen und Summation über alle Rechtecke beweist wegen

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

und Definition I.1.9 (S. 11) die Behauptung.  $\square$

Aus Satz I.2.2 (S. 17), Satz I.2.3 (S. 18), Satz I.3.6 (S. 26) und Satz II.2.1 folgt unmittelbar die folgende Fehlerabschätzung.

**SATZ II.2.2.** *Seien  $u \in H_0^1(\Omega)$  die schwache Lösung der Reaktions-Diffusionsgleichung mit homogenen Dirichlet-Randbedingungen und  $u_h \in X_h$  die Lösung des diskreten Problems (II.2.1). Es sei  $u \in H^2(\Omega)$ . Dann gilt*

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

Ist  $\Omega$  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  $\Omega$  und den Koeffizienten  $\alpha$  und  $A$  ab.

**BEMERKUNG II.2.3.** (1) Bei inhomogenen Dirichlet-Randbedingungen  $u = u_D$  auf  $\Gamma$  sucht man die Lösung  $u_h$  von (II.2.1) 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 unbekanntem Koeffizienten  $\mu_z$  an. Satz II.2.2 bleibt gültig.

(2) Bei gemischten oder Neumann-Randbedingungen muss man analog zu Definition I.3.1 (S. 22) die rechte Seite von (II.2.1) durch  $\int_{\Gamma_N} g v$  ergänzen. Satz II.2.2 bleibt gültig.

(3) Unter den Voraussetzungen von Satz I.3.3 (S. 22) gilt Satz II.2.2 auch für Konvektions-Diffusionsgleichungen (die Bilinearform  $a$  muss natürlich wie in §I.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  $\lambda_0$  ist wie in Gleichung (I.3.2). 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.  $\|A\|_{L^\infty} \sim \lambda_0 \ll \|a\|_{L^\infty}$ , ist sie dagegen unbrauchbar. Außerdem entspricht die Diskretisierung in (II.2.1) einer zentralen Differenzdiskretisierung des Konvektionstermes  $a \cdot \nabla u$ . Daher treten bei der numerischen Lösung  $u_h$  unphysikalische 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 (II.2.1), die die genannten Schwierigkeiten vermeidet und die auch auf Rechteckselemente angewendet werden kann (vgl. (II.3.5) und Satz II.3.5 (S. 46) und folgende).

### II.3. Lineare Dreieckselemente

Im Folgenden bezeichnet  $|\cdot|$  die euklidische Norm auf dem  $\mathbb{R}^d$  und  $\|\cdot\|_{\mathcal{L}}$  die zugehörige Matrixnorm.  $\Omega \subset \mathbb{R}^2$  ist ein offenes, beschränktes, zusammenhängendes Gebiet mit polygonalem Rand  $\Gamma$ . Wie in §II.2 betrachten wir zunächst die Reaktions-Diffusionsgleichung mit homogenen Dirichlet-Randbedingungen und setzen im Rahmen von §I.2

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

Sei  $\mathcal{T}_h = \{K_i : 1 \leq i \leq m_h\}$  eine Unterteilung von  $\Omega$  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  $\rho_K$  den Durchmesser des größten in  $K$  eingeschriebenen Kreises. Die Dreiecke sollen vergleichbare Größe haben, d.h.  $\frac{h_K}{\rho_K}$  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  $z_{K,1}, \dots, z_{K,3}$ . Dabei soll aus praktischen Gründen die Nummerierung 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  $\Omega$ . Wir numerieren die Punkte in  $\mathcal{N}_{h,\Omega}$  zunächst von 1 bis  $n_h$  und anschließend die Punkte in  $\mathcal{N}_h \setminus \mathcal{N}_{h,\Omega}$ , die auf dem Rand  $\Gamma$  liegen, von  $n_h + 1$  bis  $\bar{n}_h$ .

Im Rahmen von Satz I.2.2 (S. 17) setzen wir

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

Da eine Ebene im  $\mathbb{R}^3$  durch drei nicht kollineare 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

$$(II.3.1) \quad \lambda_{K,i}(x) = \frac{\det(z_{K,i+1} - x, z_{K,i+2} - z_{K,i+1})}{\det(z_{K,i+1} - z_{K,i}, z_{K,i+2} - z_{K,i})}.$$

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

$$(II.3.2) \quad \begin{aligned} \lambda_{K,i} &\in \mathbb{P}_1, \\ \lambda_{K,i}(z_{K,j}) &= \delta_{ij}, \\ \lambda_{K,1} + \lambda_{K,2} + \lambda_{K,3} &= 1. \end{aligned}$$

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

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

Man beachte, dass  $\omega_z$  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

$$(II.3.3) \quad a(u_h, v_h) = l(u_h) \quad \forall v_h \in X_h$$

äquivalent zu einem linearen Gleichungssystem mit  $n_h$  Gleichungen und Unbekannten. Wie in §II.2 ist die Steifigkeitsmatrix symmetrisch, positiv definit und dünn besetzt.

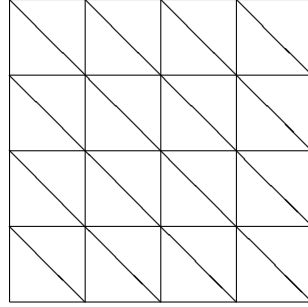


ABBILDUNG II.3.1. Courant-Triangulierung

Da die Unbekannten des zu (II.3.3) äquivalenten LGS die Werte von  $u_h$  in den Gitterpunkten  $\mathcal{N}_h$  sind, kann (II.3.3) 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 (vgl. Abbildung II.3.1). Diese besteht aus gleichschenkelig 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

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

Daher ist

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

Hieraus folgt mit leichter Rechnung, dass die linke Seite von (II.3.3) 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 (II.3.3) definieren wir wie in §II.2 einen Interpolationsoperator  $I_h$  :

$C(\bar{\Omega}) \rightarrow X_h$  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

$$(II.3.4) \quad \begin{aligned} F_K : \mathbb{R}^2 &\rightarrow \mathbb{R}^2 \\ \hat{x} &\mapsto B_K \hat{x} + b_K \end{aligned}$$

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.

**SATZ II.3.1.** *Bezeichne mit  $\hat{\pi} : C(\hat{K}) \rightarrow \mathbb{P}_1$  den linearen Interpolationsoperator zu den Eckpunkten von  $\hat{K}$ . Dann gilt für alle  $v \in H^2(\hat{K})$  die Fehlerabschätzung*

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

**BEWEIS.** Gemäß Bemerkung I.1.20(5) (S. 15) 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 I.1.6(2) (S. 9)  $C^\infty(\hat{K})$  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

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

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

Quadrieren und Integrieren dieser Gleichung liefert wegen der Cauchy-Schwarzschen Ungleichung für endliche Summen die Abschätzung

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

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

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

$$\begin{aligned}
\int_{\hat{K}} |I_2|^2 &\leq \int_{\hat{K}} \left\{ \int_0^x \int_0^y dt ds \right\} \left\{ \int_0^x \int_0^y \left| \frac{\partial^2 v}{\partial x \partial y}(s, t) \right|^2 dt ds \right\} \\
&= \int_0^1 \int_0^{1-x} \left\{ xy \int_0^x \int_0^y \left| \frac{\partial^2 v}{\partial x \partial y}(s, t) \right|^2 dt ds \right\} dy dx \\
&\leq \int_0^1 \int_0^{1-x} \int_0^1 \int_0^{1-x} \left| \frac{\partial^2 v}{\partial x \partial y}(s, t) \right|^2 dt ds dy dx \\
&\leq \frac{1}{2} \left\| \frac{\partial^2 v}{\partial x \partial y} \right\|_{L^2(\hat{K})}^2 \\
\int_{\hat{K}} |I_3|^2 &\leq \int_{\hat{K}} \left\{ \int_x^1 \left| \int_{1-s}^y dt \right| ds \right\} \left\{ \int_x^1 \left| \int_{1-s}^y \left| \frac{\partial^2 v}{\partial x \partial y}(x, t) \right|^2 dt \right| ds \right\} \\
&\leq \int_{\hat{K}} \left\{ \int_x^1 \int_0^{1-x} \left| \frac{\partial^2 v}{\partial x \partial y}(x, t) \right|^2 dt ds \right\} \\
&= \int_0^1 \int_0^{1-x} \int_0^1 \int_0^{1-x} \left| \frac{\partial^2 v}{\partial x \partial y}(x, t) \right|^2 dt ds dy dx \\
&\leq \frac{1}{2} \left\| \frac{\partial^2 v}{\partial x \partial y} \right\|_{L^2(\hat{K})}^2 \\
\int_{\hat{K}} |I_4|^2 &\leq \int_{\hat{K}} \left\{ \int_x^1 \int_x^s d\sigma ds \right\} \left\{ \int_x^1 \int_x^s \left| \frac{\partial^2 v}{\partial x^2}(\sigma, 1-s) \right|^2 d\sigma ds \right\} \\
&= \int_{\hat{K}} \left\{ \frac{1}{2} (1-x)^2 \int_0^{1-x} \int_x^{1-t} \left| \frac{\partial^2 v}{\partial x^2}(\sigma, t) \right|^2 d\sigma dt \right\} \\
&\leq \frac{1}{2} \int_{\hat{K}} \int_0^1 \int_0^{1-t} \left| \frac{\partial^2 v}{\partial x^2}(\sigma, t) \right|^2 d\sigma dt \\
&\leq \frac{1}{4} \left\| \frac{\partial^2 v}{\partial x^2} \right\|_{L^2(\hat{K})}^2 \\
\int_{\hat{K}} |I_5|^2 &\leq \int_{\hat{K}} \left\{ \int_x^1 \int_0^{1-s} dt ds \right\} \left\{ \int_x^1 \int_0^{1-s} \left| \frac{\partial^2 v}{\partial x \partial y}(s, t) \right|^2 dt ds \right\} \\
&= \int_{\hat{K}} \left\{ \frac{1}{2} (1-x)^2 \int_x^1 \int_0^{1-s} \left| \frac{\partial^2 v}{\partial x \partial y}(s, t) \right|^2 dt ds \right\} \\
&\leq \frac{1}{2} \int_{\hat{K}} \int_0^1 \int_0^{1-s} \left| \frac{\partial^2 v}{\partial x \partial y}(s, t) \right|^2 dt ds \\
&= \frac{1}{4} \left\| \frac{\partial^2 v}{\partial x \partial y} \right\|_{L^2(\hat{K})}^2.
\end{aligned}$$

Insgesamt ergibt sich

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

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

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

Hieraus folgt die Behauptung.  $\square$

Als nächstes schätzen wir die Spektralnormen  $\|B_K\|_{\mathcal{L}}$  und  $\|B_K^{-1}\|_{\mathcal{L}}$  der Transformationsmatrizen in (II.3.4) ab.

SATZ II.3.2. *Für jedes  $K \in \mathcal{T}_h$  gilt für die Matrix  $B_K$  aus (II.3.4)*

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

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  $F_K : \hat{K} \rightarrow K$  bijektiv ist, folgt

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

Also ist

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

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

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

$\square$

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

SATZ II.3.3. (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  $F_K : \hat{K} \rightarrow K$  aus (II.3.4) 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

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

Ebenso ergibt sich

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

Aus diesen Abschätzungen und den Sätzen II.3.1 und II.3.2 folgt

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

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  $c_{\mathcal{T}} = \max_{K \in \mathcal{T}_h} \frac{h_K}{\rho_K}$  aus (1) durch Quadrieren und Aufsummieren.  $\square$

Aus Satz I.2.2 (S. 17), Satz I.2.3 (S. 18), Satz I.3.6 (S. 26) und Satz II.3.3 folgt unmittelbar die folgende Fehlerabschätzung.

SATZ II.3.4. *Seien  $u \in H_0^1(\Omega)$  die schwache Lösung der Reaktions-Diffusionsgleichung mit homogenen Dirichlet-Randbedingungen und  $u_h \in X_h$  die Lösung von (II.3.4). Es sei  $u \in H^2(\Omega)$ . Dann gilt die Fehlerabschätzung*

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

*Ist zusätzlich  $\Omega$  konvex, so ist*

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

*Die Konstanten  $c_1$  und  $c_2$  hängen nur von  $c_T$ ,  $\Omega$  und den Koeffizienten  $\alpha$  und  $A$  ab.*

Bemerkung II.2.3 (S. 36) zur Behandlung von inhomogenen Dirichlet-Randbedingungen, von gemischten und Neumann-Randbedingungen und von Konvektions-Diffusionsgleichungen 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 Probleme 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  $A = \varepsilon I$ ,  $a \in \mathbb{R}^2 \setminus \{0\}$ ,  $\alpha = 0$  und homogener Dirichlet-Randbedingungen. Dabei ist die Skalierung so gewählt, dass

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

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

Im Rahmen von Satz I.2.1 (S. 16) 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 (II.3.3). Wir definieren auf  $X_h$  eine gitterabhängige Bilinearform  $a_h$ , eine Linearform  $l_h$  und eine Norm  $|\cdot|_{1,h}$

durch

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

Das neue diskrete Problem lautet dann

$$(II.3.5) \quad a_h(u_h, v_h) = l_h(v_h) \quad \forall v_h \in X_h.$$

Dabei sind die  $\delta_K$  nicht negative Parameter, die wir später bestimmen werden. Die  $\delta_K$ -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.

**SATZ II.3.5.** *Die Bilinearform  $a_h$  ist koerziv bzgl.  $|\cdot|_{1,h}$ , 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 I.3.6(1) (S. 26) 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  $|\cdot|_{1,h}$  folgt sofort die Behauptung.  $\square$

**BEMERKUNG II.3.6.** (1) Aus Satz II.3.5 folgt, dass das diskrete Problem (II.3.5) eindeutig lösbar ist.

(2) Satz II.3.5 benötigt außer der Annahme  $\delta_K \geq 0$  keine weiteren Voraussetzungen an die  $\delta_K$ .

(3) Im allgemeinen Fall bleibt Satz II.3.5 gültig, wenn wie in Satz I.3.6 (S. 26)  $-\frac{1}{2} \operatorname{div} a + \alpha \geq 0$  und  $h_K^{-1} \|A\|_{L^\infty(K)} \delta_K^{1/2} \leq 1$  ist.

Für die Fehlerabschätzung der Diskretisierung (II.3.5) benötigen wir eine Abschätzung der  $L^2$ -Norm des Interpolationsfehlers.

SATZ II.3.7. (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 II.3.1 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

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

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

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

Mit der Cauchy-Schwarzschen Ungleichung für Integrale erhalten wir für die ersten beiden Summanden

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

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

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

Aus diesen Abschätzungen und dem Beweis von Satz II.3.1 folgt insgesamt

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

AD (2): Wir gehen wir im Beweis von Satz II.3.3 vor und erhalten mit Teil (1) und  $\hat{v} = v \circ F_K$

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

□

Nach diesen Vorbereitungen können wir jetzt eine Fehlerabschätzung für die Diskretisierung (II.3.5) beweisen.

SATZ II.3.8. Sei  $u \in H_0^1(\Omega)$  die Lösung der Konvektions-Diffusionsgleichung 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 (II.3.5). Es sei  $u \in H^2(\Omega)$ . Dann gilt die Fehlerabschätzung

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

Dabei ist

$$c_{\mathcal{T}} = \max_{K \in \mathcal{T}_h} \frac{h_K}{\rho_K}.$$

Die Fehlerabschätzung ist optimal für die Wahl

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

BEWEIS. Wegen der Dreiecksungleichung gilt

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

Aus Satz II.3.3 folgt

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

Setze zur Abkürzung  $w_h = u_h - I_h u$ . Aus Satz II.3.5 folgt

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

Da  $u \in H^2(\Omega)$  ist, folgt aus der Definition von  $a_h, l_h$  und (II.3.5)

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

Hierbei haben wir die Cauchy-Schwarzsche Ungleichung zunächst für Integrale und danach für endliche Summen ausgenutzt. Mittels partieller Integration für den Konvektionsterm erhalten wir mit Satz II.3.3

und Satz II.3.7 weiterhin

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

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

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

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  $\delta_K$ -Terme und die  $\delta_K^{-1}$ -Terme gleich sind. Hieraus folgt die Behauptung über die optimale Wahl von  $\delta_K$ .  $\square$

**BEMERKUNG II.3.9.** Es gilt  $\delta_K \sim h_K$ , wenn  $\varepsilon \ll h_K$  ist. In diesem Fall liefert Satz II.3.8 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 (II.3.3) gut funktioniert, ist  $\delta_K \sim h_K^2 \varepsilon^{-1}$ , und Satz II.3.8 liefert eine Fehlerabschätzung der Form

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

die vergleichbar ist zu derjenigen von Satz II.3.4.

## II.4. Finite Elemente höherer Ordnung

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

Die Grundidee ist folgende:

- Unterteile  $\Omega$  in Teilgebiete  $K_1, \dots, K_{m_h}$  mit einfacher geometrischer Struktur. In §II.2 waren dies Rechtecke, in §II.3 Dreiecke.
- Approximiere die Sobolevräume  $W^{k,p}(\Omega)$  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 §II.2 waren dies die bilinearen Polynome in  $Q_1$ , in §II.3 die linearen Polynome in  $P_1$ .
- Konstruiere eine Basis von  $X_h$ , so dass jede Basisfunktion eine einfache Struktur und einen kleinen Träger hat. In §§II.2 und II.3 waren die Basisfunktionen die stetigen, stückweise bilinearen bzw. linearen Funktionen, die in einem Elementeckpunkt den Wert 1 haben und in allen anderen Elementeckpunkten verschwinden.
- Um den Fehler der Finite Element Diskretisierung abzuschätzen, konstruiere einen einfachen Interpolationsoperator  $I_h : W^{k,p} \rightarrow X_h$  und schätze den Interpolationsfehler ab. Letzteres geschah in §II.3 zunächst auf einem Referenzelement.

Die 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  $F_{K_i} : \hat{K} \rightarrow K_i$  eine einfache Gestalt hat. Wie in der Praxis üblich, werden wir zwei Typen von Referenzelementen betrachten:

- den REFERENZ D-SIMPLEX  $\hat{K} = \{\hat{x} \in \mathbb{R}^d : x_1 + \dots + x_d \leq 1, x_i \geq 0, 1 \leq i \leq d\}$  und
- 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  $\Omega$ , die folgende Bedingungen erfüllt:

- AFFINE ÄQUIVALENZ: zu jedem  $K \in \mathcal{T}_h$  gibt es einen affinen Diffeomorphismus  $F_K$  des Referenz-Simplexes oder -Würfels  $\hat{K}$  auf  $K$ .
- 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.
- REGULARITÄT: Der Quotient  $h_K/\rho_K$  ist durch eine Konstante  $c_{\mathcal{T}}$ , die nicht von  $K$  oder  $h$  abhängt, nach oben beschränkt.

Dabei ist ähnlich wie in §II.3  $h_K$  der Durchmesser von  $K$  und  $\rho_K$  der Durchmesser des größten, in  $K$  eingeschriebenen  $d$ -Balles.

Wie in den §§II.2 und II.3 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  $\Omega$ .

DEFINITION II.4.1. (1) Bezeichne die Eckpunkte des Referenz-Simplexes mit  $\hat{z}_1 = e_1, \dots, \hat{z}_d = e_d$  und  $\hat{z}_{d+1} = 0$ . Dann ist für  $k \in \mathbb{N}^*$

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

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

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

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) Setze

$$\mathcal{G}_h = \bigcup_{K \in \mathcal{T}_h} \Sigma_k(K), \quad \mathcal{G}_{h,\Omega} = \mathcal{G}_h \cap \Omega.$$

(4) Setze  $Q_0 = \mathbb{P}_0 = \mathbb{R}$  und definiere für  $k \in \mathbb{N}^*$

$$Q_k = \text{span}\{x^\alpha : \alpha \in \mathbb{N}^d, \max_{1 \leq i \leq d} \alpha_i \leq k\}$$

$$\mathbb{P}_k = \text{span}\{x^\alpha : \alpha \in \mathbb{N}^d, \alpha_1 + \dots + \alpha_d \leq k\}.$$

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

$$R_k = R_k(\hat{K}) = \begin{cases} Q_k & \text{falls } \hat{K} \text{ der Referenz-Würfel,} \\ \mathbb{P}_k & \text{falls } \hat{K} \text{ der Referenz-Simplex.} \end{cases}$$

BEMERKUNG II.4.2. (1) Sei  $K$  ein allgemeiner Simplex. Bezeichne die Eckpunkte von  $K$  mit  $z_1, \dots, z_{d+1}$ . Dann ist

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

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 (vgl. Abbildung II.4.1).

(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  $\varphi \mapsto \varphi \circ F_K$  und  $\psi \mapsto \psi \circ F_K^{-1}$  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.

