Numerik II
Finite Elemente

Vorlesungsskriptum Sommersemester 2019

R. Verfürth

Fakultät für Mathematik, Ruhr-Universität Bochum

Inhaltsverzeichnis

Motivation

Zur Motivation betrachten wir die Finite Element Diskretisierung des eindimensionalen Sturm-Liouville Problems

$$-(Pu')' + Qu = F \text{ in } (0,1), \quad u(0) = u(1) = 0. \tag{1}$$

Im Vergleich zu der Differenzendiskretisierung aus [16, §II.4] machen wir hier die schwächeren Annahmen

$$\begin{aligned} &F \in L^2([0,1],\mathbb{R}), \\ &Q \in C([0,1],\mathbb{R}), \quad \min_{0\le t\le 1} Q(t) \ge 0, \\ &P \in C([0,1],\mathbb{R}), \quad \min_{0\le t\le 1} P(t) = \underline{p} > 0. \end{aligned}$$

Dennoch erhalten wir allgemeinere und bessere Fehlerabschätzungen.

Die Finite Element Diskretisierung beruht auf einer geeigneten Variationsformulierung von (1). Zu deren Motivation multiplizieren wir (1) mit einer Funktion $v \in C_0^\infty((0,1),\mathbb{R})$, integrieren das Ergebnis von 0 bis 1 und benutzen partielle Integration für die Ableitungsterme. Dies liefert

$$\begin{aligned} \int_0^1 Fv &= -\int_0^1 (Pu')'v + \int_0^1 Quv = -Pu'v\Big|_0^1 + \int_0^1 (Pu')v' + \int_0^1 Quv \\ &= \int_0^1 (Pu')v' + \int_0^1 Quv. \end{aligned}$$

Daher hat eine mögliche Variationsformulierung von (1) die Struktur:

> *Finde eine Funktion u in einem geeigneten Funktionenraum X, so dass für alle Funktionen v in X der erste und der letzte Term in obiger Gleichungskette übereinstimmen.*

Um diesen Ansatz in eine mathematisch fundierte Form zu bringen, müssen wir zuerst den Raum X sauber definieren. Eine Mindestanforderung ist dabei natürlich, dass die entsprechenden Integrale endlich sind. Wegen unserer Annahmen an F, P und Q und der Cauchy-Schwarzschen Ungleichung bedeutet dies, dass die Funktionen in X quadrat-integrierbar mit quadrat-integrierbarer Ableitung sein müssen. Außerdem müssen die Randbedingungen $u(0) = u(1) = 0$ in einem geeigneten Sinn erfüllt sein.

Die folgende Definition präzisiert diese Vorstellungen.

DEFINITION 1 (Absolut stetige Funktion; Sobolev-Raum). (1) Eine Funktion $\varphi : [a,b] \to \mathbb{R}$ heißt *absolut stetig* auf $[a,b]$, wenn es zu jedem $\varepsilon > 0$ ein $\delta > 0$ gibt, so dass für jedes endliche System von Intervallen $[a_i, b_i]$ mit

$$a \le a_1 < b_1 \le a_2 < b_2 \le \ldots \le a_n < b_n \le b \quad \text{und} \quad \sum_{i=1}^{n} (b_i - a_i) < \delta$$

gilt

$$\sum_{i=1}^{n} |\varphi(b_i) - \varphi(a_i)| < \varepsilon.$$

(2) Für $m \in \mathbb{N}^*$ ist der *Sobolev-Raum* $H^m(a,b)$ definiert durch

$$H^m(a,b) = \big\{ \varphi \in C^{m-1}([a,b],\mathbb{R}) : \varphi^{(m-1)} \text{ ist absolut stetig, } \varphi^{(m)} \text{ existiert fast überall und } \varphi^{(m)} \in L^2([a,b],\mathbb{R}) \big\}.$$

Er wird versehen mit der Norm

$$\|\varphi\|_m = \left\{ \sum_{k=0}^{m} |\varphi|_k^2 \right\}^{\frac{1}{2}}$$

mit

$$|\varphi|_0 = \|\varphi\|_0 = \left\{ \int_a^b |\varphi|^2 \right\}^{\frac{1}{2}},$$

$$|\varphi|_k = \left\{ \int_a^b \left|\varphi^{(k)}\right|^2 \right\}^{\frac{1}{2}}, \ k \in \mathbb{N}^*.$$

(3) $H_0^1(a,b) = \{\varphi \in H^1(a,b) : \varphi(a) = \varphi(b) = 0\}$.

BEMERKUNG 2. (1) Jede absolut stetige Funktion ist gleichmäßig stetig. Die Umkehrung gilt i.a. nicht.
(2) $H^m(a,b)$ ist ein Hilbert-Raum mit dem Skalarproduktt

$$(\varphi\,,\,\psi)_m = \sum_{k=0}^{m} \int_a^b \varphi^{(k)} \psi^{(k)}.$$

(3) $H_0^1(a,b)$ ist die Vervollständigung von $C_0^\infty((a,b),\mathbb{R})$ bzgl. $\|\cdot\|_1$.

Für den Nachweis, dass die Variationsformulierung von (1) eine eindeutige Lösung besitzt, und für die Fehlerabschätzungen der Finite Element Diskretisierung benötigen wir das folgende Hilfsresultat.

LEMMA 3 (Friedrichsche Ungleichungen). *(1) Zu $u \in H^1(a,b)$ gebe es ein $t^* \in [a,b]$ mit $u(t^*) = 0$. Dann gilt*

$$\max_{a \le t \le b} |u(t))| \le (b-a)^{\frac{1}{2}} |u|_1, \quad \|u\|_0 \le (b-a) |u|_1.$$

(2) Für alle $u \in H_0^1(a,b)$ gilt

$$\left\{1+(b-a)^2\right\}^{-\frac{1}{2}} \|u\|_1 \leq |u|_1 \leq \|u\|_1 .$$

BEWEIS. *ad (1):* Für $t \in [a,b]$ folgt mit der Cauchy-Schwarzschen Ungleichung

$$\begin{aligned}
|u(t)| &= |u(t)-u(t^*)| \\
&= \left| \int_{t^*}^{t} u'(s) \mathrm{d}s \right| \\
&\leq \int_a^b \chi_{[\min(t,t^*),\max(t,t^*)]}(s)\, |u'(s)|\, \mathrm{d}s \\
&\leq |t-t^*|^{\frac{1}{2}} \left\{ \int_a^b |u'(s)|^2 \, \mathrm{d}s \right\}^{\frac{1}{2}} \\
&= |t-t^*|^{\frac{1}{2}} \, |u|_1 \\
&\leq (b-a)^{\frac{1}{2}} \, |u|_1 .
\end{aligned}$$

Wegen

$$\|u\|_0 \leq (b-a)^{\frac{1}{2}} \max_{a\leq t\leq b} |u(t)|$$

folgt hieraus die Behauptung.
ad (2): Die Ungleichung

$$|u|_1 \leq \|u\|_1$$

ist offensichtlich. Aus Teil (1) folgt

$$\|u\|_1 = \left\{\|u\|_0^2 + |u|_1^2\right\}^{\frac{1}{2}} \leq \left\{(b-a)^2+1\right\}^{\frac{1}{2}} |u|_1 . \qquad \square$$

Nach diesen Vorbereitungen können wir jetzt die gesuchte Variationsformulierung angeben und den Begriff einer *schwachen Lösung* von (1) einführen.

DEFINITION 4 (Schwache Lösung). Eine Funktion $u \in H_0^1(0,1)$ heißt *schwache Lösung* von (1), wenn für alle $v \in H_0^1(0,1)$ gilt

$$\int_0^1 \{Pu'v' + Quv\} = \int_0^1 Fv. \tag{2}$$

Unsere Überlegungen zu Beginn dieses Abschnittes zeigen die folgende fundamentale Beziehung zwischen schwachen und klassischen Lösungen von (1).

SATZ 5 (Schwache und klassische Lösung). *Jede klassische Lösung von (1) ist auch eine schwache Lösung. Umgekehrt ist jede zweimal stetig differenzierbare schwache Lösung von (1) auch eine klassische Lösung.*

BEMERKUNG 6 (Regularität). Satz 5 zeigt, dass in geeignetem Sinne die Probleme (1) und (2) äquivalent sind. Eine Aussage der Form

> „Jede schwache Lösung von (1) ist aus C^2 (und damit klassische Lösung).“

nennt man einen *Regularitätssatz*. Für Sturm-Liouville Probleme kann man derartige Sätze aus dem Regularitätssatz [16, Satz I.1.16] für gewöhnliche Differentialgleichungen ableiten. Für partielle Differentialgleichungen ist der Beweis von Regularitätssätzen wesentlich aufwändiger und erfordert zusätzliche Annahmen an das Gebiet, vgl. [16, Beispiel III.1.12].

Der folgende Satz zeigt, dass (1) eine eindeutige schwache Lösung besitzt.

Satz 7 (Schwache Lösbarkeit des Sturm-Liouville Problems). *Problem (1) besitzt eine eindeutige schwache Lösung. Diese ist das eindeutige Minimum des Funktionals*

$$H_0^1(0,1) \ni u \mapsto \frac{1}{2}\int_0^1 \{Pu'^2 + Qu^2\} - \int_0^1 Fu.$$

Beweis. Wir wenden den Satz von Lax-Milgram, Satz I.1.1 (S. 15), an mit $X = H_0^1(0,1)$, $\|\cdot\|_X = |\cdot|_1$ und

$$B(u,v) = \int_0^1 \{Pu'v' + Quv\}, \quad \ell(v) = \int_0^1 Fv.$$

Die Symmetrie und Bilinearität von B und die Linearität von ℓ sind offensichtlich. Die Stetigkeit von B und ℓ folgt aus unseren Annahmen an P, Q und F, der Cauchy-Schwarzschen Ungleichung und Lemma 3. Die Koerzivität von B schließlich folgt aus

$$B(u,u) \geq \int_0^1 P\,|u'|^2 \geq \underline{p}\,|u|_1^2\,. \qquad \square$$

Für die Diskretisierung von Problem (1) ersetzen wir in (2) den Raum $H_0^1(0,1)$ durch einen endlich dimensionalen Unterraum $X_\mathcal{T}$. Das Céa-Lemma, Satz I.1.2 (S. 17), zeigt dann, dass das diskrete Problem eine eindeutige Lösung $u_\mathcal{T}$ hat und dass der Fehler $|u - u_\mathcal{T}|_1$ durch die Approximationsgüte $\inf_{v_\mathcal{T} \in X_\mathcal{T}} |u - v_\mathcal{T}|_1$ bestimmt wird. Der Satz von Aubin-Nitsche, Satz I.1.5 (S. 18), mit $H = L^2(0,1)$ schließlich liefert für die L^2-Norm des Fehlers eine (hoffentlich) verbesserte Abschätzung.

Die Funktionen in $X_\mathcal{T}$ sollen stückweise Polynome sein. Die Forderung $X_\mathcal{T} \subset H_0^1(0,1)$ und die Definition 1 von $H_0^1(0,1)$ implizieren, dass diese Funktionen stetig sein müssen. Dieser Ansatz führt auf folgende Definition.

Definition 8 (Finite Element Räume). Sei $\mathcal{T} = \{I_j : 0 \leq j \leq n\}$ mit $I_j = [t_j, t_{j+1}]$ und $0 = t_0 < t_1 < \ldots < t_{n+1} = 1$ eine Unterteilung von $[0,1]$ in $n+1$ Teilintervalle. Setze

$$h_j = t_{j+1} - t_j,\ 0 \leq j \leq n, \quad \text{und} \quad h = \max_{0 \leq j \leq n} h_j.$$

Für $k \in \mathbb{N}$ und $m \in \mathbb{N}$ sei

$$\begin{aligned} S^{k,-1}(\mathcal{T}) &= \left\{ \varphi : [0,1] \to \mathbb{R} : \varphi\big|_{I_j} \in \mathbb{P}_k \ \forall 0 \le j \le n \right\}, \\ S^{k,m}(\mathcal{T}) &= S^{k,-1}(\mathcal{T}) \cap C^m([0,1],\mathbb{R}), \\ S_0^{k,0}(\mathcal{T}) &= \left\{ \varphi \in S^{k,0}(\mathcal{T}) : \varphi(0) = \varphi(1) = 0 \right\}. \end{aligned}$$

Dabei bezeichnet $\mathbb{P}_k$ den Raum der Polynome vom Grad $\le k$.

Bemerkung 9. (1) $S^{0,m}(\mathcal{T}) = \mathbb{R}$ für alle $m \in \mathbb{N}$.
(2) $S^{k,m}(\mathcal{T}) \subset H^{m+1}(0,1)$ für alle $k \in \mathbb{N}^*$, $m \in \mathbb{N}$.
(3) $\dim S^{k,0}(\mathcal{T}) = (n+1)(k-1)+n+2$, $\dim S_0^{k,0}(\mathcal{T}) = (n+1)(k-1)+n$.
(4) Die Funktionen in $S^{k,m}(\mathcal{T})$ heißen (eindimensionale) *Finite Elemente*. Ist speziell $m = k-1$, so spricht man auch von *Splines*, vgl. [15, §I.3].

Nach Wahl eines Polynomgrades $k \in \mathbb{N}^*$ lautet die Finite Element Diskretisierung von (1):

Finde $u_{\mathcal{T}} \in S_0^{k,0}(\mathcal{T})$, so dass für alle $v_{\mathcal{T}} \in S_0^{k,0}(\mathcal{T})$ gilt

$$\int_0^1 \{ P u_{\mathcal{T}}' v_{\mathcal{T}}' + Q u_{\mathcal{T}} v_{\mathcal{T}} \} = \int_0^1 F v_{\mathcal{T}}. \tag{3}$$

Problem (3) ist ein lineares Gleichungssystem mit $(n+1)(k-1)+n$ Gleichungen und Unbekannten. Die Matrix dieses LGS, die sog. *Systemsteifigkeitsmatrix* oder kurz *Steifigkeitsmatrix*, ist wegen der Symmetrie und Koerzivität der Bilinearform B symmetrisch positiv definit. Das LGS kann daher mit einer Cholesky Zerlegung [15, Algorithmen IV.2.3, IV.2.4] oder, insbesondere für große n, mit einem CG-Verfahren [15, Algorithmus IV.7.2] gelöst werden.

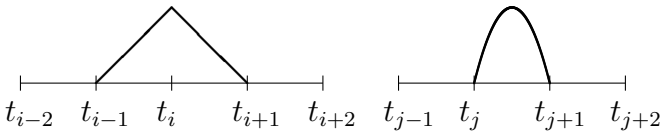

Abbildung 1. Die Basisfunktionen v_i und w_j

Beispiel 10 (Lineare und quadratische Basisfunktionen). (1) Es ist

$$S^{1,0}(\mathcal{T}) = \operatorname{span}\{v_1, \dots, v_n\}$$

mit

$$v_i(t) = \begin{cases} 0 & \text{für } t \le t_{i-1} \text{ oder } t \ge t_{i+1} \\ \frac{t - t_{i-1}}{h_{i-1}} & \text{für } t_{i-1} \le t \le t_i \\ \frac{-t + t_{i+1}}{h_i} & \text{für } t_i \le t \le t_{i+1} \end{cases}$$

(vgl. Abbildung 1). Dann ist (3) äquivalent zu

$$\begin{aligned}\int_{t_{i-1}}^{t_{i+1}} Fv_i = &\int_{t_{i-1}}^{t_{i+1}} Qv_iu \\ &+ \frac{1}{h_{i-1}^2}(u_i - u_{i-1}) \int_{t_{i-1}}^{t_i} P \\ &+ \frac{1}{h_i^2}(u_{i+1} - u_i) \int_{t_i}^{t_{i+1}} P \quad \forall 1 \le i \le n.\end{aligned}$$

Approximiert man die Integrale durch

$$\int_{t_{i-1}}^{t_{i+1}} Fv_i \approx \frac{1}{2}(h_{i-1} + h_i)F_i \quad \text{(Trapezregel)}$$

$$\int_{t_{i-1}}^{t_{i+1}} Qv_iu \approx \frac{1}{2}(h_{i-1} + h_i)Q_iu_i \quad \text{(Trapezregel)}$$

$$\int_{t_{i-\mu}}^{t_{i-\mu+1}} P \approx h_{i-\mu}P_{i-\mu+\frac{1}{2}} \quad (\mu = 0, 1) \quad \text{(Mittelpunktsregel)}$$

und wählt man $h_i = h = \frac{1}{n+1}$ für alle $1 \le i \le n$, so erhält man bis auf Skalierung mit dem Faktor h die Differenzendiskretisierung aus [16, §II.4].

(2) Es ist

$$S^{2,0}(\mathcal{T}) = \operatorname{span}\{v_1, \dots, v_n, w_0, \dots, w_n\}$$

mit $v_1, \dots, v_n$ wie in Teil (1) und

$$w_i(x) = \begin{cases} 0 & \text{für } t \le t_i \text{ oder } t \ge t_{i+1} \\ 4\frac{(t-t_i)(t_{i+1}-t)}{h_i^2} & \text{für } t_i \le t \le t_{i+1} \end{cases}$$

(vgl. Abbildung 1). Die Matrix des LGS (3) hat nun die Form

$$A_Q = \begin{pmatrix} A_L & A_{LQ} \\ A_{LQ}^T & A_{QQ} \end{pmatrix},$$

wobei A_L die Matrix von (3) zu $S^{1,0}(\mathcal{T})$ ist und A_{QQ} diagonal ist. Spaltet man den Lösungsvektor und die rechte Seite entsprechend auf, hat (3) die Form

$$\begin{pmatrix} A_L & A_{LQ} \\ A_{LQ}^T & A_{QQ} \end{pmatrix} \begin{pmatrix} u_L \\ u_Q \end{pmatrix} = \begin{pmatrix} f_L \\ f_Q \end{pmatrix}$$

und ist somit äquivalent zu

$$\begin{aligned}[A_L - A_{LQ}A_{QQ}^{-1}A_{LQ}^T]u_L &= f_L - A_{LQ}A_{QQ}^{-1}F_Q \\ u_Q &= A_{QQ}^{-1}[f_Q - A_{LQ}^T u_L].\end{aligned}$$

Man muss also wie in Teil (1) effektiv nur ein LGS der Größe $n \times n$ lösen, wobei die Matrix wieder symmetrisch, positiv definit und tridiagonal ist.

Als nächstes schätzen wir die Approximationsgüte von $S_0^{k,0}(\mathcal{T})$ in $H_0^1(0,1)$ ab.

SATZ 11 (Approximationsfehler). *Sei $u \in H^{k+1}(0,1) \cap H_0^1(0,1)$ mit $k \in \mathbb{N}^*$. Dann gilt*

$$\inf_{v_\mathcal{T} \in S_0^{k,0}(\mathcal{T})} |u - v_\mathcal{T}|_1 \leq h^k \, |u|_{k+1} .$$

BEWEIS. Für $0 \leq i \leq n$ sei $L_{k,i}$ das Lagrangesche Interpolationspolynom von u zu den Knoten $t_i + \frac{j}{k} h_i$, $0 \leq j \leq k$. Setze

$$v_\mathcal{T}^*(t) = L_{k,i}(t) \quad \forall t \in I_i, 0 \leq i \leq n.$$

Dann ist $v_\mathcal{T}^* \in S_0^{k,0}(\mathcal{T})$. Aus dem Satz von Rolle folgt für $0 \leq i \leq n$ und $0 \leq \mu \leq k$, dass $(u - v_\mathcal{T}^*)^{(\mu)}$ in I_i mindestens $k + 1 - \mu$ Nullstellen hat. Wegen Lemma 3 ist daher

$$|u - v_\mathcal{T}^*|_{\mu,I_i} \leq h_i \, |u - v_\mathcal{T}^*|_{\mu+1,I_i} \quad \forall 0 \leq i \leq n, 0 \leq \mu \leq k.$$

Da $(v_\mathcal{T}^*)^{(k+1)}$ auf jedem I_i verschwindet, folgt hieraus durch Induktion

$$|u - v_\mathcal{T}^*|_{1,I_i} \leq h_i^k \, |u|_{k+1,I_i} \quad \forall 0 \leq i \leq n$$

und damit die Behauptung. □

Satz 11 liefert zusammen mit den abstrakten Sätzen I.1.2 (S. 17) und I.1.5 (S. 18) folgende Fehlerabschätzung für die Diskretisierung (3) von Problem (1).

SATZ 12 (A priori Abschätzung). *Seien $u \in H_0^1(0,1)$ die eindeutige schwache Lösung von (1) und $u_\mathcal{T} \in S_0^{k,0}(\mathcal{T})$ die eindeutige Lösung von (3). Weiter sei $u \in H^{k+1}(0,1)$. Dann gilt*

$$\max_{a \leq t \leq b} |u(t) - u_\mathcal{T}(t)| \leq |u - u_\mathcal{T}|_1 \leq c_1 h^k \, |u|_{k+1}$$

mit $c_1 = 2\underline{p}^{-1} \max\{\|P\|_{C^0}, \|Q\|_{C^0}\}$.
Besitzt zusätzlich (1) für jede rechte Seite $\varphi \in L^2([0,1],\mathbb{R})$ eine eindeutige schwache Lösung $u_\varphi \in H_0^1(0,1) \cap H^2(0,1)$ mit

$$|u_\varphi|_2 \leq c_2 \, \|\varphi\|_0 ,$$

so gilt weiter

$$\|u - u_\mathcal{T}\|_0 \leq c_3 h^{k+1} \, |u|_{k+1}$$

mit $c_3 = 4c_2\underline{p}^{-1} \max\{\|P\|_{C^0}, \|Q\|_{C^0}\}^2$.

BEMERKUNG 13. (1) Im Vergleich mit [16, Satz II.4.2] kommt Satz 12 mit wesentlich schwächeren Regularitätsannahmen aus. Dies ist für die Übertragung auf partielle Differentialgleichungen ein großer Vorteil.
(2) Die Konstante c_2 kann grob abgeschätzt werden durch $3\underline{p}^{-2} \max\{1, \|P\|_{C^1}, \|Q\|_{C^0}\}$.

Satz 12 liefert keine Aussage über die tatsächliche Größe des Fehlers und seine Verteilung. Dies leisten nur sog. a posteriori Fehlerabschätzungen, die zusammen mit adaptiven Gitterverfeinerungen für die effiziente Diskretisierung partieller Differentialgleichungen unabdingbar sind.

SATZ 14 (A posteriori Abschätzung). *Seien $u \in H_0^1(0,1)$ die eindeutige schwache Lösung von (1) und $u_\mathcal{T} \in S_0^{k,0}(\mathcal{T})$ die eindeutige Lösung von (3). Dann gilt*

$$|u - u_\mathcal{T}|_1 \leq \underline{p}^{-1} \left\{ \sum_{j=0}^{n} \eta_j^2 \right\}^{\frac{1}{2}}$$

mit

$$\eta_j = h_j \left\| F + (P u_\mathcal{T}')' - Q u_\mathcal{T} \right\|_{0,I_j}.$$

BEWEIS. Wir benutzen die Bezeichnungen des Beweises von Satz 7 und setzen zur Abkürzung $e = u - u_\mathcal{T}$. Bezeichne mit $I_\mathcal{T} : H_0^1(0,1) \to S_0^{1,0}(\mathcal{T})$ den Operator, der jeder Funktion ihre stetige, stückweise lineare Interpolierende in den Punkten $t_0, \ldots, t_{n+1}$ zuordnet. Einsetzen von $v = I_\mathcal{T} e$ in (2) und von $v_\mathcal{T} = I_\mathcal{T} e$ in (3) und Subtraktion der resultierenden Gleichungen liefert dann die sog. Galerkin Orthogonalität

$$B(e, I_\mathcal{T} e) = 0.$$

Daher ist

$$\underline{p} \, |e|_1^2 \leq B(e,e) = B(e, e - I_\mathcal{T} e) = \ell(e - I_\mathcal{T} e) - B(u_\mathcal{T}, e - I_\mathcal{T} e).$$

Hieraus folgt durch partielle Integration auf den Teilintervallen I_j, $0 \leq j \leq n$, wegen $(e - I_\mathcal{T} e)(t_i) = 0$ für alle $0 \leq i \leq n+1$

$$\begin{aligned}
&\ell(e - I_\mathcal{T} e) - B(u_\mathcal{T}, e - I_\mathcal{T} e) \\
&\quad = \sum_{j=0}^{n} \int_{I_j} \{ F(e - I_\mathcal{T} e) - P u_\mathcal{T}' (e - I_\mathcal{T} e)' - Q u_\mathcal{T} (e - I_\mathcal{T} e) \} \\
&\quad = \sum_{j=0}^{n} \int_{I_j} \{ F + (P u_\mathcal{T}')' - Q u_\mathcal{T} \} (e - I_\mathcal{T} e).
\end{aligned}$$

Aus dem Beweis von Satz 11 ergibt sich für jedes Teilintervall I_j

$$\| e - I_\mathcal{T} e \|_{0,I_j} \leq h_j \, |e|_{1,I_j}.$$

Hieraus folgt mit der Cauchy-Schwarzschen Ungleichung

$$\begin{aligned}
\underline{p} \, |e|_1^2 &\leq \sum_{j=0}^{n} \int_{I_j} \{ F + (P u_\mathcal{T}')' - Q u_\mathcal{T} \} (e - I_\mathcal{T} e) \leq \sum_{j=0}^{n} \eta_j \, |e|_{1,I_j} \\
&\leq \left\{ \sum_{j=0}^{n} \eta_j^2 \right\}^{\frac{1}{2}} |e|_1 .
\end{aligned}$$

□

BEMERKUNG 15. (1) Mit etwas technischem Mehraufwand kann man zeigen, dass gilt

$$\eta_j \leq |u - u_{\mathcal{T}}|_{1,I_j} + \text{HOT}, \quad \forall 0 \leq j \leq n,$$

wobei HOT für einen explizit berechenbaren Term höherer Ordnung bzgl. η_j steht.
(2) Die Größe η_j nennt man einen *residuellen Fehlerindikator*. Sie kann bei bekanntem $u_{\mathcal{T}}$ explizit berechnet werden und gibt somit eine leicht berechenbare Schranke für den Fehler.
(3) Man kann die Größen η_j wie folgt zur automatischen Gitteranpassung benutzen:

(i) *Wähle ein grobes Gitter* $\mathcal{T}_0 = \{I_j^{(0)} : 0 \leq j \leq n_0\}$. *Setze* $m = 0$.
(ii) *Löse das diskrete Problem zu* $\mathcal{T}_m$ *und berechne damit die entsprechenden Fehlerindikatoren* $\eta_j^{(m)}$, $0 \leq j \leq n_m$. *Setze*

$$\eta^{(m)} = \max_{0 \leq j \leq n_m} \eta_j^{(m)}.$$

(iii) *Falls* $\eta^{(m)} \leq \varepsilon$ *ist, beende das Verfahren. Andernfalls gehe zu (iv).*
(iv) *Falls* $\eta_j^{(m)} \geq \gamma \eta^{(m)}$ *ist, halbiere das Intervall* $I_j^{(m)}$. *Andernfalls lasse es unverändert. Dies bestimmt das nächste Gitter* $\mathcal{T}_{m+1}$. *Ersetze* m *durch* $m + 1$ *und gehe nach (ii) zurück.*

Dabei ist in (iii) ε eine gegebene Toleranz und in (iv) $\gamma \in (0, 1)$ ein gegebener Parameter; typischerweise ist $\gamma = \frac{1}{2}$.

In den folgenden vier Kapiteln behandeln wir Finite Element Verfahren für lineare elliptische Differentialgleichungen zweiter Ordnung. In Kapitel I betrachten wir zunächst abstrakte Variationsprobleme und ihre Diskretisierung, führen dann die Sobolev-Räume ein und geben ihre wichtigsten Eigenschaften an und nutzen schließlich diese Ergebnisse für die Herleitung schwacher Formulierungen elliptischer Differentialgleichungen zweiter Ordnung. In Kapitel II führen wir die Finite Element Räume ein, beweisen ihre Approximationseigenschaften und leiten so a priori Fehlerabschätzungen für die Finite Element Diskretisierung elliptischer Differentialgleichungen zweiter Ordnung her. Kapitel III befasst sich mit praktischen Aspekten der Finite Element Methode: Behandlung gekrümmter Ränder, Berechnung der auftretenden Integrale, effiziente Lösung der diskreten Probleme, a posteriori Fehlerabschätzungen, adaptive Gitterverfeinerung und erforderliche Datenstrukturen. In Kapitel IV schließlich behandeln wir kurz nichtkonforme und gemischte Finite Element Diskretisierungen für elliptische Differentialgleichungen zweiter Ordnung.

KAPITEL I

Analytische Grundlagen

In diesem Kapitel betrachten wir zunächst abstrakte Variationsprobleme und deren Diskretisierung. Danach führen wir die Sobolev-Räume ein und geben einige ihrer wichtigsten Eigenschaften an. Anschließend leiten wir schwache Formulierungen elliptische Differentialgleichungen zweiter Ordnung her und zeigen einige ihrer wichtigsten Eigenschaften.

I.1. Abstrakte Variationsprobleme

Motiviert durch das einführende Kapitel und die schwache Formulierung, Definition 4 (S. 7), des Sturm-Liouville Problems (1) (S. 5), betrachten wir in diesem Paragraphen abstrakte Variationsprobleme der Form:

Finde ein $u \in X$, so dass für alle $v \in X$ gilt $B(u,v) = \ell(v)$.

Dabei ist X ein Banach-Raum, B eine Bilinearform auf X und ℓ ein lineares Funktional auf X.

Der folgende Satz sichert unter geeigneten Voraussetzungen die eindeutige Lösbarkeit derartiger Probleme und charakterisiert ihre Lösungen.

SATZ I.1.1 (Satz von Lax-Milgram). *Seien $(X, \|\cdot\|_X)$ ein Banach-Raum, $\ell \in \mathcal{L}(X, \mathbb{R})$ ein stetiges lineares Funktional und $B \in \mathcal{L}^2(X, \mathbb{R})$ eine stetige Bilinearform. Zusätzlich sei B* symmetrisch*, d.h.*

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

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

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

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

$$J(u) = \frac{1}{2} B(u,u) - \ell(u)$$

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

$$B(u^*, v) = \ell(v) \quad \forall v \in X. \tag{I.1.1}$$

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

$$DJ(u)v = B(u,v) - \ell(v) \quad \forall u, v \in X.$$

Also ist jeder kritische Punkt von J eine Lösung von (I.1.1).
2. Schritt: Seien $u_1, u_2 \in X$ zwei Lösungen von (I.1.1). Dann folgt

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

und wegen der Koerzivität von B

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

Also besitzt (I.1.1) höchstens eine Lösung.
3. Schritt: Für alle $u \in X$ gilt wegen der Koerzivität von B und der Stetigkeit von ℓ

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

Also ist J nach unten beschränkt. Sei

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

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

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

Dann folgt für $n, m \in \mathbb{N}$ wegen der Koerzivität, Symmetrie und Bilinearität von B und der Linearität von ℓ

$$\begin{aligned} \beta \|u_n - u_m\|_X^2 &\leq B(u_n - u_m, u_n - u_m) \\ &= B(u_n, u_n) - 2B(u_n, u_m) + B(u_m, u_m) \\ &= 2B(u_n, u_n) + 2B(u_m, u_m) - B(u_n + u_m, u_n + u_m) \\ &= 2B(u_n, u_n) - 4\ell(u_n) + 2B(u_m, u_m) - 4\ell(u_m) \\ &\quad - 4B\left(\frac{1}{2}(u_n + u_m), \frac{1}{2}(u_n + u_m)\right) + 8\ell\left(\frac{1}{2}(u_n + u_m)\right) \\ &= 8\left\{\frac{1}{2}J(u_n) + \frac{1}{2}J(u_m) - J\left(\frac{1}{2}(u_n + u_m)\right)\right\} \\ &\leq 8\left\{\frac{1}{2}J(u_n) + \frac{1}{2}J(u_m) - \rho\right\} \\ &\underset{n,m\to\infty}{\longrightarrow} 0. \end{aligned}$$

Also ist $(u_n)_{n\in\mathbb{N}}$ eine Cauchy-Folge und konvergiert gegen ein $u^* \in X$ mit $J(u^*) = \rho$. Also besitzt J mindestens ein Minimum. Zusammen mit den Schritten 1 und 2 folgt hieraus die Behauptung. □

Motiviert durch die eindimensionalen Finite Elemente des einführenden Kapitels betrachten wir abstrakte diskrete Probleme der Form:

Finde ein $u_\mathcal{T} \in X_\mathcal{T}$, so dass für alle $v_\mathcal{T} \in X_\mathcal{T}$ gilt $B(u_\mathcal{T}, v_\mathcal{T}) = \ell(v_\mathcal{T})$.

Dabei ist $X_\mathcal{T}$ ein geeigneter endlich dimensionaler Unterraum von X. Die diversen Finite Element Diskretisierungen unterscheiden sich dann u.a. in der Wahl der diskreten Räume $X_\mathcal{T}$.

Satz I.1.1 angewandt auf einen solchen Raum $X_\mathcal{T}$ 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 $X_\mathcal{T}$ bestimmt ist durch die Approximationsgüte des Raumes $X_\mathcal{T} \subset X$. Diese Größe ist unabhängig von den Formen B 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.1.2 (Céa-Lemma). *Die Voraussetzungen und Bezeichnungen seien wie in Satz I.1.1. Setze zur Abkürzung*

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

Sei $X_\mathcal{T} \subset X$ ein endlich dimensionaler Unterraum von X. Bezeichne mit $u \in X$ und $u_\mathcal{T} \in X_\mathcal{T}$ das eindeutige Minimum von J in X bzw. $X_\mathcal{T}$. Dann gilt

$$\|u - u_\mathcal{T}\|_X \leq \frac{\mathcal{B}}{\beta} \inf_{v_\mathcal{T} \in X_\mathcal{T}} \|u - v_\mathcal{T}\|_X . \tag{I.1.2}$$

Beweis. Wegen Satz I.1.1 besitzt J ein eindeutiges Minimum $u_\mathcal{T}$ in $X_\mathcal{T}$. Dieses ist charakterisiert durch

$$B(u_\mathcal{T}, v_\mathcal{T}) = \ell(v_\mathcal{T}) \quad \forall v_\mathcal{T} \in X_\mathcal{T}. \tag{I.1.3}$$

Aus (I.1.1) und (I.1.3) folgt wegen der Bilinearität von B die sog. *Galerkin Orthogonalität*

$$B(u - u_\mathcal{T}, v_\mathcal{T}) = 0 \quad \forall v_\mathcal{T} \in X_\mathcal{T}. \tag{I.1.4}$$

Hieraus ergibt sich für jedes $v_\mathcal{T} \in X_\mathcal{T}$ wegen der Koerzivität von B

$$\begin{aligned} \beta \|u - u_\mathcal{T}\|_X^2 &\leq B(u - u_\mathcal{T}, u - u_\mathcal{T}) \\ &= B(u - u_\mathcal{T}, u - v_\mathcal{T}) + B(u - u_\mathcal{T}, v_\mathcal{T} - u_\mathcal{T}) \\ &= B(u - u_\mathcal{T}, u - v_\mathcal{T}) \\ &\leq \mathcal{B} \|u - u_\mathcal{T}\|_X \|u - v_\mathcal{T}\|_X . \end{aligned}$$

Da $v_\mathcal{T}$ beliebig war, folgt hieraus die Behauptung. □

Wie bei dem Sturm-Liouville Problem wird bei den Anwendungen der folgenden Abschnitte $\|\cdot\|_X$ in der Regel die H^1-Norm oder eine ähnliche Norm sein. Wie bei dem eindimensionalen Problem wollen wir aber häufig auch Fehlerabschätzungen in der L^2-Norm oder einer vergleichbaren Norm herleiten. Dazu benötigen wir den Begriff der stetigen Einbettung.

DEFINITION I.1.3 (Stetige Einbettung). Seien $(X, \|\cdot\|_X)$ und $(Y, \|\cdot\|_Y)$ zwei normierte Vektorräume. Dann heißt X *stetig eingebettet* in Y, kurz $X \hookrightarrow Y$, wenn $X \subset Y$ und die kanonische Injektion $i : X \to Y$ stetig ist.

BEMERKUNG I.1.4 (Normvergleich bei stetiger Einbettung). Gilt $X \hookrightarrow Y$, so folgt aus der Definition der Stetigkeit für lineare Operatoren, dass es eine Konstante $c > 0$ gibt mit $\|\varphi\|_Y \leq c\,\|\varphi\|_X$ für alle $\varphi \in X$.

Der folgende Satz zeigt, dass wir unter bestimmten Bedingungen die Abschätzung von Satz I.1.2 verbessern können.

SATZ I.1.5 (Satz von Aubin-Nitsche). *Zusätzlich zu den Voraussetzungen der Sätze I.1.1 und I.1.2 sei H ein Hilbert-Raum 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*

$$B(v, u_\varphi) = (\varphi\,,\, v)_H \quad \forall v \in X. \tag{I.1.5}$$

Dann gilt

$$\|u - u_{\mathcal{T}}\|_H \leq \mathcal{B}\,\|u - u_{\mathcal{T}}\|_X \sup_{\varphi \in H;\|\varphi\|_H = 1} \inf_{v_{\mathcal{T}} \in X_{\mathcal{T}}} \|u_\varphi - v_{\mathcal{T}}\|_X\,. \tag{I.1.6}$$

BEWEIS. Wegen $X \hookrightarrow H$ definiert jedes $\varphi \in H$ durch $v \mapsto (\varphi\,,\, v)_H$ ein stetiges lineares Funktional auf X. Wegen Satz I.1.1 besitzt somit (I.1.5) eine eindeutige Lösung $u_\varphi \in X$. Aus (I.1.5) und (I.1.4) folgt für beliebiges $\varphi \in H$ und beliebiges $v_{\mathcal{T}} \in X_{\mathcal{T}}$

$$\begin{aligned}(u - u_{\mathcal{T}}\,,\, \varphi)_H &= B(u - u_{\mathcal{T}}, u_\varphi)\\ &= B(u - u_{\mathcal{T}}, u_\varphi - v_{\mathcal{T}})\\ &\leq \mathcal{B}\,\|u - u_{\mathcal{T}}\|_X\,\|u_\varphi - v_{\mathcal{T}}\|_X\,.\end{aligned}$$

Da wegen der Dichtheit von X in H

$$\|u - u_{\mathcal{T}}\|_H = \sup_{\varphi \in H;\|\varphi\|_H = 1} (u - u_{\mathcal{T}}\,,\, \varphi)_H$$

ist, folgt hieraus die Fehlerabschätzung. □

In den nächsten Paragraphen werden wir u.a. Konvektions-Diffusionsgleichungen betrachten. Deren Variationsformulierung und Finite Element Diskretisierung passt nicht in den Rahmen der Sätze I.1.1, I.1.2 und I.1.5, da die zugehörige Bilinearform B nicht mehr symmetrisch ist. Der folgende Satz zeigt, dass die analytischen Probleme (I.1.1) und (I.1.5) und das diskrete Problem (I.1.3) nach wie vor eindeutig lösbar sind und dass die Fehlerabschätzungen (I.1.2) und (I.1.6) gültig bleiben. Allerdings können wir wegen der fehlenden Symmetrie von B die Lösungen von (I.1.1) und (I.1.3) nicht mehr als Minimum

des Funktionals J charakterisieren. Zudem brauchen wir eine zusätzliche Bedingung an den Raum X und müssen für die eindeutige Lösbarkeit der analytischen Probleme (I.1.1) und (I.1.5) deutlich schärferes funktionalanalytisches Geschütz auffahren.