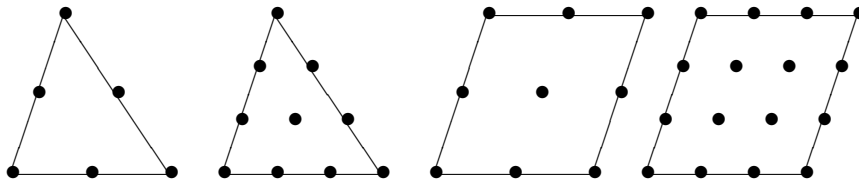


ABBILDUNG II.4.1. Die Mengen  $\Sigma_2$  und  $\Sigma_3$  für ein Dreieck und ein Parallelogramm

SATZ II.4.3. 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 Bemerkung II.4.2 (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  $\mu_i$  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  $i' \in \mathbb{N}_k^d \setminus \{i\}$ . 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 Funktionen  $\hat{\lambda}_1, \dots, \hat{\lambda}_{d+1}$  durch

$$\begin{aligned} \hat{\lambda}_i(x) &= x_i \quad 1 \leq i \leq d \\ \hat{\lambda}_{d+1}(x) &= 1 - \sum_{i=1}^d x_i. \end{aligned}$$

Die Funktionen  $\hat{\lambda}_1, \dots, \hat{\lambda}_{d+1}$  heißen die SCHWERPUNKTSKOORDINATEN von  $\hat{K}$  (vgl. §II.3) und haben offensichtlich die Eigenschaft  $\hat{\lambda}_i \in \mathbb{P}_1$ ,  $1 \leq i \leq d+1$ ,  $\hat{\lambda}_i(\hat{z}_j) = \delta_{ij}$ ,  $1 \leq i, j \leq d+1$ .

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

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

das in Eigenschaft (A) Geforderte.

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

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

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

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

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

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

das in Eigenschaft (A) Geforderte.

FALL  $k = 3$ : Definiere

$$\begin{aligned}\hat{z}_{ij} &= \frac{1}{3}(2\hat{z}_i + \hat{z}_j) \quad 1 \leq i, j \leq d+1, \quad j \neq i \\ \hat{z}_{ijk} &= \frac{1}{3}(\hat{z}_i + \hat{z}_j + \hat{z}_k) \quad 1 \leq i < j < k \leq d+1\end{aligned}$$

und

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

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

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

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

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

das in Eigenschaft (A) Geforderte.

FALL  $k \geq 4$ : Sei  $\varphi \in \mathbb{P}_k$  mit  $\varphi(z) = 0$  für alle  $z \in \hat{\Sigma}_k$ . Dann verschwindet  $\varphi$  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

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

Damit folgt die Eigenschaft (B) durch Induktion über  $k$ .  $\square$

**SATZ II.4.4.** 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  $\varphi$  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. „ $\implies$ “: Ist offensichtlich.

„ $\impliedby$ “: 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 Bemerkung II.4.2 (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  $\varphi$  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 II.4.3, dass  $q = 0$  und damit  $\varphi$  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.  $\square$

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

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

Wegen Satz I.1.7 (S. 10) 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 I.2.2 (S. 17)  $X_h = S_{h,0}^{k,0}$  mit beliebigem  $k \in \mathbb{N}^*$  wählen und dadurch die diskreten Probleme der §§II.2 und II.3 verallgemeinern. Aus den Sätzen II.4.3 und II.4.4 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  $z' \in \mathcal{G}_h \setminus \{z\}$ . 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  $\mu_z$  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 §§II.2 und II.3 definieren wir einen Interpolationsoperator  $I_h : H^2(\Omega) \rightarrow S_h^{k,0}$  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 §II.3 geschieht die Abschätzung des Interpolationsfehlers zunächst auf dem Referenzelement. Dazu definieren wir  $\hat{\Pi}_k : H^2(\hat{K}) \rightarrow R_k$  durch

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

Wegen Satz II.4.3 und Bemerkung I.1.20 (S. 15) ist diese Definition sinnvoll. Aus Bemerkung II.4.2 (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.

**SATZ II.4.5.** *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  $\|\cdot\|_{k+1}$  äquivalent ist.

**BEWEIS.** Da gemäß Satz I.1.19 (S. 14)  $H^{k+1}(\hat{K}) \hookrightarrow C(\hat{K})$  ist, ist  $[\cdot]_{k+1}$  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

$$(II.4.1) \quad \|u_n\|_{k+1} = 1 \quad \forall n \in \mathbb{N}$$

und

$$(II.4.2) \quad \lim_{n \rightarrow \infty} [u_n]_{k+1} = 0.$$

Wegen Satz I.1.19 (S. 14) und Bemerkung I.1.18 (S. 14) 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

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

Wegen (II.4.2) ist insbesondere

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

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

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

Wegen  $|u|_{k+1} = 0$  ist  $u \in \mathbb{P}_k$ . Wegen Satz I.1.19 (S. 14) gilt

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

Hieraus und aus (II.4.2) folgt aber

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

Wegen Satz II.4.3 ist also  $u = 0$  im Widerspruch zu (II.4.1).  $\square$

**SATZ II.4.6.** *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 II.4.5 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 II.4.3 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.

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

Dann folgt für beliebiges  $\varphi \in H^{k+1}(\hat{K})$  mit Satz I.1.19 (S. 14)

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

Aus dem soeben Gezeigten und dem Fall 1 folgt

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

□

SATZ II.4.7. 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  $\Omega$ ,  $k$  und der Konstanten  $c_{\mathcal{T}}$  in der Regularitätsbedingung an  $\mathcal{T}_h$  ab.

BEWEIS. Sei  $K \in \mathcal{T}_h$  und  $F_K : x \rightarrow b_K + B_K x$  eine affine Transformation des Referenzelements  $\hat{K}$  auf  $K$ . Wie im Beweis von Satz II.3.3 (S. 43) erhalten wir

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

Wie im Beweis von Satz II.3.2 (S. 43) folgt

$$\|B_K\|_{\mathcal{L}} \leq \frac{h_K}{\rho_{\hat{K}}}, \quad \|B_K^{-1}\|_{\mathcal{L}} \leq \frac{h_{\hat{K}}}{\rho_K}.$$

Aus diesen Abschätzungen und Satz II.4.6 ergibt sich wegen  $I_h u \circ F_K = \hat{\Pi}_k(u \circ F_K)$

$$\|u - I_h u\|_{L^2(K)} \leq c \left( \frac{h_K}{\rho_{\hat{K}}} \right)^{k+1} |u|_{H^{k+1}(K)},$$

$$\begin{aligned}
|u - I_h u|_{H^1(K)} &\leq c \frac{h_{\hat{K}}}{\rho_K} \left(\frac{h_K}{\rho_{\hat{K}}}\right)^{k+1} |u|_{H^{k+1}(K)} \\
&= ch_{\hat{K}} \rho_{\hat{K}}^{-k-1} \left(\frac{h_K}{\rho_K}\right) h_K^k |u|_{H^{k+1}(K)}.
\end{aligned}$$

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

BEMERKUNG II.4.8. Die Ergebnisse der Sätze II.4.5 und II.4.7 gelten mit den offensichtlichen Modifikationen für alle  $W^{k+1,p}$ -Räume mit  $1 \leq p < \infty$ .

## KAPITEL III

### Praktische Aspekte

#### III.1. Randapproximation und numerische Integration

In den §§II.2 – II.4 haben wir stets vorausgesetzt, dass  $\Omega$  ein Polyedergebiet ist, d.h. der Rand  $\Gamma$  besteht stückweise aus Hyperebenen. I.a. ist jedoch der Rand von  $\Omega$  gekrümmt. Daher ersetzt man in der Praxis  $\Omega$  durch ein approximierendes Polyeder  $\Omega_h$ , derart dass die Eckpunkte von  $\Omega_h$  auf  $\Gamma$  liegen und die Kanten von  $\Omega_h$  die Länge  $O(h)$  haben (vgl. Abbildung III.1.1). Sei  $\Gamma_h$  der Rand von  $\Omega_h$ . Wenn  $\Gamma$  stückweise  $C^2$  ist, folgt dann

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

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

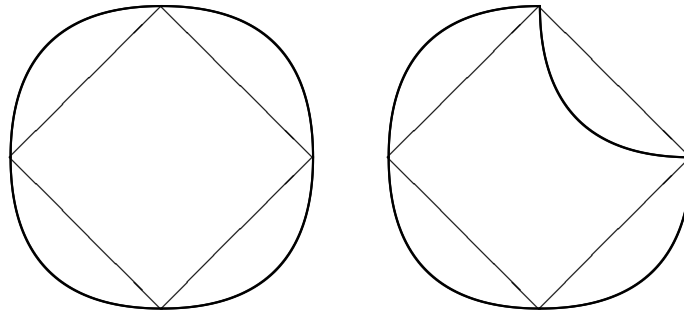


ABBILDUNG III.1.1. Randapproximation eines konvexen und eines nicht konvexen Gebietes

$\mathcal{T}_h$  ist dann eine Unterteilung von  $\Omega_h$ , die den Bedingungen der §§II.2 – II.4 genügt. Zusätzlich müssen die Eckpunkte von  $\Omega_h$  in der Menge  $\mathcal{N}_h$  der Elementeckpunkte enthalten sein. Für die Konstruktion von  $\Omega_h$  und  $\mathcal{T}_h$  muss der Benutzer den Rand  $\Gamma$  stückweise als Graph oder implizit als  $\{F(x) = 0\}$  zusammen mit  $F$  und  $\text{grad } F$  angeben.

Falls  $\Omega$  konvex ist, gilt  $\Omega_h \subset \Omega$  und die Fehlerabschätzungen der §§II.2 – II.4 können übertragen werden, wobei die Randapproximation einen zusätzlichen Konsistenzfehler von  $O(h^2)$  beiträgt.

Ist  $\Omega$  nicht konvex, gilt  $\Omega_h \setminus \Omega \neq \emptyset$ . Dann müssen die Daten  $A$ ,  $a$ ,  $\alpha$  und  $f$  auf  $\Omega_h$  fortgesetzt werden. Dies geschieht am einfachsten durch Reflexion (vgl. Abbildung III.1.2):

Ist  $x \in \Omega_h \setminus \Omega$ , setze  $f(x) = f(\tilde{x})$  und analog für  $A$ ,  $a$  und  $\alpha$ . Dabei wird  $\tilde{x}$  wie folgt bestimmt: Berechne die orthogonale Projektion  $\bar{x}$  von  $x$  auf  $\Gamma$  und wähle  $\tilde{x}$  auf der Geraden durch  $x$  und  $\bar{x}$ , so dass  $\tilde{x}$  in  $\Omega$  liegt und den gleichen Abstand zu  $\bar{x}$  hat wie  $x$ .

Mit diesen Modifikationen bleiben die Fehlerabschätzungen der §§II.2 – II.4 erhalten. Dabei müssen alle Normen auf  $\Omega \cap \Omega_h$  statt auf  $\Omega$  bezogen werden. Die Randapproximation trägt wieder einen Konsistenzfehler der Ordnung  $O(h^2)$  bei.

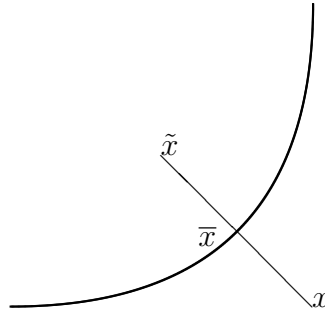


ABBILDUNG III.1.2. Reflexion am Rand  $\Gamma$

Um den Konsistenzfehler durch die Randapproximation zu verringern, kann man auch krummlinige oder isoparametrische Elemente betrachten. Dabei ist mit der Notation von §II.4 die Transformation  $F_K : \hat{K} \rightarrow K$  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 §II.3 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 erkaufte.

Die Finite Element Diskretisierungen der §§II.2 – II.4 führen auf lineare Gleichungssysteme. Zur Berechnung der rechten Seite und der Steifigkeitsmatrix müssen Integrale der Form  $\int_K \varphi$  mit  $\varphi = fv$ ,  $\varphi = \nabla u^T A \nabla v$  oder ähnlich 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 §II.4 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  $\hat{p} \in R_k(\hat{K})$  gilt

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

Sei nun  $K \in \mathcal{T}_h$  und  $F_K : \hat{K} \rightarrow K$  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

$$\begin{aligned} \omega_l &= |\det B_K| \hat{\omega}_l, \\ b_l &= F_K(\hat{b}_l) \end{aligned}$$

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

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

Man kann zeigen, dass die Fehlerabschätzungen der §§II.2 – II.4 für Finite Element Diskretisierungen der Ordnung  $m$ , d.h.  $X_h = S_{h,0}^{m,0}$ , 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 Würfelemente ( $m = 1$ ,  $\hat{K} = \hat{K}_W$ ) eine Quadraturformel der Ordnung 0 aus, d.h. nur die konstanten Funktionen müssen exakt integriert werden.

BEISPIEL III.1.1. (1) Die Mittelpunktsregel

$$Q_{\hat{K}}(\varphi) = \frac{1}{d!} \varphi\left(\frac{1}{d+1} \sum_{i=1}^{d+1} \hat{z}_i\right)$$

liefert eine Quadraturformel der Ordnung 1 für den Referenz  $d$ -Simplex.

(2) Die Formel

$$Q_{\hat{K}}(\varphi) = \frac{1}{6} \sum_{1 \leq i < j \leq 3} \varphi\left(\frac{1}{2}(\hat{z}_i + \hat{z}_j)\right)$$

hat die Ordnung 2 für das Referenzdreieck.

(3) Die Formel

$$Q_{\hat{K}}(\varphi) = \frac{1}{120} \left\{ 3 \sum_{i=1}^3 \varphi(\hat{z}_i) + 8 \sum_{1 \leq i < j \leq 3} \varphi\left(\frac{1}{2}(\hat{z}_i + \hat{z}_j)\right) + 27 \varphi\left(\frac{1}{3} \sum_{i=1}^3 \hat{z}_i\right) \right\}$$

hat die 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“

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

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

(5) Die Mittelpunktsregel liefert die Formel

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

der Ordnung 1 für das Referenzquadrat.

(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 Referenzquadrat.

(7) Die Simpsonregel liefert mit

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

die Formel

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

der Ordnung 3 für das Referenzquadrat.

### III.2. 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-Diffusionsgleichung mit homogenen Dirichlet-Randbedingungen und Dreieckselemente der Ordnung  $k \geq 1$ , d.h.  $\mathcal{T}_h$  besteht aus Dreiecken und  $X_h = S_{h,0}^{k,0}$ .

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  $(\cdot, \cdot)_h$  und eine zugehörige Norm  $\|\cdot\|_h$  auf  $X_h$  durch

$$(\varphi, \psi)_h = \sum_{K \in \mathcal{T}_h} \sum_{z \in \Sigma_k(K)} h_K^2 \varphi(z) \psi(z),$$

$$\|\varphi\|_h = (\varphi, \varphi)_h^{1/2}.$$

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

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

leicht gelöst werden.

**SATZ III.2.1.** (1)  $\|\cdot\|_h$  und  $\|\cdot\|_0$  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

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

Die Konstante  $c$  hängt von  $k$  und  $c_{\mathcal{T}}$  ab.

**BEWEIS.** AD (1): Wegen Satz II.4.3 (S. 53) 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

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

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

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

folgt aus obiger Abschätzung mit dem Transformationssatz

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

Hieraus folgt die Behauptung (1) durch Quadrieren und Summieren über alle Dreiecke. Man beachte, dass die Konstanten  $c_1$ ,  $c_2$  von  $c_T$  abhängen.

AD (2): Da  $|\cdot|_{H^1(\hat{K})}$  eine Norm auf  $\mathbb{P}_k/\mathbb{R}$  und  $\mathbb{P}_k/\mathbb{R}$  endlich dimensional ist, gibt es eine Konstante  $\hat{c}$  mit

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

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

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

Dabei hängt  $c'$  von  $c_T$  ab. Quadrieren und Summieren dieser Ungleichung beweist Teil (2).  $\square$

Aus Satz III.2.1 (1), dem Beweis des Existenzsatzes I.3.3 (S. 22) und der Friedrichschen Ungleichung I.1.15 (S. 12) folgt für alle  $v \in X_h$

$$\begin{aligned}
a(v, v) &\geq \beta |v|_1^2 \\
&\geq \beta' \|v\|_0^2 \\
&\geq \beta'' \|v\|_h^2.
\end{aligned}$$

Ebenso folgt aus dem Beweis von Satz I.3.3 (S. 22) und Satz III.2.1 für alle  $v, w \in X_h$

$$a(v, w) \leq B |v|_1 |w|_1$$

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

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“. Aufgrund unserer dortigen Erfahrungen mit Differenzenverfahren erwarten wir aber, dass Mehrgitterverfahren die besten Ergebnisse liefern.

BEISPIEL III.2.2. Wir wenden

- das CG-Verfahren,
- das PCG-Verfahren mit SSOR-Vorkonditionierung,
- das Mehrgitterverfahren mit V-Zyklus und mit einer Gauß-Seidel-Vorwärtsiteration als Vorglätter und einer Gauß-Seidel-Rückwärtsiteration als Nachglätter (MG-V( $\frac{1}{2}$ )),
- das Mehrgitterverfahren mit V-Zyklus und jeweils einer SSOR-Iteration als Vor- und Nachglättung (MG-V(1)),
- das Mehrgitterverfahren mit W-Zyklus und je einer SSOR-Iteration als Vor- und Nachglättung (MG-W(1))

auf drei diskrete Probleme an:

- (1) Die Poissongleichung im Quadrat  $(-1, 1)^2$  mit rechter Seite  $f = 1$  und homogenen Dirichlet-Randbedingungen. Die größte Triangulierung besteht aus 8 gleichschenkelig rechtwinkligen Dreiecken mit Hypotenusen in Richtung  $(1, -1)$  (Courant-Triangulierung). Die weiteren Triangulierungen werden durch uniforme Verfeinerung erzeugt. Dabei wird jedes Dreieck in vier kleinere Dreiecke unterteilt, indem die Kantenmittelpunkte miteinander verbunden werden (vgl. §III.4).
- (2) Die Poissongleichung im L-förmigen Gebiet  $(-1, 1)^2 \setminus [(0, 1) \times (-1, 0)]$  mit inhomogenen Dirichlet-Randbedingungen. Die Randfunktion  $u_D$  und die rechte Seite  $f$  sind so, dass die exakte Lösung der Differentialgleichung in Polarkoordinaten gegeben ist durch

$$u(x, y) = r^{\frac{2}{3}} \sin\left(\frac{2}{3}\varphi\right).$$

Die größte Triangulierung besteht aus 6 gleichschenkelig rechtwinkligen Dreiecken mit Hypotenusen in Richtung  $(1, -1)$ . Die weiteren Triangulierungen werden wie in (1) durch uniforme Verfeinerung erzeugt.

- (3) Die Differentialgleichung und die größte Triangulierung sind wie in (2). Die feineren Triangulierungen werden aber mit einer

adaptiven Gitterverfeinerung basierend auf einem residuellen Fehlerschätzer erzeugt (vgl. Beispiel III.3.6 (S. 94) und Abbildung III.3.5 (S. 94)).

Auf dem größten Gitter wird der Nullvektor als Startwert verwendet. Auf feineren Gittern wird die lineare Interpolierende der Näherungslösung auf dem nächst größeren Gitter als Startwert benutzt (vgl. die geschachtelte Iteration, Algorithmus III.6.1, der Vorlesung „Numerische Behandlung von Differentialgleichungen I“). Die Iteration auf dem aktuellen Gitter wird jeweils abgebrochen, wenn das Startresiduum um den Faktor 100 reduziert wurde. In den Tabellen III.2.1 – III.2.3 werden für die Verfahren jeweils die Konvergenzraten  $\kappa$  angegeben.  $N$  bezeichnet die Zahl der Unbekannten.  $T$  ist die Gesamtrechenzeit in Sekunden auf einem Macintosh G4-Powerbook.

TABELLE III.2.1. Konvergenzraten und Rechenzeiten des CG-, SSOR-PCG-, MG-V( $\frac{1}{2}$ ), MG-V(1)- und MG-W(1)-Verfahrens für das diskrete Problem aus Beispiel III.2.2 (1)

$N$	CG	SSOR-PCG	MG-V( $\frac{1}{2}$ )	MG-V(1)	MG-W(1)
	$\kappa$	$\kappa$	$\kappa$	$\kappa$	$\kappa$
49	0.592	0.496	0.183	0.150	0.150
225	0.792	0.665	0.131	0.054	0.031
961	0.899	0.801	0.136	0.034	0.027
3969	0.943	0.890	0.143	0.029	0.027
$T(s)$	0.18	0.42	0.06	0.08	0.09

TABELLE III.2.2. Konvergenzraten und Rechenzeiten des CG-, SSOR-PCG-, MG-V( $\frac{1}{2}$ ), MG-V(1)- und MG-W(1)-Verfahrens für das diskrete Problem aus Beispiel III.2.2 (2)

$N$	CG	SSOR-PCG	MG-V( $\frac{1}{2}$ )	MG-V(1)	MG-W(1)
	$\kappa$	$\kappa$	$\kappa$	$\kappa$	$\kappa$
33	0.578	0.428	0.175	0.077	0.077
161	0.789	0.617	0.181	0.073	0.066
705	0.882	0.787	0.181	0.072	0.066
2945	0.937	0.876	0.181	0.072	0.065
$T(s)$	0.13	0.22	0.09	0.06	0.07

Wir wollen nun ein Mehrgitter-Verfahren für Finite Element Diskretisierung analysieren. Dazu nehmen wir an, dass wir eine Hierarchie

TABELLE III.2.3. Konvergenzraten und Rechenzeiten des CG-, SSOR-PCG-, MG-V( $\frac{1}{2}$ ), MG-V(1)- und MG-W(1)-Verfahrens für das diskrete Problem aus Beispiel III.2.2 (3)

N	CG	SSOR-PCG	MG-V( $\frac{1}{2}$ )	MG-V(1)	MG-W(1)
	$\kappa$	$\kappa$	$\kappa$	$\kappa$	$\kappa$
16	0.511	0.385	0.136	0.066	0.066
54	0.957	0.957	0.167	0.073	0.068
221	0.966	0.966	0.293	0.139	0.099
718	0.976	0.976	0.296	0.152	0.146
$T(s)$	0.16	0.24	0.01	0.02	0.04

geschachtelter Triangulierungen  $\mathcal{T}_{h_0} \subset \mathcal{T}_{h_1} \subset \dots \subset \mathcal{T}_{h_R}$  haben. Es gilt, die Finite Element Diskretisierung zur feinsten Triangulierung numerisch zu lösen. Dazu machen wir folgende Annahmen:

- $\Omega$  ist konvex.
- 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  $c$ .

- $h_{m-1} \leq ch_m$  für alle  $m$  mit einer von  $m$  unabhängigen Konstanten  $c$ .

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

ALGORITHMUS III.2.3. (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  $i = 1, \dots, \nu_1$  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  $i = \nu_1 + 2, \dots, \nu_1 + \nu_2 + 1$  berechne  $u_m^i$  aus  $u_m^{i-1}$  als Lösung von

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

BEMERKUNG III.2.4. (1) Es ist  $l_R(v) = l(v) = \int_{\Omega} f v$ . Die rechten Seiten  $l_m$  zu den gröbereren Triangulierungen werden rekursiv berechnet.

(2) Die Dämpfungparameter  $\omega_m$  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  $\lambda_{m,1}, \dots, \lambda_{m,n_m}$ ,  $n_m = \dim X_m$ , von Eigenwerten und zugehörigen, bzgl.  $(\cdot, \cdot)_m$  orthonormierten Eigenfunktionen  $\psi_{m,1}, \dots, \psi_{m,n_m}$ :

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

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

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

Da die Triangulierungen uniform sind, folgt aus Satz III.2.1  $\Lambda_m \sim h_m^{-2}$ . 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

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

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

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

Dabei bedeutet „ $\simeq$ “ Äquivalenz von Normen mit von  $m$  unabhängigen Konstanten.

Sind  $v, w \in X_m$  mit

$$v = \sum c_\mu \psi_{m,\mu} \quad , \quad w = \sum d_\mu \psi_{m,\mu},$$

so folgt aus der Cauchy-Schwarzischen Ungleichung für Summen und der Orthogonalität der Eigenfunktionen

$$\begin{aligned} (III.2.1) \quad a(v, w) &= \sum_{\mu=1}^{n_m} \lambda_{m,\mu} c_\mu d_\mu \\ &\leq \left\{ \sum_{\mu=1}^{n_m} c_\mu^2 \right\}^{1/2} \left\{ \sum_{\mu=1}^{n_m} \lambda_{m,\mu}^2 d_\mu^2 \right\}^{1/2} \\ &= |v|_{0,m} |w|_{2,m}. \end{aligned}$$

SATZ III.2.5. *Bezeichne mit  $Q_m : X_m \rightarrow X_{m-1}$  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 (III.2.1) 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  $\Omega$  konvex ist, folgt aus Satz I.2.2 (S. 17)

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

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

Als nächstes definieren wir einen Operator  $J : X_m \rightarrow X_m$  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 III.2.3. Ist  $v \in \sum c_\mu \psi_{m,\mu}$ , so folgt

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

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

$$|v| = a(v, Jv)^{1/2}$$

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

Dann gilt offensichtlich für  $v \in X_m$

$$|v| = |J^{1/2}v|_{1,m},$$

$$0 \leq \rho(v) \leq 1.$$

SATZ III.2.6. *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  $v = \sum c_\mu \psi_{m,\mu}$  und setze zur Abkürzung  $\sigma_\mu = 1 - \frac{\lambda_{m,\mu}}{\Lambda_m}$ . Dann folgt mit der Hölderschen Ungleichung

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

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

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

Hieraus folgt

$$\begin{aligned} |J^\nu v|_{1,m} &\leq \left\{ \frac{|J^\nu v|_{1,m}}{|J^\nu v|_{1,m}} \right\}^{2\nu} |v|_{1,m} \\ &= \rho^\nu |v|_{1,m}. \end{aligned}$$

□

SATZ III.2.7. *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 III.2.5.

BEWEIS. Da  $I - Q_m$  eine Projektion ist, ist

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

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

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

Hieraus und aus Satz III.2.5 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.  $\square$

Nach diesen Vorbereitungen können wir nun die Konvergenz von Algorithmus III.2.3 beweisen.

**SATZ III.2.8.** *Bezeichne mit  $\delta_m$  die Konvergenzrate von Algorithmus III.2.3 mit  $\nu_1 = \nu_2 = \nu$  auf dem  $m$ -ten Gitter gemessen in der  $|\cdot|_{1,m}$ -Norm. Dann gilt mit der Konstanten  $c$  aus Satz III.2.5*

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

**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 III.2.3. 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

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

$$= |e_m^\nu|_{1,m}^2.$$

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

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

Da  $I - Q_m$  ein Projektor bzgl. des Skalarproduktes  $a(\cdot, \cdot)$  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-Schwarzschen Ungleichung

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

Wegen der Sätze III.2.6 und III.2.7 ist

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

und

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

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

$$|e^{2\nu+1}|_{1,m} \leq |e^0|_{1,m} \max_{0 \leq \rho \leq 1} \left\{ \rho^{2\nu} [\delta_{m-1} + (1 - \delta_{m-1}) \min\{1, c(1 - \rho)\}] \right\}.$$

Also ist

$$\delta_m \leq \max_{0 \leq \rho \leq 1} \left\{ \rho^{2\nu} [\delta_{m-1} + (1 - \delta_{m-1}) \min\{1, c(1 - \rho)\}] \right\}.$$

Da auf dem größten Gitter exakt gelöst wird, gilt

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

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

$$\delta \mapsto \max_{0 \leq \rho \leq 1} \left\{ \rho^{2\nu} [\delta + (1 - \delta) \min\{1, c(1 - \rho)\}] \right\}$$

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

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

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

**SATZ III.2.9.** Sei  $\tilde{\delta}_m$  die Konvergenzrate von Algorithmus III.2.3 mit  $\nu_1 = \nu$ ,  $\nu_2 = 0$  auf dem  $m$ -ten Gitter gemessen in der  $|\cdot|_{1,m}$ -Norm. Dann gilt mit der Konstanten  $c$  aus Satz III.2.5

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

**BEWEIS.** Sei  $M$  der Operator  $e_m^0 \mapsto e_m^{\nu+1}$  des Algorithmus III.2.3 mit  $\nu_1 = \nu$ ,  $\nu_2 = 0$ . Dann beschreibt der bzgl. des Skalarproduktes  $a(\cdot, \cdot)$  adjungierte Operator  $M^*$  die Fehlerfortpflanzung von Algorithmus III.2.3 mit  $\nu_1 = 0$ ,  $\nu_2 = \nu$ . Insbesondere beschreibt  $M^*M$  die Fehlerfortpflanzung von Algorithmus III.2.3 mit  $\nu_1 = \nu_2 = \nu$ . Damit folgt aus Satz III.2.8

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

$\square$

Der Beweis von Satz III.2.8 beruht wesentlich auf den Annahmen, dass  $\Omega$  konvex ist und die Gitter uniform sind. Die Konvexität von  $\Omega$  wird in Satz III.2.5 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 III.2.3 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  $(\cdot, \cdot)$ .
- $V_1, \dots, V_N$  sind Untervektorräume von  $V$  mit  $V = \sum_{i=1}^N V_i$  (diese Zerlegung ist in der Regel weder direkt noch gar orthogonal).
- $Q_i : V \rightarrow V_i$  sind die orthogonalen Projektionen bzgl.  $(\cdot, \cdot)$ .
- $A : V \rightarrow V$  ist ein symmetrischer, positiv definit Operator.
- $A_i : V_i \rightarrow V_i$  ist die Einschränkung von  $A$  auf  $V_i$ , d.h.  $(A_i u, v) = (A u, v)$  für alle  $u, v \in V_i$ .
- $R_i : V_i \rightarrow V_i$  ist eine leicht berechenbare, symmetrisch positiv definite Approximation an  $A_i^{-1}$ .

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

Zu lösen ist das Problem

$$Au = f.$$

Dies geschieht mit folgendem Algorithmus:

ALGORITHMUS III.2.10. (TEILRAUM-KORREKTUR-METHODE)

(0) Gegeben: Startnäherung  $u_0 \in V$ .

(1) Für  $n = 0, 1, \dots$  und  $j = 1, \dots, N$  berechne

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

BEISPIEL III.2.11. (1) Sei  $V = \mathbb{R}^N$  und  $V_i = \text{span}\{e_i\}$ . Dann ist Algorithmus III.2.10 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 III.2.10 der Mehrgitteralgorithmus III.2.3 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 III.2.10 setzen wir

$$\lambda = \min_{1 \leq i \leq N} \lambda_{\min}(R_i A_i),$$

$$\Lambda = \max_{1 \leq i \leq N} \lambda_{\max}(R_i A_i),$$

$$T_i = R_i Q_i A.$$

Aus den Voraussetzungen folgt  $0 < \lambda \leq \Lambda$ . Durch entsprechende Skalierung kann man o.E. erreichen, dass  $\Lambda < 2$  ist. Wir bezeichnen mit  $|\cdot|$  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  $|\cdot|$  zur  $H^1$ -Norm äquivalent.

SATZ III.2.12. *Es gebe zwei Konstanten  $K_0$  und  $K_1$  mit*

$$\left\{ \sum_{i=1}^N |v_i|^2 \right\}^{1/2} \leq K_0 |v| \quad \forall v = \sum_{i=1}^N v_i \in V$$

und

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

Weiter sei  $\Lambda < 2$ . Dann ist die Konvergenzrate von Algorithmus III.2.10 gemessen in der Norm  $|\cdot|$  kleiner oder gleich

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

BEWEIS. Sei  $u$  die exakte Lösung von  $Au = f$ . Aus Algorithmus III.2.10 ergibt sich

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

und somit

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

Setze zur Abkürzung

$$E_0 = I$$

$$E_j = (I - T_j) \dots (I - T_1) \quad , 1 \leq j \leq N.$$

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

$$(III.2.2) \quad |E_N v| \leq \left[ 1 - \left( \frac{2}{\Lambda} - 1 \right) \left( \frac{\lambda}{\Lambda K_0 K_1} \right)^2 \right]^{1/2} |v|.$$

Aus der Definition der  $E_j$  folgt sofort

$$-E_j + E_{j-1} = T_j E_{j-1} \quad \forall 1 \leq j \leq N.$$

Bezeichne mit  $P_i : V_i \rightarrow V_i$  die orthogonale Projektion bzgl. des Skalarproduktes  $(A \cdot, \cdot)$ . 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.  $(A \cdot, \cdot)$ .  
Damit folgt für beliebiges  $v \in V$

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

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

$$\begin{aligned}(III.2.3) \quad |v|^2 - |E_N v|^2 &= \sum_{j=1}^N \left\{ |E_{j-1} v|^2 - |E_j v|^2 \right\} \\ &\geq (2 - \Lambda) \sum_{j=1}^N (AT_j E_{j-1} v, E_{j-1} v).\end{aligned}$$

Schreibe  $v = \sum_{i=1}^N v_i$ . Da  $(R_i A_i)^{-1} T_i$  die orthogonale Projektion bzgl.  $(A \cdot, \cdot)$  ist, folgt aus der Cauchy-Schwarzschen Ungleichung und der ersten Voraussetzung des Satzes [III.2.12](#)

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

Also ist

$$(III.2.4) \quad |v| \leq \lambda^{-1} K_0 \left\{ \sum_{i=1}^N |T_i v|^2 \right\}^{1/2}.$$

Aus der zweiten Voraussetzung von Satz III.2.12 folgt

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

und somit

$$\begin{aligned} \sum_{i=1}^N |T_i v|^2 &\leq \Lambda^2 K_1^2 \sum_{j=1}^N |T_j E_{j-1} v|^2 \\ (III.2.5) \quad &= \Lambda^2 K_1^2 \sum_{j=1}^N (AT_j E_{j-1} v, T_j E_{j-1} v) \\ &\leq \Lambda^3 K_1^2 \sum_{j=1}^N (AT_j E_{j-1} v, E_{j-1} v). \end{aligned}$$

Aus den Abschätzungen (III.2.3) – (III.2.5) erhalten wir

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

Hieraus folgt offensichtlich (III.2.2) und damit die Behauptung des Satzes.  $\square$

BEMERKUNG III.2.13. (1) Wegen der Cauchy-Schwarzschen Ungleichung ist die zweite Voraussetzung von Satz III.2.12 stets mit  $K_1 \leq N$  erfüllt. Für Mehrgitterverfahren bedeutet die Abschätzung  $K_1 \leq N$ ,

dass die Konvergenzrate von Algorithmus III.2.10 sich schlimmstenfalls wie  $1/\ln|h_N|$  verhält. Häufig kann man jedoch eine VERSCHÄRFTE CAUCHY-SCHWARZSCHE 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

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

linear, stetig und surjektiv. Aus dem Satz von der offenen Abbildung (§II.5 in Yosida: Functional Analysis) folgt daher, dass die erste Voraussetzung von Satz III.2.12 ebenfalls stets erfüllt ist. 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 Sobolevräume ist dies im Rahmen von Mehrgitterverfahren für Finite Element Diskretisierungen in der Tat möglich.

### III.3. Fehlerschätzer und adaptive Gitterverfeinerung

In den §§II.2 – II.4 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 unbekanntenen 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-Diffusionsgleichung 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  $X_h = S_{h,0}^{k,0}$ .

Zunächst müssen wir einige zusätzliche Notationen einführen. Die Menge aller Dreieckskanten, die im Innern von  $\Omega$  liegen, bezeichnen wir mit  $\mathcal{E}_h$ . Jedes  $E$  ist dann die gemeinsame Kante von genau zwei Dreiecken, die wir mit  $K_{E_1}$  und  $K_{E_2}$  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  $\varphi$  mit  $[\varphi]_E$  den Sprung von  $\varphi$  über  $E$  in Richtung  $n_E$ , d.h.

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

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 §§II.3, II.4 können Größen der Form  $\frac{h_K}{h_{K'}}$  und  $\frac{h_K}{h_E}$  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  $K \in \mathcal{T}_h$  (vgl. Abbildung III.3.1):

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

Dabei ist  $v_z \in S_{h,0}^{1,0}$  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  $u \in H_0^1(\Omega)$  und  $u_h \in X_h$  die schwache Lösung der Reaktions-Diffusionsgleichung und ihre Finite Element Approximation. Wegen des Existenzsatzes I.3.3 (S. 22) und der Friedrichschen Ungleichung I.1.15 (S. 12) gibt es eine Konstante  $\beta > 0$  mit

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

$\beta$  hängt von  $\Omega$  und den Koeffizienten  $A$  und  $\alpha$  des Differentialoperators ab. Aus dieser Abschätzung folgt insbesondere die STABILITÄT, d.h.

$$(III.3.1) \quad \|u - u_h\|_1 \leq \frac{1}{\beta} \sup_{\substack{v \in H_0^1(\Omega) \\ \|v\|_1=1}} a(u - u_h, v).$$

Da  $X_h \subset H_0^1(\Omega)$  ist, haben wir die GALERKIN-ORTHOGONALITÄT des Fehlers, d.h.

$$(III.3.2) \quad a(u - u_h, v_h) = 0 \quad \forall v_h \in X_h.$$

Die Abbildung

$$v \mapsto a(u - u_h, v) = l(v) - a(u_h, v)$$

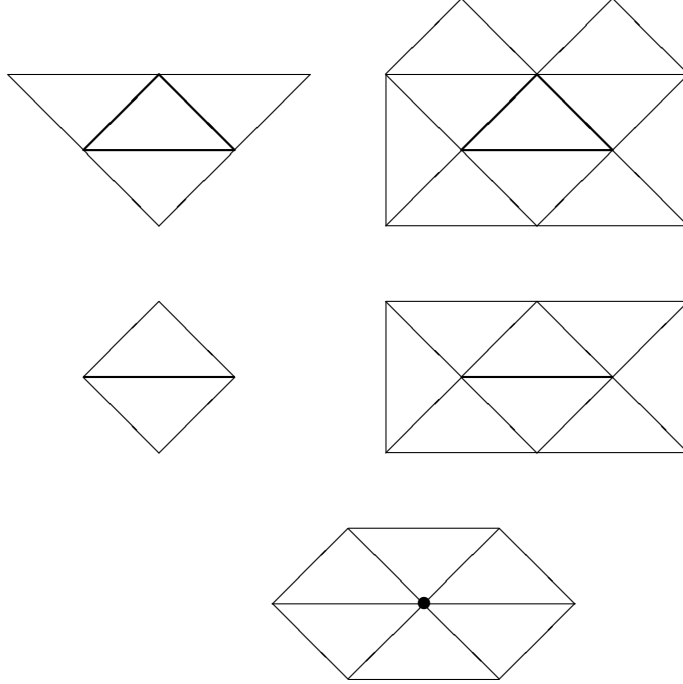


ABBILDUNG III.3.1. Gebiete  $\omega_K$ ,  $\tilde{\omega}_K$ ,  $\omega_E$ ,  $\tilde{\omega}_E$  und  $\omega_x$

definiert ein stetiges lineares Funktional auf  $H_0^1(\Omega)$ . Dies ist das RESIDUUM. Als nächstes wollen wir eine Darstellung des Residuums angeben, die praktisch handhabbar ist. Sei dazu  $v \in H_0^1(\Omega)$  mit  $\|v\|_1 = 1$  beliebig, aber fest. Anwenden des Gaußschen Integralsatzes auf jedem Element  $K \in \mathcal{T}_h$  liefert

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

$$- \sum_{E \in \mathcal{E}_h} \int_E [n_E \cdot A \nabla u_h]_E v.$$

also

$$(III.3.3) \quad \begin{aligned} a(u - u_h, v) &= \sum_{K \in \mathcal{T}_h} \int_K \{f + \nabla \cdot (A \nabla u_h) - \alpha u_h\} v \\ &\quad - \sum_{E \in \mathcal{E}_h} \int_E [n_E \cdot A \nabla u_h]_E v \quad \forall v \in H_0^1(\Omega). \end{aligned}$$

Dabei bezeichnet  $n_K$  die äußere Normale zu  $K$ . Wegen der Galerkin Orthogonalität (III.3.2) können wir auf der rechten Seite von (III.3.3) ein beliebiges Element  $v_h \in X_h$  von  $v$  subtrahieren. Aus der Cauchy-Schwarzschen Ungleichung für Integrale folgt dann

$$(III.3.4) \quad \begin{aligned} a(u - u_h, v) &\leq \sum_{K \in \mathcal{T}_h} \|f + \nabla \cdot (A \nabla u_h) - \alpha u_h\|_{L^2(K)} \\ &\quad \cdot \|v - v_h\|_{L^2(K)} \\ &\quad + \sum_{E \in \mathcal{E}_h} \|[n_E \cdot A \nabla u_h]_E\|_{L^2(E)} \\ &\quad \cdot \|v - v_h\|_{L^2(E)}. \end{aligned}$$

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 §§II.3 und II.4 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  $R_h : H_0^1(\Omega) \rightarrow S_{h,0}^{1,0}$  wie folgt:

Für jedes  $z \in \mathcal{N}_h$  bezeichnen wir mit  $\pi_z : L^2(\omega_z) \rightarrow \mathbb{R}$  die orthogonale Projektion, d.h.