SATZ I.1.6 (Koerzive, nicht symmetrische Bilinearform). *Die Voraussetzungen der Sätze I.1.1, I.1.2 und I.1.5 seien bis auf die Symmetrie der Bilinearform B erfüllt. Zudem sei X reflexiv. Dann besitzen die Probleme (I.1.1), (I.1.3) und (I.1.5) für jedes $\ell \in \mathcal{L}(X,\mathbb{R})$ bzw. $\varphi \in H$ eine eindeutige Lösung $u \in X$, $u_{\mathcal{T}} \in X_{\mathcal{T}}$ und $u_\varphi \in X$ und es gelten die Fehlerabschätzungen (I.1.2) und (I.1.6).*

BEWEIS. Im Beweis der Sätze I.1.2 und I.1.5 haben wir die Koerzivität von B, nicht aber die Symmetrie ausgenutzt. Daher müssen wir nur die eindeutige Lösbarkeit der Probleme (I.1.1), (I.1.3) und (I.1.5) zeigen.
Betrachte zunächst das diskrete Problem (I.1.3). Dieses ist ein endlich dimensionales lineares Gleichungssystem mit genauso vielen Gleichungen wie Unbekannten. Wegen der Koerzivität von B besitzt das zugehörige homogene Problem nur die triviale Lösung. Daher ist (I.1.3) für alle $\ell \in \mathcal{L}(X,\mathbb{R})$ eindeutig lösbar.
Betrachte als nächstes das analytische Problem (I.1.1). Wegen der Koerzivität von B besitzt (I.1.1) höchstens eine Lösung. Weiter definiert B durch die Vorschrift $(Lu)(v) = B(u,v)$ eine lineare Abbildung von X in seinen Dualraum $X^* = \mathcal{L}(X,\mathbb{R})$, die wegen der Stetigkeit von B ebenfalls stetig ist. Wegen der Koerzivität von B gilt $\beta \|u\|_X \leq \|Lu\|_{X^*}$ für alle $u \in X$. Daher ist L injektiv und das Bild R von L ein abgeschlossener Unterraum von X^*. Wäre $R \neq X^*$ gäbe es wegen des *Satzes von Hahn-Banach* [17, Korollar III.1.8] und der Reflexivität von X ein Element $v_0 \in X$ mit $(Lu)(v_0) = B(u,v_0) = 0$ für alle $u \in X$.[1] Dies ist ein Widerspruch zur Koerzivität von B. Also ist L auch surjektiv und damit (I.1.1) eindeutig lösbar.
Die eindeutige Lösbarkeit von Problem (I.1.5) schließlich folgt ganz analog mit dem adjungierten Operator L^* mit $(L^*u)(v) = B(v,u)$ an Stelle von L. □

BEMERKUNG I.1.7 (inf-sup Bedingung). Die Koerzivität der Bilinearform B kann zu der sog. *inf-sup Bedingung*

$$0 < \beta \leq \inf_{u \in X\setminus\{0\}} \sup_{v \in X\setminus\{0\}} \frac{B(u,v)}{\|u\|_X \|v\|_X}$$

abgeschwächt werden. Wegen

$$\|\ell\|_{X^*} = \sup_{v \in X\setminus\{0\}} \frac{|\ell(v)|}{\|v\|_X} \quad \forall \ell \in X^*$$

[1] Für einen Hilbert-Raum ist $v_0 = w - P_R w$ mit $w \in X \setminus R$ und der orthogonalen Projektion P_R auf R.

impliziert sie nämlich für obigen Operator L

$$\beta \|u\|_X \leq \|Lu\|_{X^*} \quad \forall u \in X.$$

Hieraus folgt dann wieder die Injektivität von L und die Abgeschlossenheit seines Bildes. Die inf-sup Bedingung ist für sog. Sattelpunktsprobleme von besonderer Bedeutung (vgl. §IV.3 (S. 120)).

BEMERKUNG I.1.8 (Allgemeine Diskretisierungsverfahren). Wir betrachten allgemeine Diskretisierungsverfahren $L_h u_h = f_h$ für abstrakte lineare Probleme $Lu = f$ wie in [16, §III.2]. Dabei ist $L : X \to Y$ eine nicht notwendig stetige lineare Abbildung zwischen zwei Banach-Räumen $(X, \|\cdot\|_X)$ und $(Y, \|\cdot\|_Y)$ und $L_h : X_h \to Y_h$ eine automatisch stetige lineare Abbildung zwischen zwei endlich dimensionalen Banach-Räumen $(X_h, \|\cdot\|_{X_h})$ und $(Y_h, \|\cdot\|_{Y_h})$. Die unendlich dimensionalen und die endlich dimensionalen Räume sind durch Restriktionsabbildungen $R_{X_h} : X \to X_h$ und $R_{Y_h} : Y \to Y_h$ verknüpft. Zusätzlich wird ein Fortsetzungsoperator $I_{X_h} : X_h \to \widetilde{X}$ in einen Banach-Raum $(\widetilde{X}, \|\cdot\|_{\widetilde{X}})$ mit einer schwächeren Topologie als X und $X \subset \widetilde{X}$ betrachtet, der Fehlerabschätzungen in der h-unabhängigen Norm $\|\cdot\|_{\widetilde{X}}$ erlaubt.
Die Variationsprobleme dieses Abschnittes und ihre Diskretisierungen passen wie folgt in diesen Rahmen. Es ist $Y = X^*$ der Dualraum von X, L ist wie im Beweis von Satz I.1.6 definiert durch $(Lu)(v) = B(u, v)$, $Y_h = X_h^*$ ist der Dualraum von $X_h = X_{\mathcal{T}}$ und L_h ist die Restriktion von L definiert durch $(L_h u_h)(v_h) = B(u_h, v_h)$ für alle $u_h, v_h \in X_h$. Wegen $X_{\mathcal{T}} \subset X$ kann für I_{X_h} die kanonische Injektion gewählt und R_{Y_h} als die kanonische Restriktion $(R_{Y_h}\ell)(v_h) = \ell(v_h)$ definiert werden; R_{X_h} hängt vom Diskretisierungverfahren ab und ist in der Regel ein Interpolations- oder Quasi-Interpolationsoperator (vgl. (II.2.1) (S. 41) und Definition III.3.1 (S. 82)). Üblicherweise ist $\widetilde{X}$ der Raum H aus dem Satz von Aubin-Nitsche I.1.5. Aus der Koerzivität von B bzw. der inf-sup Bedingung aus Bemerkung I.1.7 folgt die Stabilität $\left\|L_h^{-1}\right\|_{\mathcal{L}(Y_h, X_h)} \leq \beta^{-1}$. Aus der Stetigkeit der Bilinearform B folgt die Konsistenzfehlerabschätzung $\|L_h R_{X_h} u - R_{Y_h} L u\|_{Y_h} \leq \mathcal{B} \|R_{X_h} u - u\|_X$.

I.2. Sobolev-Räume

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

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

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

und alle $\alpha \in \mathbb{N}^d$

$$\int_\Omega \varphi D^\alpha \psi = (-1)^{|\alpha|} \int_\Omega \psi D^\alpha \varphi. \tag{I.2.1}$$

Gleichung (I.2.1) motiviert die folgende Definition der schwachen Ableitung. Dieser Begriff verallgemeinert denjenigen der klassischen Ableitung.

DEFINITION I.2.1 (Schwache Ableitung). Seien $\varphi, \psi \in L^1(\Omega)$ und $\alpha \in \mathbb{N}^d$. Dann heißt ψ die *α-te schwache Ableitung* von φ, kurz $\psi = D^\alpha \varphi$, wenn für alle $\rho \in C_0^\infty(\Omega)$ gilt

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

BEMERKUNG I.2.2 (Eigenschaften schwacher Ableitungen). (1) Die α-te schwache Ableitung ist, sofern sie existiert, eindeutig im Sinne von L^1-Funktionen.
(2) Ist $\varphi \in C^{|\alpha|}(\Omega)$, so stimmen die α-te schwache Ableitung und die klassische α-te Ableitung überein.

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

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

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.2.1)). □

BEISPIEL I.2.3 (Schwach, aber nicht klassisch differenzierbare Funktion). Sei $\Omega = (-1, 1)$ und $\varphi(x) = |x|$. Dann ist φ im Sinne von Definition I.2.1 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 Sobolev-Räume. Sie verallgemeinern die klassischen $C^k(\Omega)$-Räume und sind ein unverzichtbares Hilfsmittel für die Variationsrechnung (vgl. §I.3), auf der die Finite Element Methoden aufbauen.

DEFINITION I.2.4 (Sobolev-Räume $W^{k,p}(\Omega)$). (1) Für $k \in \mathbb{N}$ und $p \in [1,\infty)$ definieren wir den *Sobolev-Raum* $W^{k,p}(\Omega)$ und seine Norm $\|\cdot\|_{k,p}$ durch

$$W^{k,p}(\Omega) = \{\varphi \in L^p(\Omega) : D^\alpha\varphi \in L^p(\Omega)\ \forall\, |\alpha| \le k\}\,,$$

$$\|\varphi\|_{k,p} = \left\{ \sum_{|\alpha|\le k} \|D^\alpha\varphi\|_p^p \right\}^{1/p}$$

mit

$$\|\psi\|_p = \left\{ \int_\Omega |\psi|^p \right\}^{\frac{1}{p}}.$$

(2) Für $k \in \mathbb{N}^*$ definieren wir durch

$$|\varphi|_{k,p} = \left\{ \sum_{|\alpha|=k} \|D^\alpha\varphi\|_p^p \right\}^{\frac{1}{p}}$$

eine Semi-Norm auf $W^{k,p}(\Omega)$ und setzen zur Abkürzung $|\cdot|_{0,p} = \|\cdot\|_p$.
(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.2.5 (Unbeschränkte Sobolev-Funktionen). (1) Seien φ und Ω wie in Beispiel I.2.3. Dann gilt $\varphi \in W^{1,p}(\Omega)$ für alle $p \in [1,\infty)$.
(2) Seien $d \ge 2$, $\Omega = B(0,\frac{1}{2})$ und $\varphi(x) = |x|^s$ mit $s \in \mathbb{R}$, wobei $|\cdot|$ die euklidische Norm in $\mathbb{R}^d$ bezeichnet. Dann gilt

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

und

$$\begin{aligned} \|D^\alpha\varphi\|_p &\approx \int_0^{\frac{1}{2}} r^{(s-|\alpha|)p} r^{d-1}\mathrm{d}r < \infty \\ \iff \quad & p(s-|\alpha|) + d - 1 > -1 \\ \iff \quad & s > |\alpha| - \frac{d}{p}. \end{aligned}$$

(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

$$\int_\Omega \sum_{i=1}^{2} \left|\frac{\partial\varphi}{\partial x_i}\right|^2 = 2\pi \int_0^{\frac{1}{2}} \frac{1}{r^2(\ln r)^2} r\mathrm{d}r = 2\pi \lim_{\varepsilon\to 0} \left[-\frac{1}{\ln r}\right]_{r=\varepsilon}^{r=\frac{1}{2}}$$
$$= \frac{2\pi}{\ln 2}.$$

Also ist $\varphi \in H^1(\Omega)$. Man beachte, dass $|\varphi(x)| \to \infty$ für $|x| \to 0$ gilt. Somit zeigt dieses Beispiel, dass für Raumdimensionen $d \geq 2$ Funktionen in H^1 im allgemeinen keine Punktwerte besitzen.

Satz I.2.6 (Eigenschaften der Sobolev-Räume). *(1)* $(W^{k,p}(\Omega), \|\cdot\|_{k,p})$ *ist ein Banach-Raum.*
(2) $C^\infty(\Omega)$ *ist dicht in* $W^{k,p}(\Omega)$.
(3) $H^k(\Omega)$ *ist ein Hilbert-Raum mit Skalarprodukt*

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

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

$$\int_\Omega \varphi_n D^\alpha\rho = (-1)^{|\alpha|} \int_\Omega D^\alpha\varphi_n\rho. \tag{I.2.2}$$

Wegen des Lebesgueschen Konvergenzsatzes können wir in (I.2.2) den Grenzübergang $n \to \infty$ durchführen und erhalten

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

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

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

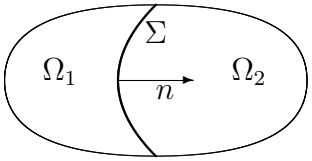

ABBILDUNG I.2.1. Zerlegung von Ω in Teilgebiete

SATZ I.2.7 (Stückweise glatte Funktionen). *Seien Ω_1, Ω_2 zwei nicht leere, offene, beschränkte und disjunkte Teilmengen von Ω mit stückweise glattem Rand und $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2$. Weiter sei $\varphi \in L^p(\Omega)$ so, dass $\varphi|_{\Omega_i} \in C^k(\Omega_i)$ ist für $i \in \{1,2\}$ und $k \geq 1$. 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 $\mathbf{n}$ die äußere Normale an Ω_1 und $\mathbb{J}_\Sigma(\varphi)$ der Sprung von φ über $\Sigma = \partial\Omega_1 \cap \partial\Omega_2$ in Richtung $\mathbf{n}$, d.h.

$$\mathbb{J}_\Sigma(\varphi)(x) = \lim_{t\to 0+} \varphi(x + t\mathbf{n}) - \lim_{t\to 0+} \varphi(x - t\mathbf{n}) \quad \forall x \in \Sigma.$$

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

$$\begin{aligned} -\int_\Omega \varphi \frac{\partial \rho}{\partial x_i} &= -\int_{\Omega_1} \varphi \frac{\partial \rho}{\partial x_i} - \int_{\Omega_2} \varphi \frac{\partial \rho}{\partial x_i} \\ &= \int_{\Omega_1} \frac{\partial \varphi}{\partial x_i}\rho - \int_{\partial\Omega_1} \varphi\rho\mathbf{n}_i + \int_{\Omega_2} \frac{\partial \varphi}{\partial x_i}\rho + \int_{\partial\Omega_2} \varphi\rho\mathbf{n}_i \\ &= \int_\Omega \frac{\partial \varphi}{\partial x_i}\rho + \int_\Sigma \mathbb{J}_\Sigma(\varphi)\rho\mathbf{n}_i. \end{aligned}$$

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

$$\int_\Sigma \mathbb{J}_\Sigma(\varphi)\rho\mathbf{n}_i = 0 \quad \forall \rho \in C_0^\infty(\Omega), i \in \{1, \ldots, d\}.$$

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

BEMERKUNG I.2.8 ($C_0^\infty(\Omega)$ nicht dicht in $H^1(\Omega)$). Gemäß Satz I.2.6 (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

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

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

Bemerkung I.2.8 führt auf folgende Definition.

DEFINITION I.2.9 (Sobolev-Räume $W_0^{k,p}(\Omega)$). $W_0^{k,p}(\Omega)$, $k \geq 1$, ist die Vervollständigung von $C_0^\infty(\Omega)$ bzgl. $\|\cdot\|_{k,p}$; $H_0^k(\Omega) = W_0^{k,2}(\Omega)$.

Wie wir aus [16, §III.1] wissen, spielt der Rand von Ω 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.

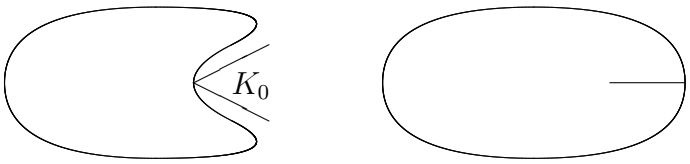

ABBILDUNG I.2.2. links: Lipschitz-Gebiet mit Kegel K_0; rechts: nicht Lipschitz-Gebiet

DEFINITION I.2.10 (Lipschitz-Rand). Ω hat einen *Lipschitz-Rand* bzw. Ω ist ein *Lipschitz-Gebiet*, wenn es ein $N \in \mathbb{N}^*$ und offene Mengen $U_1, \ldots, U_N \in \mathbb{R}^d$ 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.2.11 (Einheitsnormalenfeld; Kegelbedingung). (1) Ω sei ein Lipschitz-Gebiet. Dann existiert fast überall auf $\partial\Omega$ das äußere Einheitsnormalenfeld $\mathbf{n}$ zu Ω.
(2) Ω 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 Ω ein Lipschitz-Gebiet.

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

SATZ I.2.12 (Spursatz). *Seien Ω ein Lipschitz-Gebiet und $k \in \mathbb{N}^*$, $\ell \in \{0, \ldots, k-1\}$. Dann gibt es eine stetige lineare Abbildung $\gamma_\ell : W^{k,p}(\Omega) \to L^p(\partial\Omega)$ mit der Eigenschaft*

$$\gamma_\ell(\varphi) = \frac{\partial^\ell \varphi}{\partial \mathbf{n}^\ell}\Big|_{\partial\Omega} \quad \forall \varphi \in C^k(\overline{\Omega}).$$

BEWEISIDEE. ([1], [2, §§A5.7, A5.14]) Man zeigt zunächst, dass die Restriktionen von $C_0^\infty(\mathbb{R}^d)$-Funktionen auf Ω dicht sind in $W^{k,p}(\Omega)$. Dann führt man eine Überdeckung von $\partial\Omega$ wie in Definition I.2.10 ein und rechnet die Eigenschaft von γ_ℓ auf den Karten $\partial\Omega \cap U_i$ für $C_0^\infty(\mathbb{R}^d)$-Funktionen nach. □

BEMERKUNG I.2.13 ($W^{k-\ell-\frac{1}{p},p}(\partial\Omega)$-Räume). Die Bezeichnungen und Voraussetzungen seien wie in Satz I.2.12. Wegen des *Satzes vom abgeschlossenen Bild* (engl. closed range theorem) [17, Theorem IV.5.1] ist $\gamma_\ell(W^{k,p}(\Omega))$ ein abgeschlossener Unterraum von $L^p(\partial\Omega)$. Dieser wird üblicherweise mit $W^{k-\ell-\frac{1}{p},p}(\partial\Omega)$ bezeichnet. Für unsere Anwendungen sind die Fälle $\ell = 0$ und $\ell = 1$ besonders wichtig. Für eine alternative Charakterisierung der Räume $W^{k-\ell-\frac{1}{p},p}(\partial\Omega)$ analog zu Definition I.2.4 mit Ω ersetzt durch $\partial\Omega$ verweisen wir auf [1].

Der folgende Satz charakterisiert $W_0^{k,p}$-Funktionen als die $W^{k,p}$-Funktionen, die auf dem Rand von Ω in einem geeigneten Sinne verschwinden.

SATZ I.2.14 (Kern des Spuroperators). $W_0^{k,p}(\Omega) = \{\varphi \in W^{k,p}(\Omega) : \gamma_\ell(\varphi) = 0 \; \forall 0 \le \ell \le k-1\}$.

BEWEISIDEE. ([1], [2, §§A5.7, A5.14]) Da die Abbildungen γ_ℓ stetig sind, ist $\bigcap_{0\le\ell\le k-1} \ker\gamma_\ell$ ein abgeschlossener Unterraum von $W^{k,p}(\Omega)$, der $C_0^\infty(\Omega)$ enthält. Hieraus folgt mit Definition I.2.9 die Behauptung. □

Der folgende Satz ist ein wichtiges Hilfsmittel.

SATZ I.2.15 (Friedrichsche Ungleichung). $\|\cdot\|_{k,p}$ *und* $|\cdot|_{k,p}$ *sind äquivalente Normen auf* $W_0^{k,p}(\Omega)$.

BEWEIS. Offensichtlich gilt $|\varphi|_{k,p} \le \|\varphi\|_{k,p}$ für alle $\varphi \in W^{k,p}(\Omega)$. Für die umgekehrte Abschätzung wähle $R \in \mathbb{R}_+^*$ so, dass Ω 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

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

Integration über Ω liefert

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

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

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

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

Bemerkung I.2.16. Aus Satz I.2.15 folgt $\|\varphi\|_{k,p} \leq c_k(\Omega) |\varphi|_{k,p}$ für alle $\varphi \in W_0^{k,p}(\Omega)$. Die Konstante $c_k(\Omega)$ hängt nur von k und dem Durchmesser von Ω ab. Eine analoge Abschätzung gilt für alle Funktionen, die auf einem Teil des Randes verschwinden, der positives $(d-1)$-dimensionales Maß hat.

Als nächstes wollen wir Teilmengenbeziehungen zwischen $W^{k,p}$- und L^q-Räumen untersuchen. Dazu benötigen wir neben dem Begriff der stetigen Einbettung, Definition I.1.3 (S. 18), auch denjenigen der kompakten Einbettung.

Definition I.2.17 (Kompakte Einbettung). Seien $(X, \|\cdot\|_X)$ und $(Y, \|\cdot\|_Y)$ zwei normierte Vektorräume. Dann heißt X *kompakt eingebettet* in Y, kurz $X \overset{c}{\hookrightarrow} Y$, wenn die Einheitskugel $B_{\|\cdot\|_X}(0,1)$ bzgl. der Norm $\|\cdot\|_X$ eine kompakte Teilmenge von Y ist bzgl. der Norm $\|\cdot\|_Y$.

Bemerkung I.2.18 (Eigenschaften kompakter Einbettungen). (1) Aus $X \overset{c}{\hookrightarrow} Y$ folgt $X \hookrightarrow Y$.
(2) Ist $X \overset{c}{\hookrightarrow} Y$ und $(\varphi_n)_{n\in\mathbb{N}} \subset X$ eine beschränkte Folge bzgl. der Norm $\|\cdot\|_X$, so besitzt $(\varphi_n)_{n\in\mathbb{N}}$ eine in Y bzgl. der Norm $\|\cdot\|_Y$ konvergente Teilfolge.

Satz I.2.19 (Sobolevscher Einbettungssatz). *(1) Sei $p < d$. Dann gilt $W^{k,p}(\Omega) \hookrightarrow W^{k-1,q}(\Omega)$ für alle $q \in [1, \frac{pd}{d-p}]$ und $W^{k,p}(\Omega) \overset{c}{\hookrightarrow} W^{k-1,q}(\Omega)$ für alle $q \in [1, \frac{pd}{d-p})$.*
(2) Sei $p = d$. Dann gilt $W^{k,p}(\Omega) \overset{c}{\hookrightarrow} W^{k-1,q}(\Omega)$ für alle $q \in [1, \infty)$.
(3) Sei $k > \frac{d}{p}$. Dann gilt $W^{k,p}(\Omega) \overset{c}{\hookrightarrow} C^\ell(\overline{\Omega})$ für alle $\ell \in \mathbb{N}$ mit $0 \leq \ell < k - \frac{d}{p}$.

Beweis. [1], [2, §§A 5.1, A5.4, A8.2] □

Bemerkung I.2.20. (1) Sei $d = 2$, $p = 2$ und $\Omega = B(0; \frac{1}{2})$. Dann zeigt Beispiel I.2.5 (3), dass die Schranke an q in Satz I.2.19 (2) scharf ist.
(2) Sei $d \geq 3$, $p = 2$ und $\Omega = B(0; \frac{1}{2})$. Dann zeigt Beispiel I.2.5 (2), dass die Schranke an q in Satz I.2.19 (1) scharf ist.

(3) Sei $p = 2$ und $d = 2$. Dann ist $H^1(\Omega) \overset{c}{\hookrightarrow} L^q(\Omega)$ für jedes $q \in [1, \infty)$.
(4) Sei $p = 2$ und $d = 3$. Dann ist $H^1(\Omega) \overset{c}{\hookrightarrow} L^q(\Omega)$ für jedes $q \in [1, 6)$ und $H^1(\Omega) \hookrightarrow L^6(\Omega)$.
(5) Sei $p = 2$ und $d \in \{2, 3\}$. Dann ist $H^2(\Omega) \hookrightarrow C^0(\overline{\Omega})$. Für $H^1(\Omega)$-Funktionen sind Punktwerte dagegen nicht definiert (vgl. Beispiel I.2.5 (3)).

Der folgende Satz ist ein Analogon der Friedrichschen Ungleichung.

SATZ I.2.21 (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.2.15 müssen wir nur zeigen, dass es eine Konstante $C > 0$ gibt mit

$$\|\varphi\|_1 \leq C\,|\varphi|_1 \quad \forall \varphi \in V. \tag{I.2.3}$$

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

$$\|\varphi_n\|_1 = 1 \quad \forall n \in \mathbb{N} \tag{I.2.4}$$

und

$$\lim_{n\to\infty} |\varphi_n|_1 = 0. \tag{I.2.5}$$

Wegen Satz I.2.19 und Bemerkung I.2.18 (2) gibt es eine Teilfolge $(\varphi_{n_k})_{k\in\mathbb{N}}$ von $(\varphi_n)_{n\in\mathbb{N}}$ und eine Funktion $\varphi \in L^2(\Omega)$ mit

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

Wegen (I.2.5) konvergiert $(\varphi_{n_k})_{k\in\mathbb{N}}$ sogar in $H^1(\Omega)$. Mithin ist $\varphi \in H^1(\Omega)$ und $|\varphi|_1 = 0$. Daher ist φ konstant. Da V ein abgeschlossener Unterraum von $H^1(\Omega)$ ist, gilt aber $\int_\Omega \varphi = 0$. Also ist $\varphi = 0$ im Widerspruch zu (I.2.4). □

BEMERKUNG I.2.22. (1) Satz I.2.21 kann für $H^1(\Omega)$ nicht gelten, da die rechte Seite von (I.2.3) für die konstante Funktion $\varphi = 1$ verschwindet.
(2) Der Beweis von Satz I.2.21 ist nicht konstruktiv. Mit anderen Techniken kann man zeigen, dass die Konstante C in (I.2.3) proportional zum Durchmesser $\operatorname{diam}(\Omega) = \sup_{x,y\in\Omega} |x - y|$ von Ω ist. Ist insbesondere Ω konvex, ergibt sich $C \leq \frac{1}{\pi} \operatorname{diam}(\Omega)$ [13, §3.4].
(3) Analoge Aussagen zu Satz I.2.21 gelten für $W^{1,p}(\Omega)$ und $\{\varphi \in W^{1,p}(\Omega) : \int_\Omega \varphi = 0\}$ mit $p \in (1, \infty)$ [13, §3.4].

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 $\mathbf{n}$. Wir betrachten skalare, lineare, elliptische Differentialgleichungen 2. Ordnung.

Ihre allgemeine Form lautet [16, §III.1]

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

Dabei ist $f\in L^2(\Omega)$, $\alpha\in C(\Omega,\mathbb{R}_+)$, $\mathbf{a}=(\mathbf{a}_1,\dots,\mathbf{a}_d)\in C^1(\Omega,\mathbb{R}^d)$ und $A=(A_{ij})_{1\le i,j\le d}\in C^1(\Omega,\mathbb{R}^{d\times d})$ mit $A_{ij}(x)=A_{ji}(x)$ für alle $x\in\Omega$, $1\le i,j\le d$ und

$$\lambda_0=\inf_{x\in\Omega}\inf_{z\in\mathbb{R}^d\setminus\{0\}}\frac{z^TA(x)z}{z^Tz}>0.$$

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

- einer *Konvektions-Diffusionsgleichung*, wenn α, $\mathbf{a}$ und A beliebig sind,
- einer *Reaktions-Diffusionsgleichung*, wenn $\mathbf{a}=0$ ist,
- einer *Membrangleichung*, wenn $\alpha=0$ und $\mathbf{a}=0$ ist,
- einer *Poissongleichung*, wenn $\alpha=0$, $\mathbf{a}=0$ 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 Γ,
- (inhomogene) *Neumann-Randbedingungen*: $\mathbf{n}\cdot A\nabla u=g$ auf Γ,
- *gemischte Dirichlet-Neumann-Randbedingungen*: $u=0$ auf Γ_D und $\mathbf{n}\cdot A\nabla u=g$ auf Γ_N.

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

Sei nun $u\in C^2(\Omega)$ eine Lösung von (I.3.1) mit homogenen Dirichlet-Randbedingungen und $v\in C_0^\infty(\Omega)$. Multiplikation von (I.3.1) mit v, Integration über Ω und Anwenden des Gaußschen Integralsatzes liefert

$$\begin{aligned}\int_\Omega fv&=-\sum_{1\le i,j\le d}\int_\Omega\frac{\partial}{\partial x_i}\left(A_{ij}\frac{\partial u}{\partial x_j}\right)v\\&\quad+\sum_{i=1}^{d}\int_\Omega\mathbf{a}_i\frac{\partial u}{\partial x_i}v+\int_\Omega\alpha uv\\&=\sum_{1\le i,j\le d}\int_\Omega A_{ij}\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i}+\sum_{i=1}^{d}\int_\Omega\mathbf{a}_i\frac{\partial u}{\partial x_i}v+\int_\Omega\alpha uv\\&=\int_\Omega\left\{\nabla u\cdot A\nabla v+(\mathbf{a}\cdot\nabla u)\,v+\alpha uv\right\}.\end{aligned} \tag{I.3.2}$$

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

$$\int_\Omega \{\nabla u \cdot A\nabla v + (\mathbf{a}\cdot\nabla u)\, v + \alpha uv\} = \int_\Omega fv \tag{I.3.3}$$

für alle $v \in H_0^1(\Omega)$ erfüllt. Umgekehrt folgt aus (I.3.2), dass eine Lösung von (I.3.3) die Differentialgleichung (I.3.1) erfüllt, sofern sie hinreichend glatt, d.h. in $C^2(\Omega)$ ist. In diesem Sinne ist (I.3.3) 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.3.2) zusätzlich Randterme $-\int_\Gamma \mathbf{n}\cdot A\nabla uv$ auf. Erfüllt u Neumann-Randbedingungen, so gilt für diesen Randterm

$$-\int_\Gamma \mathbf{n}\cdot A\nabla uv = -\int_\Gamma gv.$$

Wir werden daher in diesem Fall (I.3.3) durch den zusätzlichen Term $\int_\Gamma gv$ auf der rechten Seite modifizieren. Diese Überlegungen führen auf folgende Definition.

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

$$\int_\Omega \{\nabla u \cdot A\nabla v + (\mathbf{a}\cdot\nabla u)\, v + \alpha uv\} = \int_\Omega fv.$$

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

$$\int_\Omega \{\nabla u \cdot A\nabla v + (\mathbf{a}\cdot\nabla u)\, v + \alpha uv\} = \int_\Omega fv + \int_{\Gamma_N} gv.$$

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

$$\int_\Omega \{\nabla u \cdot A\nabla v + (\mathbf{a}\cdot\nabla u)\, v + \alpha uv\} = \int_\Omega fv + \int_\Gamma gv.$$

Bemerkung I.3.2 (Eigenschaften schwacher Lösungen). (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 $\alpha \in L^\infty(\Omega)$, $\alpha \geq 0$, $\mathbf{a} \in L^\infty(\Omega, \mathbb{R}^d)$, $A \in L^\infty(\Omega, \mathbb{R}^{d\times d})$.
(3) Bei inhomogenen Dirichlet-Randbedingungen $u = u_D$ auf Γ bzw. Γ_D muss in Definition 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* $-\frac{1}{2}\operatorname{div}\mathbf{a}+\alpha\geq 0$, *so besitzt die Konvektions-Diffusionsgleichung mit homogenen Dirichlet-Randbedingungen eine eindeutige schwache Lösung.*
(2) Ist $-\frac{1}{2}\operatorname{div}\mathbf{a}+\alpha\geq 0$ *und* $\mathbf{a}\cdot\mathbf{n}\geq 0$ *auf* Γ_N, *so besitzt die Konvektions-Diffusionsgleichung mit gemischten Randbedingungen eine eindeutige schwache Lösung.*
(3) Ist $-\frac{1}{2}\operatorname{div}\mathbf{a}+\alpha\geq\alpha_0>0$ *und* $\mathbf{a}\cdot\mathbf{n}\geq 0$ *auf* Γ, *so besitzt die Konvektions-Diffusionsgleichung mit Neumann-Randbedingungen eine eindeutig schwache Lösung.*
(4) Ist $\alpha=0$, $\operatorname{div}\mathbf{a}=0$ *und* $\mathbf{a}\cdot\mathbf{n}=0$ *auf* Γ *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.1.6 (S. 19) an. In allen Fällen ist X ein abgeschlossener Unterraum von $H^1(\Omega)$ und damit reflexiv und

$$\begin{aligned}\ell(v)&=\int_\Omega fv+\int_{\Gamma_N}gv\\ B(u,v)&=\int_\Omega\{\nabla u\cdot A\nabla v+(\mathbf{a}\cdot\nabla u)v+\alpha uv\}\end{aligned}$$

mit der offensichtlichen Modifikation für $\Gamma_N=\emptyset$. Die Linearität von ℓ und die Bilinearität von B sind offensichtlich. Aus der Cauchy-Schwarzschen Ungleichung und dem Spursatz, Satz I.2.12 (S. 25), folgt

$$\begin{aligned}|\ell(v)|&\leq\|f\|\,\|v\|+\|g\|_{\Gamma_N}\|v\|_{\Gamma_N}\\&\leq\left\{\|f\|+c\,\|g\|_{\Gamma_N}\right\}\|v\|_1\\|B(u,v)|&\leq\|A\|_\infty|u|_1|v|_1+\|\mathbf{a}\|_\infty|u|_1\|v\|+\|\alpha\|_\infty\|u\|\,\|v\|\\&\leq\max\left\{\|A\|_\infty,\|\mathbf{a}\|_\infty,\|\alpha\|_\infty\right\}\|u\|_1\|v\|_1\end{aligned}$$

und damit die Stetigkeit von ℓ und B. Daher müssen wir nur noch die Koerzivität von B nachweisen. Partielle Integration des Konvektionstermes in $B(u,u)$ ergibt für alle $u\in H^1(\Omega)$

$$\int_\Omega(\mathbf{a}\cdot\nabla u)u=\int_\Omega\frac{1}{2}\mathbf{a}\cdot\nabla(u^2)=-\int_\Omega\frac{1}{2}(\operatorname{div}\mathbf{a})u^2+\int_\Gamma\frac{1}{2}\mathbf{n}\cdot\mathbf{a}u^2$$

und damit

$$\begin{aligned}&B(u,u)\\&=\int_\Omega\left\{\nabla u\cdot A\nabla u+\left(\alpha-\frac{1}{2}\operatorname{div}\mathbf{a}\right)u^2\right\}+\int_\Gamma\frac{1}{2}\mathbf{n}\cdot\mathbf{a}u^2\\&\geq\lambda_0|u|_1^2+\inf_{x\in\Omega}\left\{\alpha(x)-\frac{1}{2}\operatorname{div}\mathbf{a}(x)\right\}\|u\|^2+\int_\Gamma\frac{1}{2}\mathbf{n}\cdot\mathbf{a}u^2.\end{aligned}\tag{I.3.4}$$

Daher müssen wir in allen vier Fällen nur noch die rechte Seite von (I.3.4) geeignet abschätzen.

ad (1): In diesem Fall ist $X = H_0^1(\Omega)$. Damit folgt aus (I.3.4) für alle $u \in X$

$$B(u,u) \geq \lambda_0 \, |u|_1^2 \,. \tag{I.3.5}$$

Zusammen mit der Friedrichschen Ungleichung, Satz I.2.15 (S. 26), beweist dies die Koerzivität von B.
ad (2): Jetzt ist $X = H_D^1(\Omega)$. Wegen $\mathbf{a} \cdot \mathbf{n} \geq 0$ auf Γ_N folgt aus (I.3.4) wieder (I.3.5) und wegen der Friedrichschen Ungleichung die Koerzivität von B.
ad (3): Nun ist $X = H^1(\Omega)$. Wegen $\alpha \geq \alpha_0 > 0$ und $\mathbf{a} \cdot \mathbf{n} \geq 0$ auf Γ folgt aus (I.3.4)

$$B(u,u) \geq \lambda_0 \, |u|_1^2 + \alpha_0 \, \|u\|^2 \geq \min\{\lambda_0, \alpha_0\} \, \|u\|_1^2$$

und damit die Koerzivität von B.
ad (4): Alle Größen sind wie in (3). Wegen $\alpha = 0$, $\operatorname{div} \mathbf{a} = 0$ und $\mathbf{n} \cdot \mathbf{a} = 0$ auf Γ folgt aus (I.3.4) wieder (I.3.5). Hieraus und aus der Poincaréschen Ungleichung, Satz I.2.21 (S. 28), folgt die Koerzivität von B auf dem abgeschlossenen Unterraum $V = \{\varphi \in H^1(\Omega) : \int_\Omega \varphi = 0\}$ von X. Damit liefert Satz I.1.6 (S. 19) die Existenz einer eindeutigen Lösung $u^* \in V$ von

$$B(u^*, v) = \ell(v) \quad \forall v \in V.$$

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

$$m_v = \frac{1}{|\Omega|} \int_\Omega v \quad \text{und} \quad \overline{v} = v - m_v.$$

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

$$\begin{aligned} B(u^*, v) - \ell(v) &= B(u^*, \overline{v}) - \ell(\overline{v}) + B(u^*, m_v) - \ell(m_v) \\ &= B(u^*, m_v) - \ell(m_v). \end{aligned}$$

Wegen der Kompatibilitätsbedingung an ℓ ist $\ell(m_v) = 0$. Mit dem Gaußschen Integralsatz folgt

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

Also ist u^* eine schwache Lösung der Konvektios-Diffusionsgleichung. □

Bemerkung I.3.4. (1) Im Fall der Reaktions-Diffusionsgleichung, d.h. $\mathbf{a} = 0$, reduzieren sich die Voraussetzung von Satz I.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* $\beta u + \mathbf{n} \cdot A \nabla u = g_R$ auf $\Gamma_R \subset \Gamma$ auf. In diesem Fall muss ℓ durch $\int_{\Gamma_R} g_R v$ und B durch $\int_{\Gamma_R} \beta u v$ ergänzt werden. Die Koerzivität von B bleibt erhalten, wenn entweder $\beta \geq 0$ und $\Gamma_D \neq \emptyset$ oder $\beta \geq \beta_0 > 0$ ist.

Das folgende Beispiel aus [16, §III.1] zeigt, dass wir eine Regularitätsaussage der Form $u \in H^2(\Omega)$ für schwache Lösungen nur unter zusätzlichen Annahmen an den Rand Γ erwarten können.

BEISPIEL I.3.5 (Einspringende Ecken). Sei $0 < \alpha < 2\pi$ und Ω_α das Kreissegment $\Omega_\alpha = \{x \in \mathbb{R}^2 : x = (r\cos\varphi, r\sin\varphi), 0 < r < 1, 0 < \varphi < \alpha\}$. Definiere die Funktion $v : \Omega_\alpha \to \mathbb{R}$ durch $v(x) = r^{\frac{\pi}{\alpha}} \sin(\frac{\pi}{\alpha}\varphi)$ für $x = (r\cos\varphi, r\sin\varphi)$. 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 $\operatorname{supp} w \subset B(0, \frac{2}{3})$ und $w = 1$ auf $\overline{B(0, \frac{1}{3})}$. Definiere $u = wv$ und $f = \Delta[(1-w)v]$. 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 für $k \geq 1$ $u \in C^k(\overline{\Omega}_\alpha)$ genau dann, wenn $0 < \alpha \leq \frac{\pi}{k}$ ist, und für $k \geq 2$ $D^k u \in L^2(\Omega_\alpha)$ genau dann, wenn $0 < \alpha < \frac{\pi}{k-1}$ ist. Wir können also bei gegebenem α i.a. keine Abschätzung der Form $\|u\|_{C^{k+2}(\overline{\Omega}_\alpha)} \leq c_k \|f\|_{C^k(\overline{\Omega}_\alpha)}$ oder $\|u\|_{k+1;\Omega_\alpha} \leq c_k' \|f\|_{k;\Omega_\alpha}$ erwarten, wie sie für gewöhnliche Differentialgleichungen gelten würde.

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

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

Die Konstante c hängt nur von Ω und den Koeffizienten α, $\mathbf{a}$ und A ab.