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

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

$$(III.3.5) \quad R_h \varphi = \sum_{z \in \mathcal{N}_{h,\Omega}} (\pi_z \varphi) v_z.$$

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

**SATZ III.3.1.** *Es gibt zwei Konstanten  $c_1, c_2$  die nur von der Konstanten  $c_T$  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*

$$\|v - R_h v\|_{L^2(K)} \leq c_1 h_K |v|_{H^1(\bar{\omega}_K)},$$

$$\|v - R_h v\|_{L^2(E)} \leq c_2 h_E^{1/2} |v|_{H^1(\tilde{\omega}_E)}.$$

BEWEIS. 1. SCHRITT: Aus der Poincaréschen Ungleichung I.1.21 (S. 15) folgt für jedes  $z \in \mathcal{N}_h$  und jedes  $\varphi \in H^1(\omega_z)$

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

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

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

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 I.1.12 (S. 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  $F_K : \hat{K} \rightarrow K$  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

$$\begin{aligned} h_E^{1/2} |\det DF_K|^{-1/2} &\leq c h_K^{-1/2}, \\ h_E^{1/2} |\det DF_K|^{-1/2} \|DF_K^{-1}\|_{\mathcal{L}} &\leq c h_K^{1/2}. \end{aligned}$$

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

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

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

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

Damit folgt aus Schritt 1

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

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

4. SCHRITT: Betrachte nun ein Dreieck  $K$ , das mindestens einen Eckpunkt auf dem Rand  $\Gamma$  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

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

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  $\Gamma$  liegt. Dann ist

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

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

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

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  $\Gamma$  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

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

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

6. SCHRITT: Betrachte nun eine Kante  $E$ , die einen Endpunkt auf dem Rand  $\Gamma$  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  $E$ .  $\square$

Wir greifen nun die Abschätzung (III.3.4) wieder auf. Wir setzen  $v_h = R_h v$ , benutzen Satz III.3.1 und wenden die Cauchy-Schwarzsche Ungleichung für endliche Summen an:

$$\begin{aligned}
a(u - u_h, v) &\leq c_1 \sum_{K \in \mathcal{T}_h} h_K \|f + \nabla \cdot (A \nabla u_h) - \alpha u_h\|_{L^2(K)} |v|_{H^1(\tilde{\omega}_K)} \\
&\quad + c_2 \sum_{E \in \mathcal{E}_h} h_E^{1/2} \|[n_E \cdot A \nabla u_h]_E\|_{L^2(E)} |v|_{H^1(\tilde{\omega}_E)} \\
&\leq c_1 \left\{ \sum_{K \in \mathcal{T}_h} h_K^2 \|f + \nabla \cdot (A \nabla u_h) - \alpha u_h\|_{L^2(K)}^2 \right\}^{1/2} \\
&\quad \cdot \left\{ \sum_{K \in \mathcal{T}_h} |v|_{H^1(\tilde{\omega}_K)}^2 \right\}^{1/2} \\
&\quad + c_2 \left\{ \sum_{E \in \mathcal{E}_h} h_E \|[n_E \cdot A \nabla u_h]_E\|_{L^2(E)}^2 \right\}^{1/2}
\end{aligned}$$

$$\begin{aligned}
 & \cdot \left\{ \sum_{E \in \mathcal{E}_h} |v|_{H^1(\bar{\omega}_E)}^2 \right\}^{1/2} \\
 & \leq c' |v|_1 \left\{ \sum_{K \in \mathcal{T}_h} h_K^2 \|f + \nabla \cdot (A \nabla u_h) - \alpha u_h\|_{L^2(K)}^2 \right. \\
 & \quad \left. + \sum_{E \in \mathcal{E}_h} h_E \|[n_E \cdot A \nabla u_h]_E\|_{L^2(E)}^2 \right\}^{1/2}.
 \end{aligned}$$

Dabei haben wir im letzten Schritt die Regularitätsbedingung an  $\mathcal{T}_h$  ausgenutzt. Hieraus und aus (III.3.1) folgt insgesamt

$$(III.3.6) \quad \|u - u_h\|_1 \leq c\eta$$

mit

$$\begin{aligned}
 \eta &= \left\{ \sum_{K \in \mathcal{T}_h} \eta_K^2 \right\}^{1/2}, \\
 (III.3.7) \quad \eta_K &= \left\{ h_K^2 \|f + \nabla \cdot (A \nabla u_h) - \alpha u_h\|_{L^2(K)}^2 \right. \\
 & \quad \left. + \frac{1}{2} \sum_{E \subset \partial K \setminus \Gamma} h_E \|[n_E \cdot A \nabla u_h]_E\|_{L^2(E)}^2 \right\}^{1/2}.
 \end{aligned}$$

Der Faktor  $\frac{1}{2}$  vor der zweiten Summe in  $\eta_K$  berücksichtigt, dass bei Summation über alle Dreiecke jede innere Kante doppelt gezählt wird.

Ungleichung (III.3.6) ist eine A POSTERIORI FEHLERABSCHÄTZUNG. Die Größe  $\eta$  kann aus den gegebenen Daten  $f$ ,  $A$ ,  $\alpha$  und der berechneten numerischen Lösung  $u_h$  a posteriori berechnet werden. Sie heißt daher auch A POSTERIORI FEHLERSCHÄTZER. Ungleichung (III.3.6) zeigt, dass der Fehlerschätzer ZUVERLÄSSIG ist, d.h. ist  $\eta \leq \varepsilon$ , so ist der Fehler ebenfalls (bis auf einen Faktor) nicht größer als  $\varepsilon$ . Die Kontrolle von  $\eta$  erlaubt also, eine vorgegebene Toleranz zu erreichen. Um dies mit einem minimalen Aufwand zu erreichen, reicht die obere Schranke (III.3.6) nicht aus. Wir müssen zusätzlich garantieren, dass  $\eta$  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  $\eta$  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

$$\psi_K = 27v_{z_1}v_{z_2}v_{z_3} \quad \text{auf } K.$$

Wie man leicht nachprüft, hat  $\psi_K$  folgende Eigenschaften

$$\begin{aligned}
 \psi_K &\in \mathbb{P}_3, \\
 \psi_K &\geq 0 \quad \text{auf } K,
 \end{aligned}$$

$$\begin{aligned}\psi_K &= 0 \quad \text{auf } \partial K, \\ \max_{x \in K} \psi_K(x) &= 1.\end{aligned}$$

Insbesondere kann also  $\psi_K$  durch Null zu einer Funktion aus  $H_0^1(\Omega)$  fortgesetzt werden.

Für eine Kante  $E \in \mathcal{E}_h$  nummerieren wir die Eckpunkte der angrenzenden Dreiecke  $K_{E1}$  und  $K_{E2}$  so, dass die Endpunkte von  $E$  zuerst nummeriert 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

$$\psi_E = 4v_{z_1,i}v_{z_2,i} \quad \text{auf } K_{Ei}, \quad i = 1, 2.$$

Offensichtlich hat  $\psi_E$  folgende Eigenschaften

$$\begin{aligned}\psi_E &\in C(\omega_E) \\ \psi_E|_{K_{Ei}} &\in \mathbb{P}_2, \quad i = 1, 2, \\ \psi_E &\geq 0 \quad \text{auf } \omega_E, \\ \psi_E &= 0 \quad \text{auf } \partial\omega_E, \\ \max_{x \in E} \psi_E(x) &= 1.\end{aligned}$$

Für eine Dreiecksseite  $E \in \mathcal{E}_h$  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  $\varphi$ .

**SATZ III.3.2.** *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*

$$\begin{aligned}c_3 \|v\|_{L^2(K)} &\leq \left\{ \int_K \psi_K v^2 \right\}^{1/2} \\ &\leq \|v\|_{L^2(K)}, \\ c_4 h_K^{-1} \|\psi_K v\|_{L^2(K)} &\leq |\psi_K v|_{H^1(K)} \\ &\leq c_5 h_K^{-1} \|\psi_K v\|_{L^2(K)}, \\ c_6 \|\sigma\|_{L^2(E)} &\leq \left\{ \int_E \psi_E \sigma^2 \right\}^{1/2} \\ &\leq \|\sigma\|_{L^2(E)}, \\ c_7 h_E^{-1} \|\psi_E \sigma\|_{L^2(\omega_E)} &\leq |\psi_E \sigma|_{H^1(\omega_E)} \\ &\leq c_8 h_E^{-1} \|\psi_E \sigma\|_{L^2(\omega_E)}, \\ |\psi_E \sigma|_{L^2(\omega_E)} &\leq c_9 h_E^{1/2} \|\sigma\|_{L^2(E)}.\end{aligned}$$

Die Konstanten  $c_3, \dots, c_9$  hängen nur von  $k$  und der Größe  $c_T$  in der Regularitätsannahme an  $\mathcal{T}_h$  ab.

**BEWEIS.** Die obere Schranke in der ersten Abschätzung folgt aus der Cauchy-Schwarzschen 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

$$\hat{c} \|w\|_{L^2(\hat{K})} \leq \left\{ \int_{\hat{K}} \psi_K \circ F_K w^2 \right\}^{1/2} \quad \forall w \in \mathbb{P}_k.$$

Anwenden des Transformationssatzes liefert

$$\begin{aligned} \hat{c} \|v\|_{L^2(K)} &= \hat{c} |\det DF_K|^{1/2} \|v \circ F_K\|_{L^2(\hat{K})} \\ &\leq |\det DF_K|^{1/2} \left\{ \int_{\hat{K}} \psi_K \circ F_K (v \circ F_K)^2 \right\}^{1/2} \\ &= \left\{ \int_K \psi_K v^2 \right\}^{1/2} \end{aligned}$$

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 die zweite Abschätzung.

Die Abschätzungen für  $\sigma$  folgen mit den gleichen Argumenten wie diejenigen für  $v$ .  $\square$

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 III.3.2 ist

$$c_3^2 \|f_h + \nabla \cdot (A \nabla u_h) - \alpha u_h\|_{L^2(K)}^2 \leq \int_K w_K (f_h + \nabla \cdot (A \nabla u_h) - \alpha u_h).$$

Setzen wir  $w_K$  als Testfunktion  $v$  in III.3.3 ein und berücksichtigen, dass  $w_K$  auf  $\partial K$  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 III.3.2 folgt weiter

$$\begin{aligned} \|w_K\|_{H^1(K)} &= \left\{ \|w_K\|_{L^2(K)}^2 + |w_K|_{H^1(K)}^2 \right\}^{1/2} \\ &\leq \left\{ 1 + c_5^2 h_K^{-2} \right\}^{1/2} \|w_K\|_{L^2(K)} \\ &\leq c h_K^{-1} \|f_h + \nabla \cdot (A \nabla u_h) - \alpha u_h\|_{L^2(K)}. \end{aligned}$$

Aus diesen Abschätzungen und der Dreiecksungleichung ergibt sich

$$(III.3.8) \quad \begin{aligned} &h_K \|f + \nabla \cdot (A \nabla u_h) - \alpha u_h\|_{L^2(K)} \\ &\leq c \left\{ \|u - u_h\|_{H^1(K)} + h_K \|f - f_h\|_{L^2(K)} \right\}. \end{aligned}$$

Sei nun  $E \in \mathcal{E}_h$  eine Kante. Setze

$$w_E = \psi_E [n_E \cdot A \nabla u_h]_E.$$

Wegen Satz III.3.2 ist

$$c_6^2 \| [n_E \cdot A \nabla u_h]_E \|_{L^2(E)}^2 \leq \int_E w_E [n_E \cdot A \nabla u_h]_E.$$

Setzen wir  $w_E$  als Testfunktion  $v$  in (III.3.3) ein und berücksichtigen, dass  $w_E$  auf  $\partial\omega_E$  und außerhalb  $\omega_E$  verschwindet, so folgt

$$\begin{aligned} &\int_E w_E [n_E \cdot A \nabla u_h]_E \\ &= \int_{\omega_E} (f + \nabla \cdot (A \nabla u_h) - \alpha u_h) w_E - a(u - u_h, w_E) \\ &\leq \|f + \nabla \cdot (A \nabla u_h) - \alpha u_h\|_{L^2(\omega_E)} \|w_E\|_{L^2(\omega_E)} \\ &\quad + \|a\|_{\mathcal{L}^2} \|u - u_h\|_{H^1(\omega_E)} \|w_E\|_{H^1(\omega_E)} \end{aligned}$$

Wegen Satz III.3.2 ist

$$\begin{aligned} \|w_E\|_{L^2(\omega_E)} &\leq c_9 h_E^{1/2} \|w_E\|_{L^2(E)} \\ &\leq c_7 h_E^{1/2} \| [n_E \cdot A \nabla u_h]_E \|_{L^2(E)} \end{aligned}$$

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 (III.3.8) ergibt sich insgesamt

$$(III.3.9) \quad \begin{aligned} &h_E^{1/2} \| [n_E \cdot A \nabla u_h]_E \|_{L^2(E)} \\ &\leq c \left\{ \|u - u_h\|_{H^1(\omega_E)} + h_E \|f - f_h\|_{L^2(\omega_E)} \right\}. \end{aligned}$$

Aus den Abschätzungen (III.3.8), (III.3.9) und der Definition (III.3.7) erhalten wir folgende LOKALE UNTERE FEHLERSCHRANKE

$$(III.3.10) \quad \eta_K \leq c \left\{ \|u - u_h\|_{H^1(\omega_K)} + h_K \|f - f_h\|_{L^2(\omega_K)} \right\}.$$

Summation über alle Dreiecke liefert zudem die GLOBALE UNTERE FEHLERSCHRANKE

$$(III.3.11) \quad \eta \leq c \left\{ \|u - u_h\|_1 + \left\{ \sum_{K \in \mathcal{T}_h} h_K^2 \|f - f_h\|_{L^2(K)}^2 \right\}^{1/2} \right\}.$$

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.

Wir fassen unsere Ergebnisse zusammen:

**SATZ III.3.3.** *Bezeichne mit  $u \in H_0^1(\Omega)$  und  $u_h \in X_h$  die schwache Lösung der Reaktions-Diffusionsgleichung und ihre Finite Element Approximation. Für jedes Element  $K \in \mathcal{T}_h$  definiere den Fehlerschätzer  $\eta_K$  durch*

$$\eta_K = \left\{ h_K^2 \|f + \nabla \cdot (A \nabla u_h) - \alpha u_h\|_{L^2(K)}^2 + \frac{1}{2} \sum_{E \subset \partial K \setminus \Gamma} h_E \|[n_E \cdot A \nabla u_h]_E\|_{L^2(E)}^2 \right\}^{1/2}.$$

Dann gibt es drei Konstanten  $C_1$ ,  $C_2$  und  $C_3$ , die nur von dem Polynomgrad  $k$  des Finite Element Raumes  $X_h$  und der Regularitätskonstanten  $c_{\mathcal{T}}$  der Unterteilung  $\mathcal{T}_h$  abhängen, so dass die folgenden a posteriori Fehlerabschätzungen gelten

$$\begin{aligned} \|u - u_h\|_1 &\leq C_1 \left\{ \sum_{K \in \mathcal{T}_h} \eta_K^2 \right\}^{1/2}, \\ \eta_K &\leq C_2 \left\{ \|u - u_h\|_{H^1(\omega_K)} + h_K \|f - f_h\|_{L^2(\omega_K)} \right\}, \\ \left\{ \sum_{K \in \mathcal{T}_h} \eta_K^2 \right\}^{1/2} &\leq C_3 \left\{ \|u - u_h\|_1 + \left\{ \sum_{K \in \mathcal{T}_h} h_K^2 \|f - f_h\|_{L^2(K)}^2 \right\}^{1/2} \right\}. \end{aligned}$$

**BEMERKUNG III.3.4.** (1) Die Abschätzungen (III.3.6), (III.3.10) und (III.3.11) zeigen, dass der Fehlerschätzer  $\eta$  zuverlässig und effizient ist. Es ist nicht verwunderlich, dass wir nur in einer Richtung lokale Schranken erhalten. Denn die Abschätzung (III.3.6) benötigt den inversen Differentialoperator, der ein globaler Operator ist. Die Abschätzung (III.3.10) 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-Diffusionsgleichung. A priori hängen die Konstanten in den Fehlerabschätzungen von der

Péclet-Zahl ab und wachsen für abnehmende Diffusion an. Dieser Effekt kann durch eine verbesserte Analyse 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 (III.3.6), (III.3.10) und (III.3.11) analoge Abschätzungen beweisen.

Wir wenden uns nun dem Problem der adaptiven Gitterverfeinerung zu. Zunächst könnte man versuchen, den geschätzten Fehler  $\eta$  über alle Unterteilungen  $\mathcal{T}_h$  mit einer gegebenen Elementzahl zu minimieren. Dies ist jedoch ein hochgradig nicht lineares, extrem aufwändiges Optimierungsproblem. Einfache heuristische Argumente zeigen andererseits, dass bei einer optimalen Triangulierung alle Elemente etwa den gleichen Beitrag zu  $\eta$  liefern. Dies legt es nahe, Elemente, die einen zu großen Beitrag  $\eta_K$  liefern, zu unterteilen, und führt auf folgenden Algorithmus.

ALGORITHMUS III.3.5. (ADAPTIVE GITTERVERFEINERUNG)

- (0) Bestimme eine grobe Triangulierung  $\mathcal{T}_{h_0}$  von  $\Omega$ . 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, beende das Verfahren, 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 III.3.5 ist  $\gamma$  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 III.3.5 kann auch ggf. durch eine Vergrößerungsstrategie ergänzt werden. Dies ist für zeitabhängige Probleme wichtig.

Man kann zeigen, dass Algorithmus III.3.5 linear konvergiert, d.h. es gibt eine Konstante  $0 \leq \kappa < 1$ , die nur von den Daten  $A$  und  $\alpha$  abhängt, so dass mit jeder Iteration von Algorithmus III.3.5 der Fehler zwischen der schwachen Lösung der Differentialgleichung und der Finite Element Lösung auf dem aktuellen Gitter gemessen in der  $H^1$ -Norm mindestens um den Faktor  $\kappa$  reduziert wird.

Für die praktische Realisierung von Algorithmus III.3.5 müssen wir noch beschreiben, wie Dreiecke verfeinert werden und wie die Zulässigkeit der verfeinerten Triangulierung gesichert werden kann. Dabei müssen wir beachten, dass alle Triangulierungen die Regularitätsbedingung erfüllen sollen, d.h. die Dreieckswinkel sollen nicht zu klein oder zu groß werden. Dazu führen wir folgende Sprechweise ein (vgl. Abbildung III.3.2 und III.3.3):

- 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.

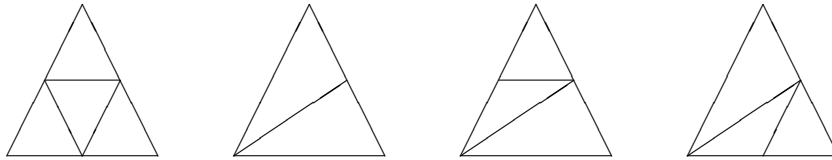


ABBILDUNG III.3.2. Rote, grüne und blaue Unterteilung eines Dreieckes

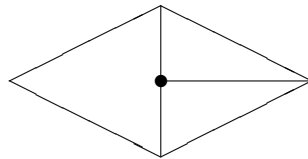


ABBILDUNG III.3.3. Beispiel für einen hängenden Knoten •

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 III.3.5 wird nun gemäß folgender Regeln durchgeführt:

- Ist  $\eta_K \geq \gamma\eta_k$ , unterteile  $K$  rot.
- Hat  $K$  drei hängende Knoten, unterteile  $K$  rot.
- Hat  $K$  zwei hängende Knoten, von denen keiner auf der längsten Kante liegt, unterteile  $K$  rot.
- Hat  $K$  zwei hängende Knoten, von denen einer auf der längsten Kante liegt, unterteile  $K$  blau.
- 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.

Die bisherigen Ergebnisse, insbesondere Satz III.3.3 und Algorithmus III.3.5, können direkt auf Unterteilungen in Parallelogramme übertragen werden. Dabei können (und müssen bei lokaler Unterteilung) Dreiecke und Parallelogramme gemischt werden. Bei der Unterteilung der Parallelogramme muss man offensichtlich neben der roten Unterteilung in vier neue ähnliche Parallelogramme zusätzliche Unterteilungen für Elemente mit 1, 2 und 3 hängenden Knoten vorsehen. Diese Regeln sind in Abbildung III.3.4 veranschaulicht.

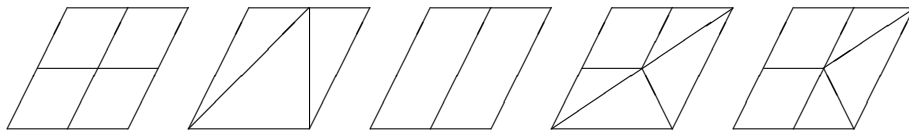


ABBILDUNG III.3.4. Rote, grüne und blaue Unterteilung eines Parallelogramms

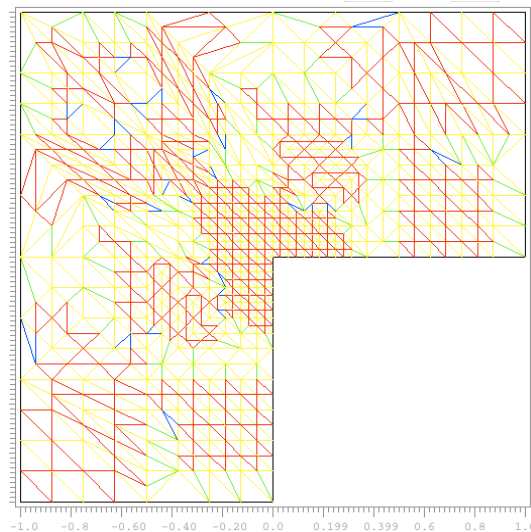


ABBILDUNG III.3.5. Feinstes adaptives Dreiecksgitter für das Problem aus Beispiel III.3.6(1)

BEISPIEL III.3.6. (1) Wir betrachten wie in Beispiel III.2.2 (S. 67) die Poissongleichung mit inhomogenen Dirichlet-Randbedingungen in dem L-förmigen Gebiet  $(-1, 1)^2 \setminus [(0, 1) \times (-1, 1)]$  mit exakter Lösung

$$u = r^{\frac{2}{3}} \sin\left(\frac{2}{3}\varphi\right).$$

Als größtes Gitter verwenden wir einmal eine Triangulierung mit 6 gleichschenkelig rechtwinkligen Dreiecken mit Hypotenusen in Richtung

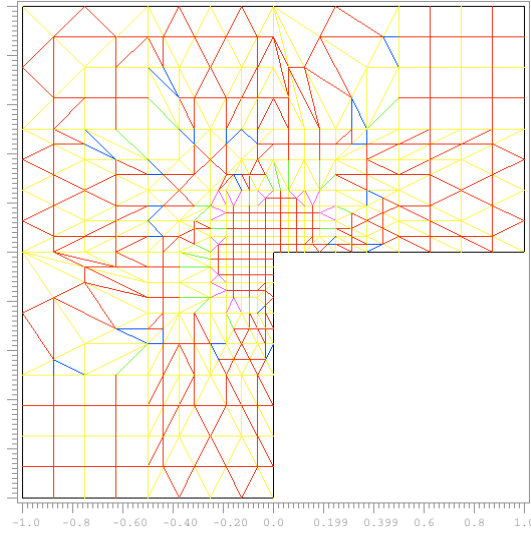


ABBILDUNG III.3.6. Feinstes adaptives Vierecksgitter für das Problem aus Beispiel III.3.6(1)

$(1, -1)$  und einmal eine Unterteilung in 3 Quadrate der Kantenlänge 1. Wir lösen die Differentialgleichung näherungsweise einmal mit einer uniformen Verfeinerung und einmal mit einer adaptiven Verfeinerung gemäß Algorithmus III.3.5 basierend auf dem Fehlerschätzer aus Satz III.3.3. Dabei wird die uniforme Verfeinerung so lange durchgeführt, bis die Speicherkapazität des benutzten Rechners (ein Macintosh G4-Powerbook) erschöpft ist. Der adaptive Algorithmus wird so lange durchgeführt, bis er eine Näherungslösung liefert, die etwa die gleiche Genauigkeit hat wie die Finite Element Lösung auf dem feinsten uniformen Gitter. In Tabelle III.3.1 geben wir für diese Beispiele jeweils die Zahl  $L$  der Gitter, die Zahl  $N$  der Unbekannten auf dem feinsten Gitter, den relativen Fehler gemessen in der  $H^1$ -Norm in Prozent auf dem feinsten Gitter und die Gesamtrechenzeit  $T$  in Sekunden an. Abbildungen III.3.5 und III.3.6 zeigen die feinsten bei der adaptiven Verfeinerung erzeugten Gitter.