BEWEIS. [5, §9.2], [9, 11] □

KAPITEL II

Theoretische Aspekte

In diesem Kapitel führen wir die Finite Element Räume ein, beweisen ihre Approximationseigenschaften und leiten daraus a priori Fehlerabschätzungen für die Finite Element Diskretisierung elliptischer Differentialgleichungen zweiter Ordnung ab. Dabei ist stets $\Omega \subset \mathbb{R}^d$ mit $d \in \{2, 3\}$ ein offenes, beschränktes, zusammenhängendes Polyedergebiet, d.h., der Rand Γ von Ω besteht stückweise aus Hyperebenen.

II.1. Finite Element Räume

Die Finite Element Methode basiert auf folgender Grundidee:

- Unterteile Ω in Teilgebiete $K_1, \ldots, K_{m_\mathcal{T}}$ mit einfacher geometrischer Struktur.
- Approximiere die Sobolev-Räume $W^{k,p}(\Omega)$ durch endlich dimensionale Räume $X_\mathcal{T}$, so dass jedes $v \in X_\mathcal{T}$ eingeschränkt auf ein beliebiges Element K_i eine einfache Struktur hat.
- Konstruiere eine Basis von $X_\mathcal{T}$, so dass jede Basisfunktion eine einfache Struktur und einen kleinen Träger hat.

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

- den *Referenz d-Simplex* $\widehat{K} = \widehat{K}_S = \{\widehat{x} \in \mathbb{R}^d : x_1 + \ldots + x_d \leq 1, x_i \geq 0, 1 \leq i \leq d\}$ und
- den *Referenz d-Würfel* $\widehat{K} = \widehat{K}_W = [0, 1]^d$.

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

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

- *Affine Äquivalenz*: Zu jedem $K \in \mathcal{T}$ gibt es einen affinen Diffeomorphismus F_K des Referenz-Simplexes oder -Würfels $\widehat{K}$ auf K.
- *Zulässigkeit*: Je zwei Elemente $K_1, K_2 \in \mathcal{T}$ sind entweder disjunkt oder haben einen Eckpunkt, oder eine Kante oder, falls $d = 3$ ist, eine Seitenfläche gemeinsam (vgl. Abbildung II.1.1).
- *Regularität*: Der Quotient $\frac{h_K}{\rho_K}$ ist durch eine Konstante $C_{\mathcal{T}}$, die nicht von K oder h abhängt, nach oben beschränkt.

Dabei ist h_K der Durchmesser von K und ρ_K der Durchmesser des größten, in K eingeschriebenen d-Balles.

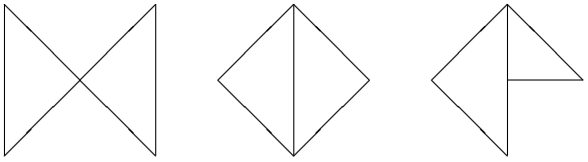

Abbildung II.1.1. Zulässige Elementpaare (links und Mitte) und nicht zulässiges Elementpaar (rechts)

Wir bezeichnen mit $\mathcal{N}$ die Menge aller Eckpunkte und mit $\mathcal{E}$ die Menge aller Kanten ($d = 2$) bzw. Seitenflächen ($d = 3$) aller $K \in \mathcal{T}$. $\mathcal{N}_\Omega$ und $\mathcal{E}_\Omega$ sind die Mengen aller Eckpunkte bzw. Kanten oder Seitenflächen im Innern von Ω. Für ein Element $K \in \mathcal{T}$ bezeichnen schließlich $\mathcal{N}_K$ und $\mathcal{E}_K$ die Menge aller Eckpunkte von K und die Menge aller Kanten bzw. Seitenflächen von K.

Definition II.1.1 ($\widehat{\Sigma}_k$, Σ_k, $\mathcal{G}$, $\widehat{R}_k$, R_k). (1) Bezeichne die Eckpunkte des Referenz-Simplexes mit $\widehat{z}_1 = e_1, \ldots, \widehat{z}_d = e_d$ und $\widehat{z}_{d+1} = 0$. Dann ist für $k \in \mathbb{N}^*$

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

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

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

falls $\widehat{K}$ der Referenz-Würfel ist.
(2) Für $K = F_K(\widehat{K}) \in \mathcal{T}$ und $k \in \mathbb{N}^*$ ist $\Sigma_k = \Sigma_k(K) = F_K(\widehat{\Sigma}_k)$.
(3) Setze $\mathcal{G} = \bigcup_{K \in \mathcal{T}} \Sigma_k(K)$ und $\mathcal{G}_\Omega = \mathcal{G} \cap \Omega$.
(4) Setze $Q_0 = \mathbb{P}_0 = \mathbb{R}$ und definiere für $k \in \mathbb{N}^*$

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

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

mit $x^\alpha = x_1^{\alpha_1} \cdot \ldots \cdot x_d^{\alpha_d}$. Setze

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

und

$$R_k = R_k(K) = \left\{ \widehat{p} \circ F_K^{-1} : \widehat{p} \in \widehat{R}_k \right\}.$$

BEMERKUNG II.1.2. (1) Sei K ein allgemeiner Simplex. Bezeichne die Eckpunkte von K mit $z_1, \ldots, z_{d+1}$. Dann ist

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

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.1.2).
(2) Die Menge $\mathcal{G}$ heißt auch *Gitter*. Die Punkte von $\mathcal{G}$ werden 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 $C(\widehat{K})$ bzw. von $C(\widehat{K})$ auf $C(K)$, die für jedes $k \in \mathbb{N}$ den Polynomraum R_k in $\widehat{R}_k$ bzw. den Polynomraum $\widehat{R}_k$ in R_k abbilden.

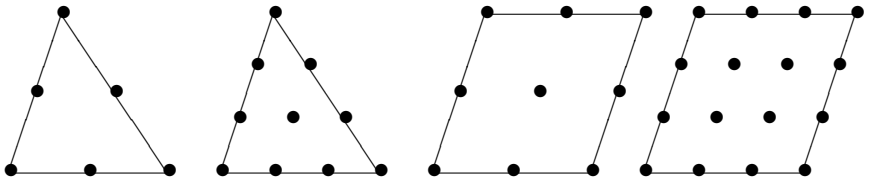

ABBILDUNG II.1.2. Die Mengen Σ_2 und Σ_3 für ein Dreieck und ein Parallelogramm

SATZ II.1.3 (Unisolvenz von $\widehat{\Sigma}_k$ und Σ_k). *Sei $k \in \mathbb{N}^*$. Dann ist jedes $\widehat{p} \in \widehat{R}_k$ und jedes $p \in R_k(K)$ mit $K \in \mathcal{T}$ eindeutig bestimmt durch seine Werte auf $\widehat{\Sigma}_k$ bzw. auf $\Sigma_k(K)$.*

BEWEIS. Wegen der Definition von $\Sigma_k(K)$ und Bemerkung II.1.2 (3) reicht es, die Behauptung für $\widehat{\Sigma}_k$ zu zeigen. Eine leichte Rechnung zeigt, dass $\dim \widehat{R}_k = \#\widehat{\Sigma}_k$ ist. Daher reicht es, eine der folgenden Aussagen (a) oder (b) zu zeigen:

(a) Zu jedem Vektor $(b_z)_{z \in \widehat{\Sigma}_k}$ existiert ein $\varphi \in \widehat{R}_k$ mit $\varphi(z) = b_z$ für alle $z \in \widehat{\Sigma}_k$.
(b) Ist $\varphi \in \widehat{R}_k$ und $\varphi(z) = 0$ für alle $z \in \widehat{\Sigma}_k$, so ist $\varphi = 0$.

Fall 1: $\widehat{K}$ ist der Referenz-Würfel: Definiere für $\mathbf{i} \in \mathbb{N}_k^d$ die Funktion $\mu_{\mathbf{i}}$ durch

$$\mu_{\mathbf{i}}(x) = \prod_{j=1}^{d} \prod_{\substack{\ell_j=0 \\ \ell_j \neq \mathbf{i}_j}}^{k} \frac{kx_j - \ell_j}{\mathbf{i}_j - \ell_j}.$$

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

$$\varphi(x) = \sum_{\mathbf{i}\in\mathbb{N}_k^d} b_{\mathbf{i}}\mu_{\mathbf{i}}(x)$$

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

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

Die Funktionen $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{d+1}$ heißen die *Schwerpunktskoordinaten* von $\widehat{K}$ und haben offensichtlich die Eigenschaft $\widehat{\lambda}_i \in \mathbb{P}_1$ für $1 \le i \le d+1$ und $\widehat{\lambda}_i(\widehat{z}_j) = \delta_{ij}$ für $1 \le i, j \le d+1$.
Fall $k = 1$: Zu gegebenem $(b_i)_{1\le i\le d+1}$ leistet offensichtlich

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

das in Eigenschaft (a) Geforderte.
Fall $k = 2$: Bezeichne mit $\widehat{z}_{ij} = \frac{1}{2}(\widehat{z}_i + \widehat{z}_j)$, $1 \le i < j \le d+1$, die Kantenmittelpunkte von $\widehat{K}$. Definiere

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

Dann gilt offensichtlich $\mu_k, \mu_{ij} \in \mathbb{P}_2$ und für alle $\forall i, j, k, \ell, m$

$$\begin{aligned} \mu_i(\widehat{z}_j) &= \delta_{ij}, & \mu_i(\widehat{z}_{k\ell}) &= 0, \\ \mu_{ij}(\widehat{z}_{k\ell}) &= \delta_{ik}\delta_{j\ell}, & \mu_{ij}(\widehat{z}_m) &= 0. \end{aligned}$$

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

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

das in Eigenschaft (a) Geforderte.
Fall $k = 3$: Definiere

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

und

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

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

$$\mu_i(\widehat{z}_j) = \delta_{ij}, \qquad \mu_i(\widehat{z}_{jjk}) = 0, \qquad \mu_i(\widehat{z}_{jk\ell}) = 0,$$
$$\mu_{iij}(\widehat{z}_{kk\ell}) = \delta_{ik}\delta_{j\ell}, \qquad \mu_{iij}(\widehat{z}_k) = 0, \quad \mu_{iij}(\widehat{z}_{k\ell m}) = 0,$$
$$\mu_{ijk}(\widehat{z}_{\ell mn}) = \delta_{i\ell}\delta_{jm}\delta_{kn}, \quad \mu_{ijk}(\widehat{z}_\ell) = 0, \quad \mu_{ijk}(\widehat{z}_{\ell\ell m}) = 0.$$

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

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

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

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

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

Satz II.1.4 (Stetigkeit stückweiser Polynome). *Seien $k \in \mathbb{N}^*$, K_1, $K_2 \in \mathcal{T}$ mit $K_1 \cap K_2 \ne \emptyset$ und $p_1 \in R_k(K_1)$, $p_2 \in R_k(K_2)$. Definiere die Funktion φ auf $K_1 \cup K_2$ stückweise durch $\varphi|_{K_i} = p_i$ für $i = 1, 2$. Dann ist $\varphi \in C(K_1 \cup K_2)$ genau dann, wenn für alle $x \in F_{K_1}(\widehat{\Sigma}_k) \cap F_{K_2}(\widehat{\Sigma}_k)$ gilt $p_1(x) = p_2(x)$.*

Beweis. „$\Longrightarrow$“: Ist offensichtlich.
„$\Longleftarrow$“: 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 $\widehat{E} = \widehat{K} \cap \{x_2 = 0\}$. Dann ist $\widehat{E}$ das Einheitsintervall, d.h. der Standard 1-Simplex. Die Transformationen F_{K_1} und F_{K_2} können so gewählt werden, dass $E = F_{K_1}(\widehat{E}) = F_{K_2}(\widehat{E})$ ist. Definiere $q = p_1 \circ F_{K_1}|_{\widehat{E}} - p_2 \circ F_{K_2}|_{\widehat{E}}$. Wegen Bemerkung II.1.2 (3)

ist $q \in R_k$ ein Polynom einer Veränderlichen. Nach Voraussetzung verschwindet q in allen Punkten von $\widehat{\Sigma}_k \cap \widehat{E}$. Also ist $q = 0$ und damit φ stetig.
Betrachte nun den Fall, dass $d = 3$ ist und K_1 und K_2 eine Seitenfläche gemeinsam haben. Setze $\widehat{E} = \widehat{K} \cap \{x_3 = 0\}$ und definiere q wie oben. Dann ist $q \in R_k$ ein Polynom in zwei Veränderlichen, das in den Punkten von $\widehat{\Sigma}_k \cap \widehat{E}$ verschwindet. Da $\widehat{E}$ der Standard 2-Simplex bzw. Standard 2-Würfel ist, folgt aus Satz II.1.3, dass $q = 0$ und damit φ stetig ist.
Ist schließlich $d = 3$ und haben K_1 und K_2 eine Kante gemeinsam, setzen wir $\widehat{E} = \widehat{K} \cap \{x_3 = x_2 = 0\}$ und gehen ansonsten wie im Fall $d = 2$ vor. □

Wegen Satz II.1.4 ist folgende Definition sinnvoll.

Definition II.1.5 (Finite Element Räume). Die Finite Element Räume zu $\mathcal{T}$ sind definiert durch

$$\begin{aligned} S^{k,-1}(\mathcal{T}) &= \{\varphi \in L^1(\Omega) : \varphi|_K \in R_k(K)\ \forall K \in \mathcal{T}\},\\ S^{k,0}(\mathcal{T}) &= S^{k,-1}(\mathcal{T}) \cap C(\overline{\Omega}),\\ S_0^{k,0}(\mathcal{T}) &= \{\varphi \in S^{k,0}(\mathcal{T}) : \varphi = 0 \text{ auf } \Gamma\},\\ S_D^{k,0}(\mathcal{T}) &= \{\varphi \in S^{k,0}(\mathcal{T}) : \varphi = 0 \text{ auf } \Gamma_D\} \end{aligned}$$

Bemerkung II.1.6 (Eigenschaften der Finite Element Räume). (1) Wegen Satz I.2.7 (S. 24) ist $S^{k,0}(\mathcal{T}) \subset W^{1,p}(\Omega)$ und $S_0^{k,0}(\mathcal{T}) \subset W_0^{1,p}(\Omega)$ für jedes $1 \le p < \infty$.
(2) Aus den Sätzen II.1.3 und II.1.4 folgt, dass die Funktionen in $S^{k,0}(\mathcal{T})$ und $S_0^{k,0}(\mathcal{T})$ eindeutig bestimmt sind durch ihre Werte in den Gitterpunkten $\mathcal{G}$ bzw. $\mathcal{G}_\Omega$. Dabei ist die Zulässigkeit von $\mathcal{T}$ wesentlich. Insbesondere gibt es zu jedem $z \in \mathcal{G}$ eine eindeutige Funktion $v_z \in S^{k,0}(\mathcal{T})$ mit $v_z(z) = 1$ und $v_z(z') = 0$ für alle $z' \in \mathcal{G} \setminus \{z\}$. Die Funktionen v_z, $z \in \mathcal{G}$, heißen *nodale Basis* von $S^{k,0}(\mathcal{T})$. Jedes $u_\mathcal{T} \in S^{k,0}(\mathcal{T})$ lässt sich eindeutig darstellen als $u_\mathcal{T} = \sum_{z\in\mathcal{G}} \mu_z v_z$, und die Koeffizienten μ_z sind die Werte $u_\mathcal{T}(z)$.
(3) Wenn $\mathcal{T}$ wie im rechten Bild von Abbildung II.1.1 zkizziert nicht zulässig ist, bleibt Definition II.1.5 gültig, aber die Konstruktion einer nodalen Basis muss abgändert werden und ist dann wesentlich aufwändiger.

II.2. Approximationseigenschaften

Motiviert durch das Céa-Lemma, Satz I.1.2 (S. 17), und den Satz von Aubin-Nitsche, Satz I.1.5 (S. 18), wollen wir die Approximationsfehler $\inf_{v_\mathcal{T}\in S^{k,0}(\mathcal{T})} \|v - v_\mathcal{T}\|_\ell$ und $\inf_{w_\mathcal{T}\in S_0^{k,0}(\mathcal{T})} \|w - w_\mathcal{T}\|_\ell$ für $\ell \in \{0, 1\}$ und $v \in H^{k+1}(\Omega)$ bzw. $w \in H^{k+1}(\Omega) \cap H_0^1(\Omega)$ abschätzen.

Dazu definieren wir einen *Interpolationsoperator* $I_{\mathcal{T}} : H^2(\Omega) \to S^{k,0}(\mathcal{T})$ durch

$$(I_{\mathcal{T}}u)(z) = u(z) \quad \forall z \in \mathcal{G} \tag{II.2.1}$$

oder äquivalent

$$I_{\mathcal{T}}u = \sum_{z\in\mathcal{G}} u(z)v_z.$$

Man beachte, dass gemäß Bemerkung I.2.20(5) (S. 27) Funktionen in $H^2(\Omega)$ stetig sind und dass $I_{\mathcal{T}}(H^1_0(\Omega) \cap H^2(\Omega)) \subset S^{k,0}_0(\mathcal{T})$ ist.

Offensichtlich ist

$$\inf_{v_{\mathcal{T}}\in S^{k,0}(\mathcal{T})} \|v - v_{\mathcal{T}}\|_\ell \le \|v - I_{\mathcal{T}}v\|_\ell$$

und

$$\inf_{w_{\mathcal{T}}\in S^{k,0}_0(\mathcal{T})} \|w - w_{\mathcal{T}}\|_\ell \le \|w - I_{\mathcal{T}}w\|_\ell .$$

Wegen der Additivität des Lebesgue Integrals müssen wir für die Abschätzung der rechten Seiten dieser Ungleichungen für jedes Element $K \in \mathcal{T}$ Größen der Form $\|u - I_{\mathcal{T}}u\|_K$ und $|u - I_{\mathcal{T}}u|_{1;K}$ kontrollieren. Dies geschieht durch Transformation auf das Referenzelement $\widehat{K}$ und Abschätzung des Interpolationsfehlers auf $\widehat{K}$.

Dazu definieren wir auf dem Referenzelement einen *lokalen Interpolationsoperator* $\widehat{\Pi}_k : H^2(\widehat{K}) \to R_k$ durch

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

Wegen Satz II.1.3 und Bemerkung I.2.20(5) (S. 27) ist diese Definition sinnvoll. Aus Bemerkung II.1.2(3) (S. 37) folgt außerdem für jedes $u \in H^2(\Omega)$ und jedes $K \in \mathcal{T}$ die Identität (vgl. Abbildung II.2.1)

$$(I_{\mathcal{T}}u)|_K = \left[\widehat{\Pi}_k \left(u|_K \circ F_K\right)\right] \circ F_K^{-1}.$$

$$\begin{array}{ccc} H^2(K) \ni u & \xrightarrow{I_{\mathcal{T}}} & I_{\mathcal{T}}u \in R_k(K) \\ {\scriptstyle \circ F_K}\downarrow & & \uparrow{\scriptstyle \circ F_K^{-1}} \\ H^2(\widehat{K}) \ni u \circ F_K & \xrightarrow{\widehat{\Pi}_K} & \widehat{\Pi}_K(u \circ F_K) \in R_k(\widehat{K}) \end{array}$$

Abbildung II.2.1. Interpolation auf K und $\widehat{K}$

Im Folgenden versehen wir Größen wie $\widehat{K}$, $\widehat{\Sigma}_k$ oder $\widehat{\Pi}_k$ 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.

LEMMA II.2.1 (Normäquivalenz). *Sei $k \in \mathbb{N}^*$. Dann wird durch*

$$|||u|||_{k+1} = |u|_{k+1;\widehat{K}} + \sum_{z \in \widehat{\Sigma}_{k,S}} |u(z)|$$

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

BEWEIS. Da gemäß Satz I.2.19 (S. 27) $H^{k+1}(\widehat{K}) \hookrightarrow C(\widehat{K})$ ist, ist $|||\cdot|||_{k+1}$ wohldefiniert, und es gibt eine Konstante $c_1 \in \mathbb{R}_+^*$ mit

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

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

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

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

$$\|u_n\|_{k+1;\widehat{K}} = 1 \quad \forall n \in \mathbb{N} \tag{II.2.2}$$

und

$$\lim_{n\to\infty} |||u_n|||_{k+1} = 0. \tag{II.2.3}$$

Wegen Satz I.2.19 (S. 27) und Bemerkung I.2.18 (S. 27) gibt es eine Teilfolge $(u_{n_m})_{m\in\mathbb{N}}$ von $(u_n)_{n\in\mathbb{N}}$ und ein $u \in H^k(\widehat{K})$ mit

$$\lim_{m\to\infty} \|u_{n_m} - u\|_{k;\widehat{K}} = 0.$$

Wegen (II.2.3) ist insbesondere

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

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

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

Wegen $|u|_{k+1;\widehat{K}} = 0$ ist $u \in \mathbb{P}_k$. Wegen Satz I.2.19 (S. 27) gilt

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

Hieraus und aus (II.2.3) folgt aber

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

Wegen Satz II.1.3 ist also $u = 0$ im Widerspruch zu (II.2.2). □

Mit Hilfe von Lemma II.2.1 können wir nun den Interpolationsfehler $u - \widehat{\Pi}_k u$ auf dem Referenzelement abschätzen.

SATZ II.2.2 (Interpolationsfehler auf dem Referenzelement). *Zu jedem $k \in \mathbb{N}^*$ gibt es eine Konstante c, die von k abhängt, mit*

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

BEWEIS. *Fall 1:* $\widehat{K} = \widehat{K}_S$*:* Aus Lemma II.2.1 folgt für beliebiges $u \in H^{k+1}(\widehat{K})$

$$\left\|u - \widehat{\Pi}_k u\right\|_{k+1;\widehat{K}} \le c_2 \left|\left|\left|u - \widehat{\Pi}_k u\right|\right|\right|_{k+1} = c_2 \left|u\right|_{k+1;\widehat{K}},$$

da $\widehat{\Pi}_k u \in \mathbb{P}_k$ und $u(z) - \widehat{\Pi}_k u(z) = 0$ ist für alle $z \in \widehat{\Sigma}_{k,S}$.
Fall 2: $\widehat{K} = \widehat{K}_W$*:* Sei $u \in H^{k+1}(\widehat{K})$ beliebig. Da $\widehat{K}_S \subset \widehat{K}_W$ ist, folgt aus der Dreiecksungleichung

$$\left\|u - \widehat{\Pi}_{k,W} u\right\|_{k+1;\widehat{K}} \le \left\|u - \widehat{\Pi}_{k,S} u\right\|_{k+1;\widehat{K}} + \left\|\widehat{\Pi}_{k,S} u - \widehat{\Pi}_{k,W} u\right\|_{k+1;\widehat{K}}.$$

Wegen $\mathbb{P}_k \subset Q_k$ und Satz II.1.3 gilt

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

und somit

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

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

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

Dann folgt für beliebiges $\varphi \in H^{k+1}(\widehat{K})$ mit Satz I.2.19 (S. 27)

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

Aus dem soeben Gezeigten und dem Fall 1 folgt

$$\begin{aligned}
\left\|u - \widehat{\Pi}_{k,W} u\right\|_{k+1;\widehat{K}} &\le \left\|u - \widehat{\Pi}_{k,S} u\right\|_{k+1;\widehat{K}} + \left\|\widehat{\Pi}_{k,W}(\widehat{\Pi}_{k,S} u - u)\right\|_{k+1;\widehat{K}} \\
&\le (1 + c') \left\|u - \widehat{\Pi}_{k,S} u\right\|_{k+1;\widehat{K}} \\
&\le (1 + c') c_2 \left|u\right|_{k+1;\widehat{K}}.
\end{aligned}$$

□

Da der Beweis von Lemma II.2.1 nicht konstruktiv ist, haben wir keine explizite Information über den Wert der Konstanten c in Satz II.2.2. Für den Spezialfall $k = 1$ geben die folgenden beiden Beispiele eine derartige Information.

Beispiel II.2.3 (Bilineare Elemente). Sei $\widehat{K} = [0,1]^2$ das Einheitsquadrat. Dann gilt für alle $u \in H^2(\widehat{K})$

$$\left|u - \widehat{\Pi}_1 u\right|_{1;\widehat{K}} \leq \frac{1}{\sqrt{3}} |u|_{2;\widehat{K}} .$$

Zum Beweis beachten wir, dass wir wegen Satz I.2.6(2) (S. 23) $u \in C^\infty(K)$ annehmen können. Bezeichne mit $\iota : C([0,1],\mathbb{R}) \to \mathbb{P}_1$ den linearen Interpolationsoperator in den Punkten 0 und 1. Mit dem Hauptsatz der Differential- und Integralrechnung und partieller Integration erhalten wir für alle $\varphi \in C^2([0,1],\mathbb{R})$ und alle $t \in \mathbb{R}$

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

und

$$\begin{aligned}\varphi(t) &= \varphi(1) - \int_t^1 \varphi'(s)\mathrm{d}s \\ &= \varphi(1) - (1-t)\varphi'(t) - \int_t^1 (1-s)\varphi''(s)\mathrm{d}s.\end{aligned}$$

Multiplikation der ersten Gleichung mit $1-t$ und der zweiten Gleichung mit t und anschließende Addition der resultierenden Gleichungen liefert

$$\begin{aligned}\varphi(t) &= \iota\varphi(t) + (1-t)\int_0^t \varphi'(s)\mathrm{d}s - t\int_t^1 \varphi'(s)\mathrm{d}s \\ &= \iota\varphi(t) - (1-t)\int_0^t s\varphi''(s)\mathrm{d}s - t\int_t^h (1-s)\varphi''(s)\mathrm{d}s.\end{aligned} \tag{II.2.4}$$

Anwenden der Gleichung (II.2.4) auf die Variable x liefert für alle $(x,y) \in \widehat{K}$

$$\begin{aligned}u(x,y) = (\iota u(\cdot,y))(x) &\\ - (1-x)\int_0^x s\frac{\partial^2 u}{\partial x^2}(s,y)\mathrm{d}s &\\ - x\int_x^1 (1-s)\frac{\partial^2 u}{\partial x^2}(s,y)\mathrm{d}s.&\end{aligned}$$

Wenden wir Gleichung (II.2.4) für festes x auf die Variable y und $\varphi(y) = (\iota u(\cdot, y))(x)$ an, erhalten wir weiter

$$\begin{aligned}
&(\iota u(\cdot, y))(x)\\
&\quad = (1-x)\left\{(1-y)u(0,0) + yu(0,1) + (1-y)\int_0^y \frac{\partial u}{\partial y}(0,t)\mathrm{d}t \right.\\
&\qquad\qquad \left. -y\int_y^1 \frac{\partial u}{\partial y}(0,t)\mathrm{d}t\right\}\\
&\qquad + x\left\{(1-y)u(1,0) + yu(1,1) + (1-y)\int_0^y \frac{\partial u}{\partial y}(1,t)\mathrm{d}t \right.\\
&\qquad\qquad \left. -y\int_y^1 \frac{\partial u}{\partial y}(1,t)\mathrm{d}t\right\}\\
&\quad = \widehat{\Pi}_1 u(x,y) + (1-y)\int_0^y \frac{\partial u}{\partial y}(0,t)\mathrm{d}t - y\int_y^1 \frac{\partial u}{\partial y}(0,t)\mathrm{d}t\\
&\qquad + x(1-y)\int_0^y \left\{\frac{\partial u}{\partial y}(1,t) - \frac{\partial u}{\partial y}(0,t)\right\}\mathrm{d}t\\
&\qquad - xy\int_y^1 \left\{\frac{\partial u}{\partial y}(1,t) - \frac{\partial u}{\partial y}(0,t)\right\}\mathrm{d}t.
\end{aligned}$$

Da für alle $t \in [0,1]$

$$\frac{\partial u}{\partial y}(1,t) - \frac{\partial u}{\partial y}(0,t) = \int_0^1 \frac{\partial^2 u}{\partial x \partial y}(s,t)\mathrm{d}s$$

ist, erhalten wir insgesamt die Darstellung

$$\begin{aligned}
u(x,y) = \widehat{\Pi}_1 u(x,y) &+ (1-y)\int_0^y \frac{\partial u}{\partial y}(0,t)\mathrm{d}t - y\int_y^1 \frac{\partial u}{\partial y}(0,t)\mathrm{d}t\\
&+ x(1-y)\int_0^y\int_0^1 \frac{\partial^2 u}{\partial x\partial y}(s,t)\mathrm{d}s\mathrm{d}t\\
&- xy\int_y^1\int_0^1 \frac{\partial^2 u}{\partial x\partial y}(s,t)\mathrm{d}s\mathrm{d}t - (1-x)\int_0^x s\frac{\partial^2 u}{\partial x^2}(s,y)\mathrm{d}s\\
&- x\int_x^1 (1-s)\frac{\partial^2 u}{\partial x^2}(s,y)\mathrm{d}s.
\end{aligned}$$

Differentiation bzgl. x liefert

$$\begin{aligned}
\frac{\partial}{\partial x}(u - \widehat{\Pi}_1 u)(x,y) = &\int_0^1\int_0^1 K_1(t,y)\frac{\partial^2 u}{\partial x\partial y}(s,t)\mathrm{d}s\mathrm{d}t\\
&+ \int_0^1 K_2(s,x)\frac{\partial^2 u}{\partial x^2}(s,y)\mathrm{d}s
\end{aligned}$$

mit

$$K_1(t,y) = \begin{cases} (1-y) & \text{für } 0 \le t < y \\ -y & \text{für } y < t \le 1 \end{cases}$$

$$K_2(s,x) = \begin{cases} s & \text{für } 0 \le s < x \\ -(1-s) & \text{für } x < s \le 1. \end{cases}$$

Quadrieren dieser Identität und Anwenden der Cauchy-Schwarzschen Ungleichung ergibt

$$\begin{aligned} &\left| \frac{\partial}{\partial x}(u - \widehat{\Pi}_1 u)(x,y) \right|^2 \\ &\quad \le 2 \int_0^1 \int_0^1 |K_1(t,y)|^2 \,\mathrm{d}s\mathrm{d}t \int_0^1 \int_0^1 \left| \frac{\partial^2 u}{\partial x \partial y}(s,t) \right|^2 \mathrm{d}s\mathrm{d}t \\ &\qquad + 2 \int_0^1 |K_2(s,x)|^2 \,\mathrm{d}s \int_0^1 \left| \frac{\partial^2 u}{\partial x^2}(s,y) \right|^2 \mathrm{d}s. \end{aligned}$$

Integration über $\widehat{K}$ liefert

$$\begin{aligned} &\int_{\widehat{K}} \left| \frac{\partial}{\partial x}(u - \widehat{\Pi}_1 u \right|^2 \\ &\quad \le 2 \int_0^1 \int_0^1 \int_0^1 \int_0^1 |K_1(t,y)|^2 \,\mathrm{d}s\mathrm{d}t\mathrm{d}x\mathrm{d}y \left\| \frac{\partial^2 u}{\partial x \partial y} \right\|_{\widehat{K}}^2 \\ &\qquad + 2 \int_0^1 \int_0^1 |K_2(s,x)|^2 \,\mathrm{d}s\mathrm{d}x \left\| \frac{\partial^2 u}{\partial x^2} \right\|_{\widehat{K}}^2. \end{aligned}$$

Wegen

$$\begin{aligned} 2 \int_0^1 \int_0^1 \int_0^1 \int_0^1 |K_1(t,y)|^2 \,\mathrm{d}s\mathrm{d}t\mathrm{d}x\mathrm{d}y &= \frac{1}{3} \\ 2 \int_0^1 \int_0^1 |K_2(s,x)|^2 \,\mathrm{d}s\mathrm{d}x &= \frac{1}{3} \end{aligned}$$

erhalten wir insgesamt die Abschätzung

$$\left\| \frac{\partial}{\partial x}(u - \widehat{\Pi}_1 u) \right\|_{\widehat{K}}^2 \le \frac{1}{3} \left\| \frac{\partial^2 u}{\partial x \partial y} \right\|_{\widehat{K}}^2 + \frac{1}{3} \left\| \frac{\partial^2 u}{\partial x^2} \right\|_{\widehat{K}}^2.$$

Vertauschen der Rollen von x und y liefert mit der gleichen Rechnung

$$\left\| \frac{\partial}{\partial y}(u - \widehat{\Pi}_1 u) \right\|_{\widehat{K}}^2 \le \frac{1}{3} \left\| \frac{\partial^2 u}{\partial x \partial y} \right\|_{\widehat{K}}^2 + \frac{1}{3} \left\| \frac{\partial^2 u}{\partial y^2} \right\|_{\widehat{K}}^2.$$

Addition dieser beiden Abschätzungen beweist wegen

$$|v|_{2;\widehat{K}}^2 = \left\| \frac{\partial^2 v}{\partial x^2} \right\|_{\widehat{K}}^2 + 2 \left\| \frac{\partial^2 v}{\partial x \partial y} \right\|_{\widehat{K}}^2 + \left\| \frac{\partial^2 v}{\partial y^2} \right\|_{\widehat{K}}^2$$

die Behauptung.

BEISPIEL II.2.4 (Lineare Elemente). Sei $\widehat{K}$ das Referenzdreieck. Dann gilt für alle $u \in H^2(\widehat{K})$

$$\left|u - \widehat{\Pi}_1 u\right|_{1;\widehat{K}} \leq 2.5\,|u|_{2;\widehat{K}}$$

und

$$\left\|u - \widehat{\Pi}_1 u\right\|_{\widehat{K}} \leq \sqrt{10}\,|u|_{2;\widehat{K}}\,.$$

Zum Beweis können wir wie in Beispiel II.2.3 $u \in C^\infty(\widehat{K})$ annehmen. Da $\widehat{\Pi}_1 u$ linear ist, ist $\nabla(\widehat{\Pi}_1 u)$ konstant und

$$\nabla(\widehat{\Pi}_1 u) = \begin{pmatrix} u(1,0) - u(0,0) \\ u(0,1) - u(0,0) \end{pmatrix}.$$

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

$$\begin{aligned}
\frac{\partial}{\partial x}(u - \widehat{\Pi}_1 u)(x,y) &= \int_0^1 \left\{ \frac{\partial u}{\partial x}(x,y) - \frac{\partial u}{\partial x}(s,0) \right\} \mathrm{d}s \\
&= \int_0^x \left\{ \frac{\partial u}{\partial x}(x,y) - \frac{\partial u}{\partial x}(s,y) \right\} \mathrm{d}s + \int_0^x \left\{ \frac{\partial u}{\partial x}(s,y) - \frac{\partial u}{\partial x}(s,0) \right\} \mathrm{d}s \\
&\quad + \int_x^1 \left\{ \frac{\partial u}{\partial x}(x,y) - \frac{\partial u}{\partial x}(x,1-s) \right\} \mathrm{d}s \\
&\quad + \int_x^1 \left\{ \frac{\partial u}{\partial x}(x,1-s) - \frac{\partial u}{\partial x}(s,1-s) \right\} \mathrm{d}s \\
&\quad + \int_x^1 \left\{ \frac{\partial u}{\partial x}(s,1-s) - \frac{\partial u}{\partial x}(s,0) \right\} \mathrm{d}s \\
&= \int_0^x \int_s^x \frac{\partial^2 u}{\partial x^2}(\sigma,y)\mathrm{d}\sigma\mathrm{d}s + \int_0^x \int_0^y \frac{\partial^2 u}{\partial x \partial y}(s,t)\mathrm{d}t\mathrm{d}s \\
&\quad + \int_x^1 \int_{1-s}^y \frac{\partial^2 u}{\partial x \partial y}(x,t)\mathrm{d}t\mathrm{d}s - \int_x^1 \int_x^s \frac{\partial^2 u}{\partial x^2}(\sigma,1-s)\mathrm{d}\sigma\mathrm{d}s \\
&\quad + \int_x^1 \int_0^{1-s} \frac{\partial^2 u}{\partial x \partial y}(s,t)\mathrm{d}t\mathrm{d}s \\
&= \sum_{i=1}^5 I_i(x,y).
\end{aligned}$$

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

$$\left\| \frac{\partial}{\partial x}(u - \widehat{\Pi}_1 u) \right\|_{\widehat{K}}^2 \leq 5 \sum_{i=1}^5 \int_{\widehat{K}} |I_i|^2\,.$$

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

$$\begin{aligned}
\int_{\widehat{K}} |I_1|^2 &\le \int_{\widehat{K}} \left\{ \int_0^x \int_s^x \mathrm{d}\sigma\mathrm{d}s \right\} \left\{ \int_0^x \int_s^x \left| \frac{\partial^2 u}{\partial x^2}(\sigma, y) \right|^2 \mathrm{d}\sigma\mathrm{d}s \right\} \\
&\le \frac{1}{2} \left\| \frac{\partial^2 u}{\partial x^2} \right\|_{\widehat{K}}^2 \\
\int_{\widehat{K}} |I_2|^2 &\le \int_{\widehat{K}} \left\{ \int_0^x \int_0^y \mathrm{d}t\mathrm{d}s \right\} \left\{ \int_0^x \int_0^y \left| \frac{\partial^2 u}{\partial x \partial y}(s, t) \right|^2 \mathrm{d}t\mathrm{d}s \right\} \\
&\le \frac{1}{2} \left\| \frac{\partial^2 u}{\partial x \partial y} \right\|_{\widehat{K}}^2 \\
\int_{\widehat{K}} |I_3|^2 &\le \int_{\widehat{K}} \left\{ \int_x^1 \left| \int_{1-s}^y \mathrm{d}t \right| \mathrm{d}s \right\} \left\{ \int_x^1 \left| \int_{1-s}^y \left| \frac{\partial^2 u}{\partial x \partial y}(x, t) \right|^2 \mathrm{d}t \right| \mathrm{d}s \right\} \\
&\le \frac{1}{2} \left\| \frac{\partial^2 u}{\partial x \partial y} \right\|_{\widehat{K}}^2 \\
\int_{\widehat{K}} |I_4|^2 &\le \int_{\widehat{K}} \left\{ \int_x^1 \int_x^s d\sigma\mathrm{d}s \right\} \left\{ \int_x^1 \int_x^s \left| \frac{\partial^2 u}{\partial x^2}(\sigma, 1-s) \right|^2 \mathrm{d}\sigma\mathrm{d}s \right\} \\
&\le \frac{1}{4} \left\| \frac{\partial^2 v}{\partial x^2} \right\|_{\widehat{K}}^2 \\
\int_{\widehat{K}} |I_5|^2 &\le \int_{\widehat{K}} \left\{ \int_x^1 \int_0^{1-s} \mathrm{d}t\mathrm{d}s \right\} \left\{ \int_x^1 \int_0^{1-s} \left| \frac{\partial^2 u}{\partial x \partial y}(s, t) \right|^2 \mathrm{d}t\mathrm{d}s \right\} \\
&\le \frac{1}{4} \left\| \frac{\partial^2 u}{\partial x \partial y} \right\|_{\widehat{K}}^2 .
\end{aligned}$$

Insgesamt ergibt sich

$$\left\| \frac{\partial}{\partial x}(u - \widehat{\Pi}_1 u) \right\|_{\widehat{K}}^2 \le \frac{25}{4} \left\{ \left\| \frac{\partial^2 u}{\partial x^2} \right\|_{\widehat{K}}^2 + \left\| \frac{\partial^2 u}{\partial x \partial y} \right\|_{\widehat{K}}^2 \right\}. \tag{II.2.5}$$

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

$$\left\| \frac{\partial}{\partial y}(u - \widehat{\Pi}_1 u) \right\|_{\widehat{K}}^2 \le \frac{25}{4} \left\{ \left\| \frac{\partial^2 u}{\partial y^2} \right\|_{\widehat{K}}^2 + \left\| \frac{\partial^2 u}{\partial x \partial y} \right\|_{\widehat{K}}^2 \right\}.$$

Hieraus folgt die erste Ungleichung.
Zum Beweis der zweiten Ungleichung sei $(x, y) \in \widehat{K}$ beliebig. Wegen

$u(0,0) = (\widehat{\Pi}_1 u)(0,0)$ gilt

$$\begin{aligned}
&(u - \widehat{\Pi}_1 u)(x,y) \\
&\quad = (u - \widehat{\Pi}_1 u)(x,y) - (u - \widehat{\Pi}_1 u)(x,0) \\
&\qquad + (u - \widehat{\Pi}_1 u)(x,0) - (u - \widehat{\Pi}_1 u)(0,0) \\
&\quad = \int_0^y \frac{\partial}{\partial y}(u - \widehat{\Pi}_1 u)(x,t)\mathrm{d}t + \int_0^x \frac{\partial}{\partial x}(u - \widehat{\Pi}_1 u)(s,0)\mathrm{d}s \\
&\quad = \int_0^y \frac{\partial}{\partial y}(u - \widehat{\Pi}_1 u)(x,t)\mathrm{d}t + \int_0^x \frac{\partial}{\partial x}(u - \widehat{\Pi}_1 u)(s,y)\mathrm{d}s \\
&\qquad - \int_0^x \int_0^y \frac{\partial^2}{\partial x \partial y}(u - \widehat{\Pi}_1 u)(s,t)\mathrm{d}t\mathrm{d}s \\
&\quad = \sum_{i=1}^{3} I_i(x,y).
\end{aligned}$$

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

$$\left\| u - \widehat{\Pi}_1 u \right\|_{\widehat{K}}^2 \leq 3 \sum_{i=1}^{3} \int_{\widehat{K}} I_i^2.$$

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

$$\int_{\widehat{K}} I_1^2 \leq \frac{1}{2} \left\| \frac{\partial}{\partial y}(u - \widehat{\Pi}_1 u) \right\|_{\widehat{K}}^2, \quad \int_{\widehat{K}} I_2^2 \leq \frac{1}{2} \left\| \frac{\partial}{\partial x}(u - \widehat{\Pi}_1 u) \right\|_{\widehat{K}}^2.$$

Wegen $\widehat{\Pi}_1 u \in \mathbb{P}_1$ ist $\frac{\partial^2}{\partial x \partial y}(\widehat{\Pi}_1 u) = 0$. Daher gilt für den dritten Summanden

$$\int_{\widehat{K}} I_3^2 \leq \int_{\widehat{K}} xy \int_0^x \int_0^y \left| \frac{\partial^2 u}{\partial x \partial y}(s,t) \right|^2 \mathrm{d}t\mathrm{d}s \leq \frac{1}{24} \left\| \frac{\partial^2 u}{\partial x \partial y} \right\|_{\widehat{K}}^2.$$

Aus diesen Abschätzungen und Ungleichung (II.2.5) folgt insgesamt

$$\begin{aligned}
&\left\| u - \widehat{\Pi}_1 u \right\|_{\widehat{K}}^2 \\
&\quad \leq \frac{3}{2} \left\| \frac{\partial}{\partial x}(u - \widehat{\Pi}_1 u) \right\|_{\widehat{K}}^2 + \frac{3}{2} \left\| \frac{\partial}{\partial y}(u - \widehat{\Pi}_1 u) \right\|_{\widehat{K}}^2 + \frac{1}{8} \left\| \frac{\partial^2 u}{\partial x \partial y} \right\|_{\widehat{K}}^2 \\
&\quad \leq \frac{45}{8} \left\| \frac{\partial^2 u}{\partial x^2} \right\|_{\widehat{K}}^2 + \frac{45}{8} \left\| \frac{\partial^2 u}{\partial y^2} \right\|_{\widehat{K}}^2 + \left(\frac{75}{4} + \frac{1}{8} \right) \left\| \frac{\partial^2 u}{\partial x \partial y} \right\|_{\widehat{K}}^2 \\
&\quad \leq 10 \left| u \right|_{2;\widehat{K}}^2 .
\end{aligned}$$

Damit wir Satz II.2.2 auf ein allgemeines Element $K \in \mathcal{T}$ übertragen können, müssen wir zunächst das Verhalten der Sobolev-Normen unter der affinen Transformation des Referenzelementes auf K und der

inversen Transformation untersuchen und die Operatornormen der Jacobi Matrizen dieser Transformationen abschätzen. Dazu bezeichnet im Folgenden K ein beliebiges Element in $\mathcal{T}$, $F_K : x \mapsto b_K + B_K x$ eine affine Transformation des Referenzelements $\widehat{K}$ auf K, $|\cdot|$ die euklidische Norm in $\mathbb{R}^d$ und $\|\cdot\|_{\mathcal{L}}$ die zugehörige Operatornorm, d.h. die Spektralnorm.

LEMMA II.2.5 (Transformation von Sobolev-Normen). *Für jedes $\ell \in \mathbb{N}$ und jedes $u \in H^\ell(K)$ gilt*

$$|u \circ F_K|_{\ell;\widehat{K}} \leq |\det B_K|^{-\frac{1}{2}} \|B_K\|_{\mathcal{L}}^{\ell} |u|_{\ell;K} ,$$
$$|u|_{\ell;K} \leq |\det B_K|^{\frac{1}{2}} \left\|B_K^{-1}\right\|_{\mathcal{L}}^{\ell} |u \circ F_K|_{\ell;\widehat{K}} .$$

BEWEIS. Für $\ell = 0$ sind die beiden Ungleichungen Gleichungen und folgen aus dem Transformationssatz für Integrale. Für $\ell \geq 1$ folgt aus der Transformationsformel für Ableitungen

$$\left(D^\ell(u \circ F_K)\right)(z_1, \ldots, z_\ell) = \left(\left(D^\ell u\right) \circ F_K\right)(B_K z_1, \ldots, B_K z_\ell)$$

und der Definition der entsprechenden Operatornormen

$$\left\|D^\ell(u \circ F_K)\right\|_{\mathcal{L}^\ell} \leq \|B_K\|_{\mathcal{L}}^{\ell} \left\|(D^\ell u) \circ F_K\right\|_{\mathcal{L}^\ell} .$$

Mit dem Transformationssatz für Integrale liefert dies

$$\begin{aligned} |u \circ F_K|_{\ell;\widehat{K}}^2 &= \int_{\widehat{K}} \left\|D^\ell(u \circ F_K)\right\|_{\mathcal{L}^\ell}^2 \\ &\leq \int_{\widehat{K}} \|B_K\|_{\mathcal{L}}^{2\ell} \left\|(D^\ell u) \circ F_K)\right\|_{\mathcal{L}^\ell}^2 \\ &= \|B_K\|_{\mathcal{L}}^{2\ell} |\det B_K|^{-1} \int_K \left\|D^\ell u\right\|_{\mathcal{L}^\ell}^2 \\ &= \|B_K\|_{\mathcal{L}}^{2\ell} |\det B_K|^{-1} |u|_{\ell;\widehat{K}}^2 . \end{aligned}$$

Dies beweist die erste Abschätzung. Die zweite Abschätzung folgt aus der ersten durch Vertauschen der Rollen von F_K und F_K^{-1} und Ausnutzen von $D\left(F_K^{-1}\right) = B_K^{-1}$. □

LEMMA II.2.6 (Normen von Transformationsmatrizen). *Es ist*

$$\|B_K\|_{\mathcal{L}} \leq \frac{h_K}{\rho_{\widehat{K}}} \quad \textit{und} \quad \left\|B_K^{-1}\right\|_{\mathcal{L}} \leq \frac{h_{\widehat{K}}}{\rho_K}.$$

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

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

Also ist

$$\|B_K\|_{\mathcal{L}} = \frac{1}{\rho_{\widehat{K}}} \sup_{\widehat{z} \in \mathbb{R}^d;\ |\widehat{z}| = \rho_{\widehat{K}}} |B_K \widehat{z}| \leq \frac{h_K}{\rho_{\widehat{K}}}.$$

Vertauschen der Rollen von K und $\widehat{K}$ beweist die Abschätzung für $\left\|B_K^{-1}\right\|_{\mathcal{L}}$. □

Nach diesen Vorbereitungen können wir jetzt den Interpolationsfehler $u - I_{\mathcal{T}} u$ auf einem allgemeinen Element K abschätzen.

Satz II.2.7 (Interpolationsfehler auf einem allgemeinen Element). *Für alle $k \in \mathbb{N}^*$, $0 \leq \ell \leq k+1$, $K \in \mathcal{T}$ und $u \in H^{k+1}(K)$ gelten die Interpolationsfehlerabschätzungen*

$$|u - I_{\mathcal{T}} u|_{\ell;K} \leq c h_K^{k+1-\ell} |u|_{k+1;K} .$$

Die Konstante c hängt nur von k, ℓ und $\mathrm{C}_{\mathcal{T}} = \max_K \frac{h_K}{\rho_K}$ ab.

Beweis. Aus Lemma II.2.5 folgt

$$|u - I_{\mathcal{T}} u|_{\ell;K} \leq |\det B_K|^{\frac{1}{2}} \left\|B_K^{-1}\right\|_{\mathcal{L}}^{\ell} |(u - I_{\mathcal{T}} u) \circ F_K|_{\ell;\widehat{K}} ,$$
$$|u \circ F_K|_{k+1;\widehat{K}} \leq |\det B_K|^{-\frac{1}{2}} \|B_K\|_{\mathcal{L}}^{k+1} |u|_{k+1;\widehat{K}} .$$

Aus diesen Abschätzungen, Lemma II.2.6 und Satz II.2.2 ergibt sich wegen $I_{\mathcal{T}} u \circ F_K = \widehat{\Pi}_k (u \circ F_K)$

$$\begin{aligned} |u - I_{\mathcal{T}} u|_{\ell;K} &\leq c \left(\frac{h_{\widehat{K}}}{\rho_K}\right)^{\ell} \left(\frac{h_K}{\rho_{\widehat{K}}}\right)^{k+1} |u|_{k+1;K} \\ &= c \frac{h_{\widehat{K}}^{\ell}}{\rho_{\widehat{K}}^{k+1}} \left(\frac{h_K}{\rho_K}\right)^{\ell} h_K^{k+1-\ell} |u|_{k+1;K} . \end{aligned}$$ □

Für $\ell \in \{0, 1\}$ folgen aus Satz II.2.7 durch Quadrieren und Summieren über alle Elemente die folgenden globalen Interpolationsfehlerabschätzungen. Man beachte, dass für $\ell \geq 2$ analoge Abschätzungen für $|u - I_{\mathcal{T}} u|_{\ell}$ nicht gelten können, da $I_{\mathcal{T}} u$ nicht in $H^2(\Omega)$ ist.

Satz II.2.8 (Interpolationsfehlerabschätzung). *Für alle $k \in \mathbb{N}^*$ und $u \in H^{k+1}(\Omega)$ gelten die globalen Interpolationsfehlerabschätzungen*

$$\|u - I_{\mathcal{T}} u\| \leq c_1 h^{k+1} |u|_{k+1} , \quad |u - I_{\mathcal{T}} u|_1 \leq c_2 h^k |u|_{k+1} .$$

Dabei ist $h = \max_{K \in \mathcal{T}} h_K$. Die Konstanten c_1 und c_2 hängen von k und $\mathrm{C}_{\mathcal{T}} = \max_K \frac{h_K}{\rho_K}$ ab.

Der folgende Satz erlaubt es uns, für Finite Element Funktionen elementweise verschiedene Sobolev-Normen gegeneinander abzuschätzen. Das nachfolgende Beispiel zeigt, dass ein derartiges Ergebnis nicht für beliebige Sobolev-Funktionen gelten kann.

Satz II.2.9 (Inverse Abschätzungen). *Für alle $k \in \mathbb{N}^*$, $m \geq 2$, $0 \leq \ell \leq m-1$, $u_{\mathcal{T}} \in S^{k,0}(\mathcal{T})$ und $K \in \mathcal{T}$ gelten die inversen Abschätzungen*

$$|u_{\mathcal{T}}|_{m;K} \leq c \frac{h_K^{\ell} h_{\widehat{K}}^{m}}{\rho_K^{m} \rho_{\widehat{K}}^{\ell}} |u_{\mathcal{T}}|_{\ell;K} .$$

Die Konstante c hängt nur von k, m und ℓ ab.

BEWEIS. Für $\ell = 0$ ist $|\cdot|_{\ell;\widehat{K}}$ eine Norm auf dem endlich dimensionalen Raum $R_k(\widehat{K})$ und für $\ell \geq 1$ eine Semi-Norm, die auf $\mathbb{P}_\ell$ verschwindet. Andererseits ist $|\cdot|_{m;\widehat{K}}$ eine Semi-Norm auf $R_k(\widehat{K})$, die auf $\mathbb{P}_m$ verschwindet. Wegen $\ell < m$ ist insbesondere $|\widehat{v}|_{m;\widehat{K}} = 0$ für alle $\widehat{v} \in R_k(\widehat{K})$, für die $|\widehat{v}|_{\ell;\widehat{K}} = 0$ ist. Daher gibt es eine nur von k, ℓ und m abhängige Konstante c mit

$$|\widehat{v}|_{m;\widehat{K}} \leq c\,|\widehat{v}|_{\ell;\widehat{K}} \quad \forall \widehat{v} \in R_k(\widehat{K}).$$

Transformation von $|u_\mathcal{T}|_{m;K}$ auf $\widehat{K}$, Ausnutzen dieser Abschätzung für $\widehat{v} = u_\mathcal{T} \circ F_K$ und Rücktransformation von $|u_\mathcal{T} \circ F_K|_{\ell;\widehat{K}}$ auf K liefert wegen Lemma II.2.5

$$\begin{aligned} |u_\mathcal{T}|_{m;K} &\leq |\det B_K|^{\frac{1}{2}} \left\|B_K^{-1}\right\|_\mathcal{L}^m |u_\mathcal{T} \circ F_K|_{m;\widehat{K}} \\ &\leq c\,|\det B_K|^{\frac{1}{2}} \left\|B_K^{-1}\right\|_\mathcal{L}^m |u_\mathcal{T} \circ F_K|_{\ell;\widehat{K}} \\ &\leq c \left\|B_K^{-1}\right\|_\mathcal{L}^m \|B_K\|_\mathcal{L}^\ell |u_\mathcal{T}|_{\ell;K}\,. \end{aligned}$$

Hieraus folgt die Behauptung mit Lemma II.2.6. □

BEISPIEL II.2.10 (Gegenbeispiel zu inverser Abschätzung). Sei $K = [0,1]$ und $v_n = \frac{\sqrt{2}}{n\pi}\sin(n\pi x)$, $n \in \mathbb{N}^*$. Dann gilt

$$|v_n|_{1;K} = 1\ \forall n \in \mathbb{N}^* \quad \text{und} \quad \|v_n\|_K \leq \frac{1}{n} \underset{n\to\infty}{\longrightarrow} 0.$$

BEMERKUNG II.2.11 (L^p-Normen). Die Ergebnisse der Lemmata II.2.1 und II.2.5 und der Sätze II.2.2, II.2.7, II.2.8 und II.2.9 gelten mit den offensichtlichen Modifikationen für alle $W^{k+1,p}$-Räume mit $1 \leq p < \infty$.

II.3. A priori Fehlerabschätzungen

Wir betrachten zunächst die Reaktions-Diffusionsgleichung mit homogenen Dirichlet-Randbedingungen

$$\begin{aligned} -\operatorname{div}(A\nabla u) + \alpha u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \Gamma \end{aligned} \tag{II.3.1}$$

mit $f \in L^2(\Omega)$, $\alpha \in C(\Omega, \mathbb{R}_+)$, $A \in C^1(\Omega, \mathbb{R}^{d\times d}_{\text{symm}})$ und

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

Im Rahmen von §I.1 (S. 15) ist gemäß §I.3 (S. 28)

$$X = H_0^1(\Omega), \quad \ell(v) = \int_\Omega fv, \quad B(u,v) = \int_\Omega \{\nabla u \cdot A\nabla v + \alpha uv\}\,.$$

Im Rahmen von Satz I.1.2 (S. 17) setzen wir $X_\mathcal{T} = S_0^{k,0}(\mathcal{T})$ mit $k \in \mathbb{N}^*$, so dass die diskrete Lösung $u_\mathcal{T} \in S_0^{k,0}(\mathcal{T})$ bestimmt ist durch

$$B(u_\mathcal{T}, v_\mathcal{T}) = \ell(v_\mathcal{T}) \quad \forall v_\mathcal{T} \in S_0^{k,0}(\mathcal{T}). \tag{II.3.2}$$

Problem (II.3.2) ist ein lineares Gleichungssystem mit $\#\mathcal{G}_\Omega$ Gleichungen und Unbekannten. Wegen der Symmetrie und Koerzivität der Bilinearform B ist die Matrix des Gleichungssystems, die sog. *Steifigkeitsmatrix*, symmetrisch positiv definit. Bei Verwendung der nodalen Basis aus Bemerkung II.1.6(2) (S. 40) sind die Koeffizienten des Lösungsvektors die Werte von $u_\mathcal{T}$ in den Gitterpunkten $\mathcal{G}_\Omega$. Da die Träger der nodalen Basisfunktionen aus wenigen Elementen bestehen, ist die Steifigkeitsmatrix dünn besetzt.

Aus Satz I.1.2 (S. 17), Satz I.1.5 (S. 18), Satz I.3.6 (S. 33) und Satz II.2.7 (S. 51) folgt unmittelbar die folgende Fehlerabschätzung.

Satz II.3.1 (A priori Fehlerabschätzung). *Seien $u \in H^1_0(\Omega)$ die schwache Lösung der Reaktions-Diffusionsgleichung (II.3.1) und $u_\mathcal{T} \in S^{k,0}_0(\mathcal{T})$ die Lösung der Finite Element Diskretisierung (II.3.2). Weiter sei $u \in H^{k+1}(\Omega)$. Dann gilt die Fehlerabschätzung*

$$|u - u_\mathcal{T}|_1 \le c_1 h^k \, |u|_{k+1} \, .$$

Ist zusätzlich Ω konvex, so ist

$$\|u - u_\mathcal{T}\| \le c_2 h^{k+1} \, |u|_{k+1} \, .$$

Die Konstanten c_1 und c_2 hängen nur von $\mathrm{C}_\mathcal{T} = \max_K \frac{h_K}{\rho_K}$, Ω und den Koeffizienten α und A ab.

Bemerkung II.3.2 (Andere Randbedingungen). (1) Bei inhomogenen Dirichlet-Randbedingungen $u = u_D$ auf Γ sucht man die Lösung $u_\mathcal{T}$ des diskreten Problems (II.3.2) statt in $S^{k,0}_0(\mathcal{T})$ in $I_\mathcal{T} u_D + S^{k,0}_0(\mathcal{T})$, d.h., man setzt $u_\mathcal{T}$ in der Form

$$u_\mathcal{T} = \sum_{z \in \mathcal{N}_\Omega} \mu_z v_z + \sum_{z \in \mathcal{N} \setminus \mathcal{N}_\Omega} u_D(z) v_z$$

mit unbekannten Koeffizienten μ_z an. Satz II.3.1 bleibt gültig.
(2) Bei gemischten oder Neumann-Randbedingungen muss man analog zu Definition I.3.1 (S. 30) auf der rechten Seite des diskreten Problems (II.3.2) den Term $\int_{\Gamma_N} g v_\mathcal{T}$ addieren und die Lösung $u_\mathcal{T}$ in dem Raum $S^{k,0}_D(\mathcal{T})$ suchen. Satz II.3.1 bleibt gültig.

Unter den Voraussetzungen von Satz I.3.3 (S. 31) gilt Satz II.3.1 auch für Konvektions-Diffusionsgleichungen, wobei die Bilinearform B wie in §I.3 beschrieben angepasst werden muss. Die Konstanten c_1 und c_2 verhalten sich dann wie $\lambda_0^{-1} \max\{\|A\|_\infty, \|\mathbf{a}\|_\infty, \|\alpha\|_\infty\}$. Für Probleme mit dominanter Diffusion, d.h. $\|\mathbf{a}\|_\infty \approx \|A\|_\infty \approx \lambda_0$, ist diese Abschätzung gut. Für Probleme mit dominanter Konvektion, d.h. $\|A\|_\infty \approx \lambda_0 \ll \|\mathbf{a}\|_\infty$, ist sie dagegen unbrauchbar. Außerdem entspricht das diskrete Problem (II.3.2) einer zentralen Differenzendiskretisierung des Konvektionstermes $\mathbf{a} \cdot \nabla u$. Daher treten bei der numerischen Lösung $u_\mathcal{T}$ unphysikalische Oszillationen auf, wenn die *Péclet-Zahl* $\lambda_0^{-1} \|\mathbf{a}\|_\infty h$ groß ist.

Im Folgenden wollen wir ein Verfahren beschreiben, das diese Probleme vermeidet und sowohl bei dominanter Konvektion wie auch bei dominanter Diffusion gute Ergebnisse liefert. Dieses Verfahren firmiert in der Literatur unter den Bezeichnungen *Stromlinien-Diffusions Methode* (engl. *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 $A = \varepsilon I$, $\mathbf{a} \in \mathbb{R}^d \setminus \{0\}$, $\alpha = 0$ und homogener Dirichlet-Randbedingungen, d.h.

$$\begin{aligned} -\varepsilon \Delta u + \mathbf{a} \cdot \nabla u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \Gamma. \end{aligned} \tag{II.3.3}$$

Dabei ist die Skalierung so gewählt, dass $|\mathbf{a}| = 1$ und $0 < \varepsilon \ll 1$ ist. Andere Randbedingungen und variable Koeffizienten werden im Prinzip genauso behandelt, erfordern aber größeren technischen Aufwand.

Im Rahmen von Satz I.1.1 (S. 15) ist jetzt

$$X = H_0^1(\Omega), \quad \ell(v) = \int_\Omega f v, \quad B(u,v) = \int_\Omega \{\varepsilon \nabla u \cdot \nabla v + (\mathbf{a} \cdot \nabla u)\, v\}.$$

Der diskrete Raum $X_\mathcal{T}$ bleibt unverändert $S_0^{k,0}(\mathcal{T})$. Wir definieren auf $S_0^{k,0}(\mathcal{T})$ eine gitterabhängige Bilinearform $B_\mathcal{T}$, Linearform $\ell_\mathcal{T}$ und Norm $|\cdot|_{1,\mathcal{T}}$ durch

$$\begin{aligned} B_\mathcal{T}(u_\mathcal{T}, v_\mathcal{T}) &= B(u_\mathcal{T}, v_\mathcal{T}) + \sum_{K \in \mathcal{T}} \delta_K \int_K (-\varepsilon \Delta u_\mathcal{T} + \mathbf{a} \cdot \nabla u_\mathcal{T})\, \mathbf{a} \cdot \nabla v_\mathcal{T}, \\ \ell_\mathcal{T}(u_\mathcal{T}) &= \ell(v_\mathcal{T}) + \sum_{K \in \mathcal{T}} \delta_K \int_K f \mathbf{a} \cdot \nabla v_\mathcal{T}, \\ |u_\mathcal{T}|_{1,\mathcal{T}} &= \left\{ \varepsilon\, |u_\mathcal{T}|_1^2 + \sum_{K \in \mathcal{T}} \delta_K \, \|\mathbf{a} \cdot \nabla u_\mathcal{T}\|_K^2 \right\}^{\frac{1}{2}}. \end{aligned}$$

Das neue diskrete Problem lautet dann

$$B_\mathcal{T}(u_\mathcal{T}, v_\mathcal{T}) = \ell_\mathcal{T}(v_\mathcal{T}) \quad \forall v_\mathcal{T} \in S_0^{k,0}(\mathcal{T}). \tag{II.3.4}$$

Dabei sind die δ_K nicht negative Parameter, die wir später bestimmen werden. Die δ_K-Terme werden eine zusätzliche Stabilisierung ergeben. Der zusätzliche Term in $B_\mathcal{T}$ entspricht einer Diskretisierung von $-\frac{\partial^2 u}{\partial \mathbf{a}^2}$. 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-Diffusionsgleichung statt mit $v_\mathcal{T}$ mit $v_\mathcal{T} + \sum_{K \in \mathcal{T}} \delta_K \mathbf{a} \cdot \nabla v_\mathcal{T}$ getestet wird.

LEMMA II.3.3 (Koerzivität von $B_\mathcal{T}$). *Für alle Elemente $K \in \mathcal{T}$ sei $\delta_K h_K^{-2} \varepsilon \le c_I^{-1}$, wobei die Konstante c_I nur von k und $\mathrm{C}_\mathcal{T} = \max_K \frac{h_K}{\rho_K}$*

abhängt. Dann gilt für alle $u_{\mathcal{T}} \in S_0^{k,0}(\mathcal{T})$

$$B_{\mathcal{T}}(u_{\mathcal{T}}, u_{\mathcal{T}}) \geq \frac{1}{2} |u_{\mathcal{T}}|_{1,\mathcal{T}}^2 .$$

Beweis. Aus dem Beweis von Satz I.3.6(1) (S. 33) ergibt sich für alle $v \in H_0^1(\Omega)$

$$B(v, v) \geq \varepsilon |v|_1^2 .$$

Hieraus folgt mit der Cauchy-Schwarzschen Ungleichung für jedes $u_{\mathcal{T}} \in S_0^{k,0}(\mathcal{T})$

$$\begin{aligned} &B_{\mathcal{T}}(u_{\mathcal{T}}, u_{\mathcal{T}}) \\ &\quad = B(u_{\mathcal{T}}, u_{\mathcal{T}}) + \sum_{K\in\mathcal{T}} \delta_K \|\mathbf{a}\cdot\nabla u_{\mathcal{T}}\|_K^2 - \sum_{K\in\mathcal{T}} \delta_K \varepsilon \int_K \Delta u_{\mathcal{T}} \mathbf{a}\cdot\nabla u_{\mathcal{T}} \\ &\quad \geq \varepsilon |u_{\mathcal{T}}|_1^2 + \sum_{K\in\mathcal{T}} \delta_K \|\mathbf{a}\cdot\nabla u_{\mathcal{T}}\|_K^2 - \sum_{K\in\mathcal{T}} \delta_K \varepsilon \|\Delta u_{\mathcal{T}}\|_K \|\mathbf{a}\cdot\nabla u_{\mathcal{T}}\|_K . \end{aligned}$$

Wegen Satz II.2.9(1) (S. 51) gibt es eine Konstante c_I, die nur von k und $\mathrm{C}_{\mathcal{T}}$ abhängt, mit

$$\|\Delta u_{\mathcal{T}}\|_K \leq c_I h_K^{-1} |u_{\mathcal{T}}|_{1;K} \quad \forall K \in \mathcal{T}.$$

Wegen $-xy \geq -\frac{1}{2}x^2 - \frac{1}{2}y^2$ für $x, y \in \mathbb{R}$ folgt aus diesen Abschätzungen

$$\begin{aligned} &B_{\mathcal{T}}(u_{\mathcal{T}}, u_{\mathcal{T}}) \\ &\quad \geq \left(1 - \max_{K\in\mathcal{T}} \frac{1}{2} c_I^2 \delta_K h_K^{-2} \varepsilon\right) \varepsilon |u_{\mathcal{T}}|_1^2 + \frac{1}{2} \sum_{K\in\mathcal{T}} \delta_K \|\mathbf{a}\cdot\nabla u_{\mathcal{T}}\|_K^2 \\ &\quad \geq \frac{1}{2} |u_{\mathcal{T}}|_{1,\mathcal{T}}^2 . \end{aligned}$$

□

Aus Lemma II.3.3 folgt, dass das diskrete Problem (II.3.4) eindeutig lösbar ist. Der folgende Satz gibt eine Fehlerabschätzung für diese Lösung.

Satz II.3.4 (A priori Fehlerabschätzung). *Die Bedingung von Lemma II.3.3 an die Stabilisierungsparameter* δ_K *sei erfüllt. Bezeichne mit* $u \in H_0^1(\Omega)$ *die schwache Lösung der Konvektions-Diffusionsgleichung* (II.3.3) *und mit* $u_{\mathcal{T}} \in S_0^{k,0}(\mathcal{T})$ *die Lösung der SDFEM Diskretisierung* (II.3.4). *Es sei* $u \in H^{k+1}(\Omega)$. *Dann gilt die Fehlerabschätzung*

$$|u - u_{\mathcal{T}}|_{1,\mathcal{T}} \leq c \left\{ \sum_{K\in\mathcal{T}} \left[\varepsilon + \delta_K + \varepsilon^2 \delta_K h_K^{-2} + \delta_K^{-1} h_K^2\right] h_K^{2k} |u|_{k+1;K}^2 \right\}^{\frac{1}{2}} .$$

Die Konstante c *hängt nur von* k *und* $\mathrm{C}_{\mathcal{T}} = \max_K \frac{h_K}{\rho_K}$ *ab. Die Fehlerabschätzung ist optimal für die Wahl* $\delta_K = \frac{h_K^2}{\sqrt{\varepsilon^2 + h_K^2}}$.

BEWEIS. Wegen der Dreiecksungleichung gilt

$$|u - u_{\mathcal{T}}|_{1,\mathcal{T}} \leq |u - I_{\mathcal{T}} u|_{1,\mathcal{T}} + |I_{\mathcal{T}} u - u_{\mathcal{T}}|_{1,\mathcal{T}}\,.$$

Aus Satz II.2.7 (S. 51) folgt wegen $|\mathbf{a}| = 1$

$$|u - I_{\mathcal{T}} u|_{1,\mathcal{T}} \leq c \left\{ \sum_{K \in \mathcal{T}} (\varepsilon + \delta_K) h_K^{2k} \, |u|^2_{k+1;K} \right\}^{\frac{1}{2}}.$$

Setze zur Abkürzung $w_{\mathcal{T}} = u_{\mathcal{T}} - I_{\mathcal{T}} u$. Wegen Satz II.3.3 ist

$$\begin{aligned} \frac{1}{2} \, |w_{\mathcal{T}}|^2_{1,\mathcal{T}} &\leq B_{\mathcal{T}}(w_{\mathcal{T}}, w_{\mathcal{T}}) \\ &= B_{\mathcal{T}}(u - I_{\mathcal{T}} u, w_{\mathcal{T}}) + B_{\mathcal{T}}(u_{\mathcal{T}} - u, w_{\mathcal{T}}). \end{aligned}$$

Da $u \in H^{k+1}(\Omega) \subset H^2(\Omega)$ ist, folgt aus der Definition von $B_{\mathcal{T}}$, $\ell_{\mathcal{T}}$ und (II.3.4)

$$\begin{aligned} &B_{\mathcal{T}}(u_{\mathcal{T}} - u, w_{\mathcal{T}}) \\ &\quad = \ell_{\mathcal{T}}(w_{\mathcal{T}}) - B_{\mathcal{T}}(u, w_{\mathcal{T}}) \\ &\quad = \ell(w_{\mathcal{T}}) + \sum_{K \in \mathcal{T}} \delta_K \int_K f \mathbf{a} \cdot \nabla w_{\mathcal{T}} \\ &\qquad - B(u, w_{\mathcal{T}}) - \sum_{K \in \mathcal{T}} \delta_K \int_K (-\varepsilon \Delta u + \mathbf{a} \cdot \nabla u) \, \mathbf{a} \cdot \nabla w_{\mathcal{T}} \\ &\quad = \sum_{K \in \mathcal{T}} \delta_K \int_K \{ f + \varepsilon \Delta u - \mathbf{a} \cdot \nabla u \} \, \mathbf{a} \cdot \nabla w_{\mathcal{T}} \\ &\quad = 0. \end{aligned}$$

Mittels partieller Integration für den Konvektionsterm erhalten wir wegen $\operatorname{div} \mathbf{a} = 0$

$$\int_\Omega (\mathbf{a} \cdot \nabla (u - I_{\mathcal{T}} u)) \, w_{\mathcal{T}} = - \int_\Omega (\mathbf{a} \cdot \nabla w_{\mathcal{T}}) \, (u - I_{\mathcal{T}} u)$$

und somit

$$\begin{aligned} &B_{\mathcal{T}}(u - I_{\mathcal{T}} u, w_{\mathcal{T}}) \\ &\quad = \varepsilon \int_\Omega \nabla (u - I_{\mathcal{T}} u) \cdot \nabla w_{\mathcal{T}} - \int_\Omega (\mathbf{a} \cdot \nabla w_{\mathcal{T}}) \, (u - I_{\mathcal{T}} u) \\ &\qquad + \sum_{K \in \mathcal{T}} \delta_K \int_K (-\varepsilon \Delta (u - I_{\mathcal{T}} u) + \mathbf{a} \cdot \nabla (u - I_{\mathcal{T}} u)) \, \mathbf{a} \cdot \nabla w_{\mathcal{T}}. \end{aligned}$$

Hieraus folgt mit der Cauchy-Schwarzschen Ungleichung wegen $|\mathbf{a}| = 1$

$$\begin{aligned}
&B_{\mathcal{T}}(u - I_{\mathcal{T}}u, w_{\mathcal{T}}) \\
&\quad\le \varepsilon \,|u - I_{\mathcal{T}}u|_1 \,|w_{\mathcal{T}}|_1 + \sum_{K\in\mathcal{T}} \|u - I_{\mathcal{T}}u\|_K \,\|\mathbf{a}\cdot\nabla w_{\mathcal{T}}\|_K \\
&\qquad + \sum_{K\in\mathcal{T}} \delta_K \left(\varepsilon \,|u - I_{\mathcal{T}}u|_{2;K} + \|\mathbf{a}\cdot\nabla(u - I_{\mathcal{T}}u)\|_K\right) \|\mathbf{a}\cdot\nabla w_{\mathcal{T}}\|_K \\
&\quad\le \sqrt{3}\,|w_{\mathcal{T}}|_{1,\mathcal{T}} \left\{ \sum_{K\in\mathcal{T}} [\varepsilon + \delta_K] \,|u - I_{\mathcal{T}}u|^2_{1;K} + \sum_{K\in\mathcal{T}} \varepsilon^2\delta_K \,|u - I_{\mathcal{T}}u|^2_{2;K} \right. \\
&\qquad\qquad \left. + \sum_{K\in\mathcal{T}} \delta_K^{-1} \,\|u - I_{\mathcal{T}}u\|^2_K \right\}^{\frac12}.
\end{aligned}$$

Hieraus ergibt sich mit Satz II.2.7 (S. 51)

$$\begin{aligned}
&B_{\mathcal{T}}(u - I_{\mathcal{T}}u, w_{\mathcal{T}}) \\
&\quad\le c\,|w_{\mathcal{T}}|_{1,\mathcal{T}} \left\{ \sum_{K\in\mathcal{T}} [\varepsilon + \delta_K] h_K^{2k} \,|u|^2_{k+1;K} + \sum_{K\in\mathcal{T}} \varepsilon^2\delta_K h_K^{2k-2} \,|u|^2_{k+1;K} \right. \\
&\qquad\qquad \left. + \sum_{K\in\mathcal{T}} \delta_K^{-1} h_K^{2k+2} \,|u|^2_{k+1;K} \right\}^{\frac12} \\
&\quad= c\,|w_{\mathcal{T}}|_{1,\mathcal{T}} \left\{ \sum_{K\in\mathcal{T}} \left[\varepsilon + \delta_K + \varepsilon^2\delta_K h_K^{-2} + \delta_K^{-1}h_K^2\right] h_K^{2k} \,|u|^2_{k+1;K} \right\}^{\frac12}.
\end{aligned}$$

Aus diesen Abschätzungen folgt

$$\begin{aligned}
&|u_{\mathcal{T}} - I_{\mathcal{T}}u|_{1,\mathcal{T}} = |w_{\mathcal{T}}|_{1,\mathcal{T}} \\
&\qquad\qquad \le c' \left\{ \sum_{K\in\mathcal{T}} \left[\varepsilon + \delta_K + \varepsilon^2\delta_K h_K^{-2} + \delta_K^{-1}h_K^2\right] h_K^{2k} \,|u|^2_{k+1;K} \right\}^{\frac12}.
\end{aligned}$$

Zusammen mit der bereits bewiesenen Abschätzung für $|u - I_{\mathcal{T}}u|_{1,\mathcal{T}}$ folgt hieraus die Fehlerabschätzung für $|u - u_{\mathcal{T}}|_{1,\mathcal{T}}$. Offensichtlich ist sie optimal, wenn die δ_K- und δ_K^{-1}-Terme gleich sind. Hieraus folgt die Behauptung über die optimale Wahl von δ_K. □

Bemerkung II.3.5. Es gilt $\delta_K \approx h_K$, wenn $\varepsilon \ll h_K$ ist. In diesem Fall liefert Satz II.3.4 eine Fehlerabschätzung der Form $\|\mathbf{a}\cdot\nabla(u - u_{\mathcal{T}})\| \le ch^k \,|u|_{k+1}$ in Stromrichtung. Im Fall $h_K \le \varepsilon$, in dem auch die Standard-Diskretisierung gut funktioniert, ist $\delta_K \approx h_K^2\varepsilon^{-1}$, und Satz II.3.4 liefert eine Fehlerabschätzung der Form $|u - u_{\mathcal{T}}|_1 \le ch^k \,|u|_{k+1}$, die vergleichbar ist zu derjenigen von Satz II.3.1.

KAPITEL III

Praktische Aspekte

In diesem Kapitel befassen wir uns mit einigen praktischen Aspekten der Finite Element Methode. Zuerst beschreiben wir die Vorgehensweise bei gekrümmten Rändern und untersuchen den Effekt einer näherungsweisen numerischen Berechnung der Integrale, die beim Aufstellen des diskreten Problems auftreten. Anschließend befassen wir uns in §III.2 mit der effizienten numerischen Lösung der diskreten Probleme. Da diese in der Regel sehr groß, die Steifigkeitsmatrizen aber dünn besetzt sind, scheiden direkte Lösungsverfahren wie die Gauß-Elimination oder Cholesky-Zerlegung wegen des Speicherbedarfes und des Rechenaufwandes aus. Stattdessen werden die diskreten Probleme nur näherungsweise mit einem iterativen Verfahren gelöst. Dabei nutzen wir aus, dass die diskreten Probleme nur mit einer Genauigkeit gelöst werden müssen, die dem Approximationsfehler der Finite Element Räume entspricht, und dass wir im Rahmen eines adaptiven Gitterverfeinerungsprozesses mit der Näherungslösung des gröberen Gitters einen guten Startwert für die Iteration auf dem aktuellen feineren Gitter haben. In §III.2 betrachten wir daher Mehrgitterverfahren, geben einen dem Finite Element Rahmen angemessenen Konvergenzbeweis und betten diese Verfahren in die größere Klasse der sog. Teilraumkorrekturmethoden ein. In den §§III.3 und III.4 befassen uns mit adaptiven Finite Element Methoden basierend auf a posteriori Fehlerschätzern. Wir untersuchen beispielhaft einen sog. residuellen Fehlerschätzer und beweisen unter vereinfachten Annahmen die Konvergenz des adaptiven Verfahrens. In §III.5 stellen wir schließlich kurz geeignete Datenstrukturen vor und geben einige numerische Beispiele.

III.1. Randapproximation und numerische Integration

In Kapitel II haben wir stets vorausgesetzt, dass Ω ein Polyedergebiet ist, d.h. der Rand Γ besteht stückweise aus Hyperebenen. Im allgemeinen ist jedoch der Rand von Ω gekrümmt. Daher ersetzt man in der Praxis Ω durch ein approximierendes Polyeder Ω_h, derart dass die Eckpunkte von Ω_h auf Γ liegen und die Kanten von Ω_h die Länge $O(h)$ haben (vgl. Abbildung III.1.1). $\mathcal{T}$ ist dann eine Unterteilung von Ω_h, die den Bedingungen von §II.1 (S. 35) genügt. Zusätzlich müssen die Eckpunkte von Ω_h in der Menge $\mathcal{N}$ der Elementeckpunkte enthalten sein. Für die Konstruktion von Ω_h und $\mathcal{T}$ muss der Benutzer den Rand

Γ stückweise als Graph oder implizit als Nullstellenmenge $\{F(x) = 0\}$ zusammen mit F und grad F angeben.