(2) Wir betrachten die Reaktions-Diffusionsgleichung mit inhomogenen Dirich-Randbedingungen auf dem Quadrat  $(-1, 1)^2$ . Die Randfunktion  $u_D$ , die rechte Seite  $f$  und der Reaktionsterm  $\alpha$  sind so gewählt, dass die exakte Lösung gegeben ist durch

$$u = \tanh\left(100\left(x^2 + y^2 - \frac{1}{4}\right)\right) - 1.$$

Die exakte Lösung hat eine scharfe innere Grenzschicht entlang des Randes des Kreises um Null mit Radius  $\frac{1}{2}$ . Als größtes Gitter verwenden wir einmal eine Triangulierung mit 8 gleichschenkelig rechtwinkligen Dreiecken mit Hypotenusen in Richtung  $(1, -1)$  und einmal eine Unterteilung in 4 Quadrate der Kantenlänge 1. Ansonsten gehen wir wie

in Teil (1) vor. Die Ergebnisse sind in Tabelle III.3.2 wiedergegeben. Abbildungen III.3.7 und III.3.8 zeigen die feinsten bei der adaptiven Verfeinerung erzeugten Gitter.

TABELLE III.3.1. Vergleich uniformer und adaptiver Verfeinerung für das Problem aus Beispiel III.3.6 (1)

	Dreiecke		Vierecke	
	uniform	adaptiv	uniform	adaptiv
$L$	5	5	5	5
$N$	2945	718	2945	405
$\varepsilon(\%)$	1.3	1.5	3.6	3.9
$T(s)$	0.69	0.48	0.7	0.33

TABELLE III.3.2. Vergleich uniformer und adaptiver Verfeinerung für das Problem aus Beispiel III.3.6 (2)

	Dreiecke		Vierecke	
	uniform	adaptiv	uniform	adaptiv
$L$	5	8	5	6
$N$	3969	2934	3969	2466
$\varepsilon(\%)$	3.8	2.8	6.1	4.5
$T(s)$	0.93	0.72	0.92	0.65

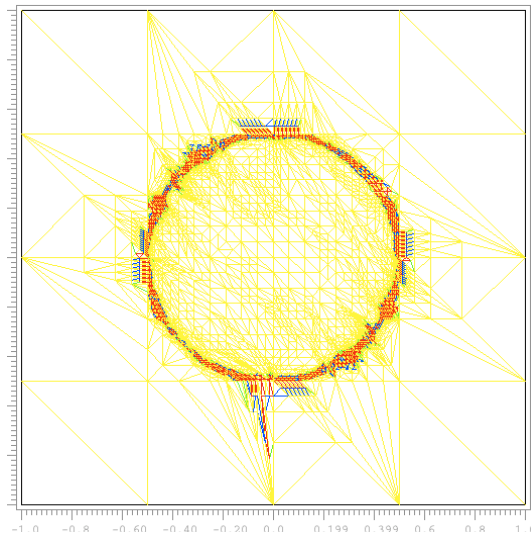


ABBILDUNG III.3.7. Feinstes adaptives Dreiecksgitter für das Problem aus Beispiel III.3.6(2)

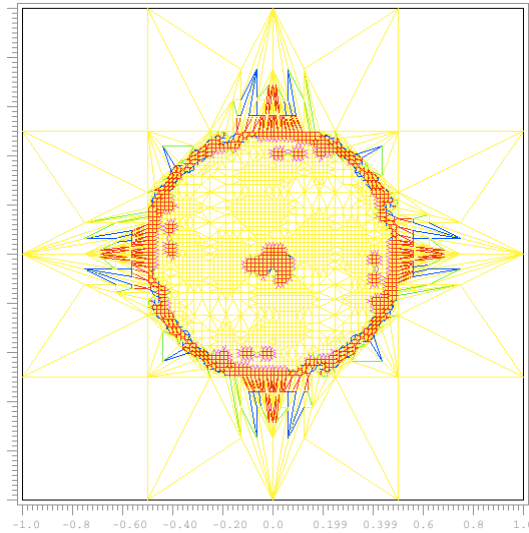


ABBILDUNG III.3.8. Feinstes adaptives Vierecksgitter für das Problem aus Beispiel III.3.6(2)

### III.4. Implementierung

In diesem Abschnitt gehen wir kurz auf die Implementierung der Methoden der letzten Paragraphen und die dafür benötigten Java-Datenstrukturen ein. Um die Notationen und den technischen Aufwand möglichst gering zu halten, beschränken wir uns auf zweidimensionale Probleme und lineare Dreiecks- und bilineare affine Viereckselemente. Wir erinnern daran, dass die beschriebenen Methoden als Java-Demonstrationsapplet implementiert sind und unter der Adresse

<http://www.rub.de/num1/demo/alf/ALF.html>

zur Verfügung stehen. Eine Benutzeranleitung ist unter der Adresse

<http://www.rub.de/num1/demo/alf/ALFUserGuide.htm>

online verfügbar bzw. kann unter der Adresse

<http://www.rub.de/num1/demo/alf/ALFUserGuide.pdf>

als pdf-Datei heruntergeladen werden.

Die Java-Klasse `node` realisiert das Konzept eines Knotens, d.h. eines Elementeckpunktes. Sie hat folgende Mitglieder:

- `c`: Dies ist ein zweidimensionales Feld vom Typ `double` und speichert die euklidischen Koordinaten des betreffenden Knotens.

- **t**: Diese Größe vom Typ `integer` beschreibt den Typ des Knotens. Es ist

$$t = \begin{cases} 0 & \text{falls der Knoten im Innern des} \\ & \text{Gebietes liegt,} \\ k & k > 0, \text{ falls der Knoten auf dem} \\ & k\text{-ten Dirichletrand liegt,} \\ -k & k > 0, \text{ falls der Knoten auf dem} \\ & k\text{-ten Neumannrand liegt.} \end{cases}$$

- **d**: Diese Größe vom Typ `integer` gibt die Nummer des Freiheitsgrades des betreffenden Knotens an. Sie ist gleich  $-1$ , falls der Knoten kein Freiheitsgrad ist.

Das Attribut **d** berücksichtigt, dass nicht jeder Knoten ein Freiheitsgrad ist, z.B. weil er auf einem Dirichletrand liegt. Für die Effizienz der Lösungsalgorithmen ist es aber wichtig, Freiheitsgrade fortlaufend zu nummerieren, damit z.B. Vektoradditionen und Skalarprodukte durch einfache Schleifen realisiert werden können. Durch das Attribut **d** wird eine einfache Verbindung zwischen der Menge der Knoten und der Teilmenge der Freiheitsgrade hergestellt.

Die Java-Klasse `element` realisiert das Konzept eines Elementes. Sie hat folgende Mitglieder:

- **nv**: Diese Größe vom Typ `integer` gibt die Zahl der Elementeckpunkte an und legt damit den Elementtyp, d.h. Dreieck oder Viereck, fest.
- **v**: Dies ist ein Feld der Länge 4 vom Typ `integer`. Es gibt die globalen Nummern der Elementeckpunkte an. Lokal, d.h. innerhalb des Elementes, werden die Eckpunkte fortlaufend beginnend mit 0 im mathematisch positiven Sinn durchnummeriert. Falls das Element ein Dreieck ist, ist  $v[3] = -1$ .
- **e**: Dies ist ein Feld der Länge 4 vom Typ `integer`. Es beschreibt die Nachbarschaftsbeziehungen zwischen den Elementen. Es gilt

$$e[i] = \begin{cases} j & \text{falls die } i\text{-te Kante des} \\ & \text{Elementes an das Element} \\ & \text{mit der Nummer } j \text{ grenzt,} \\ -1 & \text{falls die } i\text{-te Kante des} \\ & \text{Elementes Teil eines geraden} \\ & \text{Randstückes des Gebietes ist,} \\ -k - 2 & k > 0, \text{ falls die } i\text{-te Kante des} \\ & \text{Elementes Teil des } k\text{-ten} \\ & \text{gekrümmten Randstückes des} \\ & \text{Gebietes ist.} \end{cases}$$

Dabei werden die Kanten so nummeriert, dass die Kante mit Nummer  $i$  die Eckpunkte  $i + 1$  und  $i + 2$  als Endpunkte hat (vgl. Abbildungen III.4.1 und III.4.2), wobei Ausdrücke der Form  $i + 1$ ,  $i + 2$  usw. immer modulo der Größe  $nv$  zu nehmen sind.

- **p**: Diese Größe vom Typ `integer` gibt die Nummer des Eltern-elementes des betreffenden Elementes an.
- **t**: Diese Größe vom Typ `integer` beschreibt den Verfeinerungstyp des Elementes. Es ist (vgl. Abbildungen III.4.1 und III.4.2)

$$t = \begin{cases} 0 & \text{falls das Element nicht} \\ & \text{unterteilt wird,} \\ 5 & \text{falls das Element rot} \\ & \text{unterteilt wird,} \\ i + 1 & \text{falls die } i\text{-te Kante} \\ & \text{unterteilt wird,} \\ i + 1 + 5j & \text{falls die Kanten } i \text{ und } i + j, \\ & j > 0, \text{ unterteilt werden,} \\ i + 35 & \text{falls die Kanten } i, i + 1 \text{ und} \\ & i + 2 \text{ unterteilt werden,} \end{cases}$$

- **c**: Diese Größe vom Typ `integer` gibt die Nummer des ersten Kindes des Elementes an. Die weiteren Kinder werden wie in Abbildungen III.4.1 und III.4.2 angegeben fortlaufend nummeriert.

Auf den ersten Blick mag es ungewöhnlich erscheinen, die Informationen über Knoten und Elemente in verschiedenen Klassen zu halten. Dieser Ansatz bietet aber einige Vorteile:

- Er reduziert den Speicherbedarf. Die Koordinaten eines Knotens müssen so nur einmal gespeichert werden. Wenn Knoten und Elemente in einer gemeinsamen Klasse gehalten werden, müssen die Knotenkoordinaten etwa 4 bis 6 mal gespeichert werden.
- Die Elemente repräsentieren die Topologie des Gitters, die unabhängig ist von der aktuellen Positionierung der Knoten. Diese ändert sich z.B. nicht, wenn Knoten verschoben werden. Daher ist es bei dem bestehenden Ansatz wesentlich einfacher, sogenannte Gitterglättungsalgorithmen (mesh smoothing algorithms), die nur die Position der Knoten nicht aber die Gittertopologie beeinflussen, zu implementieren.

Bei der Gitterverfeinerung sind die Knoten komplett hierarchisch, d.h. ein Knoten der Unterteilung  $\mathcal{T}_i$  ist auch Knoten jeder Unterteilung  $\mathcal{T}_j$  mit  $j > i$ . Daher können alle Knoten in einem Feld vom Typ `node`

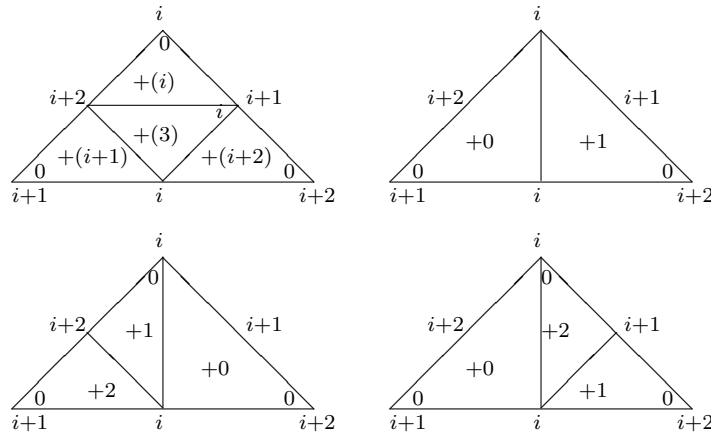


ABBILDUNG III.4.1. Nummerierung der Eckpunkte und Kanten eines Dreieckes und seiner Nachfahren. Die Zahlen  $i, \dots, i+2$  sind modulo 3 zu nehmen. Fehlende Nummern von Eckpunkten innerhalb eines Elementes sind durch fortlaufende Nummerierung im mathematisch positiven Sinn zu ergänzen. Ein Elementeintrag  $+(k)$  bedeutet, dass das Element die Nummer  $j+k$  erhält, wenn  $j$  der Wert des Attributes  $c$  des Elternelementes ist.

konsekutiv gespeichert werden. Ein Knoten wird zu dem Feld hinzugefügt, sobald er bei einer Elementunterteilung erzeugt wird.

Da die Gitter i.a. nur teilweise verfeinert werden, sind die Elemente nicht komplett hierarchisch. Daher wird die Information über alle Elemente aller Unterteilungen gespeichert. Dies geschieht mittels eines Feldes vom Typ `level`. Seine Länge ist die maximal mögliche Zahl von Verfeinerungsstufen. Die Java-Klasse `level` hat folgende Mitglieder:

- `nn`: die Zahl der Knoten des aktuellen Gitters,
- `cn`: die kumulative Zahl der Knoten des aktuellen und aller größeren Gitter,
- `first`: die Adresse des ersten Elementes des aktuellen Gitters,
- `last`: die Zahl der Elemente des aktuellen Gitters,
- `nt`: die Zahl der Dreiecke des aktuellen Gitters,
- `nq`: die Zahl der Vierecke des aktuellen Gitters,
- `ne`: die Zahl der Kanten des aktuellen Gitters,
- `dof`: die Zahl der Freiheitsgrade des aktuellen Gitter.

Alle Größen haben den Typ `integer`.

Bezeichnet `ll` das Feld vom Typ `level`, das die Gitterinformation speichert, hat daher eine Schleife über alle Knoten des Gitters  $k$  die Form

```
for(i = 0; i < ll[k].nn; i++){...}
```

Eine Schleife über alle Elemente des Gitters  $k$  dagegen hat die Form

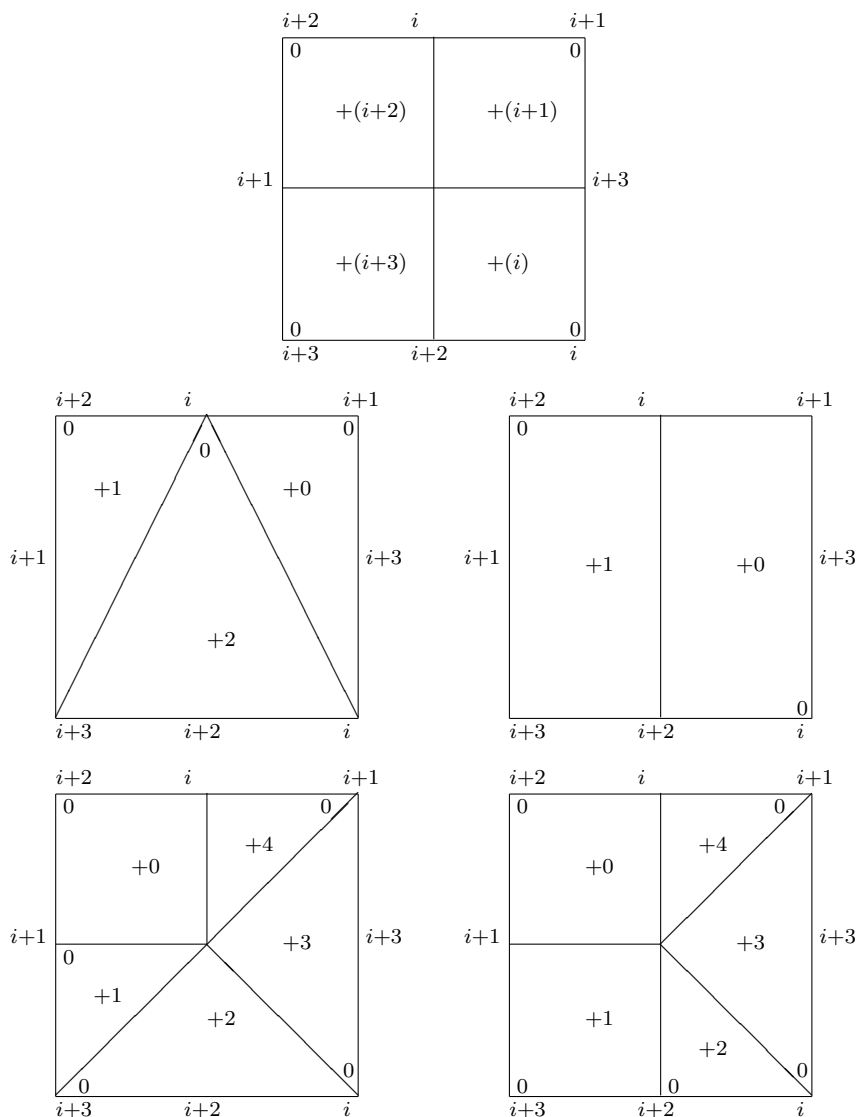


ABBILDUNG III.4.2. Nummerierung der Eckpunkte und Kanten eines Viereckes und seiner Nachfahren. Die Zahlen  $i, \dots, i+3$  sind modulo 4 zu nehmen. Fehlende Nummern von Eckpunkten innerhalb eines Elementes sind durch fortlaufende Nummerierung im mathematisch positiven Sinn zu ergänzen. Ein Elementeintrag  $+(k)$  bedeutet, dass das Element die Nummer  $j+k$  erhält, wenn  $j$  der Wert des Attributes  $c$  des Elternelementes ist.

```
for(i = ll[[k].first; i < ll[k].last; i++){...}
```

Für das Aufstellen des diskreten Problems muss man die Größen

$$b_i = l(v_i) \text{ und } a_{ij} = a(v_i, v_j)$$

berechnen. Dabei ist  $v_i$  die nodale Basisfunktion zum  $i$ -ten Knoten. Die Berechnung geschieht elementweise. Außerdem ist zu beachten, dass Knoten auf dem Dirichletrand, die kein Freiheitsgrad sind, ausgeblendet werden. Bezeichne für ein Element  $K$  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, und mit `vertices` das Feld vom Typ `node` das die Knoteninformation enthält. Dann erfolgt die Berechnung der rechten Seite und der Steifigkeitsmatrix nach folgendem Schema:

für alle Elemente  $K$  des aktuellen Gitters  
 für alle Knotennummern  $i$  von  $K$   
 $j = \text{vertices}[K.v[i]].d$   
 $b_j = b_j + l_K(v_i)$   
 für alle Knotennummern  $l > i$  von  $K$   
 $m = \text{vertices}[K.v[l]].d$   
 $a_{mj} = a_{mj} + a_K(v_l, v_i)$

Dabei wird  $l_K(v_i)$  mit einer Quadraturformel berechnet. Für ein Dreieck und die Formel aus Beispiel III.1.1(2) (S. 63) ergibt sich z.B.

$$l_K(v_i) = \frac{1}{12} D \left\{ f\left(\frac{1}{2}(z_i + z_{i+1})\right) + f\left(\frac{1}{2}(z_i + z_{i+2})\right) \right\}.$$

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  $i + 2$  modulo 3 zu verstehen. Ähnliche Formeln ergeben für Vierecke.

Ganz analog geht man bei der Berechnung der Größen  $a_K(v_l, v_i)$  vor. Dabei sind zusätzlich folgende Punkte zu beachten:

- Bei Dreieckselementen sind die Ableitungen der Basisfunktionen stückweise konstant.
- Bei Viereckselementen ist die partielle Ableitung nach  $x$  bzw.  $y$  einer Basisfunktion eine lineare Funktion nur der Variablen  $y$  bzw.  $x$ .
- Die Steifigkeitsmatrix ist dünn besetzt; das Element  $i, j$  ist höchstens dann von Null verschieden, wenn die entsprechenden Eckpunkte auf einer gemeinsamen Elementkante liegen.