Sei Γ_h der Rand von Ω_h. Wenn Γ stückweise C^2 ist, folgt dann

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

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

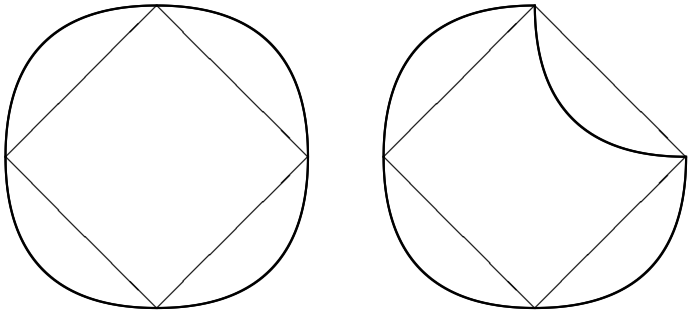

ABBILDUNG III.1.1. Randapproximation eines konvexen und eines nicht konvexen Gebietes

Falls Ω konvex ist, gilt $\Omega_h \subset \Omega$ und die Fehlerabschätzungen von §II.3 (S. 52) können übertragen werden, wobei die Randapproximation einen zusätzlichen Konsistenzfehler von $\mathrm{O}(h^2)$ beiträgt.
Ist Ω nicht konvex, gilt $\Omega_h \setminus \Omega \neq \emptyset$. Dann müssen die Daten A, $\mathbf{a}$, α und f auf Ω_h fortgesetzt werden. Dies geschieht am einfachsten durch Reflexion (vgl. Abbildung III.1.2):

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

Mit diesen Modifikationen bleiben die Fehlerabschätzungen von §II.3 (S. 52) erhalten. Dabei müssen alle Normen auf $\Omega \cap \Omega_h$ statt auf Ω bezogen werden. Die Randapproximation trägt wieder einen Konsistenzfehler der Ordnung $\mathrm{O}(h^2)$ bei.

Um den Konsistenzfehler durch die Randapproximation zu verringern, kann man auch krummlinige oder *isoparametrische Elemente* betrachten. Dabei ist mit der Notation von §II.1 (S. 35) die Transformation $F_K : \widehat{K} \to K$ nicht mehr affin, sondern ein allgemeiner Diffeomorphismus (krummlinige Elemente) oder komponentenweise eine Funktion aus R_k (isoparametrische Elemente). In diesem Sinne sind lineare simpliziale Elemente ($k = 1, \widehat{K} = \widehat{K}_S$) isoparametrische Elemente der Ordnung 1. Bei Verwendung isoparametrischer Elemente der Ordnung $k \geq 2$ kann man eine Randapproximation der Ordnung $\mathrm{O}(h^{k+1})$

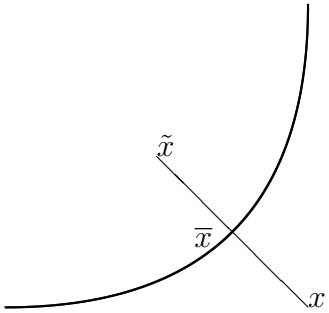

ABBILDUNG III.1.2. Reflexion am Rand Γ

erreichen. Dementsprechend ist der Konsistenzfehler durch die Randapproximation auch von der Ordnung $O(h^{k+1})$. Die größere Genauigkeit wird natürlich durch einen erhöhten Aufwand erkauft.

Die Finite Element Diskretisierungen von §II.3 (S. 52) führen auf lineare Gleichungssysteme. Zur Berechnung der rechten Seite und der Steifigkeitsmatrix müssen Integrale der Form $\int_K \varphi$ mit $\varphi = fv$, $\varphi = \nabla u \cdot A \nabla v$ 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.2 (S. 40) am besten aus solchen für das Referenzelement hergeleitet. Sei also

$$Q_{\widehat{K}}(\widehat{\varphi}) = \sum_{\ell=1}^{L} \widehat{\omega}_\ell \widehat{\varphi}(\widehat{b}_\ell)$$

eine Quadraturformel für $\int_{\widehat{K}} \widehat{\varphi}$. Sie hat die Ordnung m, wenn für alle $\widehat{p} \in R_m(\widehat{K})$ gilt

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

Sei nun $K \in \mathcal{T}$ und $F_K : \widehat{K} \to K$ eine affine Transformation mit $B_K = DF_K$. Da F_K affin ist, folgt aus [16, Satz II.1.4], dass

$$Q_K(\varphi) = \sum_{\ell=1}^{L} \omega_\ell \varphi(b_\ell) \quad \text{mit} \quad \omega_\ell = |\det B_K|\, \widehat{\omega}_\ell,\ b_\ell = F_K(\widehat{b}_\ell)$$

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

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

Man kann zeigen, dass die Fehlerabschätzungen von §II.3 (S. 52) für Finite Element Diskretisierungen der Ordnung k, d.h. $X_{\mathcal{T}} = S_0^{k,0}(\mathcal{T})$, gültig bleiben, wenn die Quadraturformel $Q_{\widehat{K}}$ mindestens die Ordnung $2k-2$ hat [7, §§25-29]. Insbesondere reicht also für lineare simpliziale

Elemente ($k = 1, \widehat{K} = \widehat{K}_S$) und d-lineare Würfelelemente ($k = 1, \widehat{K} = \widehat{K}_W$) eine Quadraturformel der Ordnung 0 aus, d.h. nur die konstanten Funktionen müssen exakt integriert werden.

Beispiel III.1.1 (Quadraturformeln). (1) Die Mittelpunktsregel

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

liefert eine Quadraturformel der Ordnung 1 für den Referenz d-Simplex.
(2) Die Formel

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

hat die Ordnung 2 für das Referenzdreieck.
(3) Die Formel

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

hat die Ordnung 3 für das Referenzdreieck.
(4) Sei

$$\widetilde{Q}(\psi) = \sum_{\ell=1}^{L}\widetilde{\omega}_\ell\psi(\widetilde{x}_\ell)$$

eine Quadraturformel der Ordnung m für $\int_0^1 \psi(x)\mathrm{d}x$. Dann ist gemäß [15, Beispiel II.1.3 (3)]

$$Q_{\widehat{K}}(\widehat{\varphi}) = \sum_{\ell_1=1}^{L}\cdots\sum_{\ell_d=1}^{L}\widetilde{\omega}_{\ell_1}\cdot\ldots\cdot\widetilde{\omega}_{\ell_d}\widehat{\varphi}(\widetilde{x}_{\ell_1},\ldots,\widetilde{x}_{\ell_d})$$

eine Quadraturformel der Ordnung m für den Referenz d-Würfel.
(5) Die Mittelpunktsregel liefert die Formel

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

der Ordnung 1 für das Referenzquadrat.
(6) Die Trapezregel liefert die Formel

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

der Ordnung 1 für das Referenzquadrat.
(7) Die Simpson-Regel liefert mit

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

die Formel

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

der Ordnung 3 für das Referenzquadrat.

III.2. Lösung der diskreten Probleme

Wir betrachten zunächst die Reaktions-Diffusionsgleichung mit homogenen Dirichlet-Randbedingungen (II.3.1) (S. 52) und das zugehörige diskrete Problem (II.3.2) (S. 52). Auf Konvektions-Diffusionsgleichungen gehen wir kurz in Bemerkung III.2.12 ein.

Da die Träger der nodalen Basisfunktionen nur aus wenigen Elementen 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 Elemente 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 $(\cdot\,,\,\cdot)_{\mathcal{T}}$ und eine zugehörige Norm $\|\cdot\|_{\mathcal{T}}$ auf $S_0^{k,0}(\mathcal{T})$ durch

$$(\varphi\,,\,\psi)_{\mathcal{T}} = \sum_{K\in\mathcal{T}} \sum_{z\in\Sigma_k(K)} h_K^d \varphi(z)\psi(z), \quad \|\varphi\|_{\mathcal{T}} = (\varphi\,,\,\varphi)_{\mathcal{T}}^{\frac{1}{2}}.$$

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

$$(u\,,\,v)_{\mathcal{T}} = \ell(v) \quad \forall v \in S_0^{k,0}(\mathcal{T})$$

leicht gelöst werden.

Lemma III.2.1 (Normäquivalenz). *$\|\cdot\|_{\mathcal{T}}$ und $\|\cdot\|$ sind äquivalente Normen auf $S_0^{k,0}(\mathcal{T})$. Die entsprechenden Konstanten hängen nur von k und $\mathrm{C}_{\mathcal{T}} = \max_K \frac{h_K}{\rho_K}$ ab.*

Beweis. Wegen Satz II.1.3 (S. 37) ist $\left\{\sum_{z\in\widehat{\Sigma}_k} |\varphi(z)|^2\right\}^{\frac{1}{2}}$ eine Norm auf $\widehat{R}_k$. Da $\widehat{R}_k$ endlich dimensional ist, gibt es zwei Konstanten $\widehat{c}_1$, $\widehat{c}_2$,

die nur von k abhängen, mit

$$\widehat{c}_1 \|\varphi\|_{\widehat{K}} \leq \left\{ \sum_{z \in \widehat{\Sigma}_k} \varphi(z)^2 \right\}^{\frac{1}{2}} \leq \widehat{c}_2 \|\varphi\|_{\widehat{K}} \quad \forall \varphi \in \widehat{R}_k.$$

Sei nun $K \in \mathcal{T}$ beliebig und $F_K : \widehat{K} \to K$ eine affine Transformation. Bezeichnet $|D|$ das d-dimensionale Lebesgue-Maß einer messbaren Menge $D \subset \mathbb{R}^d$, gilt dann

$$|\det DF_K| = \frac{|K|}{\left|\widehat{K}\right|} \leq \widehat{c}_3 h_K^d,$$
$$|\det DF_K| = \frac{|K|}{\left|\widehat{K}\right|} \geq \widehat{c}_4 \rho_K^d \geq \widehat{c}_5 h_K^d.$$

Aus diesen Abschätzungen und Lemma II.2.5 (S. 50) folgt

$$\|\varphi\|_K = |\det DF_K|^{\frac{1}{2}} \|\varphi \circ F_K\|_{\widehat{K}} \leq c_1 h_K^{\frac{d}{2}} \left\{ \sum_{z \in \widehat{\Sigma}_k} |\varphi \circ F_K(z)|^2 \right\}^{\frac{1}{2}}$$
$$= c_1 \left\{ \sum_{z \in \Sigma_k(K)} h_K^d |\varphi(z)|^2 \right\}^{\frac{1}{2}}$$

und

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

Hieraus folgt die Behauptung durch Quadrieren und Summieren über alle Dreiecke. Man beachte, dass die Konstanten c_1, c_2 von $C_{\mathcal{T}}$ abhängen. □

Aus Lemma III.2.1, dem Beweis des Existenzsatzes I.3.3 (S. 31) und der Friedrichschen Ungleichung I.2.15 (S. 26) folgt für alle $v \in S_0^{k,0}(\mathcal{T})$

$$B(v,v) \geq \beta |v|_1^2 \geq \beta' \|v\|^2 \geq \beta'' \|v\|_{\mathcal{T}}^2 .$$

Ebenso folgt aus dem Beweis von Satz I.3.3 (S. 31), der inversen Abschätzung von Satz II.2.9 (S. 51) und Lemma III.2.1 für alle $v, w \in$

$S_0^{k,0}(\mathcal{T})$

$$B(v,w) \le \mathcal{B}\,|v|_1\,|w|_1 \le c^2\mathcal{B}\left(\min_{K\in\mathcal{T}} h_K\right)^{-2}\|v\|\,\|w\|$$
$$\le c'\left(\min_{K\in\mathcal{T}} h_K\right)^{-2}\|v\|_{\mathcal{T}}\,\|w\|_{\mathcal{T}}\,.$$

Also hat die Steifigkeitsmatrix die Kondition $\mathrm{O}\left((\min_{K\in\mathcal{T}} h_K)^{-2}\right)$. Daher scheiden das Jacobi und Gauß-Seidel-Verfahren als Löser aus. Da die Bilinearform B symmetrisch und koerziv ist, ist die Steifigkeitsmatrix symmetrisch positiv definit, und ein CG-Verfahren oder besser ein PCG-Verfahren können als Löser benutzt werden. Als Vorkonditionierer eignet sich z.B. der SSOR-Vorkonditionierer aus [16, §III.6]. Aufgrund unserer dortigen Erfahrungen mit Differenzenverfahren erwarten wir aber, dass Mehrgitterverfahren die besten Ergebnisse liefern.

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)

	CG	SSOR-PCG	MG-V($\frac{1}{2}$)	MG-V(1)	MG-W(1)
N	κ	κ	κ	κ	κ
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)

	CG	SSOR-PCG	MG-V($\frac{1}{2}$)	MG-V(1)	MG-W(1)
N	κ	κ	κ	κ	κ
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

BEISPIEL III.2.2 (Verfahrensvergleich). 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 gröbste Unterteilung besteht aus 8 gleichschenklig rechtwinkligen Dreiecken mit Hypotenusen in Richtung $(1,-1)$ (*Courant-Triangulierung*). Die weiteren Unterteilungen werden durch uniforme Verfeinerung erzeugt. Dabei wird jedes Dreieck in vier kleinere Dreiecke unterteilt, indem die Kantenmittelpunkte miteinander verbunden werden (vgl. §III.4 (S. 91)).

(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 $u(x,y) = r^{\frac{2}{3}} \sin(\frac{2}{3}\varphi)$. Die gröbste Unterteilung besteht aus 6 gleichschenklig rechtwinkligen Dreiecken mit Hypotenusen in Richtung $(1,-1)$. Die weiteren Unterteilungen werden wie in (1) durch uniforme Verfeinerung erzeugt.

(3) Die Differentialgleichung und die gröbste Unterteilung sind wie in (2). Die feineren Unterteilungen werden aber mit einer adaptiven Gitterverfeinerung basierend auf einem residuellen Fehlerschätzer erzeugt (vgl. Beispiel III.4.1 (S. 99) und Abbildung III.4.5 (S. 100)).

Auf dem gröbsten Gitter wird der Nullvektor als Startwert verwendet. Auf feineren Gittern wird die lineare Interpolierende der Näherungslösung auf dem nächst gröberen Gitter als Startwert benutzt (vgl. die geschachtelte Iteration [16, Algorithmus III.6.1]). 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 κ angegeben. N bezeichnet die Zahl der Unbekannten. T ist die Gesamtrechenzeit in Sekunden auf einem Macintosh G4-Powerbook.

Wir wollen ein Mehrgitter-Verfahren, Algorithmus III.2.1, für die Finite Element Diskretisierung analysieren. Dazu nehmen wir an, dass wir eine Hierarchie von Unterteilungen $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_R$ mit zugehörigen *geschachtelten* Finite Element Räumen $S_0^{k,0}(\mathcal{T}_0) \subset S_0^{k,0}(\mathcal{T}_1) \subset \ldots \subset$

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)

	CG	SSOR-PCG	MG-V($\frac{1}{2}$)	MG-V(1)	MG-W(1)
N	κ	κ	κ	κ	κ
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

Algorithmus III.2.1 MG-Verfahren mit V-Zyklus und Jacobi Glättung

Gegeben: Näherung $u_m^0 \in X_m$ für die Lösung des diskreten Problems.
Gesucht: Verbesserte Näherung $u_m^{\nu_1+\nu_2+1} \in X_m$.

1: **if** $m = 0$ **then**
2: Löse das Problem
$$B(u_0^*, v) = \ell_0(v) \quad \forall v \in X_0.$$
3: $u_0^{\nu_1+\nu_2+1} = u_0^*$, `stop`
4: **end if**
5: **for** $i = 1, \ldots, \nu_1$ **do** ▷ Vor-Glättung
6: Berechne u_m^i aus u_m^{i-1} als Lösung von
$$\left(u_m^i - u_m^{i-1}\,,\, v\right)_m = \omega_m^{-1}\left\{\ell_m(v) - B(u_m^{i-1}, v)\right\} \quad \forall v \in X_m.$$
7: **end for**
8: Berechne ▷ Grobgitterkorrektur
$$\ell_{m-1}(v) = \ell_m(v) - B(u_m^{\nu_1}, v) \quad \forall v \in X_{m-1}.$$
9: Wende das MG-Verfahren mit Startwert $u_{m-1}^0 = 0$ auf das Problem
$$B(u_{m-1}^*, v) = \ell_{m-1}(v) \quad \forall v \in X_{m-1}$$
an. Das Ergebnis sei $\widetilde{u}_{m-1}$. Setze
$$u_m^{\nu_1+1} = u_m^{\nu_1} + \widetilde{u}_{m-1}.$$
10: **for** $i = \nu_1 + 2, \ldots, \nu_1 + \nu_2 + 1$ **do** ▷ Nach-Glättung
11: Berechne u_m^i aus u_m^{i-1} als Lösung von
$$\left(u_m^i - u_m^{i-1}\,,\, v\right)_m = \omega_m^{-1}\left\{\ell_m(v) - B(u_m^{i-1}, v)\right\} \quad \forall v \in X_m.$$
12: **end for**

$S_0^{k,0}(\mathcal{T}_R)$ haben. Es gilt, die Finite Element Diskretisierung zur feinsten Unterteilung zu berechnen. Dazu machen wir folgende Annahmen:

- Ω ist konvex.
- Jede Unterteilung $\mathcal{T}_m$ ist *uniform*, d.h.
$$h_m = \max_{K\in\mathcal{T}_m} h_K \le c \min_{K\in\mathcal{T}_m} h_K$$
mit einer von m unabhängigen Konstanten c.
- $h_{m-1} \le c h_m$ für alle m mit einer von m unabhängigen Konstanten c.

Zur Vereinfachung der Notation ersetzen wir einen Index $\mathcal{T}_m$ in der Regel durch m und setzen $X_m = S_0^{k,0}(\mathcal{T}_m)$.

Bemerkung III.2.3. (1) Es ist $\ell_R(v) = \ell(v) = \int_\Omega fv$. Die rechten Seiten ℓ_m zu den gröberen Unterteilungen werden rekursiv berechnet.
(2) Die Dämpfungsparameter ω_m werden später bestimmt.
(3) Die Gleichungssysteme in den Glättungsschritten haben eine diagonale Koeffizientenmatrix.
(4) Die Grobgitterkorrektur nutzt aus, dass die Finite Element Räume geschachtelt sind, d.h. $X_{m-1} \subset X_m$. In der Praxis stellt man u_m und u_{m-1} als Vektoren dar, deren Komponenten die Werte in den entsprechenden Gitterpunkten sind. Insbesondere muss u_{m-1} vom Gitter $\mathcal{G}_{m-1}$ auf das Gitter $\mathcal{G}_m$ interpoliert werden.

Da die Bilinearform B symmetrisch ist, besitzt sie auf jedem X_m einen vollständigen Satz $\lambda_{m,1}, \ldots, \lambda_{m,n_m}$, $n_m = \dim X_m$, von Eigenwerten und zugehörigen, bzgl. $(\cdot\,,\cdot)_m$ orthonormierten Eigenfunktionen $\psi_{m,1}, \ldots, \psi_{m,n_m}$, d.h.
$$B(\psi_{m,\mu}, v) = \lambda_{m,\mu}\,(\psi_{m,\mu}\,,\, v)_m \quad \forall v \in X_m$$
$$(\psi_{m,\mu}\,,\, \psi_{m,\nu})_m = \delta_{\mu\nu}.$$
O.E. können wir die Eigenwerte der Größe nach ordnen
$$0 < \lambda_{m,1} < \ldots < \lambda_{m,n_m} = \Lambda_m.$$
Da die Unterteilungen uniform sind, folgt aus Satz II.2.9 (S. 51) und Lemma III.2.1 $\Lambda_m \approx h_m^{-2}$. Wir nehmen im Folgenden an, dass $\omega_m = \Lambda_m$ ist. Es reicht jedoch, wenn $\omega_m \ge \Lambda_m$ und $\omega_m \approx h_m^{-2}$ 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
$$|v|_{s,m} = \left\{ \sum_{\mu=1}^{n_m} \lambda_{m,\mu}^s c_\mu^2 \right\}^{\frac{1}{2}}$$
eine Norm auf X_m definieren. Aus den Voraussetzungen folgt für alle $v \in X_m$
$$|v|_{0,m} = \|v\|_m \approx \|v\|\,, \quad |v|_{1,m} = B(v,v)^{\frac{1}{2}} \approx \|v\|_1\,.$$

Dabei bedeutet „≈“ Äquivalenz von Normen mit von m unabhängigen Konstanten. Sind $v, w \in X_m$ mit $v = \sum c_\mu \psi_{m,\mu}$, $w = \sum d_\mu \psi_{m,\mu}$, so folgt aus der Cauchy-Schwarzschen Ungleichung für Summen und der Orthogonalität der Eigenfunktionen

$$\text{(III.2.1)} \qquad \begin{aligned} B(v,w) &= \sum_{\mu=1}^{n_m} \lambda_{m,\mu} c_\mu d_\mu \le \left\{ \sum_{\mu=1}^{n_m} c_\mu^2 \right\}^{\frac{1}{2}} \left\{ \sum_{\mu=1}^{n_m} \lambda_{m,\mu}^2 d_\mu^2 \right\}^{\frac{1}{2}} \\ &= |v|_{0,m} \, |w|_{2,m} \, . \end{aligned}$$

Satz III.2.4 (Ritz-Projektion). *Bezeichne mit $Q_m : X_m \to X_{m-1}$ die* Ritz-Projektion*, d.h. $Q_m v \in X_{m-1}$ und*

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

Dann gilt für alle $v \in X_m$

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

Die Konstante c hängt nicht von m ab.

Beweis. Aus der Definition der Ritz-Projektion und (III.2.1) folgt

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

Da Ω konvex ist, folgt aus Satz I.1.2 (S. 17)

$$\|v - Q_m v\| \le c h_{m-1} \, \|v - Q_m v\|_1 \le c' h_{m-1} \, |v - Q_m v|_{1,m} \, .$$

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

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

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

J beschreibt die Fehlerfortpflanzung in den Glättungsschritten von Algorithmus III.2.1. 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 $B(\cdot, \cdot)$. Definiere für $v \in X_m$

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

Dann gilt offensichtlich für $v \in X_m$

$$|v| = \left| J^{\frac{1}{2}} v \right|_{1,m}, \quad 0 \le \rho(v) \le 1.$$

SATZ III.2.5 (Glättungseigenschaft). *Sei $v \in X_m$ und $\rho = \rho(J^\nu v)$. Dann gilt*

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

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

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

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

$$|J^\nu v|_{1,m}^{2\nu+1} \le \left| J^{\nu+\frac{1}{2}} v \right|_{1,m}^{2\nu} |v|_{1,m} = |J^\nu v|^{2\nu} \, |v|_{1,m} .$$

Hieraus folgt

$$|J^\nu v|_{1,m} \le \left\{ \frac{|J^\nu v|}{|J^\nu v|_{1,m}} \right\}^{2\nu} |v|_{1,m} = \rho^\nu \, |v|_{1,m} . \qquad \square$$

SATZ III.2.6 (Approximationseigenschaft). *Sei $v \in X_m$ und $\rho = \rho(v)$. Dann gilt*

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

Dabei ist c die Konstante aus Satz III.2.4.

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

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

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.2.4 folgt

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

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

Nach diesen Vorbereitungen können wir nun die Konvergenz von Algorithmus III.2.1 beweisen.

SATZ III.2.7 (Konvergenzrate des Mehrgitteralgorithmus). *Bezeichne mit δ_m die Konvergenzrate von Algorithmus III.2.1 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.2.4*

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

BEWEIS. Bezeichne mit u_m^* die Lösung der Finite Element Probleme auf dem m-ten Gitter und mit e_m^i den Fehler im i-ten Schritt von Algorithmus III.2.1. Dann gilt

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

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

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

Wegen der Galerkin Orthogonalität

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

folgt aus $u_{m-1}^* = Q_m e_m^\nu$

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

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

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

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

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

Für den zweiten Summanden gilt

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

Aus diesen beiden Abschätzungen folgt mit der Cauchy-Schwarzschen Ungleichung

$$\begin{aligned} B(e_m^{2\nu+1}, w) &\le (1-\delta_{m-1})\, |[I-Q_m]e_m^\nu|_{1,m}\, |[I-Q_m]J^\nu w|_{1,m} \\ &\quad + \delta_{m-1}\, |e_m^\nu|_{1,m}\, |J^\nu w|_{1,m} \\ &\le \left\{(1-\delta_{m-1})\, |[I-Q_m]e_m^\nu|_{1,m}^2 + \delta_{m-1}\, |e_m^\nu|_{1,m}^2\right\}^{\frac12} \\ &\quad \cdot \left\{(1-\delta_{m-1})\, |[I-Q_m]J^\nu w|_{1,m}^2 + \delta_{m-1}\, |J^\nu w|_{1,m}^2\right\}^{\frac12}. \end{aligned}$$

Wegen der Sätze III.2.5 und III.2.6 ist

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

und

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

Da

$$\left|e^{2\nu+1}\right|_{1,m} = \sup_{w\in X_m;\, |w|_{1,m}=1} B(e^{2\nu+1}, w)$$

ist, folgt hieraus

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

Also ist

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

Da auf dem gröbsten Gitter exakt gelöst wird, gilt

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

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

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

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

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

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. □

Satz III.2.8 (Konvergenzrate des Mehrgitteralgorithmus). *Sei $\widetilde{\delta}_m$ die Konvergenzrate von Algorithmus III.2.1 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.2.4*

$$\widetilde{\delta}_m \le \left[\frac{c}{c+2\nu}\right]^{\frac{1}{2}}.$$

Beweis. Sei M der Operator $e_m^0 \mapsto e_m^{\nu+1}$ des Algorithmus III.2.1 mit $\nu_1 = \nu$, $\nu_2 = 0$. Dann beschreibt der bzgl. des Skalarproduktes $B(\cdot,\cdot)$ adjungierte Operator M^* die Fehlerfortpflanzung von Algorithmus III.2.1 mit $\nu_1 = 0$, $\nu_2 = \nu$. Insbesondere beschreibt M^*M die Fehlerfortpflanzung von Algorithmus III.2.1 mit $\nu_1 = \nu_2 = \nu$. Damit folgt aus Satz III.2.7

$$\|M\|_{\mathcal{L}}^2 = \|M^*M\|_{\mathcal{L}} \le \frac{c}{c+2\nu}.$$ □

Der Beweis von Satz III.2.7 beruht wesentlich auf den Annahmen, dass Ω konvex ist und die Gitter uniform sind. Die Konvexität von Ω wird in Satz III.2.4 benötigt, da dort das Dualitätsargument von Aubin-Nitsche benutzt wird, das die H^2-Regularität der Differentialgleichung voraussetzt. Die Uniformität der Gitter wird für die inverse Abschätzung zur Kontrolle des größten Eigenwertes der Steifigkeitsmatrix benötigt. Diese Einschränkungen sind für die Praxis zu restriktiv. Ebenso hat sich gezeigt, dass man mit anderen Glättern als der simplen Jacobi Iteration in Algorithmus III.2.1 wesentlich bessere Konvergenzraten erzielen kann.

Die genannten Nachteile können mit einem allgemeineren, abstrakten Zugang, der sog. *Teilraum-Korrektur-Methode (TRK)*, 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 Hilbert-Raum mit Skalarprodukt $(\cdot\,,\cdot)$.
- $V_1, \ldots, V_N$ sind Untervektorräume von V mit $V = \sum_{i=1}^{N} V_i$, wobei diese Zerlegung in der Regel weder direkt noch gar orthogonal ist.
- $Q_i : V \to V_i$ sind die orthogonalen Projektionen bzgl. $(\cdot\,,\cdot)$.
- $L : V \to V$ ist ein symmetrischer, positiv definiter Operator.
- $L_i : V_i \to V_i$ ist die Einschränkung von L auf V_i, d.h. $(L_i u\,, v) = (Lu\,, v)$ für alle $u, v \in V_i$.
- $R_i : V_i \to V_i$ ist eine leicht berechenbare, symmetrisch positiv definite Approximation an L_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. $(\cdot\,,\cdot)_{\mathcal{T}}$ und L wird durch eine symmetrische, koerzive Bilinearform erzeugt.
Zu lösen ist das Problem

$$Lu = f.$$

Dies geschieht mit Algorithmus III.2.2.

Algorithmus III.2.2 Teilraum-Korrektur-Methode,

Gegeben: Näherung $u \in V$.
Gesucht: Verbesserte Näherung u.

```
for i = 1, ..., do
    for j = 1, ..., N do
        u ← u + R_j Q_j (f − Lu).
    end for
end for
```

Beispiel III.2.9 (Gauß-Seidel und Mehrgitterverfahren). (1) Sei $V = \mathbb{R}^N$ und $V_i = \operatorname{span}\{e_i\}$. Dann ist Algorithmus III.2.2 das Gauß-Seidel-Verfahren.
(2) Sei $V = S_0^{k,0}(\mathcal{T}_N)$, $V_i = S_0^{k,0}(\mathcal{T}_i)$, $R_0 = L_0^{-1}$ und $R_i = \omega_i^{-1} I$, $i > 0$. Dann ist Algorithmus III.2.2 der Mehrgitteralgorithmus III.2.1 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.2.2 setzen wir

$$\lambda = \min_{1 \le i \le N} \lambda_{\min}(R_i L_i), \quad \Lambda = \max_{1 \le i \le N} \lambda_{\max}(R_i L_i), \quad T_i = R_i Q_i L.$$

Aus den Voraussetzungen folgt $0 < \lambda \le \Lambda$. Durch entsprechende Skalierung kann man o.E. erreichen, dass $\Lambda < 2$ ist. Wir bezeichnen mit

$|\cdot|$ die zu L gehörige Energienorm, d.h.

$$|u| = (Lu\,,\,u)^{\frac{1}{2}} \quad \forall u \in V.$$

Für eine Finite Element Diskretisierung der Reaktions-Diffusionsgleichung ist insbesondere $|\cdot|$ zur H^1-Norm äquivalent.

SATZ III.2.10 (Konvergenzrate der Teilraum-Korrektur-Methode). *Es gebe zwei Konstanten K_0 und K_1 mit*

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

und

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

Weiter sei $\Lambda < 2$. Dann ist die Konvergenzrate von Algorithmus III.2.2 gemessen in der Norm $|\cdot|$ kleiner oder gleich

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

BEWEIS. Sei u die exakte Lösung von $Lu = f$. Aus Algorithmus III.2.2 ergibt sich

$$u - u_{n+\frac{j}{N}} = (I - T_j)\left(u - u_{n+\frac{j-1}{N}}\right)$$

und somit

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

Setze zur Abkürzung

$$E_0 = I, \quad E_j = (I - T_j) \cdot \ldots \cdot (I - T_1),\ 1 \leq j \leq N.$$

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

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

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 $P_i : V \to V_i$ die orthogonale Projektion bzgl. des Skalarproduktes $(L\cdot\,,\,\cdot)$. Dann folgt

$$Q_i L = L_i P_i$$

und damit

$$(R_i L_i)^{-1} T_i = (R_i L_i)^{-1} R_i Q_i L = (R_i L_i)^{-1} R_i L_i P_i = P_i,$$

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

$$\begin{aligned}
&|E_{j-1}v|^2 - |E_jv|^2 \\
&\qquad = |T_jE_{j-1}v + E_jv|^2 - |E_jv|^2 \\
&\qquad = (LT_jE_{j-1}v\,,\,T_jE_{j-1}v) + 2\,(LT_jE_{j-1}v\,,\,E_jv) \\
&\qquad = (LT_jE_{j-1}v\,,\,T_jE_{j-1}v) + 2\,(LT_jE_{j-1}v\,,\,(I-T_j)E_{j-1}v) \\
&\qquad = (LT_jE_{j-1}v\,,\,(2I-T_j)E_{j-1}v) \\
&\qquad \geq (2-\Lambda)\,(LT_jE_{j-1}v\,,\,E_{j-1}v)\,.
\end{aligned}$$

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

$$\begin{aligned}
|v|^2 - |E_Nv|^2 &= \sum_{j=1}^{N} \left\{|E_{j-1}v|^2 - |E_jv|^2\right\} \\
&\geq (2-\Lambda) \sum_{j=1}^{N} (LT_jE_{j-1}v\,,\,E_{j-1}v)\,.
\end{aligned} \tag{III.2.3}$$

Schreibe $v = \sum_{i=1}^{N} v_i$. Da $(R_iL_i)^{-1}T_i$ die orthogonale Projektion bzgl. $(L\cdot\,,\,\cdot)$ ist, gilt wegen der Definition von λ

$$\begin{aligned}
|v|^2 = (Lv, v) &= \sum_{i=1}^{N} (Lv\,,\,v_i) = \sum_{i=1}^{N} \left(L(R_iL_i)^{-1}T_iv\,,\,v_i\right) \\
&\leq \lambda^{-1} \sum_{i=1}^{N} (LT_iv\,,\,v_i)\,.
\end{aligned}$$

Hieraus folgt mit der Chauchy-Schwarzschen Ungleichung und der ersten Voraussetzung des Satzes

$$\begin{aligned}
|v|^2 &\leq \lambda^{-1} \sum_{i=1}^{N} |T_iv|\,|v_i| \leq \lambda^{-1} \left\{\sum_{i=1}^{N} |T_iv|^2\right\}^{\frac{1}{2}} \left\{\sum_{i=1}^{N} |v_i|^2\right\}^{\frac{1}{2}} \\
&\leq \lambda^{-1}K_0\,|v| \left\{\sum_{i=1}^{N} |T_iv|^2\right\}^{\frac{1}{2}}
\end{aligned}$$

und somit

$$|v| \leq \lambda^{-1}K_0 \left\{\sum_{i=1}^{N} |T_iv|^2\right\}^{\frac{1}{2}}. \tag{III.2.4}$$

Mit einem Teleskop-Summen-Argument ergibt sich wegen $E_0 = I$ und der Definition von Λ

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

Hieraus folgt mit der zweiten Voraussetzung des Satzes

$$\begin{aligned}\sum_{i=1}^{N} |T_i v|^2 &\le \Lambda \sum_{i=1}^{N} \sum_{j=1}^{i} (LT_i v\,,\, T_j E_{j-1} v) \\ &\le \Lambda K_1 \left\{ \sum_{i=1}^{N} |T_i v|^2 \right\}^{\frac{1}{2}} \left\{ \sum_{j=1}^{N} |T_j E_{j-1} v|^2 \right\}^{\frac{1}{2}}\end{aligned}$$

und somit

$$\sum_{i=1}^{N} |T_i v|^2 \le \Lambda^2 K_1^2 \sum_{j=1}^{N} |T_j E_{j-1} v|^2\,.$$

Wegen der Definition von Λ ist

$$\begin{aligned}\sum_{j=1}^{N} |T_j E_{j-1} v|^2 &= \sum_{j=1}^{N} (LT_j E_{j-1} v\,,\, T_j E_{j-1} v) \\ &\le \Lambda \sum_{j=1}^{N} (LT_j E_{j-1} v\,,\, E_{j-1} v)\,.\end{aligned}$$

Insgesamt erhalten wir somit

$$\sum_{i=1}^{N} |T_i v|^2 \le \Lambda^3 K_1^2 \sum_{j=1}^{N} (LT_j E_{j-1} v\,,\, E_{j-1} v)\,. \tag{III.2.5}$$

Aus (III.2.3), (III.2.4) und (III.2.5) folgt

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

und damit (III.2.2). □

Bemerkung III.2.11. (1) Aus der Cauchy-Schwarzschen Ungleichung folgt die zweite Voraussetzung von Satz III.2.10 stets mit $K_1 \leq N$. Für Mehrgitterverfahren bedeutet die Abschätzung $K_1 \leq N$, dass die Konvergenzrate von Algorithmus III.2.2 sich schlimmstenfalls wie $1/\ln|h_N|$ verhält. Häufig kann man jedoch eine *verschärfte Cauchy-Schwarzsche Ungleichung*

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

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

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

linear, stetig und surjektiv. Aus dem Satz von der offenen Abbildung [2, §6.6] folgt daher, dass die erste Voraussetzung von Satz III.2.10 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 Sobolev-Räume ist dies im Rahmen von Mehrgitterverfahren für Finite Element Diskretisierungen in der Tat möglich.

Bemerkung III.2.12 (Konvektions-Diffusionsgleichung). Bei Konvektions-Diffusionsgleichungen geht die Symmetrie der Bilinearform B und der Steifigkeitsmatrix verloren. Die Nachteile direkter Lösungsverfahren bleiben davon unberührt. Ebenso ändert sich die Kondition von $\mathrm{O}\left((\min_{K\in\mathcal{T}} h_K)^{-2}\right)$ der Steifigkeitmatrix nicht, so dass klassische iterative Verfahren wie das Gauß-Seidel Verfahren nach wie vor nicht geeignet sind. Wegen der fehlenden Symmetrie kann das CG-Verfahren nicht mehr angewandt werden. Stattdessen muss man auf das BiCG-Stab-Verfahren und seine vorkonditionierten Varianten [14, Algorithm IV.3.4] ausweichen. Beim Mehrgitterverfahren ist ein für unsymmetrische Probleme geeigneter Glätter, wie z.B. die Jacobi-Iteration für

die Normalengleichungen, zu benutzen. Die theoretischen Konvergenzresultate sind etwas schwächer, indem sie gitterunabhängige Konvergenzraten nur für hinreichend viele Glättungsschritte liefern. In der Praxis beobachtet man bei 1 bis 2 Glättungsschritten für symmetrische Probleme Konvergenzraten von etwa 0.1 und für unsymmetrische Probleme von etwa 0.4 bis 0.5.

III.3. A posteriori Fehlerabschätzungen

In §II.3 (S. 52) haben wir sog. *a priori Fehlerabschätzungen* gezeigt, d.h., wir haben den Fehler der Finite Element Diskretisierung abgeschätzt, ohne dazu das entsprechende diskrete Problem zu lösen. Die Fehlerabschätzungen sind alle *asymptotisch*, d.h., sie sagen etwas über die Konvergenzgeschwindigkeit des Fehlers aus, wenn die Elementgrößen gegen Null streben. Für ein gegebenes Problem und eine gegebene Unterteilung sind sie aber unbrauchbar, da sie u.a. von Normen der unbekannten Lösung der Differentialgleichung abhängen. Für die Praxis stellt sich aber natürlich die Frage nach dem tatsächlichen Fehler der berechneten Finite Element Lösung. Zudem will man i.a. eine bestimmte Genauigkeit mit minimalem Aufwand, d.h. möglichst wenigen Elementen erreichen. Diese Fragen werden durch *a posteriori Fehlerabschätzungen*, die wir in diesem Abschnitt betrachten, und *adaptive Gitterverfeinerungstechniken*, die wir im nächsten Abschnitt behandeln, gelöst.

Wir betrachten zunächst die Reaktions-Diffusionsgleichung mit homogenen Dirichlet-Randbedingungen (II.3.1) (S. 52) und ihre Diskretisierung (II.3.2) (S. 52). Um die Darstellung zu vereinfachen, nehmen wir zudem an, dass die Koeffizienten A und α konstant sind. In Bemerkung III.3.11 am Ende dieses Abschnittes gehen wir auf variable Koeffizienten und Konvektions-Diffusionsgleichungen ein.

Zunächst müssen wir einige zusätzliche Notationen einführen. Jede Kante ($d = 2$) bzw. Seitenfläche ($d = 3$) $E \in \mathcal{E}_\Omega$ im Innern von Ω ist die gemeinsame Grenzfläche von genau zwei Elementen, die wir mit K_{E1} und K_{E2} bezeichnen. Jedem $E \in \mathcal{E}_\Omega$ ordnen wir einen Einheitsvektor $\mathbf{n}_E$ zu, der senkrecht auf E steht und von K_{E1} nach K_{E2} zeigt. Für jede stückweise stetige Funktionen φ bezeichnen wir mit $\mathbb{J}_E(\varphi)$ den Sprung von φ über E in Richtung $\mathbf{n}_E$, d.h.

$$\mathbb{J}_E(\varphi)(x) = \lim_{t\to 0+} \varphi(x + t\mathbf{n}_E) - \lim_{t\to 0+} \varphi(x - t\mathbf{n}_E) \quad \forall x \in E.$$

Der Sprung $\mathbb{J}_E(\varphi)$ hängt von der Orientierung von $\mathbf{n}_E$ ab. Größen der Form $\mathbb{J}_E(\mathbf{n}_E \cdot \nabla\varphi)$ sind aber von der Orientierung von $\mathbf{n}_E$ unabhängig. Für $E \in \mathcal{E}$ bezeichnet h_E den Durchmesser von E. Wegen der Regularitätsbedingung von §II.1 (S. 35) 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 $\mathrm{C}_\mathcal{T} = \max_K \frac{h_K}{\rho_K}$ abgeschätzt werden.

Schließlich benötigen wir verschiedene Umgebungen von Elementeckpunkten $z \in \mathcal{N}$, Kanten bzw. Seitenflächen $E \in \mathcal{E}_\Omega$ und Elementen $K \in \mathcal{T}$ (vgl. Abbildung III.3.1):

$$\omega_K = \bigcup_{K \cap K' \in \mathcal{E}} K', \qquad \widetilde{\omega}_K = \bigcup_{K \cap K' \neq \emptyset} K',$$
$$\omega_E = \bigcup_{E \subset \partial K'} K', \qquad \widetilde{\omega}_E = \bigcup_{E \cap K' \neq \emptyset} K',$$
$$\omega_z = \operatorname{supp} \lambda_z = \bigcup_{z \in K'} K',$$

Dabei ist $\lambda_z \in S^{1,0}(\mathcal{T})$ die nodale Basisfunktion zu z, und $K \cap K' \in \mathcal{E}$ bedeutet, dass K und K' eine Kante bzw. Seitenfläche gemeinsam haben.

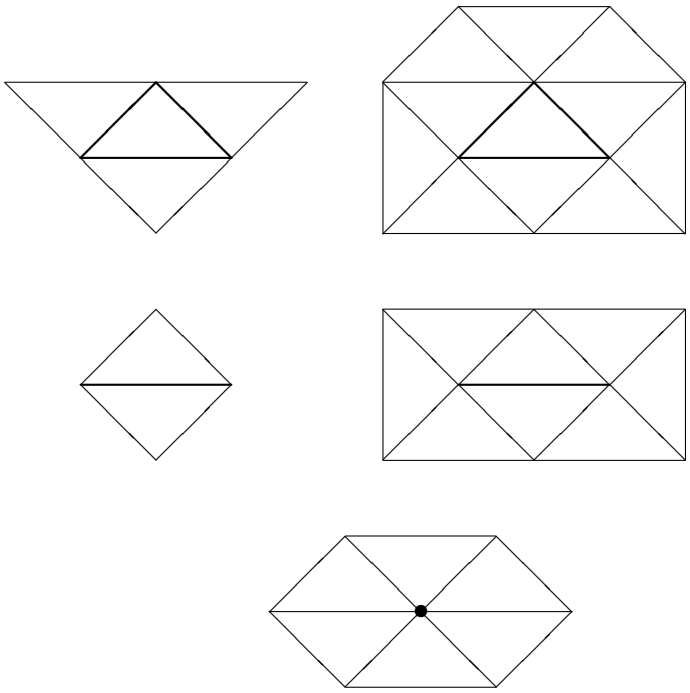

ABBILDUNG III.3.1. Gebiete ω_K, $\widetilde{\omega}_K$, ω_E, $\widetilde{\omega}_E$ und ω_z

Im Folgenden bezeichnen $u \in H^1_0(\Omega)$ und $u_\mathcal{T} \in S^{k,0}_0(\mathcal{T})$ die schwache Lösung der Reaktions-Diffusionsgleichung (II.3.1) (S. 52) und die Lösung des diskreten Problems (II.3.2) (S. 52). Die Abbildung

$$R : v \mapsto \ell(v) - B(u_\mathcal{T}, v)$$

definiert ein stetiges lineares Funktional auf $H^1_0(\Omega)$, das sog. *Residuum*. Zwischen dem Fehler und dem Residuum besteht offensichtlich die Beziehung

$$R(v) = B(u - u_\mathcal{T}, v) \quad \forall v \in H^1_0(\Omega). \tag{III.3.1}$$

Während der Fehler nicht berechenbar ist, ist das Residuum für jedes $v \in H_0^1(\Omega)$ berechenbar.
Wegen der Koerzivität und Stetigkeit der Bilinearform B gilt für jedes $v \in H_0^1(\Omega)$

$$\beta \|u - u_{\mathcal{T}}\|_1^2 \leq B(u - u_{\mathcal{T}}, u - u_{\mathcal{T}}),\ B(u - u_{\mathcal{T}}, v) \leq \mathcal{B} \|u - u_{\mathcal{T}}\|_1 \|v\|_1 .$$

Hieraus folgt die *Äquivalenz von Fehler und Residuum*

$$\beta \|u - u_{\mathcal{T}}\|_1 \leq \sup_{v \in H_0^1(\Omega)\setminus\{0\}} \frac{R(v)}{\|v\|_1} \leq \mathcal{B} \|u - u_{\mathcal{T}}\|_1 . \tag{III.3.2}$$

Obwohl das Residuum berechenbar ist, ist seine Dualnorm $\sup_v \frac{R(v)}{\|v\|_1}$ nicht berechenbar, da dies die Lösung eines unendlich dimensionalen Variationsproblems erfordern würde und damit genauso aufwändig wäre wie die Lösung der Reaktions-Diffusionsgleichung. Alle a posteriori Fehlerschätzer ersetzen diese Dualnorm durch leicht berechenbare untere und obere Schranken.

Bevor wir derartige Schranken herleiten, halten wir noch einige wichtige strukturelle Eigenschaften des Residuums fest.
Aus (III.3.1), der Definition von B und der Cauchy-Schwarzschen Ungleichung folgt für jedes $v \in H_0^1(\Omega)$

$$|R(v)| \leq \mathcal{B} \|u - u_{\mathcal{T}}\|_{1;\operatorname{supp} v} \|v\|_{1;\operatorname{supp} v} . \tag{III.3.3}$$

Da $S_0^{k,0}(\mathcal{T}) \subset H_0^1(\Omega)$ ist, haben wir die *Galerkin-Orthogonalität* des Fehlers

$$R(v_{\mathcal{T}}) = B(u - u_{\mathcal{T}}, v_{\mathcal{T}}) = 0 \quad \forall v_{\mathcal{T}} \in S_0^{k,0}(\mathcal{T}). \tag{III.3.4}$$

Schließlich wollen wir eine Darstellung des Residuums angeben, die praktisch handhabbar ist. Sei dazu $v \in H_0^1(\Omega)$ beliebig. Anwenden des Gaußschen Integralsatzes auf jedem Element $K \in \mathcal{T}$ liefert mit der äußeren Normalen $\mathbf{n}_K$ zu K

$$\begin{aligned}
R(v) &= \ell(v) - B(u_{\mathcal{T}}, v) \\
&= \int_\Omega fv - \int_\Omega \{\nabla u_{\mathcal{T}} \cdot A\nabla v + \alpha u_{\mathcal{T}} v\} \\
&= \sum_{K\in\mathcal{T}} \left\{ \int_K fv - \int_K \nabla u_{\mathcal{T}} \cdot A\nabla v - \int_K \alpha u_{\mathcal{T}} v \right\} \\
&= \sum_{K\in\mathcal{T}} \left\{ \int_K fv + \int_K \nabla\cdot(A\nabla u_{\mathcal{T}})v - \int_{\partial K} \mathbf{n}_K \cdot A\nabla u_{\mathcal{T}} v - \int_K \alpha u_{\mathcal{T}} v \right\} \\
&= \sum_{K\in\mathcal{T}} \int_K \{f + \nabla\cdot(A\nabla u_{\mathcal{T}}) - \alpha u_{\mathcal{T}}\}\, v - \sum_{E\in\mathcal{E}_\Omega} \int_E \mathbb{J}_E(\mathbf{n}_E \cdot A\nabla u_{\mathcal{T}})v.
\end{aligned}$$

Dies liefert die sog. *L^2-Darstellung des Residuums*

$$R(v) = \sum_{K\in\mathcal{T}} \int_K r_K v + \sum_{E\in\mathcal{E}_\Omega} \int_E j_E v \quad \forall v \in H_0^1(\Omega) \tag{III.3.5a}$$

mit

$$r_K = f + \nabla \cdot (A \nabla u_{\mathcal{T}}) - \alpha u_{\mathcal{T}}, \tag{III.3.5b}$$

$$j_E = -\mathbb{J}_E(\mathbf{n}_E \cdot A \nabla u_{\mathcal{T}}). \tag{III.3.5c}$$

Im folgenden sei $v \in H^1_0(\Omega)$ mit $\|v\|_1 = 1$ beliebig, aber fest. Wegen der Galerkin Orthogonalität (III.3.4) können wir auf der rechten Seite von (III.3.5a) ein beliebiges Element $v_{\mathcal{T}} \in S^{k,0}_0(\mathcal{T})$ von v subtrahieren. Aus der Cauchy-Schwarzschen Ungleichung für Integrale folgt dann

$$R(v) \leq \sum_{K \in \mathcal{T}} \|r_k\|_K \, \|v - v_{\mathcal{T}}\|_K + \sum_{E \in \mathcal{E}} \|j_E\|_E \, \|v - v_{\mathcal{T}}\|_E \,. \tag{III.3.6}$$

Als nächstes müssen wir $v_{\mathcal{T}} \in S^{k,0}_0(\mathcal{T})$ geschickt wählen, so dass die Normen $\|v - v_{\mathcal{T}}\|_K$ und $\|v - v_{\mathcal{T}}\|_E$ möglichst klein werden. Wegen der Ergebnisse von §II.2 (S. 40) sind wir versucht, $v_{\mathcal{T}} = I_{\mathcal{T}} v$ zu wählen. Dies ist aber nicht möglich, da v nur aus $H^1_0(\Omega)$ ist und damit $I_{\mathcal{T}} v$ gar nicht definiert ist. Stattdessen konstruieren wir einen sog. *Quasi-Interpolationsoperator*.

Definition III.3.1 (Quasi-Interpolationsoperator). Für jedes $z \in \mathcal{N}$ sei $\overline{\varphi}_z = \frac{1}{|\omega_z|} \int_{\omega_z} \varphi$ der Mittelwert von φ auf ω_z, wobei $|\omega_z|$ das d-dimensionale Lebesgue-Maß von ω_z ist. Dann ist der *Quasi-Interpolationsoperator* $\mathcal{J}_{\mathcal{T}} : H^1_0(\Omega) \to S^{1,0}_0(\mathcal{T})$ definiert durch

$$\mathcal{J}_{\mathcal{T}} \varphi = \sum_{z \in \mathcal{N}_\Omega} \overline{\varphi}_z \lambda_z.$$

Man beachte, dass $\mathcal{J}_{\mathcal{T}} \varphi$ für alle $\varphi \in L^1(\Omega)$ definiert ist und dass nur über die Elementeckpunkte im Innern von Ω summiert wird.

Wir wollen lokale Fehlerabschätzungen für $\mathcal{J}_{\mathcal{T}}$ beweisen. Dazu benötigen wir die folgenden beiden Hilfsergebnisse, die von eigenständigem Interesse sind.

Lemma III.3.2 (Poincarésche Ungleichung). *Es gibt eine von* $\mathrm{C}_{\mathcal{T}} = \max_K \frac{h_K}{\rho_K}$ *abhängige Konstante* c_P*, so dass für alle Elementeckpunkte* z*, alle Elemente* K *in* ω_z *und alle* $\varphi \in H^1(\omega_z)$ *gilt*

$$\|\varphi - \overline{\varphi}_z\|_{\omega_z} \leq c_P h_K \, \|\varphi\|_{1;\omega_z} \,.$$

Beweis. Aus der Poincaréschen Ungleichung, Satz I.2.21 (S. 28), und Bemerkung I.2.22(2) (S. 28) folgt für jedes $z \in \mathcal{N}$ und jedes $\varphi \in H^1(\omega_z)$

$$\|\varphi - \overline{\varphi}_z\|_{\omega_z} \leq c_{z1} \operatorname{diam}(\omega_z) \, \|\varphi\|_{1;\omega_z} \,.$$

Falls ω_z konvex ist, ist $c_{z1} = \frac{1}{\pi}$ [13, Proposition 3.10], andernfalls hängt c_{z1} von $\mathrm{C}_{\mathcal{T}}$ ab [13, Proposition 3.21]. Wegen der Regularitätsannahme an $\mathcal{T}$ gibt es weiter eine Konstante c_{z2} mit

$$\operatorname{diam}(\omega_z) \leq c_{z2} h_K \quad \forall K \subset \omega_z. \qquad \square$$

LEMMA III.3.3 (Spurungleichung). *Es gibt eine von* $\mathrm{C}_{\mathcal{T}} = \max_K \frac{h_K}{\rho_K}$ *abhängige Konstante* c_{tr}*, so dass für alle Elemente* K*, alle Kanten* ($d = 2$) *bzw. Seitenflächen* ($d = 3$) E *von* K *und alle* $\varphi \in H^1(K)$ *gilt*

$$\|\varphi\|_E \le c_{tr} \left\{ h_K^{-\frac{1}{2}} \|\varphi\|_K + h_K^{\frac{1}{2}} \|\varphi\|_{1;K} \right\}.$$

BEWEIS. Bezeichne mit $\widehat{E}$ die Kante ($d = 2$) bzw. Seitenfläche ($d = 3$) des Referenz-Elementes $\widehat{K}$ in der Hyperebene $\{x_d = 0\}$. Dann ist $\widehat{E}$ das Referenzelement in $\mathbb{R}^{d-1}$. Wegen des Spursatzes I.2.12 (S. 25) gibt es eine Konstante $\widehat{c}$, so dass für alle $\varphi \in H^1(\widehat{K})$ gilt

$$\|\varphi\|_{\widehat{E}} \le \widehat{c} \|\varphi\|_{1;\widehat{K}}.$$

Wähle die affine Transformation $F_K : \widehat{K} \to K$ so, dass $\widehat{E}$ auf E abgebildet wird. Bezeichne mit F_E die Restriktion von F_K auf $\widehat{E}$ und setze $\widehat{\varphi} = \varphi \circ F_K \in H^1(\widehat{K})$. Dann folgt aus Lemma II.2.5 (S. 50)

$$\|\varphi\|_E = |DF_E|^{\frac{1}{2}} \|\widehat{\varphi}\|_{\widehat{E}} \le \widehat{c} |DF_E|^{\frac{1}{2}} \|\widehat{\varphi}\|_{1;\widehat{K}}$$

und

$$\|\widehat{\varphi}\|_{1;\widehat{K}} = \left\{ \|\widehat{\varphi}\|_{\widehat{K}}^2 + |\widehat{\varphi}|_{1;\widehat{K}}^2 \right\}^{\frac{1}{2}} \le |DF_K|^{-\frac{1}{2}} \left\{ \|\varphi\|_K^2 + \|DF_K\|_{\mathcal{L}}^2 |\varphi|_{1;K}^2 \right\}^{\frac{1}{2}}.$$

Wegen der Regularitätsannahme an $\mathcal{T}$ ist $\frac{|DF_E|}{|DF_K|} = \frac{|E|\,|\widehat{K}|}{|\widehat{E}|\,|K|} \le c h_K^{-1}$ mit einer von $\mathrm{C}_{\mathcal{T}}$ abhängigen Konstanten. Hieraus folgt die Behauptung mit Lemma II.2.6 (S. 50). □

SATZ III.3.4 (Fehlerabschätzung für den Quasi-Interpolationsoperator). *Es gibt zwei Konstanten* c_1*,* c_2 *die nur von* $\mathrm{C}_{\mathcal{T}} = \max_K \frac{h_K}{\rho_K}$ *abhängen, so dass für alle* $v \in H_0^1(\Omega)$*,* $K \in \mathcal{T}$ *und* $E \in \mathcal{E}_\Omega$ *gilt*

$$\|v - \mathcal{J}_{\mathcal{T}} v\|_K \le c_1 h_K \|v\|_{1;\widetilde{\omega}_K},$$
$$\|v - \mathcal{J}_{\mathcal{T}} v\|_E \le c_2 h_E^{\frac{1}{2}} \|v\|_{1;\widetilde{\omega}_E}.$$

BEWEIS. *1. Schritt:* Sei $K \in \mathcal{T}$ ein Element, das keinen Eckpunkt auf dem Rand Γ hat. Dann ist $\sum_{z \in \mathcal{N}_K} \lambda_z = 1$ auf K. Hieraus folgt

$$\|v - \mathcal{J}_{\mathcal{T}} v\|_K = \left\| \sum_{z \in \mathcal{N}_K} \lambda_z (v - \overline{v}_z) \right\|_K \le \sum_{z \in \mathcal{N}_K} \|\lambda_z (v - \overline{v}_z)\|_K.$$

Wegen $\|\lambda_z\|_{\infty;K} \le 1$ und Lemma III.3.2 ist für jedes $z \in \mathcal{N}_K$

$$\|\lambda_z (v - \overline{v}_z)\|_K \le \|v - \overline{v}_z\|_K \le \|v - \overline{v}_z\|_{\omega_z} \le c_P h_K \|v\|_{1;\omega_z}.$$

Damit folgt insgesamt

$$\|v - \mathcal{J}_{\mathcal{T}} v\|_K \le \sum_{z \in \mathcal{N}_K} c_P h_K \|v\|_{1;\omega_z} \le c h_K \|v\|_{1;\widetilde{\omega}_K}.$$

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

$$\begin{aligned} v - \mathcal{J}_{\mathcal{T}} v &= \sum_{z \in \mathcal{N}_K} \lambda_z v - \sum_{z \in \mathcal{N}_K \cap \mathcal{N}_\Omega} \lambda_z \overline{v}_z \\ &= \sum_{z \in \mathcal{N}_K} \lambda_z (v - \overline{v}_z) + \sum_{z \in \mathcal{N}_K \setminus \mathcal{N}_\Omega} \lambda_z \overline{v}_z \end{aligned}$$

und somit

$$\|v - \mathcal{J}_{\mathcal{T}} v\|_K \leq \sum_{z \in \mathcal{N}_K} \|\lambda_z (v - \overline{v}_z)\|_K + \sum_{z \in \mathcal{N}_K \setminus \mathcal{N}_\Omega} \|\lambda_z \overline{v}_z\|_K .$$

Der erste Summand wurde bereits in Schritt 1 abgeschätzt. Sei also $z \in \mathcal{N}_K \setminus \mathcal{N}_\Omega$ ein Eckpunkt von K, der auf Γ liegt. Wegen $\|\lambda_z\|_{\infty;K} \leq 1$ ist

$$\|\lambda_z \overline{v}_z\|_K \leq |K|^{\frac{1}{2}} |\overline{v}_z| .$$

Da $z \in \Gamma$ ist, gibt es ein Element $K' \in \mathcal{T}$ und eine Kante bzw. Seitenfläche E' von K', so dass z ein Endpunkt von E' und $E' \subset \Gamma$ ist. Da v auf E' verschwindet, ist

$$|\overline{v}_z| = |E'|^{-\frac{1}{2}} \|\overline{v}_z\|_{E'} = |E'|^{-\frac{1}{2}} \|v - \overline{v}_z\|_{E'} .$$

Da $\overline{v}_z$ konstant ist, folgt aus Lemma III.3.3

$$\begin{aligned} \|v - \overline{v}_z\|_{E'} &\leq c_{tr} \left\{ h_{K'}^{-\frac{1}{2}} \|v - \overline{v}_z\|_{K'} + h_{K'}^{\frac{1}{2}} |v - \overline{v}_z|_{1;K'} \right\} \\ &= c_{tr} \left\{ h_{K'}^{-\frac{1}{2}} \|v - \overline{v}_z\|_{K'} + h_{K'}^{\frac{1}{2}} |v|_{1;K'} \right\} . \end{aligned}$$

Wegen der Regularität ist $\frac{|K|}{|E'|} \leq c h_K$ mit einer von $\mathrm{C}_{\mathcal{T}}$ abhängigen Konstanten. Zusammen mit Schritt 1 folgt aus diesen Abschätzungen die Behauptung für K.

3. Schritt: Sei $E \in \mathcal{E}$ eine Kante ($d = 2$) bzw. Seitenfläche ($d = 3$), die keinen Eckpunkt auf dem Rand Γ hat. Dann ist $\sum_{z \in \mathcal{N}_E} \lambda_z = 1$ auf E. Hieraus folgt wie in Schritt 1

$$\begin{aligned} \|v - \mathcal{J}_{\mathcal{T}} v\|_E &= \left\| \sum_{z \in \mathcal{N}_E} \lambda_z (v - \overline{v}_z) \right\|_E \leq \sum_{z \in \mathcal{N}_E} \|\lambda_z (v - \overline{v}_z)\|_E \\ &\leq \sum_{z \in \mathcal{N}_E} \|v - \overline{v}_z\|_E . \end{aligned}$$

Bezeichnet K_E ein Element, das E als Kante bzw. Seitenfläche hat, folgt aus Lemmata III.3.2 und III.3.3

$$\begin{aligned} \|v - \overline{v}_z\|_E &\leq c_{tr} \left\{ h_{K_E}^{-\frac{1}{2}} \|v - \overline{v}_z\|_{K_E} + h_{K_E}^{\frac{1}{2}} |v - \overline{v}_z|_{1;K_E} \right\} \\ &\leq c_P c_{tr} h_{K_E}^{\frac{1}{2}} \|v\|_{1;\omega_z} \\ &\leq c h_E^{\frac{1}{2}} \|v\|_{1;\widetilde{\omega}_E} . \end{aligned}$$

Aus diesen Abschätzungen folgt die Behauptung für E.
4. Schritt: Betrachte abschließend eine Kante bzw. Seitenfläche E, die einen Endpunkt auf dem Rand Γ hat. Dann ist auf E

$$v - \mathcal{J}_{\mathcal{T}} v = \sum_{z \in \mathcal{N}_E} \lambda_z (v - \overline{v}_z) + \sum_{z \in \mathcal{N}_E \setminus \mathcal{N}_\Omega} \lambda_z \overline{v}_z .$$

Der erste Summand wird wie in Schritt 3 abgeschätzt. Der zweite Summand wird mit den gleichen Methoden wie in Schritt 2 behandelt. Hieraus folgt dann die Behauptung für E. □

Wir greifen nun die Abschätzung (III.3.6) wieder auf. Wir setzen $v_{\mathcal{T}} = \mathcal{J}_{\mathcal{T}} v$, benutzen Satz III.3.4 und wenden die Cauchy-Schwarzsche Ungleichung für endliche Summen an:

$$\begin{aligned} R(v) &\le c_1 \sum_{K \in \mathcal{T}} h_K \|r_K\|_K \|v\|_{1;\widetilde{\omega}_K} + c_2 \sum_{E \in \mathcal{E}_\Omega} h_E^{\frac{1}{2}} \|j_E\|_E \|v\|_{1;\widetilde{\omega}_E} \\ &\le c_1 \left\{ \sum_{K \in \mathcal{T}} h_K^2 \|r_K\|_K^2 \right\}^{\frac{1}{2}} \left\{ \sum_{K \in \mathcal{T}} \|v\|_{1;\widetilde{\omega}_K}^2 \right\}^{\frac{1}{2}} \\ &\quad + c_2 \left\{ \sum_{E \in \mathcal{E}_\Omega} h_E \|j_E\|_E^2 \right\}^{\frac{1}{2}} \left\{ \sum_{E \in \mathcal{E}_\Omega} \|v\|_{1;\widetilde{\omega}_E}^2 \right\}^{\frac{1}{2}} \\ &\le c' \|v\|_1 \left\{ \sum_{K \in \mathcal{T}} h_K^2 \|r_K\|_K^2 + \sum_{E \in \mathcal{E}_\Omega} h_E \|j_E\|_E^2 \right\}^{\frac{1}{2}} . \end{aligned}$$

Dabei haben wir im letzten Schritt die Regularitätsbedingung an $\mathcal{T}$ ausgenutzt. Hieraus und aus (III.3.3) folgt insgesamt

$$\|u - u_{\mathcal{T}}\|_1 \le c\eta \tag{III.3.7a}$$

mit

$$\eta = \left\{ \sum_{K \in \mathcal{T}} \eta_K^2 \right\}^{\frac{1}{2}} , \tag{III.3.7b}$$

$$\eta_K = \left\{ h_K^2 \|r_K\|_K^2 + \frac{1}{2} \sum_{E \subset \partial K \setminus \Gamma} h_E \|j_E\|_E^2 \right\}^{\frac{1}{2}} \tag{III.3.7c}$$

und r_K, j_E wie in (III.3.5b) und (III.3.5c). Der Faktor $\frac{1}{2}$ vor der zweiten Summe in η_K berücksichtigt, dass bei Summation über alle Elemente jede innere Kante ($d = 2$) bzw. Seitenfläche ($d = 3$) doppelt gezählt wird.

Ungleichung (III.3.7a) ist eine *a posteriori Fehlerabschätzung*. Die Größen η und η_K in (III.3.7b) und (III.3.7c) können aus den gegebenen Daten f, A, α und der berechneten numerischen Lösung $u_{\mathcal{T}}$ a posteriori berechnet werden. Sie heißen daher auch *a posteriori Fehlerschätzer*.

Ungleichung (III.3.7a) zeigt, dass der Fehlerschätzer *zuverlässig* ist, d.h., ist $\eta \leq \varepsilon$, so ist der Fehler ebenfalls (bis auf einen Faktor) nicht größer als ε. Die Kontrolle von η erlaubt also, eine vorgegebene Toleranz zu erreichen. Um dies mit einem minimalen Aufwand zu erreichen, reicht die obere Schranke (III.3.7) nicht aus. Wir müssen zusätzlich garantieren, dass η den Fehler nicht überschätzt und die räumliche Verteilung des Fehlers auch richtig widerspiegelt. Dies nennt man *Effizienz*. Sie ist gegeben, wenn es gelingt, den Fehler auch nach unten durch η abzuschätzen.

Für den Beweis der unteren Schranken, benötigen wir sog. *Blasenfunktionen*.

DEFINITION III.3.5 (Blasenfunktionen). Für jedes Element $K \in \mathcal{T}$ und jede Kante ($d = 2$) bzw. Seitenfläche ($d = 3$) $E \in \mathcal{E}$ bezeichne mit $\mathcal{N}_K$ und $\mathcal{N}_E$ die Menge der Eckpunkte von K bzw. E und setze

$$\psi_K = \alpha_K \prod_{z \in \mathcal{N}_K} \lambda_z, \quad \psi_E = \alpha_E \prod_{z \in \mathcal{N}_E} \lambda_z$$

mit

$$\alpha_K = \begin{cases} (d+1)^{d+1} & \text{falls } K \text{ ein } d\text{-Simplex}, \\ \left(2^d\right)^{2^d} & \text{falls } K \text{ ein } d\text{-Parallelepiped}, \end{cases}$$

$$\alpha_E = \begin{cases} 2^2 & \text{falls } E \text{ eine Kante}, \\ d^d & \text{falls } E \text{ ein } (d-1)\text{-Simplex}, \\ \left(2^{d-1}\right)^{2^{d-1}} & \text{falls } E \text{ ein } (d-1)\text{-Parallelepiped}, \end{cases}$$

Die folgenden beiden Lemmata geben einige Eigenschaften der Blasenfunktionen an.

LEMMA III.3.6 (Eigenschaften der Blasenfunktionen). *Für alle $K \in \mathcal{T}$ und alle $E \in \mathcal{E}$ gilt*

$$\begin{aligned} &\psi_K \in C_0(K), && \psi_E \in C_0(\omega_E), \\ &\psi_K \in R_{d+1}(K), && \psi_E|_K \in R_d(K),\ \forall K \subset \omega_E, \\ &\psi_K \geq 0 \quad \textit{auf } K, && \psi_E \geq 0 \quad \textit{auf } \omega_E, \\ &\psi_K = 0 \quad \textit{auf } \partial K, && \psi_E = 0 \quad \textit{auf } \partial\omega_E, \\ &\max_{x \in K} \psi_K(x) = 1, && \max_{x \in E} \psi_E(x) = 1. \end{aligned}$$

BEWEIS. Die ersten vier Eigenschaften von ψ_K und ψ_E folgen aus den Eigenschaften der nodalen Basisfunktionen λ_z. Die Aussagen über den Maximalwert der Blasenfunktionen folgt aus der Beobachtung, dass diese aus Symmetriegründen ihr Maximum im Schwerpunkt von K bzw. E annehmen, dass dort die beteiligten Basisfunktionen λ_z alle den Wert $(\#\{z \in \mathcal{N}_K\})^{-1}$ bzw. $(\#\{z \in \mathcal{N}_E\})^{-1}$ haben und dass ein d-Simplex und ein d-Parallelepiped $d+1$ bzw. 2^d Eckpunkte hat. □

LEMMA III.3.7 (Inverse Abschätzungen). *Für jedes Element $K \in \mathcal{T}$, jede Kante ($d = 2$) bzw. Seitenfläche ($d = 3$) $E \in \mathcal{E}$, jedes $k \in \mathbb{N}$, jedes $v \in R_k(K)$ und jedes $\sigma \in R_k(E)$ gilt*

$$\begin{aligned}
c_3 \|v\|_K &\leq \left\{ \int_K \psi_K v^2 \right\}^{\frac{1}{2}} \leq \|v\|_K , \\
c_4 h_K^{-1} \|\psi_K v\|_K &\leq |\psi_K v|_{1;K} \leq c_5 h_K^{-1} \|\psi_K v\|_K , \\
c_6 \|\sigma\|_E &\leq \left\{ \int_E \psi_E \sigma^2 \right\}^{\frac{1}{2}} \leq \|\sigma\|_E , \\
c_7 h_E^{-1} \|\psi_E \sigma\|_{\omega_E} &\leq |\psi_E \sigma|_{1;\omega_E} \leq c_8 h_E^{-1} \|\psi_E \sigma\|_{\omega_E} , \\
\|\psi_E \sigma\|_{\omega_E} &\leq c_9 h_E^{\frac{1}{2}} \|\sigma\|_E .
\end{aligned}$$

Die Konstanten $c_3, \ldots, c_9$ hängen nur von k und $\mathrm{C}_{\mathcal{T}} = \max_K \frac{h_K}{\rho_K}$ ab.

BEWEIS. Die obere Schranke in der ersten Abschätzung folgt aus der Cauchy-Schwarzschen Ungleichung.
Wie man leicht nachrechnet, definiert $w \mapsto \left\{ \int_{\widehat{K}} \psi_K \circ F_K w^2 \right\}^{\frac{1}{2}}$ eine Norm auf $\widehat{R}_k$. Da auf endlich dimensionalen Räumen alle Normen äquivalent sind, gibt es eine Konstante $\widehat{c}$ mit

$$\widehat{c} \|w\|_{\widehat{K}} \leq \left\{ \int_{\widehat{K}} \psi_K \circ F_K w^2 \right\}^{\frac{1}{2}} \quad \forall w \in \widehat{R}_k.$$

Diese Abschätzung und Lemma II.2.5 (S. 50) beweisen die untere Schranke der ersten Abschätzung.
Da $\psi_K \circ F_K$ in den Eckpunkten von $\widehat{K}$ verschwindet, wird durch $w \mapsto |\psi_K \circ F_K w|_{1;\widehat{K}}$ eine Norm auf $\widehat{R}_k$ definiert. Mithin gibt es zwei Konstanten $\widehat{c}_1$ und $\widehat{c}_2$ mit

$$\widehat{c}_1 \|\psi_K \circ F_K w\|_{\widehat{K}} \leq |\psi_K \circ F_K w|_{1;\widehat{K}} \leq \widehat{c}_2 \|\psi_K \circ F_K w\|_{\widehat{K}} \quad \forall w \in \widehat{R}_k.$$

Hieraus und Lemmata II.2.5 (S. 50) und II.2.6 (S. 50) folgt die zweite Abschätzung.
Die Abschätzungen für σ folgen mit den gleichen Argumenten wie diejenigen für v. □

Bezeichne mit $f_{\mathcal{T}}$ irgendeine, im Folgenden feste Finite Element Approximation an f, z.B. die L^2-Projektion auf die stückweise konstanten Funktionen $S^{0,-1}(\mathcal{T})$. Sei $K \in \mathcal{T}$ beliebig. Setze

$$w_K = \psi_K (f_{\mathcal{T}} + \nabla \cdot (A \nabla u_{\mathcal{T}}) - \alpha u_{\mathcal{T}}).$$

Wegen Lemma III.3.7 ist

$$c_3^2 \|f_{\mathcal{T}} + \nabla \cdot (A \nabla u_{\mathcal{T}}) - \alpha u_{\mathcal{T}}\|_K^2 \leq \int_K w_K (f_{\mathcal{T}} + \nabla \cdot (A \nabla u_{\mathcal{T}}) - \alpha u_{\mathcal{T}}).$$

Setzen wir w_K als Testfunktion v in (III.3.5a) ein und berücksichtigen, dass w_K auf ∂K und außerhalb von K verschwindet, so erhalten wir wegen (III.3.5b), (III.3.1) und (III.3.3)

$$\begin{aligned}
&\int_K w_K(f_\mathcal{T} + \nabla\cdot(A\nabla u_\mathcal{T}) - \alpha u_\mathcal{T}) \\
&\quad = \int_K w_K r_K + \int_K w_K(f_\mathcal{T} - f) \\
&\quad = B(u - u_\mathcal{T}, w_K) + \int_K w_K(f_\mathcal{T} - f) \\
&\quad \le \mathcal{B}\,\|u - u_\mathcal{T}\|_{1;K}\,\|w_K\|_{1;K} + \|f_\mathcal{T} - f\|_K\,\|w_K\|_K\,.
\end{aligned}$$

Offensichtlich ist

$$\|w_K\|_K \le \|f_\mathcal{T} + \nabla\cdot(A\nabla u_\mathcal{T}) - \alpha u_\mathcal{T}\|_K\,.$$

Aus Lemma III.3.7 folgt weiter

$$\begin{aligned}
\|w_K\|_{1;K} &= \left\{\|w_K\|_K^2 + |w_K|_{1;K}^2\right\}^{\frac12} \le \left\{1 + c_5^2 h_K^{-2}\right\}^{\frac12}\|w_K\|_K \\
&\le c h_K^{-1}\,\|f_\mathcal{T} + \nabla\cdot(A\nabla u_\mathcal{T}) - \alpha u_\mathcal{T}\|_K\,.
\end{aligned}$$

Aus diesen Abschätzungen und der Dreiecksungleichung ergibt sich

$$h_K\,\|r_K\|_K \le c\left\{\|u - u_\mathcal{T}\|_{1;K} + h_K\,\|f - f_\mathcal{T}\|_K\right\}. \tag{III.3.8}$$

Sei nun $E \in \mathcal{E}_\Omega$ eine Kante ($d = 2$) bzw. Seitenfläche ($d = 3$). Setze

$$w_E = \psi_E j_E.$$

Wegen Lemma III.3.7 ist

$$c_6^2\,\|j_E\|_E^2 \le \int_E w_E j_E.$$

Setzen wir w_E als Testfunktion v in (III.3.5a) ein und berücksichtigen, dass w_E auf $\partial\omega_E$ und außerhalb ω_E verschwindet, so folgt wegen (III.3.1) und (III.3.3) mit der Cauchy-Schwarzschen Ungleichung

$$\begin{aligned}
\int_E w_E j_E &= \sum_{K\subset\omega_E}\int_K r_K w_E - B(u - u_\mathcal{T}, w_E) \\
&\le \left\{\sum_{K\subset\omega_E}\|r_K\|_K^2\right\}^{\frac12}\|w_E\|_{\omega_E} + \mathcal{B}\,\|u - u_\mathcal{T}\|_{1;\omega_E}\,\|w_E\|_{1;\omega_E}
\end{aligned}$$

Wegen Lemma III.3.7 ist

$$\|w_E\|_{\omega_E} \le c_9 h_E^{\frac12}\,\|w_E\|_E \le c_9 h_E^{\frac12}\,\|j_E\|_E$$

und

$$\|w_E\|_{1;\omega_E} \le c h_E^{-1}\,\|w_E\|_{\omega_E} \le c' h_E^{-\frac12}\,\|j_E\|_E\,.$$

Aus diesen Abschätzungen und (III.3.8) ergibt sich insgesamt

$$h_E^{\frac{1}{2}} \|j_E\|_E \leq c \left\{ \|u - u_{\mathcal{T}}\|_{1;\omega_E} + h_E \|f - f_{\mathcal{T}}\|_{\omega_E} \right\}. \tag{III.3.9}$$

Aus den Abschätzungen (III.3.8), (III.3.9) und der Definition (III.3.7c) erhalten wir folgende *lokale untere Fehlerschranke*

$$\eta_K \leq c \left\{ \|u - u_{\mathcal{T}}\|_{1;\omega_K} + h_K \|f - f_{\mathcal{T}}\|_{\omega_K} \right\}. \tag{III.3.10}$$

Summation über alle Dreiecke liefert zudem die *globale untere Fehlerschranke*

$$\eta \leq c \left\{ \|u - u_{\mathcal{T}}\|_1 + \left\{ \sum_{K \in \mathcal{T}} h_K^2 \|f - f_{\mathcal{T}}\|_K^2 \right\}^{\frac{1}{2}} \right\}. \tag{III.3.11}$$

In beiden Abschätzungen sind die $f - f_{\mathcal{T}}$-Terme Störterme, die unter zusätzlichen Regularitätsannahmen an f von höherer Ordnung sind und die zudem a priori allein aus der Kenntnis der Daten kontrolliert werden können, ohne eine Differentialgleichung oder ein diskretes Problem zu lösen.