Damit ergibt sich für den Gradienten  $\nabla_K v_i$  der Basisfunktion  $v_i$  eingeschränkt auf das Element  $K$  die Formel

$$\nabla_K v_i = \frac{1}{D} \begin{pmatrix} z_{i+2,1} - z_{i,1} - (z_{i+2,2} - z_{i,2}) \\ z_{i+1,2} - z_{i,2} - (z_{i+1,1} - z_{i,1}) \end{pmatrix}.$$

Eine analoge Formel gilt für  $\nabla_K v_l$ . Da die Ableitungen konstant sind können die Integrale mit der Mittelpunktsregel aus Beispiel III.1.1(1) (S. 63) berechnet werden.

Bei Viereckselementen geht man analog vor. Da die Ableitungen linear

sind, können sie wieder durch Differenzenquotienten und die Integrale durch die Simpsonregel aus Beispiel III.1.1(7) (S. 63) berechnet werden.

Die dünne Besetzung der Steifigkeitsmatrix wird durch folgende spezielle Speichertechnik berücksichtigt:

- Es werden nur die von Null verschiedenen Matrixelemente gespeichert.
- Die relevanten Matrixelemente werden hintereinander in einem Feld `m` vom Typ `double` abgelegt.
- Ein zusätzliches Feld `c` vom Typ `integer` speichert den Spaltenindex des entsprechenden Eintrages in `m`, d.h. `m[i]` steht in der Spalte `c[i]`.
- Ein zusätzliches Feld `r` vom Typ `integer` speichert den Beginn der Zeilen, d.h. alle Einträge `m[i]` mit `r[k] ≤ i < r[k+1]` stehen in der  $k$ -ten Zeile.
- Ein zusätzliches Feld `d` vom Typ `integer` speichert die Adressen der Diagonalelemente, d.h. `m[d[k]]` ist das  $k$ -te Diagonalelement  $a_{kk}$ .

Dabei ist berücksichtigt, dass die Steifigkeitsmatrix wegen des Konvektionstermes i.a. nicht symmetrisch ist. Wenn man sich auf Reaktions-Diffusionsgleichungen beschränkt, erhält man symmetrische Matrizen und kann sich auf die Speicherung der Matrixelemente oberhalb der Diagonalen einschließlich der Diagonalen beschränken.

Mit der beschriebenen Datenstruktur hat eine Matrix-Vektor-Multiplikation z.B. folgende Form:

```
for(i = 0; i < nn; i++)
  y[i] = 0;
  for(j = r[i]; j < r[i+1]; j++)
    y[i] += m[j]*x[c[j]];
```

Dabei ist `nn` die Zahl der Unbekannten.



## KAPITEL IV

### Ergänzungen

#### IV.1. Nichtkonforme Methoden

Bisher haben wir stets konforme Finite Element Methoden 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 §I.2 eine abstrakte Situation. Sei dazu  $(X, \|\cdot\|_X)$  ein Banachraum,  $l \in \mathcal{L}(X, \mathbb{R})$  und  $a \in \mathcal{L}^2(X, \mathbb{R})$  symmetrisch und koerziv. Zu lösen ist das Variationsproblem

$$(IV.1.1) \quad a(u, v) = l(v) \quad \forall v \in X.$$

Weiter seien  $(X_h, \|\cdot\|_{X_h})$  ein endlich dimensionaler Banachraum,  $l_h \in \mathcal{L}(X_h, \mathbb{R})$  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  $\alpha_h$  von  $a_h$  seien gleichmäßig in  $h$  nach oben bzw. – für  $\alpha_h$  – nach unten weg von 0 beschränkt. Die diskrete Approximation von (IV.1.1) lautet

$$(IV.1.2) \quad a_h(u_h, v_h) = l_h(v_h) \quad \forall v_h \in X_h.$$

Wegen Satz I.2.1 (S. 16) haben (IV.1.1) und (IV.1.2) 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  $\alpha_h$  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  $\|\cdot\|_{X_h}$  auch für Elemente von  $X$  einen Sinn machen und für solche Elemente mit  $a$ ,  $l$  bzw.  $\|\cdot\|_X$  übereinstimmen. Der folgende Satz ist unter diesen Annahmen das nicht-konforme Analogon zum Céa-Lemma I.2.2 (S. 17).

**SATZ IV.1.1 (ZWEITES STRANG-LEMMA).** *Unter den obigen Voraussetzungen gilt die Fehlerabschätzung*

$$\begin{aligned} \|u - u_h\|_{X_h} &\leq \left(1 + \frac{A_h}{\alpha_h}\right) \inf_{v_h \in X_h} \|u - v_h\|_{X_h} \\ &\quad + \frac{1}{\alpha_h} \sup_{\substack{w_h \in X_h \\ \|w_h\|_{X_h}=1}} |a_h(u, w_h) - l_h(w_h)|. \end{aligned}$$

Dabei ist  $A_h = \|a_h\|_{\mathcal{L}^2}$  und  $\alpha_h$  die Koerzivitätskonstante von  $a_h$ .

BEWEIS. Sei  $v_h \in X_h$  beliebig und  $w_h = u_h - v_h$ . Aus der Koerzitivität von  $a_h$  folgt dann

$$\begin{aligned} \alpha_h \|w_h\|_{X_h}^2 &\leq a_h(w_h, w_h) \\ &= a_h(u_h - v_h, w_h) \\ &= a_h(u - v_h, w_h) + a_h(u_h, w_h) - a_h(u, w_h) \\ &= a_h(u - v_h, w_h) + l_h(w_h) - a_h(u, w_h) \\ &\leq A_h \|u - v_h\|_{X_h} \|w_h\|_{X_h} \\ &\quad + |l_h(w_h) - a_h(u, w_h)|. \end{aligned}$$

Wegen

$$\|u - u_h\|_{X_h} \leq \|u - v_h\|_{X_h} + \|w_h\|_{X_h}$$

folgt hieraus die Behauptung.  $\square$

Der folgende Satz ist die nicht-konforme Variante des Dualitätsargumentes von Aubin-Nitsche, Satz I.2.3 (S. 18).

SATZ IV.1.2. *Die Bezeichnungen und Voraussetzungen seien wie in Satz IV.1.1. Zusätzlich sei  $H$  ein Hilbertraum mit Skalarprodukt  $(\cdot, \cdot)_H$  und Norm  $\|\cdot\|_H$ , 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

$$\begin{aligned} \|u - u_h\|_H &\leq \sup_{\substack{\varphi \in H \\ \|\varphi\|_H=1}} \left\{ A_h \|u - u_h\|_{X_h} \|u_\varphi - u_{\varphi,h}\|_{X_h} \right. \\ &\quad \left. + |a_h(u - u_h, u_\varphi) - (\varphi, u - u_h)_H| \right. \\ &\quad \left. + |a_h(u, u_\varphi - u_{\varphi,h}) - l_h(u_\varphi - u_{\varphi,h})| \right\}. \end{aligned}$$

BEWEIS. Da  $X \hookrightarrow H$  dicht und  $X_h \subset H$  ist, ist

$$(IV.1.3) \quad \|u - u_h\|_H = \sup_{\substack{\varphi \in H \\ \|\varphi\|_H=1}} (\varphi, u - u_h)_H.$$

Sei nun  $\varphi \in H$  mit  $\|\varphi\|_H = 1$  beliebig. Dann folgt

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

$$+ a_h(u - u_h, u_{\varphi,h}).$$

Für den zweiten und dritten Summanden dieser Gleichung folgt

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

und

$$\begin{aligned} a_h(u - u_h, u_{\varphi,h}) &= a_h(u, u_{\varphi,h}) - a_h(u, u_\varphi) \\ &\quad + \underbrace{a_h(u, u_\varphi)}_{=l(u_\varphi)} - \underbrace{a_h(u_h, u_{\varphi,h})}_{=l_h(u_{\varphi,h})} \\ &= -a_h(u, u_\varphi - u_{\varphi,h}) + l_h(u_\varphi - u_{\varphi,h}). \end{aligned}$$

Also ist

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

Hieraus und aus (IV.1.3) folgt die Behauptung.  $\square$

Wir wenden diese abstrakten Ergebnisse auf ein einfaches, aber typisches Modellproblem an: die CROUZEIX-RAVIART DISKRETISIERUNG der ebenen Poissongleichung mit homogenen Dirichlet Randbedingungen. Dabei ist  $\Omega \subset \mathbb{R}^2$  ein beschränktes, zusammenhängendes Gebiet mit stückweise geradem Rand  $\Gamma$ ,  $X = H_0^1(\Omega)$ ,  $H = L^2(\Omega)$ ,  $l(v) = \int_\Omega f v$  und  $a(u, v) = \int_\Omega \nabla u^T \nabla v$ .

Wie in §II.3 bezeichnet  $\mathcal{T}_h = \{K_i : 1 \leq i \leq m_h\}$  eine zulässige und reguläre Unterteilung von  $\Omega$  in Dreiecke. Wir bezeichnen die Dreiecks-kanten 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 kollineare 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

$$\begin{aligned} X_h &= \left\{ \varphi \in L^2(\Omega) : \varphi|_K \in \mathbb{P}_1 \ \forall K \in \mathcal{T}_h, \right. \\ &\quad \varphi \text{ ist stetig in } m_E, E \in \mathcal{E}_{h,\Omega}, \\ &\quad \left. \varphi(m_E) = 0 \ \forall E \in \mathcal{E}_h \setminus \mathcal{E}_{h,\Omega} \right\}. \end{aligned}$$

$X_h$  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 §II.3, 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 §III.4 die  $i$ -te Kante vereinbarungsgemäß dem  $i$ -ten Eckpunkt gegenüber und Knotennummern werden modulo 3 genommen.

Die Funktionen in  $X_h$  sind nicht stetig und verschwinden nicht identisch auf  $\Gamma$ . 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  $\|\cdot\|_{X_h}$ , die Linearform  $l_h$  und die Bilinearform  $a_h$  werden definiert durch

$$\begin{aligned} \|u_h\|_{X_h} &= \left\{ \sum_{K \in \mathcal{T}_h} |u_h|_{H^1(K)}^2 \right\}^{1/2}, \\ l_h(u_h) &= \sum_{K \in \mathcal{T}_h} \int_K f u_h, \\ a_h(u_h, v_h) &= \sum_{K \in \mathcal{T}_h} \int_K \nabla u_h^T \nabla v_h. \end{aligned}$$

Offensichtlich sind die Voraussetzungen der Sätze IV.1.1 und IV.1.2 erfüllt. Insbesondere ist  $A_h = \alpha_h = 1$ .

**SATZ IV.1.3.** *Sei  $u \in H_0^1(\Omega) \cap H^2(\Omega)$  die Lösung von (IV.1.1) und*

$$L_u(w) = a_h(u, w) - l_h(w) \quad \forall w \in X \oplus X_h.$$

*Dann gilt für alle  $w \in X \oplus X_h$*

$$|L_u(w)| \leq ch|u|_2 \|w\|_{X_h}.$$

*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 §III.3

$$\begin{aligned} L_u(w) &= \sum_{K \in \mathcal{T}_h} \int_K \nabla u^T \nabla w - \sum_{K \in \mathcal{T}_h} \underbrace{\int_K f w}_{= \int_K -\Delta u w} \\ &= \sum_{K \in \mathcal{T}_h} \int_K \nabla u^T \nabla w + \sum_{K \in \mathcal{T}_h} \underbrace{\int_K \Delta u w}_{= - \int_K \nabla u^T \nabla w + \int_{\partial K} \partial_{n_K} u w} \end{aligned}$$

$$\begin{aligned}
&= \sum_{K \in \mathcal{T}_h} \int_{\partial K} \partial_{n_K} u w \\
&= \sum_{E \in \mathcal{E}_{h,\Omega}} \int_E [\partial_{n_E} u w]_E + \sum_{E \in \mathcal{E}_h \setminus \mathcal{E}_{h,\Omega}} \int_E \partial_n u w.
\end{aligned}$$

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

$$\bar{w}_E = h_E^{-1} \int_E w,$$

so folgt

$$\begin{aligned}
L_u(w) &= \sum_{E \in \mathcal{E}_{h,\Omega}} \int_E [\partial_{n_E} u (w - \bar{w}_E)]_E \\
&\quad + \sum_{E \in \mathcal{E}_h \setminus \mathcal{E}_{h,\Omega}} \int_E \partial_n u (w - \bar{w}_E).
\end{aligned}$$

Bezeichne mit  $I_h : X \cap H^2(\Omega) \rightarrow S_{h,0}^{1,0} \subset X_h$  den nodalen Interpolationsoperator aus §II.3. Da für jede Kante  $E$

$$\int_E (w - \bar{w}_E) = 0$$

und  $\partial_{n_E}(I_h u)|_E$  konstant ist, folgt

$$\begin{aligned}
L_u(w) &= \sum_{E \in \mathcal{E}_{h,\Omega}} \int_E [\partial_{n_E} (u - I_h u) (w - \bar{w}_E)]_E \\
&\quad + \sum_{E \in \mathcal{E}_h \setminus \mathcal{E}_{h,\Omega}} \int_E \partial_n (u - I_h u) (w - \bar{w}_E).
\end{aligned} \tag{IV.1.4}$$

Wegen  $|n| = |n_E| = 1$  folgt hieraus mit der Cauchy-Schwarzschen Ungleichung

$$|L_u(w)| \leq \sum_{E \in \mathcal{E}_h} \|\nabla(u - I_h u)\|_{L^2(E)} \|w - \bar{w}_E\|_{L^2(E)}.$$

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  $F_K : \hat{K} \rightarrow K$  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

$$\bar{w}_E = h_E^{-1} \int_E w = \int_{\hat{E}} \hat{w}.$$

Also ist

$$\int_{\hat{E}} (\hat{w} - \bar{w}_E) = 0.$$

Wie im Beweis der Poincaréschen Ungleichung I.1.21 (S. 15) folgt, dass es eine Konstante  $\hat{c}$  gibt mit

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

für alle  $\varphi \in H^1(\hat{K})$  mit  $\int_{\hat{E}} \varphi = 0$ . Hieraus folgt mit dem Transformationssatz und dem Spursatz I.1.12 (S. 12)

$$\begin{aligned} \|w - \bar{w}_E\|_{L^2(E)} &= h_E^{1/2} \|\hat{w} - \bar{w}_E\|_{L^2(\hat{E})} \\ &\leq h_E^{1/2} c \left\{ \|\hat{w} - \bar{w}_E\|_{L^2(\hat{K})}^2 + |\hat{w} - \bar{w}_E|_{H^1(\hat{K})}^2 \right\}^{1/2} \\ &\leq h_E^{1/2} c (1 + \hat{c}^2)^{1/2} |\hat{w} - \bar{w}_E|_{H^1(\hat{K})} \\ &= h_E^{1/2} c (1 + \hat{c}^2)^{1/2} |\hat{w}|_{H^1(\hat{K})} \\ &\leq h_E^{1/2} c (1 + \hat{c}^2)^{1/2} |\det DF_K|^{-1/2} \|DF_K\|_{\mathcal{L}} |w|_{H^1(K)} \\ &\leq c' h_E^{1/2} |w|_{H^1(K)}. \end{aligned}$$

also

$$(IV.1.5) \quad \|w - \bar{w}_E\|_{L^2(E)} \leq c' h_E^{1/2} |w|_{H^1(K)}.$$

Hierbei haben wir im letzten Schritt ausgenutzt, dass wegen Satz II.3.2 (S. 43) und der Regularität von  $\mathcal{T}_h$  der Ausdruck  $|\det DF_K|^{-1/2} \|DF_K\|_{\mathcal{L}}$  durch eine von  $c_{\mathcal{T}}$  abhängige Konstante beschränkt ist.

Mit den gleichen Argumenten folgt aus dem Spursatz I.1.12 (S. 12), Satz II.3.1 (S. 40) und Satz II.3.7 (S. 47) durch Transformation auf das Referenzelement mit  $\hat{u} = u \circ F_K$

$$\begin{aligned} &\|\nabla(u - I_h u)\|_{L^2(E)} \\ &= h_E^{-1/2} \|\nabla(\hat{u} - \hat{\pi} \hat{u})\|_{L^2(\hat{E})} \\ &\leq c h_E^{-1/2} \|\hat{u} - \hat{\pi} \hat{u}\|_{H^2(\hat{K})} \\ &= c h_E^{-1/2} \left\{ \|\hat{u} - \hat{\pi} \hat{u}\|_{L^2(\hat{K})}^2 + |\hat{u} - \hat{\pi} \hat{u}|_{H^1(\hat{K})}^2 + |\hat{u}|_{H^2(\hat{K})}^2 \right\}^{1/2} \\ &\leq c' h_E^{-1/2} |\hat{u}|_{H^2(\hat{K})} \\ &\leq c' h_E^{-1/2} |\det DF_K|^{-1/2} \|DF_K\|_{\mathcal{L}}^2 |u|_{H^2(\hat{K})} \\ &\leq c'' h_E^{1/2} |u|_{H^2(K)}. \end{aligned}$$

also

$$(IV.1.6) \quad \|\nabla(u - I_h u)\|_{L^2(E)} \leq c'' h_E^{1/2} |u|_{H^2(K)}.$$

Hierbei haben wir im letzten Schritt ausgenutzt, dass wegen Satz II.3.2 (S. 43) und der Regularität von  $\mathcal{T}_h$  die Größe  $h_E^{-1} |\det DF_K|^{-1/2} \|DF_K\|_{\mathcal{L}}^2$  durch eine von  $c_{\mathcal{T}}$  abhängige Konstante beschränkt ist.

Aus den Abschätzungen (IV.1.4) – (IV.1.6) und der Cauchy-Schwarzschen Ungleichung für Summen folgt

$$|L_u(u)| \leq c \sum_{E \in \mathcal{E}_h} h_E |w|_{H^1(K)} |u|_{H^2(K)} \leq c' h |w|_{X_h} |u|_2.$$

□

Nach diesen Vorbereitungen können wir eine Fehlerabschätzung für die Crouzeix-Raviart Diskretisierung beweisen.

**SATZ IV.1.4.** *Seien  $u \in H_0^1(\Omega) \cap H^2(\Omega)$  die schwache Lösung der Poissongleichung und  $u_h \in X_h$  die Lösung der Crouzeix-Raviart Diskretisierung.  $\Omega$  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  $\Omega$  und der Konstanten  $c_{\mathcal{T}}$  aus der Regularitätsbedingung an  $\mathcal{T}_h$  ab.

**BEWEIS.** Da  $S_{h,0}^{1,0} \subset X_h$  ist, folgt aus Satz II.3.3 (S. 43)

$$\begin{aligned} \inf_{v_h \in X_h} \|u - v_h\|_{X_h} &\leq \inf_{w_h \in S_{h,0}^{1,0}} \|u - w_h\|_{X_h} \\ &= \inf_{w_h \in S_{h,0}^{1,0}} |u - w_h|_1 \\ &\leq |u - I_h u|_1 \\ &\leq ch |u|_2. \end{aligned}$$

Damit folgt die erste Fehlerabschätzung aus den Sätzen IV.1.1 und IV.1.3.

Satz IV.1.3 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 IV.1.3 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 IV.1.2. □

BEMERKUNG IV.1.5. (1) Sei  $J(u) = \frac{1}{2}a_h(u, u) - l_h(u)$ . Wegen  $S_{h,0}^{1,0} \subset X$  ist

$$\inf_{u \in X} J(u) \leq \inf_{u_h \in S_{h,0}^{1,0}} J(u_h).$$

Wegen  $S_{h,0}^{1,0} \subset X_h$  ist ebenso

$$\inf_{u_h \in X_h} J(u_h) \leq \inf_{u_h \in S_{h,0}^{1,0}} J(u_h).$$

Daher ist das Minimum von  $J$  bei nicht-konformen Methoden kleiner als bei konformen Methoden und häufig dichter am kontinuierlichen Minimum.

(2) Satz IV.1.4 benötigt im Gegensatz zu Satz II.3.4 (S. 45) 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  $\Omega$  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

$$L_u(w) = 0 \quad \forall w \in H_0^1(\Omega).$$

Dies gilt aber nicht für  $L^2$ -Funktionen, da  $L_u$  auf  $L^2(\Omega)$  nicht stetig fortsetzbar ist.