Wir fassen unsere Ergebnisse zusammen:

Satz III.3.8 (A posteriori Fehlerabschätzung). *Bezeichne mit $u \in H_0^1(\Omega)$ und $u_{\mathcal{T}} \in S_0^{k,0}(\mathcal{T})$ die schwache Lösung der Reaktions-Diffusionsgleichung (II.3.1) (S. 52) und ihre Finite Element Approximation (II.3.2) (S. 52). Die Koeffizienten A und α seien konstant und $f_{\mathcal{T}} \in S^{k-1}(\mathcal{T})$ sei irgendeine Finite Element Approximation an f. Für jedes Element $K \in \mathcal{T}$ definiere den Fehlerschätzer η_K durch*

$$\eta_K = \left\{ h_K^2 \|f + \nabla \cdot (A \nabla u_{\mathcal{T}}) - \alpha u_{\mathcal{T}}\|_K^2 + \frac{1}{2} \sum_{E \subset \partial K \setminus \Gamma} h_E \|\mathbb{J}_E(\mathbf{n}_E \cdot A \nabla u_{\mathcal{T}})\|_E^2 \right\}^{\frac{1}{2}}. \tag{III.3.12}$$

Dann gibt es drei Konstanten C_1, C_2 und C_3, die nur von k und $C_{\mathcal{T}} = \max_K \frac{h_K}{\rho_K}$ abhängen, so dass die folgenden a posteriori Fehlerabschätzungen gelten

$$\|u - u_{\mathcal{T}}\|_1 \leq C_1 \left\{ \sum_{K \in \mathcal{T}} \eta_K^2 \right\}^{\frac{1}{2}},$$

$$\eta_K \leq C_2 \left\{ \|u - u_{\mathcal{T}}\|_{1;\omega_K} + h_K \|f - f_{\mathcal{T}}\|_{\omega_K} \right\},$$

$$\left\{ \sum_{K \in \mathcal{T}} \eta_K^2 \right\}^{\frac{1}{2}} \leq C_3 \left\{ \|u - u_{\mathcal{T}}\|_1 + \left\{ \sum_{K \in \mathcal{T}} h_K^2 \|f - f_{\mathcal{T}}\|_K^2 \right\}^{\frac{1}{2}} \right\}.$$

Bemerkung III.3.9 (Struktur der Abschätzungen). (1) Die Größen r_K und j_E aus (III.3.5b) und (III.3.5c) heißen das *Elementresiduum* und das *Kantenresiduum*. Das Elementresiduum ist elementweise das Residuum der Finite Element Lösung bzgl. der starken Form der Differentialgleichung. Diese Größe ist global nicht definiert, da die diskrete Lösung nicht in $H^2(\Omega)$ enthalten ist, wohl aber elementweise, da auf den einzelnen Elementen die diskrete Lösung als Polynom beliebig oft differenzierbar ist. Das Kantenresiduum ist der Sprung des Spuroperators, der beim Übergang von der starken zur schwachen Form der Differentialgleichung mittels partieller Integration auftritt. Diese Struktur der Residuen überträgt sich auf allgemeine elliptische Differentialgleichungen.
(2) Die Äquivalenz (III.3.2) von Fehler und Residuum ist eine strukturelle Eigenschaft des Variationsproblems und völlig unabhängig von der speziellen Diskretisierung. Dementsprechend können die Konstanten in dieser Ungleichung nicht durch die Wahl der Diskretisierung beeinflusst werden. Physikalisch beschreiben sie, wie empfindlich die Lösung der Differentialgleichung auf Störungen der Daten reagiert.
(3) Die L^2-Darstellung (III.3.5) des Residuums ist ebenfalls eine strukturelle Eigenschaft der Differentialgleichung. Sie gilt für alle Systeme in Divergenzform.
(4) Die Galerkin-Orthogonalität ist nur eine technische Eigenschaft, die die Herleitung der oberen Fehlerschranken wesentlich vereinfacht. Für die SUPG-Diskretisierung (II.3.4) (S. 54) der Konvektions-Diffusionsgleichung z.B. ist sie verletzt. Dennoch kann man für diese Diskretisierung und Differentialgleichung Abschätzungen wie in Satz III.3.8 beweisen [**13**, §4.4].
(5) Die obere Schranke (III.3.7) ist global, da sie den inversen Differentialoperator, der ein globaler Operator ist, benötigt. Die untere Schranke (III.3.10) dagegen ist lokal, da sie nur den Differentialoperator benutzt, der ein lokaler Operator ist.

Bemerkung III.3.10 (Zuverlässigkeit, Effizienz, Robustheit). Die obere Schranke (III.3.7) beweist die *Zuverlässigkeit* des Fehlerschätzers; ist der Schätzer unterhalb einer Toleranz, gilt dies auch für den Fehler. Die untere Schranke (III.3.10) beweist die *Effizienz* des Fehlerschätzers; ist der Schätzer oberhalb einer Toleranz, gilt dies auch für den Fehler. Beide Schranken und die lokale Natur der unteren Schranke zusammen garantieren, dass der adaptive Gitterverfeinerungs-Prozess III.4.1 eine Näherungslösung der Differentialgleichung mit vorgegebener Toleranz und minimalem Aufwand liefert. Das Produkt C_1C_3 der Konstanten aus Satz III.3.8 ist ein Maß für die Güte des Fehlerschätzers ähnlich zur Kondition einer Matrix. Wegen (III.3.2) ist diese Größe proportional zu $\beta^{-1}\mathcal{B}$. Falls die Differentialgleichung und damit β und $\mathcal{B}$ von Parametern abhängen, sollte C_1C_3 gleichmäßig beschränkt sein bzgl. dieser Parameter. Das ist die sog. *Robustheit*. Die Robustheit ist eine

nicht triviale Bedingung. Um dies einzusehen, betrachte die Reaktions-Diffusionsgleichung (II.3.1) (S. 52) mit $A = \varepsilon I$, $\alpha = 1$ und $0 < \varepsilon \ll 1$. Versieht man $H_0^1(\Omega)$ mit der üblichen Norm $\|\cdot\|_1$ ist dann $\beta \approx \varepsilon$, $\mathcal{B} \approx 1$ und $C_1 C_3 \approx \varepsilon^{-1}$. Damit ist die a posteriori Fehlerabschätzung von Satz III.3.8 unbrauchbar. Um diesen unerwünschten Effekt zu vermeiden, muss man $H_0^1(\Omega)$ mit einer geeigneteren Norm versehen und Satz III.3.4 und Lemma III.3.7 verfeinern [**13**, §4.3].

Bemerkung III.3.11 (Neumann Randbedingungen, variable Koeffizienten, Konvektions-Diffusionsgleichungen). (1) Bei Neumann Randbedingungen $\mathbf{n} \cdot A\nabla u = g$ auf einem Teil Γ_N des Randes muss man bei der Definition (III.3.12) von η_K noch die Terme

$$\sum_{E \subset \partial K \cap \Gamma_N} h_E \, \|g - \mathbf{n} \cdot A\nabla u_{\mathcal{T}}\|_E^2$$

hinzufügen [**13**, §1.4].
(2) Bei variablen Koeffizienten A und α muss man wegen Lemma III.3.7 beim Beweis der unteren Schranken A und α durch diskrete Approximationen $A_{\mathcal{T}}$ und $\alpha_{\mathcal{T}}$ ersetzen. Dies führt in den unteren Schranken von Satz III.3.8 zu zusätzlichen Termen der Form

$$h_K \, \|-\operatorname{div}((A - A_{\mathcal{T}})\nabla u_{\mathcal{T}} + (\alpha - \alpha_{\mathcal{T}})u_{\mathcal{T}}\|_K$$

[**13**, Theorem 4.19].
(3) Satz III.3.8 kann auf Konvektions-Diffusionsgleichungen übertragen werden unabhängig davon, ob die Standard-Diskretisierung (II.3.2) (S. 52) oder die SUPG-Diskretisierung (II.3.4) (S. 54) verwendet wird. Die Struktur des Fehlerschätzers bleibt erhalten. Allerdings muss man die Skalierungsfaktoren h_K und h_E modifizieren, um Abschätzungen zu erhalten, die robust sind bzgl. der relativen Größe der Diffusion, Konvektion und Reaktion zueinander [**13**, §4.4].

Bemerkung III.3.12 (Andere Fehlerschätzer). Es gibt einen ganzen Zoo von Fehlerschätzern [**13**, §§1.5-1.12], die alle zu dem hier betrachteten in dem Sinne äquivalent sind, dass die verschiedenen Schätzer bis auf multiplikative Faktoren nach oben und unten gegeneinander abgeschätzt werden können. So kann man z.B. bei Verwendung von $S_0^{1,0}(\mathcal{T})$ auf die Elementresiduen verzichten [**13**, §1.6] oder den elementweise konstanten Gradienten $\nabla u_{\mathcal{T}}$ mitteln und den Fehler der Finite Element Lösung $u_{\mathcal{T}}$ durch die Differenz der Mittelung zum ursprünglichen Gradienten abschätzen [**13**, §1.9].

III.4. Adaptivität

Wir wenden uns nun dem Problem der adaptiven Gitterverfeinerung basierend auf einer posteriori Fehlerabschätzung zu und betrachten dazu Algorithmus III.4.1.

Die praktische Durchführung von Algorithmus III.4.1 erfordert in den Schritten 11 und 12 zwei Komponenten:

Algorithmus III.4.1 Adaptive Gitterverfeinerung

Gegeben: Daten der Differentialgleichung, Toleranz ε.
Gesucht: Näherungslösung mit gegebener Toleranz ε.

1: Bestimme eine grobe Unterteilung $\mathcal{T}_0$ von Ω.
2: **for** $k = 0, 1, \ldots$ **do**
3: Löse das diskrete Problem zur Unterteilung $\mathcal{T}_k$.
4: **for** $K \in \mathcal{T}_k$ **do**
5: Berechne einen a posteriori Fehlerschätzer η_K.
6: **end for**
7: $\eta \leftarrow \max_{K \in \mathcal{T}_k} \eta_K$
8: **if** $\eta \leq \varepsilon$ **then**
9: `stop` ▷ Toleranz erreicht
10: **end if**
11: Bestimme aus $(\eta_K)_K$ eine Menge $\widetilde{\mathcal{T}}_k$ von zu verfeinernden Elementen.
12: Bestimme aus $\widetilde{\mathcal{T}}_k$ eine zulässige Verfeinerung $\mathcal{T}_{k+1}$ von $\mathcal{T}_k$.
13: **end for**

- einen Algorithmus, der die zu unterteilenden Elemente bestimmt, und
- eine Vorschrift wie Elemente unterteilt werden.

Wir wenden uns zunächst dem ersten Problem zu. Man könnte versuchen, den geschätzten Fehler η über alle Unterteilungen $\mathcal{T}$ mit einer gegebenen Elementzahl zu minimieren. Dies ist jedoch ein hochgradig nichtlineares, extrem aufwändiges Optimierungsproblem. Einfache heuristische Argumente zeigen andererseits, dass bei einer optimalen Unterteilung alle Elemente etwa den gleichen Beitrag zu η liefern. Dies legt es nahe, Elemente, die einen zu großen Beitrag η_K liefern, zu unterteilen, und führt auf die Algorithmen III.4.2 und III.4.3.

Algorithmus III.4.2 Maximum Strategie

Gegeben: Unterteilung $\mathcal{T}$, Fehlerschätzer $(\eta_K)_{K \in \mathcal{T}}$, Parameter $\theta \in (0, 1)$.
Gesucht: Teilmenge $\widetilde{\mathcal{T}}$ markierter Elemente, die verfeinert werden sollen.

1: $\widetilde{\mathcal{T}} \leftarrow \emptyset$
2: $\eta \leftarrow \max_{K \in \mathcal{T}} \eta_K$
3: **for** $K \in \mathcal{T}$ **do**
4: **if** $\eta_K \geq \theta \eta$ **then**
5: $\widetilde{\mathcal{T}} \leftarrow \widetilde{\mathcal{T}} \cup \{K\}$
6: **end if**
7: **end for**

Algorithmus III.4.3 Ausgleichsstrategie; Dörfler Strategie

Gegeben: Unterteilung $\mathcal{T}$, Fehlerschätzer $(\eta_K)_{K\in\mathcal{T}}$, Parameter $\theta \in (0,1)$.

Gesucht: Teilmenge $\widetilde{\mathcal{T}}$ markierter Elemente, die verfeinert werden sollen.

1: $\widetilde{\mathcal{T}} \leftarrow \emptyset$, $\Sigma \leftarrow 0$, $\Theta \leftarrow \sum_{K\in\mathcal{T}} \eta_K^2$
2: **while** $\Sigma < \theta\Theta$ **do**
3: $\quad \eta \leftarrow \max_{K\in\mathcal{T}\setminus\widetilde{\mathcal{T}}} \eta_K$
4: $\quad$ **for** $K \in \mathcal{T} \setminus \widetilde{\mathcal{T}}$ **do**
5: $\quad\quad$ **if** $\eta_K = \eta$ **then**
6: $\quad\quad\quad \widetilde{\mathcal{T}} \leftarrow \widetilde{\mathcal{T}} \cup \{K\}$, $\Sigma \leftarrow \Sigma + \eta_K^2$
7: $\quad\quad$ **end if**
8: $\quad$ **end for**
9: **end while**

Am Ende von Algorithmus III.4.3 ist

$$\sum_{K\in\widetilde{\mathcal{T}}} \eta_K^2 \geq \theta \sum_{K\in\mathcal{T}} \eta_K^2.$$

Beide Markierungsstrategien liefern vergleichbare Ergebnisse. Die Maximum Strategie ist offensichtlich billiger als die Ausgleichsstrategie. Bei der Maximum Strategie führt ein großer Wert von θ zu einer kleinen Menge $\widetilde{\mathcal{T}}$, d.h., nur wenige Elemente werden für die Verfeinerung markiert; ein kleiner Wert von θ dagegen liefert eine große Menge $\widetilde{\mathcal{T}}$, d.h., nahezu alle Elemente werden markiert. Bei der Ausgleichsstrategie ist der Effekt umgekehrt: ein kleiner Wert von θ führt zu einer kleinen Menge $\widetilde{\mathcal{T}}$, ein großer Wert von θ liefert eine große Menge $\widetilde{\mathcal{T}}$. Bei beiden Strategien hat sich die Wahl $\theta \approx 0.5$ bewährt.

Die Algorithmen III.4.2 und III.4.3 können auch ggf. durch eine Vergröberungsstrategie ergänzt werden. Dies ist für zeitabhängige Probleme wichtig.

Als nächstes beschreiben wir, wie Elemente verfeinert werden und wie die Zulässigkeit der verfeinerten Unterteilung gesichert wird. Dabei müssen wir beachten, dass alle Unterteilungen die Regularitätsbedingung erfüllen sollen, d.h., die Elementswinkel sollen nicht zu klein oder zu groß werden. Dazu führen wir folgende Sprechweise ein (vgl. Abbildung III.4.1 und III.4.2):

- 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.

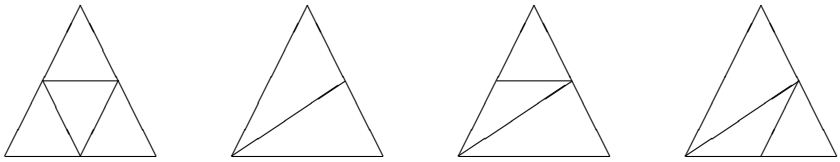

Abbildung III.4.1. Rote, grüne und blaue Unterteilung eines Dreieckes

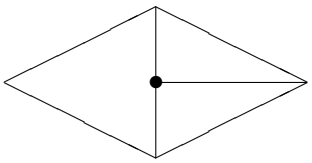

Abbildung III.4.2. Beispiel für einen hängenden Knoten •

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 Unterteilung genau dann zulässig, wenn kein Dreieck hängende Knoten hat.
In Schritt (4) von Algorithmus III.4.1 werden nun markierte Elemente rot unterteilt. Nicht markierte Elemente K werden gemäß folgender Regeln behandelt:

- 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 diese Regeln in endlich vielen Schritten eine verfeinerte Unterteilung erzeugen, die den oben diskutierten Vorgaben genügt.

Die bisherigen Ergebnisse 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.4.3 veranschaulicht.

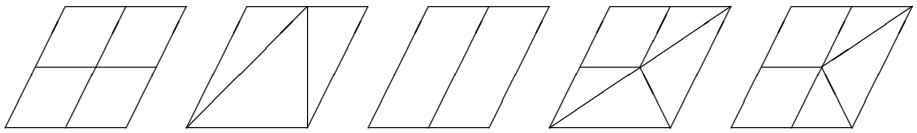

ABBILDUNG III.4.3. Rote, grüne und blaue Unterteilung eines Parallelogramms

Eine interessante Alternative zu der beschriebenen roten Unterteilung ist die *Bisektion markierter Kanten*. Diese erfordert keine Zusatzregeln für die Behandlung hängender Knoten und erleichert nötigenfalls die Vergröberung einer Unterteilung. Die Bisektion markierter Kanten erfolgt auf Basis folgender Regeln (vgl. Abbildung III.4.4):

- Die gröbste Unterteilung $\mathcal{T}_0$ wird so konstruiert, dass die längste Kante eines jeden Elementes entweder eine Randkante oder die längste Kante des angrenzenden Elementes ist.
- Die längsten Kanten der Elemente in $\mathcal{T}_0$ werden markiert.
- Ein Element wird unterteilt, indem der Mittelpunkt seiner markierten Kante mit dem gegenüberliegenden Eckpunkt verbunden wird.
- Wird ein Element durch Bisektion unterteilt, werden seine nicht unterteilten Kanten die markierten Kanten der beiden neuen Dreiecke.

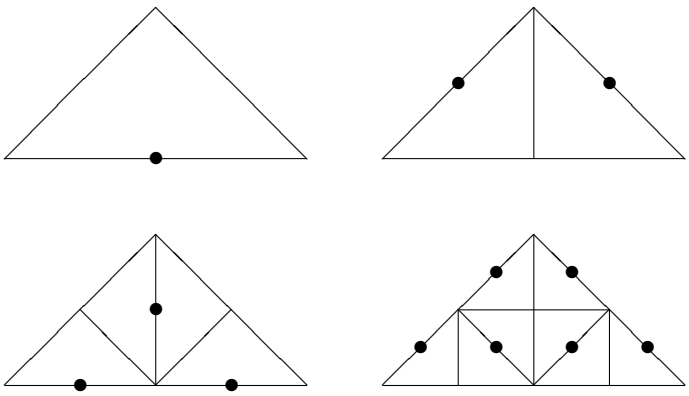

ABBILDUNG III.4.4. Sukzessive Bisektion markierter Kanten; markierte Kanten sind durch • gekennzeichnet

Wir wollen nun zeigen, dass Algorithmus III.4.1 konvergiert, wenn in Schritt (4) die Ausgleichsstrategie für die Markierung gewählt wird. Zur Vereinfachung der Darstellung betrachten wir die zweidimensionale Poissongleichung mit Dirichletrandbedingungen, d.h. (II.3.1) (S. 52) mit $A = I$, $\alpha = 0$ und $\Omega \subset \mathbb{R}^2$, lineare Dreieckselemente, d.h. (II.3.2) (S. 52) mit $k = 1$ und $\mathcal{T}$ aus Dreiecken bestehend, und den residuellen Fehlerschätzer η_K aus Satz III.3.8 (S. 89). Die Ergebnisse können mit einigem technischen Mehraufwand auf allgemeinere Differentialgleichungen, Diskretisierungen und Fehlerschätzer übertragen werden [**13**, §4.13].

In einem ersten Schritt nehmen wir an, dass die rechte Seite f stückweise konstant ist, so dass die Terme $\|f - f_{\mathcal{T}}\|$ verschwinden. Diese Einschränkung werden wir dann in einem zweiten Schritt überwinden.

Betrachte eine Unterteilung $\mathcal{T}_1$ von Ω und eine Verfeinerung $\mathcal{T}_2$ von $\mathcal{T}_1$, d.h., jedes Element in $\mathcal{T}_1$ ist die Vereinigung von Elementen in $\mathcal{T}_2$. Die zugehörigen Finite Element Räume $S_0^{1,0}(\mathcal{T}_1)$ und $S_0^{1,0}(\mathcal{T}_2)$ sind dann geschachtelt, d.h.

$$S_0^{1,0}(\mathcal{T}_1) \subset S_0^{1,0}(\mathcal{T}_2). \tag{III.4.1}$$

Bezeichne mit u die schwache Lösung der Poissongleichung und mit u_1 und u_2 die zugehörigen Finite Element Lösungen. Aus Beziehung (III.4.1) und der Galerkin-Orthogonalität (III.3.4) (S. 81) folgt

$$\|\nabla(u - u_2)\|^2 = \|\nabla(u - u_1)\|^2 - \|\nabla(u_1 - u_2)\|^2 . \tag{III.4.2}$$

Da f stückweise konstant ist, folgt aus Satz III.3.8

$$\|\nabla(u - u_1)\|^2 \leq C_1^2 \sum_{K \in \mathcal{T}_1} \eta_K^2. \tag{III.4.3}$$

Wir wählen nun einen Parameter $\theta \in (0, 1)$ und bestimmen mit Algorithmus III.4.3 eine Teilmenge $\widetilde{\mathcal{T}}_1$ von $\mathcal{T}_1$, so dass

$$\sum_{K \in \widetilde{\mathcal{T}}_1} \eta_K^2 \geq \theta \sum_{K \in \mathcal{T}_1} \eta_K^2 \tag{III.4.4}$$

ist. Aus den Abschätzungen (III.4.3) und (III.4.4) ergibt sich dann

$$\|\nabla(u - u_1)\|^2 \leq \frac{C_1^2}{\theta} \sum_{K \in \widetilde{\mathcal{T}}_1} \eta_K^2. \tag{III.4.5}$$

Wir nehmen nun zusätzlich an, dass die verfeinerte Unterteilung $\mathcal{T}_2$ folgende Bedingungen erfüllt:

- Der Mittelpunkt jeder Kante in $\widetilde{\mathcal{T}}_1$ ist ein Elementeckpunkt von $\mathcal{T}_2$.
- Zu jedem Element in $\widetilde{\mathcal{T}}_1$ gibt es einen Punkt in seinem Inneren, der ein Elementeckpunkt von $\mathcal{T}_2$ ist.

Diese Bedingungen können z.B. durch zwei Schritte der roten Unterteilung oder durch drei Schritte der Bisektion markierter Kanten der Elemente in $\widetilde{\mathcal{T}}_1$ erfüllt werden.

Wegen dieser Annahmen können wir für jedes Element K und jede Kante E in $\widetilde{\mathcal{T}}_1$ die Blasenfunktionen ψ_K und ψ_E durch die nodalen Basisfunktionen von $S_0^{1,0}(\mathcal{T}_2)$ ersetzen, die dem inneren Elementeckpunkt in K und dem Mittelpunkt von E entsprechen. Man überzeugt sich leicht, dass Lemma III.3.7 (S. 87) auch für diese Funktionen gilt. Daher bleibt der obige Beweis der unteren Fehlerschranke gültig. Da f stückweise konstant ist, erhalten wir somit die Abschätzung

$$\sum_{K\in\widetilde{\mathcal{T}}_1} \eta_K^2 \le C_2^2 \left\|\nabla(u_2-u_1)\right\|^2 . \tag{III.4.6}$$

Aus den Abschätzungen (III.4.5) und (III.4.6) folgt

$$-\left\|\nabla(u_2-u_1)\right\|^2 \le -\frac{1}{C_2^2}\sum_{K\in\widetilde{\mathcal{T}}_1}\eta_K^2 \le -\frac{\theta}{C_2^2C_1^2}\left\|\nabla(u-u_1)\right\|^2 .$$

Setzen wir dies in Ungleichung (III.4.2) ein, erhalten wir schließlich

$$\left\|\nabla(u-u_2)\right\|^2 \le \left(1-\frac{\theta}{C_2^2C_1^2}\right)\left\|\nabla(u-u_1)\right\|^2 .$$

Dies beweist die Konvergenz von Algorithmus III.4.1, sofern die rechte Seite f auf der gröbsten Unterteilung $\mathcal{T}_0$ stückweise konstant ist. Jede Iteration von Algorithmus III.4.1 reduziert dann den Fehler der Finite Element Approximation um den Faktor $\sqrt{1-\frac{\theta}{C_2^2C_1^2}}$, der nur von dem Parameter θ und dem Formparameter $\mathrm{C}_{\mathcal{T}} = \max_K \frac{h_K}{\rho_K}$ (d.h. kleinste Winkel) von $\mathcal{T}_0$ abhängt.

Nun wollen wir die Annahme einer stückweise konstanten rechten Seite f überwinden. Betrachte dazu eine beliebige Funktion $f \in L^2(\Omega)$, eine beliebige Unterteilung $\mathcal{T}$ und die L^2-Projektion $f_{\mathcal{T}}$ von f auf $S^{0,-1}(\mathcal{T})$. Bezeichne mit u die schwache Lösung der Poissongleichung mit rechter Seite f, mit $\widetilde{u}$ die schwache Lösung der Poissongleichung mit rechter Seite $f_{\mathcal{T}}$, mit $u_{\mathcal{T}}$ die Lösung des diskreten Problems mit rechter Seite f, und mit $\widetilde{u}_{\mathcal{T}}$ die Lösung des diskreten Problems mit rechter Seite $f_{\mathcal{T}}$. Da $u_{\mathcal{T}}$ die beste Approximation an u in $S_0^{1,0}(\mathcal{T})$ bzgl. der Norm $|\cdot|_1 = \|\nabla\cdot\|$ ist, folgt

$$\left\|\nabla(u-u_{\mathcal{T}})\right\| \le \left\|\nabla(u-\widetilde{u}_{\mathcal{T}})\right\|$$

Weiter folgt aus der Dreiecksungleichung

$$\left\|\nabla(u-\widetilde{u}_{\mathcal{T}})\right\| \le \left\|\nabla(u-\widetilde{u})\right\| + \left\|\nabla(\widetilde{u}-\widetilde{u}_{\mathcal{T}})\right\| .$$

Der Term $\|\nabla(\widetilde{u}-\widetilde{u}_{\mathcal{T}})\|$ kann mit obigen Techniken für stückweise konstante rechte Seiten behandelt werden. Daher müssen wir nur noch den Term $\|\nabla(u-\widetilde{u})\|$ kontrollieren.

Um dies zu erreichen, beachten wir, dass

$$\|\nabla(u-\widetilde{u})\| = \sup_{w\in H_0^1(\Omega)\setminus\{0\}} \frac{\int_\Omega \nabla(u-\widetilde{u})\cdot\nabla w}{\|\nabla w\|}$$

ist. Da u und $\widetilde{u}$ die Poissongleichung zu den rechten Seiten f und $f_\mathcal{T}$ lösen, gilt für jedes $w \in H_0^1(\Omega)$

$$\int_\Omega \nabla(u-\widetilde{u})\cdot\nabla w = \int_\Omega (f-f_\mathcal{T})w.$$

Da aber $f_\mathcal{T}$ die L^2-Projektion von f auf den Raum der stückweise konstanten Funktionen ist, gilt

$$\begin{aligned}\int_\Omega (f-f_\mathcal{T})w &= \int_\Omega (f-f_\mathcal{T})(w-w_\mathcal{T})\\ &= \sum_{K\in\mathcal{T}} \int_K (f-\overline{f}_K)(w-\overline{w}_K),\end{aligned}$$

wobei $w_\mathcal{T}$ die L^2-Projektion von w auf $S^{0,-1}(\mathcal{T})$ ist und $\overline{f}_K$ und $\overline{w}_K$ die Mittelwerte von f und w auf K bezeichnen. Aus der Cauchy-Schwarzschen Ungleichung folgt somit

$$\int_\Omega \nabla(u-\widetilde{u})\cdot\nabla w \le \sum_{K\in\mathcal{T}} \left\|f-\overline{f}_K\right\|_K \|w-\overline{w}_K\|_K .$$

Da jedes Element K konvex ist, folgt aus der Poincaréschen Ungleichung, Satz I.2.21 (S. 28), und Bemerkung I.2.22(2) (S. 28) für jedes Element

$$\|w-\overline{w}_K\|_K \le \frac{h_K}{\pi}\|\nabla w\|_K .$$

Mit Hilfe einer gewichteten Cauchy-Schwarzschen Ungleichung erhalten wir somit die Abschätzung

$$\|\nabla(u-\widetilde{u})\| \le \frac{1}{\pi}\left\{\sum_{K\in\mathcal{T}} h_K^2 \left\|f-\overline{f}_K\right\|_K^2\right\}^{\frac{1}{2}} .$$

Also müssen wir für allgemeine Daten f die rechte Seite dieser Abschätzung, die häufig als *Datenoszillation* bezeichnet wird, kontrollieren. Dies kann mit folgender Abwandlung von Algorithmus III.4.1 geschehen.

Wir betrachten wieder eine Unterteilung $\mathcal{T}_1$ und eine Verfeinerung $\mathcal{T}_2$ von $\mathcal{T}_1$. Dann wenden wir Algorithmus III.4.3 auf $\mathcal{T}_1$ mit η_K^2 ersetzt durch $h_K^2\left\|f-\overline{f}_K\right\|_K^2$ an und bestimmen so eine Teilmenge $\widetilde{\mathcal{T}}_1$ von $\mathcal{T}_1$ mit

$$\sum_{K\in\widetilde{\mathcal{T}}_1} h_K^2\left\|f-\overline{f}_K\right\|_K^2 \ge \theta \sum_{K\in\mathcal{T}_1} h_K^2\left\|f-\overline{f}_K\right\|^2 .$$

Wir nehmen nun an, dass $\mathcal{T}_2$ folgende Bedingung erfüllt: Jedes Element K in $\widetilde{\mathcal{T}}_1$ ist die Vereinigung von Elementen in $\mathcal{T}_2$, derart dass jedes dieser Elemente höchstens den Durchmesser $\frac{1}{2}h_K$ hat.
Diese Bedingung kann wieder durch zwei Schritte der roten Unterteilung oder drei Schritte der Bisektion markierter Kanten erfüllt werden.

Wir spalten nun $\mathcal{T}_2$ so in zwei disjunkte Teilmengen $\mathcal{T}_{2,R}$ and $\mathcal{T}_{2,U}$ auf, dass $\bigcup_{K\in\mathcal{T}_{2,R}} K \supset \bigcup_{K\in\widetilde{\mathcal{T}}_1} K$ ist. Dann gilt

$$\begin{aligned}\sum_{K\in\mathcal{T}_{2,U}} h_K^2 \left\|f-\overline{f}_K\right\|_K^2 &\le \sum_{K\in\mathcal{T}_1\setminus\widetilde{\mathcal{T}}_1} h_K^2 \left\|f-\overline{f}_K\right\|_K^2 \\ &= \sum_{K\in\mathcal{T}_1} h_K^2 \left\|f-\overline{f}_K\right\|_K^2 - \sum_{K\in\widetilde{\mathcal{T}}_1} h_K^2 \left\|f-\overline{f}_K\right\|_K^2\end{aligned}$$

und

$$\sum_{K\in\mathcal{T}_{2,R}} h_K^2 \left\|f-\overline{f}_K\right\|_K^2 \le \frac{1}{4}\sum_{K\in\widetilde{\mathcal{T}}_1} h_K^2 \left\|f-\overline{f}_K\right\|_K^2 .$$

Hieraus folgt

$$\begin{aligned}\sum_{K\in\mathcal{T}_2} h_K^2 \left\|f-\overline{f}_K\right\|_K^2 &= \sum_{K\in\mathcal{T}_{2,R}} h_K^2 \left\|f-\overline{f}_K\right\|_K^2 + \sum_{K\in\mathcal{T}_{2,U}} h_K^2 \left\|f-\overline{f}_K\right\|_K^2 \\ &\le \sum_{K\in\mathcal{T}_1} h_K^2 \left\|f-\overline{f}_K\right\|_K^2 - \frac{3}{4}\sum_{K\in\widetilde{\mathcal{T}}_1} h_K^2 \left\|f-\overline{f}_K\right\|_K^2 \\ &\le \left(1-\frac{3\theta}{4}\right)\sum_{K\in\mathcal{T}_1} h_K^2 \left\|f-\overline{f}_K\right\|_K^2 .\end{aligned}$$

Algorithmus III.4.1 mit $h_K \left\|f-\overline{f}_K\right\|_K$ an Stelle von η_K reduziert dann mit jeder Iteration die Größe $\|\nabla(u-\widetilde{u})\|$ mindestens um den Faktor $\sqrt{1-\frac{3\theta}{4}}$. Für jede Toleranz ε erhalten wir daher nach endlich vielen Schritten eine Unterteilung $\mathcal{T}_0$ mit $\|\nabla(u-\widetilde{u})\| \le \frac{\varepsilon}{2}$. Diese bildet dann den Ausgangspunkt für den adaptiven Algorithmus für stückweise konstante rechte Seiten f. So erhalten wir dann nach einer weiteren endlichen Zahl von Schritten eine verfeinerte Unterteilung $\mathcal{T}_1$ with $\|\nabla(\widetilde{u}-u_1)\| \le \frac{\varepsilon}{2}$ und $\|\nabla(u-u_1)\| \le \varepsilon$.

Zum Abschluss geben wir zwei Beispiele, die die Vorteile der adaptiven Gitterverfeinerung basierend auf a posteriori Fehlerschätzern illustrieren.

Beispiel III.4.1 (Einspringende Ecke). Wir betrachten wie in Beispiel III.2.2 (S. 65) 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 gröbstes Gitter verwenden wir eine Unterteilung mit 6 gleichschenklig rechtwinkligen Dreiecken mit Hypotenusen in Richtung $(1,-1)$. Wir lösen die Differentialgleichung näherungsweise einmal mit einer uniformen Verfeinerung und einmal

mit einer adaptiven Verfeinerung gemäß Algorithmus III.4.1 und der Maximum Strategie III.4.2 basierend auf dem Fehlerschätzer aus Satz III.3.8 (S. 89) und der Bisektion markierter Kanten. Dabei wird die uniforme Verfeinerung so lange durchgeführt, bis die Speicherkapazität 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.4.1 geben wir jeweils die Zahl L der Gitter, die Zahl NT der Elemente, die Zahl NV der Eckpunkte, die Zahl NN der Unbekannten auf dem feinsten Gitter und den relativen Fehler ε gemessen in der H^1-Norm in Prozent auf dem feinsten Gitter. Abbildung III.4.5 zeigt das feinste bei der adaptiven Verfeinerung erzeugte Gitter.

TABELLE III.4.1. Vergleich uniformer und adaptiver Verfeinerung für das Problem aus Beispiel III.4.1

	uniform	adaptiv
L	12	17
NT	12288	784
NV	6273	499
NN	6017	365
$\varepsilon(\%)$	0.13	0.2

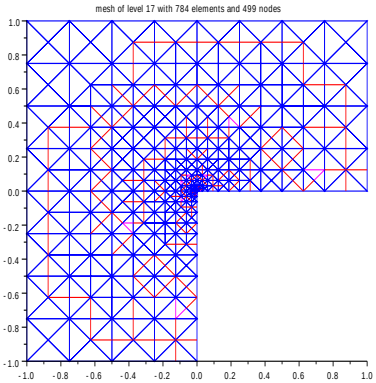

ABBILDUNG III.4.5. Feinstes adaptives Dreiecksgitter für das Problem aus Beispiel III.4.1

BEISPIEL III.4.2 (Innere Grenzschicht). Wir betrachten die Reaktions-Diffusionsgleichung mit inhomogenen Diricht-Randbedingungen auf dem Quadrat $(-1,1)^2$. Die Randfunktion u_D, die rechte Seite f und der Reaktionsterm α 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 gröbstes Gitter verwenden wir eine Unterteilung mit 8 gleichschenklig rechtwinkligen Dreiecken mit Hypotenusen in Richtung $(1,-1)$. Ansonsten gehen wir wie in Beispiel III.4.1 vor. Die Ergebnisse sind in Tabelle III.4.2 wiedergegeben. Abbildung III.4.6 zeigt das feinste bei der adaptiven Verfeinerung erzeugte Gitter.

TABELLE III.4.2. Vergleich uniformer und adaptiver Verfeinerung für das Problem aus Beispiel III.4.2

	uniform	adaptiv
`L`	11	11
`NT`	8192	1000
`NV`	4225	701
`NN`	3969	493
$\varepsilon(\%)$	2.13	2.05

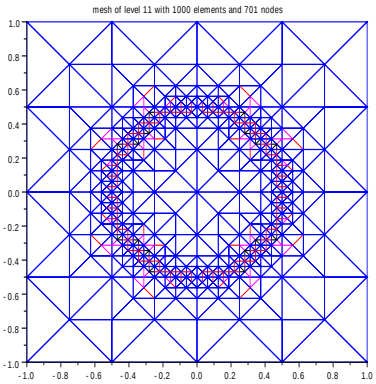

ABBILDUNG III.4.6. Feinstes adaptives Dreiecksgitter für das Problem aus Beispiel III.4.2

III.5. Implementierung

In diesem Abschnitt gehen wir kurz auf die Implementierung der Methoden der letzten Paragraphen und die dafür benötigten Datenstrukturen ein. Um die Notationen und den technischen Aufwand gering zu halten, beschränken wir uns auf zweidimensionale Probleme und lineare Dreiecks- und bilineare affine Viereckselemente.

Die 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

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

- `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 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 $\texttt{v}[3] = -1$.
- `e`: Dies ist ein Feld der Länge 4 vom Typ `integer`. Es beschreibt die Nachbarschaftsbeziehungen zwischen den Elementen. Es gilt

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

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.5.1 und III.5.2), 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 Elternelementes des betreffenden Elementes an.
- `t`: Diese Größe vom Typ `integer` beschreibt den Verfeinerungstyp des Elementes. Es ist (vgl. Abbildungen III.5.1 und III.5.2)

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

- `c`: Diese Größe vom Typ `integer` gibt die Nummer des ersten Kindes des Elementes an. Die weiteren Kinder werden wie in Abbildungen III.5.1 und III.5.2 angegeben fortlaufend nummeriert.

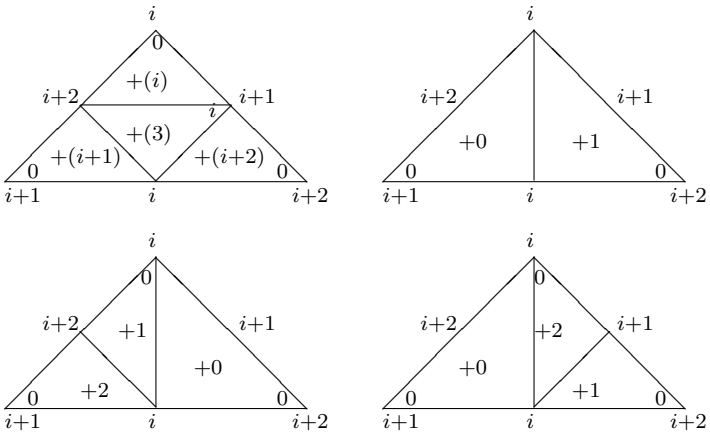

ABBILDUNG III.5.1. Nummerierung der Eckpunkte und Kanten eines Dreieckes und seiner Nachfahren. Die Zahlen $i, \ldots, i+2$ sind modulo 3 zu nehmen. Fehlende Nummern von Eckpunkten innerhalb eines Elementes sind durch fortlaufende Nummerierung im mathematisch positiven Sinn zu ergänzen. Ein Elementeintrag $+(k)$ bedeutet, dass das Element die Nummer $j+k$ erhält, wenn j der Wert des Attributes `c` des Elternelementes ist.