## IV.2. Gemischte Methoden

Um technische Schwierigkeiten zu vermeiden und die wesentlichen Aspekte besser herauszuarbeiten, beschränken wir uns im Folgenden auf ein einfaches Modellbeispiel: die Poissongleichung in einem zwei-dimensionalen, beschränkten, *konvexen* Polygon  $\Omega$  mit homogenen Dirichlet-Randbedingungen

$$(IV.2.1) \quad \begin{aligned} -\Delta u &= f && \text{in } \Omega \\ u &= 0 && \text{auf } \Gamma. \end{aligned}$$

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 Spannungen viel wichtiger. Im Rahmen dieses einfachen Modells ist dies der Gradient  $\nabla u$  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 (IV.2.1) über in das folgende Differentialgleichungssystem 1. Ordnung

$$(IV.2.2) \quad \begin{aligned} \sigma - \nabla u &= 0 && \text{in } \Omega \\ -\operatorname{div} \sigma &= f && \text{in } \Omega \\ u &= 0 && \text{auf } \Gamma. \end{aligned}$$

Multiplizieren wir die erste Gleichung von (IV.2.2) mit einem hinreichend glatten Vektorfeld  $\tau$ , integrieren über  $\Omega$ , wenden den Gaußschen Integralsatz an und nutzen die Randbedingungen für  $u$  aus, erhalten wir

$$\begin{aligned} 0 &= \int_{\Omega} \sigma \cdot \tau - \int_{\Omega} \tau \cdot \nabla u \\ &= \int_{\Omega} \sigma \cdot \tau + \int_{\Omega} u \operatorname{div} \tau - \int_{\partial\Omega} \underbrace{u}_{=0} \tau \cdot n \\ &= \int_{\Omega} \sigma \cdot \tau + \int_{\Omega} u \operatorname{div} \tau. \end{aligned}$$

Offensichtlich ist die zweite Gleichung von (IV.2.2) äquivalent zu

$$-\int_{\Omega} v \operatorname{div} \sigma = \int_{\Omega} f v \quad \forall v \in L^2(\Omega).$$

Diese Beobachtung führt auf folgende schwache Formulierung von Problem (IV.2.1):

Finde  $[\sigma, u] \in X = M \times Q$ , so dass

$$(IV.2.3) \quad \begin{aligned} \int_{\Omega} \sigma \cdot \tau + \int_{\Omega} u \operatorname{div} \tau &= 0 \quad \forall \tau \in M \\ -\int_{\Omega} v \operatorname{div} \sigma &= \int_{\Omega} f v \quad \forall v \in Q. \end{aligned}$$

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 I.1.6 (S. 9) kann man zeigen, dass  $M$  mit dieser Norm ein Banachraum ist.

Die Herleitung von (IV.2.3) zeigt, dass dieses Problem im üblichen Sinne zu (IV.2.2) äquivalent ist: Jede klassische Lösung von (IV.2.2) löst auch (IV.2.3) und jede hinreichend glatte Lösung von (IV.2.3) ist auch eine klassische Lösung von (IV.2.2).

Problem (IV.2.3) ist ein sogenanntes GEMISCHTES oder SATTEL-PUNKTSPROBLEM. 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 (IV.2.3) 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 (IV.2.3) offensichtlich äquivalent zu

$$(IV.2.4) \quad a([\sigma, u], [\tau, v]) = l([\tau, v]) \quad \forall [\tau, v] \in X.$$

Auf den ersten Blick scheint Problem (IV.2.4) in den abstrakten Rahmen von §I.2 zu passen. Aber die Bilinearform  $a$  ist nicht symmetrisch und nicht koerziv. Dies spiegelt die Sattelpunktsstruktur von (IV.2.3) wider. Statt der Koerzivität erfüllt  $a$  allerdings eine sogenannte INF-SUP BEDINGUNG:

SATZ IV.2.1 (INF-SUP BEDINGUNG). *Es gibt eine Konstante  $\alpha > 0$ , die nur von  $\Omega$  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  $\Omega$  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

$$\begin{aligned} \operatorname{div} \tau_u &= u, \\ \|\tau_u\|_0 &\leq c_0 \|u\|_0. \end{aligned}$$

Dabei hängt die Konstante  $c_0$  nur vom Durchmesser von  $\Omega$  ab. O.E. ist  $c_0 \geq 1$ . Hieraus folgt

$$a([\sigma, u], [\tau_u, 0]) = \int_{\Omega} \sigma \cdot \tau_u + \|u\|_0^2$$

$$\begin{aligned}
&\geq -\|\sigma\|_0 \|\tau_u\|_0 + \|u\|_0^2 \\
&\geq -\|\sigma\|_0 c_0 \|u\|_0 + \|u\|_0^2 \\
&\geq \frac{1}{2} \|u\|_0^2 - \frac{1}{2} c_0^2 \|\sigma\|_0^2.
\end{aligned}$$

Setze

$$\rho = c_0^2 \sigma + \tau_u, \quad w = c_0^2 u - \frac{1}{2} \operatorname{div} \sigma.$$

Dann folgt

$$\begin{aligned}
a([\sigma, u], [\rho, w]) &= c_0^2 a([\sigma, u], [\sigma, u]) + a([\sigma, u], [\tau_u, 0]) \\
&\quad + \frac{1}{2} a([\sigma, u], [0, -\operatorname{div} \sigma]) \\
&\geq c_0^2 \|\sigma\|_0^2 + \frac{1}{2} \|u\|_0^2 - \frac{1}{2} c_0^2 \|\sigma\|_0^2 + \frac{1}{2} \|\operatorname{div} \sigma\|_0^2 \\
&\geq \frac{1}{2} \|[\sigma, u]\|_X^2
\end{aligned}$$

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  $\alpha = \frac{1}{4c_0^2}$ .  $\square$

Aus Satz IV.2.1 folgt:

**SATZ IV.2.2.** *Problem (IV.2.3) besitzt eine eindeutige Lösung.*

**BEWEIS.** Sei  $u \in H_0^1(\Omega)$  die schwache Lösung von (IV.2.1) im Sinne von Definition I.3.1 (S. 22). Da  $\Omega$  konvex ist, folgt aus Satz I.3.6 (S. 26)  $u \in H^2(\Omega)$ . Also ist  $\sigma = \nabla u \in H(\operatorname{div}, \Omega)$ . Aus der Herleitung von (IV.2.3) folgt, dass  $[\sigma, u]$  eine Lösung von (IV.2.3) ist. Wir müssen also noch die Eindeutigkeit zeigen. Dazu reicht es, zu zeigen, dass das homogene Problem, d.h. (IV.2.3) mit  $f = 0$  bzw. (IV.2.4) mit  $l = 0$ , nur die triviale Lösung hat. Ist aber  $[\sigma, u]$  eine Lösung des homogenen Problems (IV.2.4), so folgt aus Satz IV.2.1

$$\alpha \|[\sigma, u]\|_X \leq \sup_{[\tau, v] \in X \setminus \{0\}} \frac{a([\sigma, u], [\tau, v])}{\|[\tau, v]\|_X} = 0.$$

Also ist  $\sigma = u = 0$ .  $\square$

Für die Diskretisierung von (IV.2.3) betrachten wir nur das einfachste Beispiel, das sogenannte RAVIART-THOMAS ELEMENT niedrigster Ordnung. Dazu bezeichnet  $\mathcal{T}_h$  eine Familie zulässiger und regulärer Triangulierungen von  $\Omega$  und  $\mathcal{E}_h$  die Menge der Dreiecksanten in  $\mathcal{T}_h$ . Für ein Dreieck  $K$  sei

$$RT(K) = \left\{ \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \gamma \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : \alpha, \beta, \gamma \in \mathbb{R} \right\}.$$

SATZ IV.2.3. Sei  $K$  ein Dreieck und  $n_K$  das äußere Einheitsnormalenfeld zu  $\partial K$ . Dann gilt für jedes  $\sigma \in RT(K)$ :

- (1)  $\sigma \cdot n_K$  ist konstant auf den Kanten von  $K$ .
- (2)  $\sigma$  ist eindeutig bestimmt durch die Werte von  $\sigma \cdot n_K$  auf den Kanten von  $K$ .

BEWEIS. AD (1): Die Funktion  $x \mapsto x \cdot n_K$  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 Gaußschen Integralsatz

$$\begin{aligned} 0 &= \int_{\partial K} \sigma \cdot n_K \\ &= \int_K \operatorname{div} \sigma \\ &= \int_K 2\gamma \\ \implies \gamma &= 0. \end{aligned}$$

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  $\alpha = \beta = 0$ .  $\square$

Wir definieren

$$\begin{aligned} RT_h^{-1} &= \{ \sigma : \Omega \longrightarrow \mathbb{R}^2 : \sigma|_K \in RT(K) \forall K \in \mathcal{T}_h \} \\ RT_h &= RT_h^{-1} \cap H(\operatorname{div}, \Omega). \end{aligned}$$

Wie im Beweis von Satz I.1.7 (S. 10) folgt, dass  $\sigma \in RT_h^{-1}$  genau dann in  $RT_h$  liegt, wenn  $\sigma \cdot n_K$  stetig ist über alle Dreiecksanten, die in  $\Omega$  liegen. Wegen Satz IV.2.3 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  $\dim RT_h = \#\mathcal{E}_h$ . Wir setzen nun

$$M_h = RT_h,$$

$$\begin{aligned} Q_h &= S_h^{0,-1}, \\ X_h &= M_h \times Q_h \end{aligned}$$

und approximieren Problem (IV.2.4) durch

Finde  $[\sigma_h, u_h] \in X_h$ , so dass

$$(IV.2.5) \quad a([\sigma_h, u_h], [\tau_h, v_h]) = l([\tau_h, v_h]) \quad \forall [\tau_h, v_h] \in X_h$$

oder in anderer Schreibweise

Finde  $[\sigma_h, u_h] \in X_h$ , so dass

$$(IV.2.6) \quad \begin{aligned} \int_{\Omega} \sigma_h \cdot \tau_h + \int_{\Omega} u_h \operatorname{div} \tau_h &= 0 \quad \forall \tau_h \in M_h \\ - \int_{\Omega} v_h \operatorname{div} \sigma_h &= \int_{\Omega} f v_h \quad \forall v_h \in Q_h. \end{aligned}$$

SATZ IV.2.4. (1) Es ist  $\operatorname{div} M_h = Q_h$ . Zu jedem  $u_h \in Q_h$  gibt es ein  $\tau_{h,u_h} \in M_h$  mit

$$\begin{aligned} \operatorname{div} \tau_{h,u_h} &= u_h, \\ \|\tau_{h,u_h}\|_0 &\leq c_1 \|u_h\|_0. \end{aligned}$$

Die Konstante  $c_1$  hängt nur vom Durchmesser von  $\Omega$  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 (IV.2.5) bzw. (IV.2.6) 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 IV.2.1. Wir definieren einen Operator

$$I_h : H(\operatorname{div}, \Omega) \rightarrow M_h$$

durch

$$(IV.2.7) \quad (I_h \tau) \cdot n_K = h_E^{-1} \int_E \tau \cdot n_K \quad \forall K \in \mathcal{T}_h, E \in \mathcal{E}_h, E \subset \partial K.$$

Dabei ist  $h_E$  die Länge von  $E$ . Wegen Satz IV.2.3 ist diese Definition sinnvoll. Für jedes  $K \in \mathcal{T}_h$  folgt mit dem Gaußschen Integralsatz

$$\int_K \operatorname{div}(I_h \tau_{u_h}) = \sum_{\substack{E \in \mathcal{E}_h \\ E \subset \partial K}} \int_E (I_h \tau_{u_h}) \cdot n_K$$

$$\begin{aligned}
&= \sum_{\substack{E \in \mathcal{E}_h \\ E \subset \partial K}} \int_E \tau_{u_h} \cdot n_K \\
&= \int_K \operatorname{div} \tau_{u_h} \\
&= \int_K u_h.
\end{aligned}$$

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 IV.2.1.

AD (2): Wegen Teil (1) können wir den Beweis von Satz IV.2.1 kopieren. Dabei übernimmt  $\tau_{h,u_h}$  aus Teil (1) die Rolle von  $\tau_u$  aus dem Beweis von Satz IV.2.1.

AD (3): Wie im Beweis von Satz IV.2.2 folgt aus Teil (2), dass das homogene Problem (IV.2.5), d.h. (IV.2.5) mit  $l = 0$ , nur die triviale Lösung besitzt. Da (IV.2.5) ein lineares Gleichungssystem mit der gleichen Anzahl von Gleichungen und Unbekannten ist, folgt hieraus die Behauptung.  $\square$

**SATZ IV.2.5.** *Seien  $[\sigma, u] \in X$  und  $[\sigma_h, u_h] \in X_h$  die eindeutigen Lösungen der Probleme (IV.2.4) und (IV.2.5). 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*

$$\begin{aligned}
&\|\sigma - \sigma_h\|_0 + \|\operatorname{div}(\sigma - \sigma_h)\|_0 \\
&\quad + \|u - u_h\|_0 \leq ch \{|\sigma|_1 + |\operatorname{div} \sigma|_1 + |u|_1\}.
\end{aligned}$$

**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 (IV.2.4) und (IV.2.5) 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 IV.2.4 (3) folgt

$$\beta \|[\tau_h - \sigma_h, v_h - u_h]\|_X \leq \sup_{[\rho_h, w_h] \in X_h \setminus \{0\}} \frac{a([\tau_h - \sigma_h, v_h - u_h], [\rho_h, w_h])}{\|[\rho_h, w_h]\|_X}$$

$$\begin{aligned}
&= \sup_{[\rho_h, w_h] \in X_h \setminus \{0\}} \frac{a([\tau_h - \sigma, v_h - u], [\rho_h, w_h])}{\|[\rho_h, w_h]\|_X} \\
&\leq \|[\tau_h - \sigma, v_h - u]\|_X.
\end{aligned}$$

Hierbei haben wir im letzten Schritt die Cauchy-Schwarzsche Ungleichung für Integrale und Summen und die Definition von  $\|\cdot\|_X$  ausgenutzt. Da  $[\tau_h, v_h] \in X_h$  beliebig war, beweisen diese Abschätzungen das folgende Analogon zum C ea-Lemma, Satz I.2.2 (S. 17),

$$\|[\sigma - \sigma_h, u - u_h]\|_X \leq \left(1 + \frac{1}{\beta}\right) \inf_{[\tau_h, v_h] \in X_h} \|[\sigma - \tau_h, u - v_h]\|_X.$$

Bezeichne mit  $\pi_h : L^2(\Omega) \longrightarrow Q_h$  die  $L^2$ -Projektion. Dann folgt aus obiger Abschätzung

$$\begin{aligned}
&\{\|\sigma - \sigma_h\|_0^2 + \|\operatorname{div}(\sigma - \sigma_h)\|_0^2 + \|u - u_h\|_0^2\}^{1/2} \\
&\leq \left(1 + \frac{1}{\beta}\right) \|[\sigma - I_h\sigma, u - \pi_h u]\|_X \\
&= \left(1 + \frac{1}{\beta}\right) \{\|\sigma - I_h\sigma\|_0^2 + \|\operatorname{div}(\sigma - I_h\sigma)\|_0^2 + \|u - \pi_h u\|_0^2\}^{1/2}.
\end{aligned}$$

Dabei ist  $I_h$  der Interpolationsoperator aus (IV.2.7). Aus der Poincar eschen Ungleichung, Satz I.1.21 (S. 15), folgt

$$\begin{aligned}
\|u - \pi_h u\|_0 &= \left\{ \sum_{K \in \mathcal{T}_h} \|u - \pi_h u\|_{L^2(K)}^2 \right\}^{1/2} \\
&= \left\{ \sum_{K \in \mathcal{T}_h} \left\| u - \frac{1}{|K|} \int_K u \right\|_{L^2(K)}^2 \right\}^{1/2} \\
&\leq \left\{ \sum_{K \in \mathcal{T}_h} c^2 h_K^2 |u|_{H^1(K)}^2 \right\}^{1/2} \\
&\leq ch|u|_1.
\end{aligned}$$

Durch Transformation auf das Referenzdreieck folgt wie im Beweis von Satz II.3.3 (S. 43) f ur 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 Poincar esche Ungleichung, Satz I.1.21 (S. 15), ausgenutzt. Quadrieren dieser Abschätzung und Summieren  uber alle Dreiecke liefert

$$\|\sigma - I_h\sigma\|_0 \leq ch|\sigma|_1.$$

Sei nun  $v \in L^2(\Omega)$  beliebig. Dann ist

$$\begin{aligned} \int_{\Omega} \operatorname{div}(\sigma - I_h\sigma)v &= \int_{\Omega} \operatorname{div}(\sigma - I_h\sigma)(v - \pi_h v) \\ &\quad + \int_{\Omega} \operatorname{div}(\sigma - I_h\sigma)\pi_h v. \end{aligned}$$

Wegen (IV.2.7) 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-Schwarzschen und der Poincaréschen Ungleichung

$$\begin{aligned} &\int_{\Omega} \operatorname{div}(\sigma - I_h\sigma)(v - \pi_h v) \\ &\leq \|\operatorname{div}(\sigma - \sigma_h)\|_0 \left\{ \sum_{K \in \mathcal{T}_h} \|v - \pi_h v\|_{L^2(K)}^2 \right\}^{1/2} \\ &\leq c \|\operatorname{div}(\sigma - \sigma_h)\|_0 \left\{ \sum_{K \in \mathcal{T}_h} h_K^2 |v|_{H^1(K)}^2 \right\}^{1/2}. \end{aligned}$$

Wählen wir speziell  $v = \operatorname{div}(\sigma - I_h\sigma)$  und beachten, dass  $\operatorname{div} I_h\sigma$  elementweise konstant ist, erhalten wir aus obiger Abschätzung

$$\begin{aligned} &\|\operatorname{div}(\sigma - I_h\sigma)\|_0^2 \\ &\leq c \|\operatorname{div}(\sigma - I_h\sigma)\|_0 \left\{ \sum_{K \in \mathcal{T}_h} h_K^2 |\operatorname{div} \sigma|_{H^1(K)}^2 \right\}^{1/2} \\ &\leq ch \|\operatorname{div}(\sigma - I_h\sigma)\|_0 |\operatorname{div} \sigma|_1 \end{aligned}$$

und damit

$$\|\operatorname{div}(\sigma - I_h\sigma)\|_0 \leq ch |\operatorname{div} \sigma|_1.$$

Hieraus folgt die Fehlerabschätzung des Satzes.  $\square$

BEMERKUNG IV.2.6. Da  $\Omega$  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 IV.2.5 erfüllt, wenn  $f \in H^1(\Omega)$  ist.

Das lineare Gleichungssystem (IV.2.6) 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 (IV.2.3) wider. Die Matrix  $A$  ist symmetrisch, positiv definit, und die Matrix  $B$  hat wegen Satz IV.2.4 (1) maximalen Rang. Die Größe des LGS (IV.2.6) ist  $\#\mathcal{E}_h + \#\mathcal{T}_h$ . 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 §IV.1 mit  $\mathcal{E}_{h,\Omega}$  die Menge aller Kanten im Innern von  $\Omega$ . Jedem  $E \in \mathcal{E}_{h,\Omega}$  ordnen wir wieder einen dazu orthogonalen Einheitsvektor  $n_E$  zu und bezeichnen mit  $[\varphi]_E$  den Sprung von  $\varphi$  über  $E$  in Richtung  $n_E$ . Setze

$$\Sigma = \bigcup_{E \in \mathcal{E}_{h,\Omega}} E$$

und bezeichne mit

$$S_{h,\Sigma}^{0,-1} = \{\lambda : \Sigma \longrightarrow \mathbb{R} : \lambda|_E \in \mathbb{R} \ \forall E \in \mathcal{E}_{h,\Omega}\}$$

die stückweise konstanten Funktionen auf  $\Sigma$ .

SATZ IV.2.7. (1) Es ist

$$RT_h = \left\{ \sigma \in RT_h^{-1} : \sum_{E \in \mathcal{E}_{h,\Omega}} \int_E [\sigma \cdot n_E]_E \lambda = 0 \ \forall \lambda \in S_{h,\Sigma}^{0,-1} \right\}.$$

(2) Sei  $\varphi \in \mathcal{L}(RT_h^{-1}, \mathbb{R})$  mit  $\varphi(\sigma) = 0$  für alle  $\sigma \in RT_h$ . 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 I.1.7 (S. 10) 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 IV.2.3 (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  $\lambda_\varphi$  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 IV.2.3 gibt es ein  $\sigma^* \in RT_h^{-1}$  mit

$$\begin{aligned} \sigma^*|_K &= 0 && \text{für alle } K \in \mathcal{T}_h \setminus \{K^*\} \\ \sigma^* \cdot n_E &= 0 && \text{für alle Kanten } E \text{ von } K^* \text{ mit } E \neq E^* \\ \sigma^* \cdot n_{E^*} &= 1. \end{aligned}$$

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:

Finde  $\tilde{\sigma}_h \in RT_h^{-1}$ ,  $\tilde{u}_h \in S_h^{0,-1}$ ,  $\tilde{\mu}_h \in S_{h,\Sigma}^{0,-1}$ , so dass

$$\begin{aligned} &\int_\Omega \tilde{\sigma}_h \cdot \tau_h + \int_\Omega \tilde{u}_h \operatorname{div} \tau_h \\ &+ \sum_{E \in \mathcal{E}_{h,\Omega}} \int_E [\tau_h \cdot n_E]_E \tilde{\mu}_h = 0 \quad \forall \tau_h \in RT_h^{-1} \\ \text{(IV.2.8)} \quad & - \int_\Omega v_h \operatorname{div} \tilde{\sigma}_h = \int_\Omega f v_h \quad \forall v_h \in S_h^{0,-1} \\ & \sum_{E \in \mathcal{E}_{h,\Omega}} \int_E [\tilde{\sigma}_h \cdot n_E]_E \lambda_h = 0 \quad \forall \lambda_h \in S_{h,\Sigma}^{0,-1}. \end{aligned}$$

SATZ IV.2.8. Die Probleme (IV.2.6) und (IV.2.8) sind äquivalent.

BEWEIS. Sei  $\tilde{\sigma}_h$ ,  $\tilde{u}_h$ ,  $\tilde{\mu}_h$  eine beliebige Lösung von (IV.2.8). Aus der dritten Gleichung von (IV.2.8) und Satz IV.2.7 (1) folgt  $\tilde{\sigma}_h \in RT_h = M_h$ . Indem wir in der ersten Gleichung von (IV.2.8) nur Vektorfelder  $\tau_h \in RT_h$  als Testfunktionen betrachten, sehen wir, dass  $\tilde{\sigma}_h$ ,  $\tilde{u}_h$  eine Lösung von (IV.2.6) ist.

Betrachte nun umgekehrt die Lösung  $\sigma_h$ ,  $u_h$  von (IV.2.6). Diese erfüllt wegen (IV.2.6) und Satz IV.2.7 (1) die zweite und dritte Gleichung von

(IV.2.8). Wegen (IV.2.6) verschwindet zudem die lineare Abbildung

$$\tau \mapsto \int_{\Omega} \sigma_h \cdot \tau + \int_{\Omega} u_h \operatorname{div} \tau$$

auf  $RT_h$ . Wegen Satz IV.2.7 (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  $\sigma_h, u_h, \mu_h$  eine Lösung von (IV.2.8).  $\square$

Problem (IV.2.8) 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  $\#\mathcal{E}_h + \#\mathcal{T}_h + \#\mathcal{E}_{h,\Omega}$  hat, haben wir auf den ersten Blick im Vergleich zu (IV.2.6) nichts gewonnen. Dieser Eindruck täuscht aber. 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  $\tilde{A}$  eine blockdiagonale Matrix; die Zahl der Blöcke ist  $\#\mathcal{T}_h$ . Wegen Satz IV.2.3 (2) ist jeder Block eine  $3 \times 3$  Matrix. Wir können daher die Unbekannte  $\sigma_h$  elementweise durch  $u_h$  und  $\mu_h$  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  $\mu_h$  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  $\mu_h$  ein, erhalten wir

$$\begin{aligned} 0 &= C\sigma_h \\ &= -C\tilde{A}^{-1}B^T u_h - C\tilde{A}^{-1}C^T \mu_h \\ &= -C\tilde{A}^{-1}B^T \{B\tilde{A}^{-1}B^T\}^{-1} \{b_h - B\tilde{A}^{-1}C^T \mu_h\} \\ &\quad - C\tilde{A}^{-1}C^T \mu_h \\ &= \{C\tilde{A}^{-1}B^T \{B\tilde{A}^{-1}B^T\}^{-1} B\tilde{A}^{-1}C^T - C\tilde{A}^{-1}C^T\} \mu_h \\ &\quad - C\tilde{A}^{-1}B^T \{B\tilde{A}^{-1}B^T\}^{-1} b_h. \end{aligned}$$

Wir können also die Unbekannten  $\sigma_h$  und  $u_h$  elementweise eliminieren und erhalten ein LGS der Form

$$(IV.2.9) \quad H\mu_h = g_h$$

für  $\mu_h$ . Es hat nur noch die Größe  $\#\mathcal{E}_{h,\Omega}$ . 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 (IV.2.8) wegen Satz IV.2.4 (3) und Satz IV.2.8 regulär ist, ist  $H$  auch positiv definit. Daher kann Problem (IV.2.9) z.B. mit einem CG- oder PCG-Verfahren gelöst werden.

Die Freiheitsgrade von  $\mu_h$  sind die Werte in den Mittelpunkten der inneren Kanten. Gleiches gilt für die nicht-konforme Crouzeix-Raviart Diskretisierung aus §IV.1. In der Tat sind das Problem (IV.2.9) und die Crouzeix-Raviart Diskretisierung eng verwandt. Aufgrund dieser Verwandtschaft kann man zudem die Unbekannte  $\mu_h$  benutzen, um eine Approximation der Ordnung  $O(h^2)$  für die Verschiebung zu erhalten.

**SATZ IV.2.9.** *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 (IV.2.3) und (IV.2.8). Bezeichne mit  $CR_h$  den Raum der Crouzeix-Raviart Elemente aus § IV.1 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  $\bar{u}_h = \pi_h u$  die  $L^2$ -Projektion von  $u$  auf  $S_h^{0,-1}$ . Dann folgt mit elementweiser partieller Integration für alle  $\tau_h \in RT_h^{-1}$

$$\begin{aligned} \int_{\Omega} \sigma \cdot \tau_h + \int_{\Omega} \bar{u}_h \operatorname{div} \tau_h &= \int_{\Omega} \sigma \cdot \tau_h + \int_{\Omega} u \operatorname{div} \tau_h \\ &= \int_{\Omega} \sigma \cdot \tau_h + \sum_{K \in \mathcal{T}_h} \int_K u \operatorname{div} \tau_h \\ &= \int_{\Omega} \sigma \cdot \tau_h \\ &\quad + \sum_{K \in \mathcal{T}_h} \left\{ - \int_K \underbrace{\nabla u}_{=\sigma} \cdot \tau_h + \int_{\partial K} u n_K \cdot \tau_h \right\} \\ &= \sum_{K \in \mathcal{T}_h} \int_{\partial K} u n_K \cdot \tau_h \end{aligned}$$

$$\begin{aligned}
&= \sum_{E \in \mathcal{E}_{h,\Omega}} \int_{\partial E} u [n_E \cdot \tau_h]_E \\
&= \sum_{E \in \mathcal{E}_{h,\Omega}} \int_{\partial E} u_h^* [n_E \cdot \tau_h]_E.
\end{aligned}$$

Hieraus und aus der ersten Gleichung von (IV.2.8) erhalten wir

$$\begin{aligned}
&\sum_{E \in \mathcal{E}_{h,\Omega}} \int_{\partial E} (\hat{u}_h - u_h^*) [n_E \cdot \tau_h]_E \\
&= \sum_{E \in \mathcal{E}_{h,\Omega}} \left\{ - \int_{\partial E} \mu_h [n_E \cdot \tau_h]_E - \int_{\partial E} u_h^* [n_E \cdot \tau_h]_E \right\} \\
&= \int_{\Omega} (\sigma_h - \sigma) \cdot \tau_h + \int_{\Omega} (u_h - \bar{u}_h) \operatorname{div} \tau_h.
\end{aligned}$$

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 Regularitätsbedingung an  $\mathcal{T}_h$  abhängt, mit

$$\begin{aligned}
&\|\hat{u}_h - u_h^*\|_0 \\
&\leq \sup_{\tau_h \in RT_h^{-1}} \frac{\left\{ \sum_{E \in \mathcal{E}_{h,\Omega}} \int_{\partial E} (\hat{u}_h - u_h^*) [n_E \cdot \tau_h]_E \right\}}{\left\{ \sum_{K \in \mathcal{T}_h} \left[ h_K^{-2} \|\tau_h\|_{L^2(K)}^2 + \|\operatorname{div} \tau_h\|_{L^2(K)}^2 \right] \right\}^{1/2}}.
\end{aligned}$$

Damit folgt aus obiger Identität

$$\begin{aligned}
\|\hat{u}_h - u_h^*\|_0 &\leq c \left\{ \sum_{K \in \mathcal{T}_h} \left[ h_K^2 \|\sigma - \sigma_h\|_{L^2(K)}^2 + \|u_h - \bar{u}_h\|_{L^2(K)}^2 \right] \right\}^{1/2} \\
&\leq c' \left[ h \|\sigma - \sigma_h\|_0 + \|u_h - \bar{u}_h\|_0 \right].
\end{aligned}$$

Gemäß Satz IV.2.5 und Bemerkung IV.2.6 ist

$$h \|\sigma - \sigma_h\|_0 \leq ch^2 \left[ |f|_1 + \|u\|_2 \right].$$

Wir wollen zeigen, dass gleiches für  $\|u_h - \bar{u}_h\|_0$  gilt. Dazu benutzen wir ein Dualitätsargument und bezeichnen mit  $z$  die schwache Lösung von

$$\begin{aligned}
\Delta z &= \bar{u}_h - u_h && \text{in } \Omega \\
z &= 0 && \text{auf } \Gamma
\end{aligned}$$

und setzen  $\varphi = \nabla z$ . Mit dem Interpolationsoperator  $I_h$  aus (IV.2.7) folgt dann

$$\begin{aligned}
\|\bar{u}_h - u_h\|_0^2 &= \int_{\Omega} (\bar{u}_h - u_h) \operatorname{div} \varphi \\
&= \int_{\Omega} (\bar{u}_h - u_h) \operatorname{div}(I_h \varphi) \\
&= \int_{\Omega} (u - u_h) \operatorname{div}(I_h \varphi) \\
&= \int_{\Omega} (\sigma_h - \sigma) I_h \varphi \\
&= \int_{\Omega} (\sigma_h - \sigma)(I_h \varphi - \varphi) + \int_{\Omega} (\sigma_h - \sigma) \varphi \\
&= \int_{\Omega} (\sigma_h - \sigma)(I_h \varphi - \varphi) + \int_{\Omega} (\sigma_h - \sigma) \cdot \nabla z \\
&= \int_{\Omega} (\sigma_h - \sigma)(I_h \varphi - \varphi) - \int_{\Omega} \operatorname{div}(\sigma_h - \sigma) z \\
&= \int_{\Omega} (\sigma_h - \sigma)(I_h \varphi - \varphi) - \int_{\Omega} \operatorname{div}(\sigma_h - \sigma)(z - \pi_h z) \\
&\leq \|\sigma - \sigma_h\|_0 \|\varphi - I_h \varphi\|_0 + \|\operatorname{div}(\sigma - \sigma_h)\|_0 \|z - \pi_h z\|_0 \\
&\leq ch \left\{ \|\sigma - \sigma_h\|_0 |\varphi|_1 + \|\operatorname{div}(\sigma - \sigma_h)\|_0 |z|_1 \right\}.
\end{aligned}$$

Wegen

$$|z|_1 + |\varphi|_1 \leq c \|\bar{u}_h - u_h\|_0$$

und Satz IV.2.5 liefert dies

$$\begin{aligned}
\|\bar{u}_h - u_h\|_0 &\leq ch \left\{ \|\sigma - \sigma_h\|_0 + \|\operatorname{div}(\sigma - \sigma_h)\|_0 \right\} \\
&\leq c'h^2 \left\{ |f|_1 + \|u\|_2 \right\}.
\end{aligned}$$

Dies beweist die Fehlerabschätzung des Satzes. □

# Zusammenfassung

## I Analytische Grundlagen

### 1. Sobolevräume

Schwache Ableitungen; Zusammenhang mit klassischen Ableitungen; Sobolevräume  $W^{k,p}(\Omega)$ ,  $H^k(\Omega)$ ,  $W_0^{k,p}(\Omega)$  und  $H_0^k(\Omega)$  und ihre Normen; Charakterisierung stückweiser glatter  $W^{k,p}$ -Funktionen; Lipschitz-Gebiete und Kegelbedingung; Spursatz; Friedrichsche Ungleichung; stetige und kompakte Einbettung; Sobolevscher Einbettungssatz; Poincarésche Ungleichung

### 2. Abstrakte Variationsprobleme

Satz von Lax-Milgram; Céa-Lemma; Satz von Aubin-Nitsche; Erweiterung auf nicht symmetrische und nicht koerzive Bilinearformen

### 3. Schwache Lösungen

Schwache Lösungen von Konvektions-Diffusionsgleichungen mit Dirichlet-, gemischten und Neumann-Randbedingungen; Existenz- und Eindeutigkeitssatz für schwache Lösungen; Regularitätssatz

## II Finite Element Räume

### 1. Eindimensionale lineare Elemente

Variationsproblem; Finite Element Raum; Basis; Fehlerabschätzung

### 2. Bilineare Rechteckselemente

Finite Element Raum; Basis; Interpretation als Differenzenverfahren; Interpolationsoperator; Abschätzung des Interpolationsfehlers; Fehlerabschätzung; Schwierigkeiten bei Konvektions-dominanten Problemen

### 3. Lineare Dreieckselemente

Finite Element Raum; Basis; Interpretation als Differenzenverfahren; Interpolationsoperator; Abschätzung des Interpolationsfehlers in der  $H^1$ -Norm auf dem Referenzelement; Abschätzung der affinen Transformation auf das Referenzelement; Fehlerabschätzung; Schwierigkeiten bei Konvektions-dominanten Problemen; SDFEM Diskretisierung; Koerzivität der Bilinearform; Fehlerabschätzung des Interpolationsfehlers in der  $L^2$ -Norm; Fehlerabschätzung für die SDFEM Diskretisierung; optimale Wahl des Stabilisierungsparameters

### 4. Finite Elemente höherer Ordnung

Grundidee; Referenz-Simplex und Referenz-Würfel; affine Äquivalenz; Zulässigkeit; Regularität; Mengen  $\hat{\Sigma}_k$  und  $\Sigma_k$ ; Polynomräume  $R_k$ ; Funktionen in  $R_k$  sind eindeutig bestimmt durch ihre Werte in  $\Sigma_k$ ; Finite Element Räume  $S_h^{k,-1}$ ,  $S_h^{k,0}$  und  $S_{h,0}^{k,0}$ ; nodale Basis; Interpolationsoperator  $I_h$ ; Abschätzung des Interpolationsfehlers

### III Praktische Aspekte

1. Randapproximation und numerische Integration  
 Polygonale Randapproximation; Fortsetzung der Daten bei nicht konvexem Gebiet; Randapproximation durch krummlinige und isoparametrische Elemente; Quadraturformeln; erforderliche Genauigkeit; Beispiele
2. Numerische Lösung der diskreten Probleme  
 Inverse Abschätzung; Kondition der diskreten Probleme; Mehrgitteralgorithmus mit Jacobi-Glättung; Ritz-Projektion; Abschätzung der Ritz-Projektion; Abschätzung der Glättung; Konvergenzbeweis für den Mehrgitteralgorithmus; Konvergenzrate  $\frac{c}{c+2\nu}$ ; Abschätzung für den Mehrgitteralgorithmus mit reiner Vorglättung; Teilraum-Korrektur-Methoden; Gauß-Seidel- und Mehrgitterverfahren sind Teilraum-Korrektur-Methoden; Konvergenzbeweis für die Teilraum-Korrektur-Methoden; Gültigkeit der Annahmen des Konvergenzbeweises; verschärfte Cauchy-Schwarzsche Ungleichung
3. Fehlerschätzer und adaptive Gitterverfeinerung  
 Nachteile asymptotischer a priori Fehlerabschätzungen; Ziele von a posteriori Fehlerabschätzungen; Stabilität; Galerkin-Orthogonalität; Residuum;  $L^2$ -Darstellung des Residuums; Quasi-Interpolationsoperator; Fehlerabschätzung für den Quasi-Interpolationsoperator; obere Fehlerschranke; Funktionen  $\psi_K, \psi_E$  und ihre Eigenschaften; untere Fehlerschranken; adaptiver Algorithmus; Verfeinerung von Dreiecken und Vierecken
4. Implementierung  
 Datenstrukturen; Berechnung der rechten Seite und der Steifigkeitsmatrix; Speichertechnik für dünn besetzte Matrizen

### IV Ergänzungen

1. Nichtkonforme Methoden  
 Zweites Strang-Lemma; Übertragung des Aubin-Nitsche Argumentes; Crouzeix-Raviart Diskretisierung; Abschätzung des Konsistenzfehlers; Fehlerabschätzung; Vor- und Nachteile gegenüber konformen Methoden
2. Gemischte Methoden  
 Gemischte Formulierung der Poissongleichung; inf-sup Bedingung für das Sattelpunktsproblem; Lösbarkeit des Sattelpunktsproblems; Raviart-Thomas Element; Freiheitsgrade des Raviart-Thomas Elementes; diskretes Problem; inf-sup Bedingung für das diskrete Problem; Lösbarkeit des diskreten Problems; Fehlerabschätzung; Abschwächung der Stetigkeitsbedingung für  $\sigma \cdot n$ ; Äquivalenz zur Raviart-Thomas Diskretisierung; elementweise Elimination von Freiheitsgraden; Zusammenhang mit der Crouzeix-Raviart Diskretisierung; Verbesserte Fehlerabschätzung für die Verschiebungen

## Index

- $\hookrightarrow^c$ , 14
- $\hookrightarrow$ , 14
- $|\cdot|_{k,p}$ , 8
- $\|\cdot\|_h$ , 65
- $\|\cdot\|_{H(\operatorname{div})}$ , 113
- $\|\cdot\|_{k,p}$ , 8
- $(\cdot, \cdot)_h$ , 65
- $(\cdot, \cdot)_k$ , 9
- $[\cdot]_E$ , 81, 121
- $|\alpha|$ , 7
- $B_K, b_K$ , 40
- $c_{\mathcal{T}}$ , 37
- $D^\alpha$ , 7
- $\mathcal{E}_h$ , 80, 107, 116
- $\mathcal{E}_{h,\Omega}$ , 107, 121
- $F_K$ , 40
- $\mathcal{G}_h$ , 52
- $\mathcal{G}_{h,\Omega}$ , 52
- $H(\operatorname{div}, \Omega)$ , 113
- $h_E$ , 81
- $H_D^1(\Omega)$ , 22
- $H^k(\Omega)$ , 8
- $h_K$ , 29, 37, 52
- $H_0^k(\Omega)$ , 11
- $I_h$ , 32, 39, 57
- $\hat{K}$ , 40
- $K_{E1}, K_{E2}$ , 80
- $\lambda_{K,i}$ , 38
- $m_E$ , 107
- $n_E$ , 81, 121
- $\mathcal{N}_h$ , 29, 37, 52
- $\mathcal{N}_{h,\Omega}$ , 29, 37, 52
- $\omega_z$ , 38
- $\hat{\pi}$ , 40
- $\hat{\Pi}_k$ , 57
- $Q_1$ , 29
- $Q_k$ , 52
- $Q_m$ , 71
- $R_h$ , 83
- $R_k$ , 53
- $\rho_K$ , 37, 52
- $RT_h$ , 116
- $RT_h^{-1}$ , 116
- $RT(K)$ , 116
- $S_{h,0}^{1,0}$ , 27
- $S_{h,\Sigma}^{0,-1}$ , 121
- $\Sigma_k$ , 52
- $\hat{\Sigma}_k$ , 52
- $S_h^{k,-1}$ , 56
- $S_h^{k,0}$ , 56
- $S_{h,0}^{k,0}$ , 56
- $\Sigma$ , 121
- $\mathcal{T}_h$ , 27, 29, 37, 52, 116
- $v_z$ , 38, 56
- $W^{k,p}(\Omega)$ , 8
- $W_0^{k,p}(\Omega)$ , 11
- a posteriori Fehlerabschätzung, 80, 87
- a posteriori Fehlerschätzer, 87
- a priori Fehlerabschätzung, 80
- adaptive Gitterverfeinerung, 80, 92
- affin äquivalent, 40, 51
- affine Äquivalenz, 52
- asymptotisch, 80
- blaue Unterteilung, 93
- Céa-Lemma, 17
- Courant-Triangulierung, 39
- Crouzeix-Raviart Diskretisierung, 107
- Crouzeix-Raviart Raum, 107
- Dirichlet-Randbedingungen, 21
- Effizienz, 87
- Existenz- und Eindeigkeitssatz für schwache Lösungen, 22
- Friedrichsche Ungleichung, 12

- Galerkin-Orthogonalität, 81
- gemischte Finite Element Methode, 113
- gemischte Randbedingungen, 21
- gemischtes Problem, 114
- Gitter, 53
- globale untere Fehlerschranke, 91
- grüne Unterteilung, 93
- hängender Knoten, 93
- inf-sup Bedingung, 114
- inverse Abschätzung, 65
- isoparametrische Elemente, 52
- Kegelbedingung, 12
- Knoten, 53
- koerziv, 16
- kompakt eingebettet, 14
- kompakte lineare Abbildung, 14
- Konvektions-Diffusionsgleichung, 21
- Lipschitz-Gebiet, 11
- Lipschitz-Rand, 11
- lokale untere Fehlerschranke, 90
- Membrangleichung, 21
- Neumann-Randbedingungen, 21
- nicht-konforme Methode, 105
- nodale Basis, 38, 56
- Péclet-Zahl, 36
- Poincarésche Ungleichung, 15
- Poissongleichung, 21
- Quadraturformel, 63
- Quasi-Interpolationsoperator, 83
- Raviart-Thomas Element, 116
- Reaktions-Diffusionsgleichung, 21
- Referenz Simplex, 51
- Referenz Würfel, 51
- Referenzelement, 40
- Reflexion, 62
- Regularität, 37, 52
- Regularitätssatz, 26
- Residuum, 82
- Ritz-Projektion, 71
- Robin-Randbedingungen, 25
- rote Unterteilung, 92
- Sattelpunktsproblem, 114
- Satz von Aubin-Nitsche, 18
- Satz von Lax-Milgram, 16
- schwache Ableitung, 7
- schwache Lösung, 22
- Schwerpunktskoordinaten, 38, 54
- SDFEM, 45
- Sobolevraum, 8
- Sobolevscher Einbettungssatz, 14
- Spursatz, 12
- Stabilität, 81
- Steifigkeitsmatrix, 30
- stetig eingebettet, 14
- streamline upwind Petrov-Galerkin Verfahren, 45
- streamline-diffusion finite element method, 45
- Stromlinien-Diffusions Methode, 45
- SUPG, 45
- symmetrisch, 16
- Teilraum-Korrektur-Methode, 76
- TRK, 76
- uniforme Triangulierung, 69
- verschärfte Cauchy-Schwarzsche Ungleichung, 80
- Winkelbedingung, 37
- Zulässigkeit, 29, 37, 52
- zuverlässig, 87
- zweites Strang-Lemma, 105