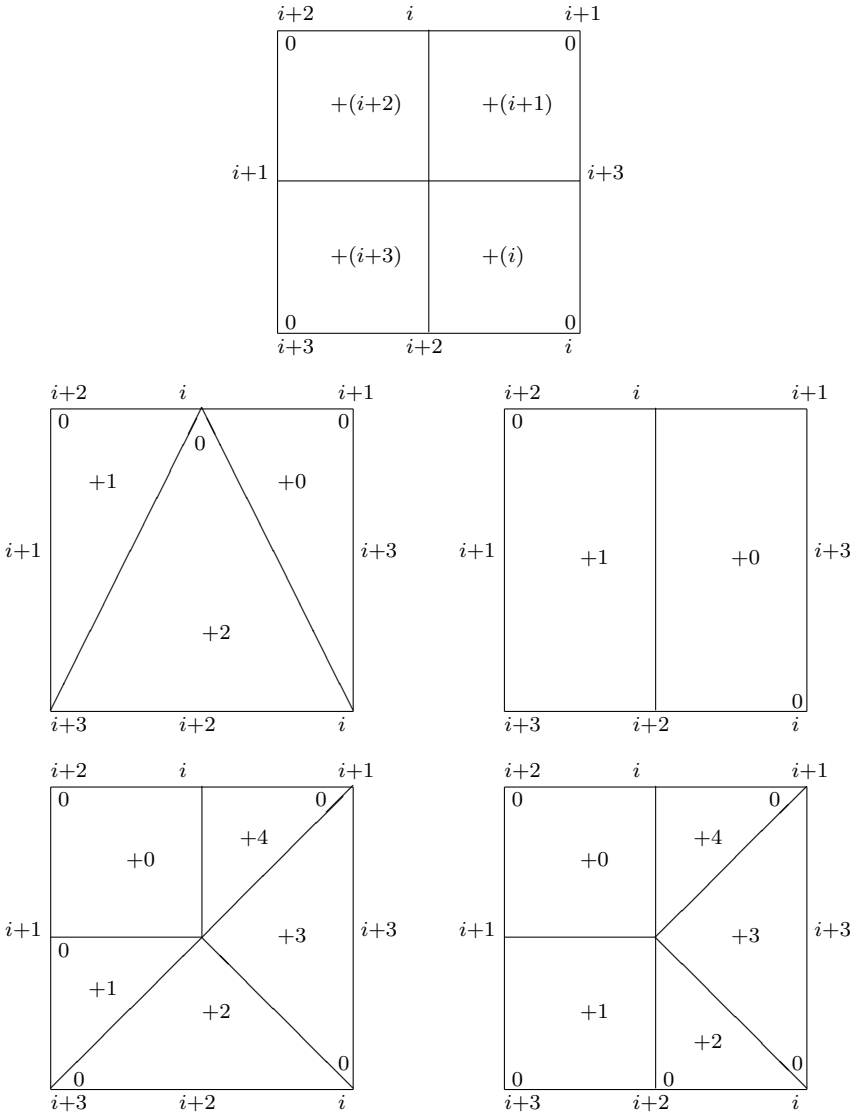

ABBILDUNG III.5.2. Nummerierung der Eckpunkte und Kanten eines Viereckes und seiner Nachfahren. Die Zahlen $i, \ldots, i+3$ sind modulo 4 zu nehmen. Fehlende Nummern von Eckpunkten innerhalb eines Elementes sind durch fortlaufende Nummerierung im mathematisch positiven Sinn zu ergänzen. Ein Elementeintrag $+(k)$ bedeutet, dass das Element die Nummer $j+k$ erhält, wenn j der Wert des Attributes `c` des Elternelementes ist.

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` 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 Klasse `level` hat folgende Mitglieder:

- `nn`: die Zahl der Knoten des aktuellen Gitters,
- `cn`: die kumulative Zahl der Knoten des aktuellen und aller gröberen 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

$$\texttt{for(i = 0; i < ll[}k\texttt{].nn; i++)}\{\ldots\}.$$

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

$$\texttt{for(i = ll[[}k\texttt{].first; i < ll[}k\texttt{].last; i++)}\{\ldots\}.$$

Für das Aufstellen des diskreten Problems muss man die Größen

$$b_i = \ell(v_i) \quad \text{und} \quad a_{ij} = B(v_i, v_j)$$

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. B_K die Einschränkung von ℓ bzw. B 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
    $j =$ `vertices`$[K.\mathtt{v}[i]].\mathtt{d}$
    $b_j = b_j + \ell_K(v_i)$
    für alle Knotennummern $l \geq i$ von K
      $m =$ `vertices`$[K.\mathtt{v}[l]].\mathtt{d}$
      $a_{mj} = a_{mj} + B_K(v_l, v_i)$

Dabei wird $\ell_K(v_i)$ mit einer Quadraturformel berechnet. Für ein Dreieck und die Formel aus Beispiel III.1.1(2) (S. 62) ergibt sich z.B.

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

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 $B_K(v_l, v_i)$ 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. 62) 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. 62) 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] $\leq i <$ `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

Als Ergänzung der bisherigen Ergebnisse betrachten wir in diesem Kapitel nicht-konforme und gemischte Finite Elemente, sowie Discontinuous Galerkin und Finite Volumen Methoden. Bei ersteren ist der Finite Element Raum nicht mehr in dem Raum, in dem das zugrundeliegende Variationsproblem formuliert ist, enthalten. Dies gibt mehr Flexibilität für die Diskretisierung, führt aber auch zu zusätzlichen Konsistenzfehlern, die sorgfältig kontrolliert werden müssen. Die gemischten Finite Element Methoden sind zugeschnitten auf sog. Sattelpunktsprobleme. Bei diesen wird ein geeignetes Funktional bzgl. eines Anteils der gesuchten Lösung minimiert und bzgl. eines anderen Anteils maximiert. Als einfachstes Beispiel betrachten wir die Satelpunktsformulierung der Poissongleichung. Hier sind die erwähnten Lösungskomponenten die Funktion u und ihr Gradient ∇u. Discontinuous Galerkin Methoden sind spezielle nicht-konforme Methoden. Bei ihnen ist die diskrete Lösung komplett unstetig und verletzt in der Regel auch eventuelle Dirichlet Randbedingungen. Die Lösbarkeit des diskreten Problems wird durch geschickte Stabilisierungsterme erreicht, die zudem so gewählt werden, dass keine Konsistenzfehler auftreten. Finite Volumen Methoden sind auf sogenannte Systeme in Divergenzform zugeschnitten. Sie fallen konzeptionell zunächst völlig aus dem bisherigen Rahmen, haben aber dennoch strukturelle Beziehungen zu Finite Element Methoden. Geeignete Discontinuous Galerkin Methoden für Systeme in Divergenzform stellen zudem einen gemeinsamen Rahmen für Finite Element und Finite Volumen Methoden her.

IV.1. Nicht-konforme Finite Elemente

Bisher haben wir stets konforme Finite Element Methoden betrachtet, bei denen der diskrete Raum $X_{\mathcal{T}}$ 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.1 eine abstrakte Situation. Sei dazu $(X, \|\cdot\|_X)$ ein Banach-Raum, $\ell \in \mathcal{L}(X, \mathbb{R})$ und $B \in \mathcal{L}^2(X, \mathbb{R})$ symmetrisch und koerziv. Zu lösen ist das Variationsproblem

$$B(u, v) = \ell(v) \quad \forall v \in X. \tag{IV.1.1}$$

Weiter seien $(X_\mathcal{T}, \|\cdot\|_{X_\mathcal{T}})$ ein endlich dimensionaler Banach-Raum, $\ell_\mathcal{T} \in \mathcal{L}(X_\mathcal{T}, \mathbb{R})$ und $B_\mathcal{T} \in \mathcal{L}^2(X_\mathcal{T}, \mathbb{R})$ symmetrisch und koerziv. Die Normen von $\ell_\mathcal{T}$ und $B_\mathcal{T}$ sowie die Koerzivitätskonstante $\beta_\mathcal{T}$ von $B_\mathcal{T}$ seien gleichmäßig bzgl. $\mathcal{T}$ nach oben bzw. für $\beta_\mathcal{T}$ nach unten weg von 0 beschränkt. Die diskrete Approximation von (IV.1.1) lautet

$$B_\mathcal{T}(u_\mathcal{T}, v_\mathcal{T}) = \ell_\mathcal{T}(v_\mathcal{T}) \quad \forall v_\mathcal{T} \in X_\mathcal{T}. \tag{IV.1.2}$$

Wegen Satz I.1.1 (S. 15) haben (IV.1.1) und (IV.1.2) jeweils eine eindeutige Lösung u bzw. $u_\mathcal{T}$. Man beachte, dass hierfür die gleichmäßige Beschränktheit von $\|\ell_\mathcal{T}\|_\mathcal{L}$, $\|B_\mathcal{T}\|_{\mathcal{L}^2}$ und $\beta_\mathcal{T}$ nicht benötigt wird. Da wir an der nicht-konformen Situation interessiert sind, ist $X_\mathcal{T} \not\subset X$. Allerdings setzen wir voraus, dass $B_\mathcal{T}$, $\ell_\mathcal{T}$ und $\|\cdot\|_{X_\mathcal{T}}$ auch für Elemente von X einen Sinn machen und für solche Elemente mit B, ℓ bzw. $\|\cdot\|_X$ übereinstimmen. Der folgende Satz ist unter diesen Annahmen das nicht-konforme Analogon zum Céa-Lemma, Satz I.1.2 (S. 17). Sein Beweis ist analog zu dem von Satz II.3.4 (S. 55).

Satz IV.1.1 (Zweites Strang-Lemma). *Unter den obigen Voraussetzungen gilt die Fehlerabschätzung*

$$\begin{aligned}\|u - u_\mathcal{T}\|_{X_\mathcal{T}} \leq \left(1 + \frac{\mathcal{B}_\mathcal{T}}{\beta_\mathcal{T}}\right) \inf_{v_\mathcal{T} \in X_\mathcal{T}} \|u - v_\mathcal{T}\|_{X_\mathcal{T}} \\ + \frac{1}{\beta_\mathcal{T}} \sup_{w_\mathcal{T} \in X_\mathcal{T};\, \|w_\mathcal{T}\|_{X_\mathcal{T}} = 1} |B_\mathcal{T}(u, w_\mathcal{T}) - \ell_\mathcal{T}(w_\mathcal{T})|\,.\end{aligned}$$

Dabei ist $\mathcal{B}_\mathcal{T} = \|B_\mathcal{T}\|_{\mathcal{L}^2}$ und $\beta_\mathcal{T}$ die Koerzivitätskonstante von $B_\mathcal{T}$.

Beweis. Sei $v_\mathcal{T} \in X_\mathcal{T}$ beliebig und $w_\mathcal{T} = u_\mathcal{T} - v_\mathcal{T}$. Aus der Koerzivität von $B_\mathcal{T}$ folgt dann

$$\begin{aligned}\beta_\mathcal{T} \|w_\mathcal{T}\|_{X_\mathcal{T}}^2 &\leq B_\mathcal{T}(w_\mathcal{T}, w_\mathcal{T}) = B_\mathcal{T}(u_\mathcal{T} - v_\mathcal{T}, w_\mathcal{T}) \\ &= B_\mathcal{T}(u - v_\mathcal{T}, w_\mathcal{T}) + B_\mathcal{T}(u_\mathcal{T}, w_\mathcal{T}) - B_\mathcal{T}(u, w_\mathcal{T}) \\ &= B_\mathcal{T}(u - v_\mathcal{T}, w_\mathcal{T}) + \ell_\mathcal{T}(w_\mathcal{T}) - B_\mathcal{T}(u, w_\mathcal{T}) \\ &\leq \mathcal{B}_\mathcal{T} \|u - v_\mathcal{T}\|_{X_\mathcal{T}} \|w_\mathcal{T}\|_{X_\mathcal{T}} + |\ell_\mathcal{T}(w_\mathcal{T}) - B_\mathcal{T}(u, w_\mathcal{T})|\,.\end{aligned}$$

Hieraus folgt die Behauptung mit der Dreiecksungleichung

$$\|u - u_\mathcal{T}\|_{X_\mathcal{T}} \leq \|u - v_\mathcal{T}\|_{X_\mathcal{T}} + \|w_\mathcal{T}\|_{X_\mathcal{T}}\,. \qquad \square$$

Der folgende Satz ist die nicht-konforme Variante des Dualitätsargumentes von Aubin-Nitsche, Satz I.1.5 (S. 18).

Satz IV.1.2 (Satz von Aubin-Nitsche für nicht-konforme Diskretisierungen). *Die Bezeichnungen und Voraussetzungen seien wie in Satz IV.1.1. Zusätzlich sei H ein Hilbert-Raum mit Skalarprodukt $(\cdot\,,\cdot)_H$ und Norm $\|\cdot\|_H$, so dass $X \hookrightarrow H$ dicht und $X_\mathcal{T} \subset H$ ist. Für jedes $\varphi \in H$ seien $u_\varphi \in X$ die eindeutige Lösung von*

$$B(v, u_\varphi) = (\varphi\,,\, v)_H \quad \forall v \in X$$

und $u_{\varphi,\mathcal{T}} \in X_{\mathcal{T}}$ die eindeutige Lösung von

$$B_{\mathcal{T}}(v_{\mathcal{T}}, u_{\varphi,\mathcal{T}}) = (\varphi\,,\, v_{\mathcal{T}})_H \quad \forall v_{\mathcal{T}} \in X_{\mathcal{T}}.$$

Dann gilt die Fehlerabschätzung

$$\begin{aligned}\|u - u_{\mathcal{T}}\|_H \le \sup_{\varphi\in H;\,\|\varphi\|_H=1} \Big\{ & \mathcal{B}_{\mathcal{T}}\, \|u - u_{\mathcal{T}}\|_{X_{\mathcal{T}}}\, \|u_\varphi - u_{\varphi,\mathcal{T}}\|_{X_{\mathcal{T}}} \\ & + |B_{\mathcal{T}}(u - u_{\mathcal{T}}, u_\varphi) - (\varphi\,,\, u - u_{\mathcal{T}})_H| \\ & + |B_{\mathcal{T}}(u, u_\varphi - u_{\varphi,\mathcal{T}}) - \ell_{\mathcal{T}}(u_\varphi - u_{\varphi,\mathcal{T}})| \Big\}.\end{aligned}$$

Beweis. Da $X \hookrightarrow H$ dicht und $X_{\mathcal{T}} \subset H$ ist, ist

$$\|u - u_{\mathcal{T}}\|_H = \sup_{\varphi\in H;\,\|\varphi\|_H=1} (\varphi\,,\, u - u_{\mathcal{T}})_H\,. \tag{IV.1.3}$$

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

$$\begin{aligned}&(\varphi\,,\, u - u_{\mathcal{T}})_H \\ &\quad = B(u, u_\varphi) - B_{\mathcal{T}}(u_{\mathcal{T}}, u_{\varphi,\mathcal{T}}) = B_{\mathcal{T}}(u, u_\varphi) - B_{\mathcal{T}}(u_{\mathcal{T}}, u_{\varphi,\mathcal{T}}) \\ &\quad = B_{\mathcal{T}}(u - u_{\mathcal{T}}, u_\varphi - u_{\varphi,\mathcal{T}}) + B_{\mathcal{T}}(u_{\mathcal{T}}, u_\varphi - u_{\varphi,\mathcal{T}}) + B_{\mathcal{T}}(u - u_{\mathcal{T}}, u_{\varphi,\mathcal{T}}).\end{aligned}$$

Wegen

$$B_{\mathcal{T}}(u, u_\varphi) = (\varphi\,,\, u)_H \quad \text{und} \quad B_{\mathcal{T}}(u_{\mathcal{T}}, u_{\varphi,\mathcal{T}}) = (\varphi\,,\, u_{\mathcal{T}})_H$$

folgt für den zweiten und dritten Summanden dieser Gleichung

$$\begin{aligned}&B_{\mathcal{T}}(u_{\mathcal{T}}, u_\varphi - u_{\varphi,\mathcal{T}}) \\ &\quad = B_{\mathcal{T}}(u_{\mathcal{T}}, u_\varphi) - B_{\mathcal{T}}(u, u_\varphi) + B_{\mathcal{T}}(u, u_\varphi) - B_{\mathcal{T}}(u_{\mathcal{T}}, u_{\varphi,\mathcal{T}}) \\ &\quad = -B_{\mathcal{T}}(u - u_{\mathcal{T}}, u_\varphi) + (\varphi\,,\, u - u_{\mathcal{T}})_H\end{aligned}$$

und

$$\begin{aligned}&B_{\mathcal{T}}(u - u_{\mathcal{T}}, u_{\varphi,\mathcal{T}}) \\ &\quad = B_{\mathcal{T}}(u, u_{\varphi,\mathcal{T}}) - B_{\mathcal{T}}(u, u_\varphi) + \underbrace{B_{\mathcal{T}}(u, u_\varphi)}_{=\ell(u_\varphi)} - \underbrace{B_{\mathcal{T}}(u_{\mathcal{T}}, u_{\varphi,\mathcal{T}})}_{=\ell_{\mathcal{T}}(u_{\varphi,\mathcal{T}})} \\ &\quad = -B_{\mathcal{T}}(u, u_\varphi - u_{\varphi,\mathcal{T}}) + \ell_{\mathcal{T}}(u_\varphi - u_{\varphi,\mathcal{T}}).\end{aligned}$$

Also ist

$$\begin{aligned}(\varphi\,,\, u - u_{\mathcal{T}})_H = {} & B_{\mathcal{T}}(u - u_{\mathcal{T}}, u_\varphi - u_{\varphi,\mathcal{T}}) \\ & - \{B_{\mathcal{T}}(u - u_{\mathcal{T}}, u_\varphi) - (\varphi\,,\, u - u_{\mathcal{T}})_H\} \\ & - \{B_{\mathcal{T}}(u, u_\varphi - u_{\varphi,\mathcal{T}}) - \ell_{\mathcal{T}}(u_\varphi - u_{\varphi,\mathcal{T}})\}.\end{aligned}$$

Hieraus und aus (IV.1.3) folgt die Behauptung. □

Wir wenden diese abstrakten Ergebnisse auf ein einfaches, aber typisches Modellproblem an: die *Crouzeix-Raviart Diskretisierung* der zwei-dimensionalen Poissongleichung mit homogenen Dirichlet Randbedingungen. Dabei ist $\Omega \subset \mathbb{R}^2$ ein beschränktes, zusammenhängendes

Gebiet mit stückweise geradem Rand Γ, $X = H_0^1(\Omega)$, $H = L^2(\Omega)$, $\ell(v) = \int_\Omega fv$ und $B(u,v) = \int_\Omega \nabla u \cdot \nabla v$.

Wie in §II.3 (S. 52) bezeichnet $\mathcal{T} = \{K_i : 1 \le i \le m_\mathcal{T}\}$ eine zulässige und reguläre Unterteilung von Ω in Dreiecke. Wir bezeichnen die Dreieckskanten in $\mathcal{T}$ mit $\mathcal{E}$ und die inneren Kanten mit $\mathcal{E}_\Omega$. Für $E \in \mathcal{E}$ ist m_E der Kantenmittelpunkt. Da jede Ebene durch drei nicht ko-lineare Punkte eindeutig bestimmt ist, ist für jedes $K \in \mathcal{T}$ jedes lineare Polynom $p \in \mathbb{P}_1$ eindeutig bestimmt durch seine Werte in den Kantenmittelpunkten von K. Daher ist folgende Definition sinnvoll

$$\begin{aligned} X_\mathcal{T} = CR(\mathcal{T}) = \big\{\varphi \in L^2(\Omega) : \varphi|_K \in \mathbb{P}_1 \ \forall K \in \mathcal{T}, \\ \varphi \text{ ist stetig in } m_E, \forall E \in \mathcal{E}_\Omega, \\ \varphi(m_E) = 0 \ \forall E \in \mathcal{E} \setminus \mathcal{E}_\Omega\big\}. \end{aligned}$$

$CR(\mathcal{T})$ heißt der Finite Element Raum von *Crouzeix-Raviart*.

Jedes $u_\mathcal{T} \in CR(\mathcal{T})$ ist eindeutig bestimmt durch seine Werte in den Kantenmittelpunkten m_E, $E \in \mathcal{E}_\Omega$. Bezeichnen z_0, z_1, z_2 die Eckpunkte von K und λ_0, λ_1, λ_2 die zugehörigen nodalen Basisfunktionen in $S^{1,0}(\mathcal{T})$ aus §II.3 (S. 52), so folgt mit einer leichten Rechnung $u_\mathcal{T}|_K = \sum_{i=0}^2 u_\mathcal{T}(m_i) w_i$, wobei m_i der Mittelpunkt der i-ten Kante von K und $w_i = \lambda_{i+1} + \lambda_{i+2} - \lambda_i$ ist. Dabei liegt wie in §III.5 (S. 101) die i-te Kante vereinbarungsgemäß dem i-ten Eckpunkt gegenüber und Knotennummern werden modulo 3 genommen.

Die Funktionen in $CR(\mathcal{T})$ sind nicht stetig und verschwinden nicht identisch auf Γ. Daher ist $X_\mathcal{T} \not\subset X$. Es ist aber $X_\mathcal{T} \subset H = L^2(\Omega)$. Außerdem ist offensichtlich $S_0^{1,0}(\mathcal{T}) \subset CR(\mathcal{T})$. Die Norm $\|\cdot\|_{X_\mathcal{T}}$, die Linearform $\ell_\mathcal{T}$ und die Bilinearform $B_\mathcal{T}$ werden definiert durch

$$\begin{aligned} \|u_\mathcal{T}\|_{X_\mathcal{T}} &= \left\{ \sum_{K\in\mathcal{T}} |u_\mathcal{T}|_{1;K}^2 \right\}^{\frac{1}{2}}, \\ \ell_\mathcal{T}(u_\mathcal{T}) &= \sum_{K\in\mathcal{T}} \int_K f u_\mathcal{T}, \\ B_\mathcal{T}(u_\mathcal{T}, v_\mathcal{T}) &= \sum_{K\in\mathcal{T}} \int_K \nabla u_\mathcal{T} \cdot \nabla v_\mathcal{T}. \end{aligned}$$

Offensichtlich sind die Voraussetzungen der Sätze IV.1.1 und IV.1.2 erfüllt. Insbesondere ist $\mathcal{B}_\mathcal{T} = \beta_\mathcal{T} = 1$.

Satz IV.1.3 (Abschätzung des Konsistenzfehlers). *Sei $u \in H_0^1(\Omega) \cap H^2(\Omega)$ die Lösung von (IV.1.1) und*

$$L_u(w) = B_\mathcal{T}(u,w) - \ell_\mathcal{T}(w) \quad \forall w \in H_0^1(\Omega) \oplus CR(\mathcal{T}).$$

Dann gilt für alle $w \in H_0^1(\Omega) \oplus CR(\mathcal{T})$ mit einer nur von $\mathrm{C}_\mathcal{T} = \max_K \frac{h_K}{\rho_K}$ abhängigen Konstanten

$$|L_u(w)| \le c h_\mathcal{T} \|u\|_2 \|w\|_{X_\mathcal{T}}.$$

BEWEIS. Sei $w \in H^1_0(\Omega) \oplus CR(\mathcal{T})$ beliebig. Durch elementweise partielle Integration folgt mit den gleichen Notationen wie in §III.3 (S. 79)

$$\begin{aligned} L_u(w) &= \sum_{K\in\mathcal{T}} \int_K \nabla u \cdot \nabla w - \sum_{K\in\mathcal{T}} \int_K fw \\ &= \sum_{K\in\mathcal{T}} \int_K \nabla u \cdot \nabla w + \sum_{K\in\mathcal{T}} \int_K \Delta u w \\ &= \sum_{K\in\mathcal{T}} \int_{\partial K} (\mathbf{n}_K \cdot \nabla u) w \\ &= \sum_{E\in\mathcal{E}_\Omega} \int_E \mathbb{J}_E((\mathbf{n}_E \cdot \nabla u) w) + \sum_{E\in\mathcal{E}\setminus\mathcal{E}_\Omega} \int_E (\mathbf{n} \cdot \nabla u) w. \end{aligned}$$

Da $u \in H^2(\Omega)$ ist, gilt für jede innere Kante $E \in \mathcal{E}_\Omega$

$$\int_E \mathbb{J}_E(\mathbf{n}_E \cdot \nabla u) = 0.$$

Ist dagegen $E \in \mathcal{E} \setminus \mathcal{E}_\Omega$ 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 $w \in CR(\mathcal{T})$ ist. Definieren wir daher für jede Kante $E \in \mathcal{E}$ den Mittelwert von w auf E durch

$$\overline{w}_E = h_E^{-1} \int_E w,$$

so folgt

$$L_u(w) = \sum_{E\in\mathcal{E}_\Omega} \int_E \mathbb{J}_E((\mathbf{n}_E \cdot \nabla u)(w - \overline{w}_E)) + \sum_{E\in\mathcal{E}\setminus\mathcal{E}_\Omega} \int_E (\mathbf{n} \cdot \nabla u)(w - \overline{w}_E).$$

Bezeichne mit $I_\mathcal{T} : X \cap H^2(\Omega) \to S^{1,0}_0(\mathcal{T}) \subset CR(\mathcal{T})$ den nodalen Interpolationsoperator aus §II.3 (S. 52). Da für jede Kante E

$$\int_E (w - \overline{w}_E) = 0$$

und $(\mathbf{n}_E \cdot \nabla(I_\mathcal{T} u))|_E$ konstant ist, folgt

$$\begin{aligned} L_u(w) = &\sum_{E\in\mathcal{E}_\Omega} \int_E \mathbb{J}_E((\mathbf{n}_E \cdot \nabla(u - I_\mathcal{T} u))(w - \overline{w}_E)) \\ &+ \sum_{E\in\mathcal{E}\setminus\mathcal{E}_\Omega} \int_E (\mathbf{n} \cdot \nabla(u - I_\mathcal{T} u))(w - \overline{w}_E). \end{aligned}$$

Wegen $|\mathbf{n}| = |\mathbf{n}_E| = 1$ folgt hieraus mit der Cauchy-Schwarzschen Ungleichung

$$|L_u(w)| \le 2 \sum_{E\in\mathcal{E}} \|\nabla(u - I_{\mathcal{T}} u)\|_E \, \|w - \overline{w}_E\|_E \, .$$

Betrachte nun eine beliebige Kante $E \in \mathcal{E}$ und ein Dreieck $K \in \mathcal{T}$, das E als Kante hat. Bezeichne mit $\widehat{E}$ die horizontale Kathete des Referenz-Dreieckes $\widehat{K}$ und mit $F_K : \widehat{K} \to K$ eine affine Transformation von $\widehat{K}$ auf K, die $\widehat{E}$ auf E abbildet. Setze $\widehat{w} = w \circ F_K$. Aus dem Transformationssatz folgt

$$\overline{w}_E = h_E^{-1} \int_E w = \int_{\widehat{E}} \widehat{w} \quad \text{und daher} \quad \int_{\widehat{E}} (\widehat{w} - \overline{w}_E) = 0.$$

Wie im Beweis der Poincaréschen Ungleichung, Satz I.2.21 (S. 28), folgt, dass es eine Konstante $\widehat{c}$ gibt mit $\|\varphi\|_{\widehat{K}} \le \widehat{c}\,|\varphi|_{1;\widehat{K}}$ für alle $\varphi \in H^1(\widehat{K})$ mit $\int_{\widehat{E}} \varphi = 0$. Hieraus folgt mit dem Spursatz, Satz I.2.12 (S. 25), und Lemmata II.2.5 (S. 50) und II.2.6 (S. 50)

$$\begin{aligned}
\|w - \overline{w}_E\|_E &= h_E^{\frac12} \, \|\widehat{w} - \overline{w}_E\|_{\widehat{E}} \\
&\le h_E^{\frac12} c \left\{ \|\widehat{w} - \overline{w}_E\|_{\widehat{K}}^2 + |\widehat{w} - \overline{w}_E|_{1;\widehat{K}}^2 \right\}^{\frac12} \\
&\le h_E^{\frac12} c (1 + \widehat{c}^{\,2})^{\frac12} \, |\widehat{w} - \overline{w}_E|_{1;\widehat{K}} \\
&= h_E^{\frac12} c (1 + \widehat{c}^{\,2})^{\frac12} \, |\widehat{w}|_{1;\widehat{K}} \\
&\le c' h_E^{\frac12} \, |w|_{1;K} \, .
\end{aligned}$$

Mit den gleichen Argumenten folgt aus dem Spursatz I.2.12 (S. 25), Satz II.2.2 (S. 42) und Satz II.2.7 (S. 51) durch Transformation auf das Referenzelement mit $\widehat{u} = u \circ F_K$

$$\begin{aligned}
\|\nabla(u - I_{\mathcal{T}} u)\|_E &= h_E^{-\frac12} \, \|\nabla(\widehat{u} - \widehat{\pi}\widehat{u})\|_{\widehat{E}} \\
&\le c h_E^{-\frac12} \, \|\widehat{u} - \widehat{\pi}\widehat{u}\|_{2;\widehat{K}} \\
&= c h_E^{-\frac12} \left\{ \|\widehat{u} - \widehat{\pi}\widehat{u}\|_{\widehat{K}}^2 + |\widehat{u} - \widehat{\pi}\widehat{u}|_{1;\widehat{K}}^2 + |\widehat{u}|_{2;\widehat{K}}^2 \right\}^{\frac12} \\
&\le c' h_E^{-\frac12} \, |\widehat{u}|_{2;\widehat{K}} \\
&\le c'' h_E^{\frac12} \, |u|_{2;K} \, .
\end{aligned}$$

Aus diesen Abschätzungen und der Cauchy-Schwarzschen Ungleichung für Summen folgt

$$|L_u(u)| \le c \sum_{E\in\mathcal{E}} h_E \, |w|_{1;K} \, |u|_{2;K} \le c' h \, \|w\|_{X_{\mathcal{T}}} \, |u|_2 \, . \qquad \square$$

Nach diesen Vorbereitungen können wir eine Fehlerabschätzung für die Crouzeix-Raviart Diskretisierung beweisen.

SATZ IV.1.4 (A priori Fehlerabschätzung). *Seien Ω konvex, $u \in H_0^1(\Omega) \cap H^2(\Omega)$ die schwache Lösung der Poissongleichung und $u_\mathcal{T} \in CR(\mathcal{T})$ die Lösung der Crouzeix-Raviart Diskretisierung. Dann gelten die Fehlerabschätzungen*

$$\|u - u_\mathcal{T}\|_{X_\mathcal{T}} \leq c_1 h_\mathcal{T} |u|_2 , \quad \|u - u_\mathcal{T}\| \leq c_2 h_\mathcal{T}^2 |u|_2 .$$

Die Konstanten c_1, c_2 hängen von Ω und $\mathrm{C}_\mathcal{T} = \max_K \frac{h_K}{\rho_K}$ ab.

BEWEIS. Da $S_0^{1,0}(\mathcal{T}) \subset CR(\mathcal{T})$ ist, folgt aus Satz II.2.7 (S. 51)

$$\begin{aligned}\inf_{v_\mathcal{T} \in CR(\mathcal{T})} \|u - v_\mathcal{T}\|_{X_\mathcal{T}} &\leq \inf_{w_\mathcal{T} \in S_0^{1,0}(\mathcal{T})} \|u - w_\mathcal{T}\|_{X_\mathcal{T}} = \inf_{w_\mathcal{T} \in S_0^{1,0}(\mathcal{T})} |u - w_\mathcal{T}|_1 \\ &\leq |u - I_\mathcal{T} u|_1 \\ &\leq c h_\mathcal{T} |u|_2 .\end{aligned}$$

Damit folgt die erste Fehlerabschätzung aus den Sätzen IV.1.1 und IV.1.3.
Satz IV.1.3 mit $w = u - u_\mathcal{T}$ liefert andererseits

$$\begin{aligned}|B_\mathcal{T}(u - u_\mathcal{T}, u_\varphi) - (\varphi\,,\, u - u_\mathcal{T})| &\leq c h_\mathcal{T} |u_\varphi|_2 \|u - u_\mathcal{T}\|_{X_\mathcal{T}} \\ &\leq c' h_\mathcal{T} \|\varphi\| \|u - u_\mathcal{T}\|_{X_\mathcal{T}} .\end{aligned}$$

Satz IV.1.3 mit $w = u_\varphi - u_{\varphi,\mathcal{T}}$ und der erste Teil des Beweises ergeben weiter

$$\begin{aligned}|B_\mathcal{T}(u, u_\varphi - u_{\varphi,\mathcal{T}}) - \ell_\mathcal{T}(u_\varphi - u_{\varphi,\mathcal{T}})| &\leq c h_\mathcal{T} \|u_\varphi - u_{\varphi,\mathcal{T}}\|_{X_\mathcal{T}} |u|_2 \\ &\leq c' h_\mathcal{T}^2 |u_\varphi|_2 |u|_2 \\ &\leq c'' h_\mathcal{T}^2 \|\varphi\| |u|_2 .\end{aligned}$$

Damit folgt die zweite Fehlerabschätzung des Satzes aus der ersten Abschätzung und Satz IV.1.2. □

BEMERKUNG IV.1.5. (1) Sei $J(u) = \frac{1}{2} B_\mathcal{T}(u, u) - \ell_\mathcal{T}(u)$. Wegen $S_0^{1,0}(\mathcal{T}) \subset X$ ist

$$\inf_{u \in X} J(u) \leq \inf_{u_\mathcal{T} \in S_0^{1,0}(\mathcal{T})} J(u_\mathcal{T}).$$

Wegen $S_0^{1,0}(\mathcal{T}) \subset CR(\mathcal{T})$ ist ebenso

$$\inf_{u_\mathcal{T} \in CR(\mathcal{T})} J(u_\mathcal{T}) \leq \inf_{u_\mathcal{T} \in S_0^{1,0}(\mathcal{T})} J(u_\mathcal{T}).$$

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.3.1 (S. 53) nicht nur die H^2-Regularität der aktuellen Lösung u sondern die H^2-Regularität für jede rechte Seite in $L^2(\Omega)$. Daher muss Ω als konvex vorausgesetzt werden. Diese zusätzliche Einschränkung ist nicht Beweistechnik. Nicht-konforme Methoden reagieren häufig empfindlicher als konforme Methoden auf mangelnde H^2-Regularität z.B. aufgrund einspringender

Ecken.
(3) Konstruktionsgemäß gilt $L_u(w) = 0$ für alle $w \in H^1_0(\Omega)$. Dies gilt aber nicht für L^2-Funktionen, da L_u auf $L^2(\Omega)$ nicht stetig fortsetzbar ist.
(4) Aus der Eulerschen Polyederformel $\sharp\mathcal{T} - \sharp\mathcal{E} + \sharp\mathcal{N} = 1$ und der Identität $3\sharp\mathcal{T} = 2\sharp\mathcal{E} - \sharp\mathcal{E}_\Gamma$ folgt $\sharp\mathcal{T} \approx 2\sharp\mathcal{N}$ und $\sharp\mathcal{E} \approx 3\sharp\mathcal{N}$. Daher ist $\dim CR(\mathcal{T}) \approx 3 \dim S^{1,0}_0(\mathcal{T})$.

IV.2. Discontinuous Galerkin Methoden

In diesem Abschnitt betrachten wir wie im vorigen die Poissongleichung mit homogenen Dirichlet Randbedingungen. Im abstrakten Rahmen von §IV.1 ist wieder $\Omega \subset \mathbb{R}^2$ ein beschränktes, zusammenhängendes Gebiet mit stückweise geradem Rand Γ, $X = H^1_0(\Omega)$, $H = L^2(\Omega)$, $\ell(v) = \int_\Omega fv$ und $B(u,v) = \int_\Omega \nabla u \cdot \nabla v$.

Wir wollen jetzt für das diskrete Problem (IV.1.2) (S. 110) den diskreten Raum $X_\mathcal{T}$ und die diskreten Bilinear- und Linearformen $B_\mathcal{T}$ und $\ell_\mathcal{T}$ so konstruieren, dass folgende Wünsche erfüllt sind:

- *Unstetigkeit:* Die Ansatz- und Testfunktionen sind komplett unstetig und müssen die Dirichlet Randbedingung nicht erfüllen, d.h. $S^{k,-1}(\mathcal{T}) \subset X_\mathcal{T} \subset L^2(\Omega)$,
- *Konsistenz:* Für die schwache Lösung u der Poissongleichung und jede Funktion $v \in L^2(\Omega)$, die elementweise in $H^1(K)$ ist, gilt $B_\mathcal{T}(u,v) = \ell_\mathcal{T}(v)$.
- *Symmetrie:* $B_\mathcal{T}$ ist symmetrisch.
- *Beschränktheit:* Die Bilinearform $B_\mathcal{T}$ ist stetig und die Norm $\mathcal{B}_\mathcal{T}$ von $B_\mathcal{T}$ ist bezüglich $\mathcal{T}$ und eventueller weiterer Parameter der Diskretisierung gleichmäßig nach oben beschränkt.
- *Koerzivität:* Die Bilinearform $B_\mathcal{T}$ ist koerziv und die entsprechende Größe $\beta_\mathcal{T}$ ist bezüglich $\mathcal{T}$ und eventueller weiterer Parameter der Diskretisierung gleichmäßig positiv.

Bevor wir $B_\mathcal{T}$ und $\ell_\mathcal{T}$ konstruieren, erinnern wir an den Sprung $\mathbb{J}_E(\cdot)$ über eine innere Kante E in Richtung eines Einheitsnormalenvektors $\mathbf{n}_E$. Das Vorzeichen von $\mathbb{J}_E(\cdot)$ hängt von der Orientierung von $\mathbf{n}_E$ ab. Dies ist aber für die folgenden Argumente unerheblich. Zusätzlich definieren wir für eine innere Kante E den Mittelwert $\mathbb{A}_E(\cdot)$ durch

$$\mathbb{A}_E(\varphi)(x) = \frac{1}{2}\left(\lim_{t\to 0+} \varphi(x + t\mathbf{n}_E) + \lim_{t\to 0+} \varphi(x - t\mathbf{n}_E)\right) \quad \forall x \in E.$$

Dann gilt für alle inneren Kanten und alle Funktionen φ, ψ

$$\mathbb{J}_E(\varphi\psi) = \mathbb{J}_E(\varphi)\mathbb{A}_E(\psi) + \mathbb{A}_E(\varphi)\mathbb{J}_E(\psi).$$

Für Randkanten $E \subset \Gamma$ setzen wir zudem

$$\mathbb{J}_E(\varphi)(x) = \mathbb{A}_E(\varphi)(x) = \lim_{t\to 0+} \varphi(x - t\mathbf{n}_E) \quad \forall x \in E,$$

wobei jetzt $\mathbf{n}_E$ die äußere Normale zu Γ ist.

Im Folgenden nehmen wir an, dass die rechte Seite f in $L^2(\Omega)$ und die schwache Lösung u der Poissongleichung in $H^2(\Omega)$ ist. Wegen der Dirichlet Randbedingung gilt dann für jede Kante $\mathbb{J}_E(u) = 0$. Außerdem ist $\nabla u \in H(\mathrm{div};\Omega)$, so dass für jede innere Kante $\mathbb{J}_E(\mathbf{n}_E \cdot \nabla u) = 0$ ist.

Seien nun v und w beliebige Funktionen in $L^2(\Omega)$, die für jedes Element K in $H^1(K)$ sind. Wir multiplizieren die Poissongleichung auf einem beliebigen Element K mit w, integrieren über K und integrieren die Ableitungsterme partiell. Dies liefert mit der äußeren Normalen $\mathbf{n}_K$ zu ∂K

$$\int_K fw = \int_K (-\Delta u)w = \int_K \nabla u \cdot \nabla w - \int_{\partial K} \mathbf{n}_K \cdot \nabla u w$$

Summation über alle Elemente K ergibt

$$\begin{aligned}\sum_{K\in\mathcal{T}} \int_K fw = \sum_{K\in\mathcal{T}} \int_K \nabla u \cdot \nabla w \\ - \sum_{E\in\mathcal{E}_\Omega} \int_E \mathbb{J}_E(\mathbf{n}_E \cdot \nabla u w) - \sum_{E\in\mathcal{E}_\Gamma} \int_E \mathbf{n}_E \cdot \nabla u w.\end{aligned}$$

Für jede innere Kante ist wegen $\mathbb{J}_E(\mathbf{n}_E \cdot \nabla u) = 0$

$$\begin{aligned}\int_E \mathbb{J}_E(\mathbf{n}_E \cdot \nabla u w) &= \int_E \mathbb{J}_E(\mathbf{n}_E \cdot \nabla u)\mathbb{A}_E(w) + \int_E \mathbb{A}_E(\mathbf{n}_E \cdot \nabla u)\mathbb{J}_E(w) \\ &= \int_E \mathbb{A}_E(\mathbf{n}_E \cdot \nabla u)\mathbb{J}_E(w),\end{aligned}$$

und für jede Randkante ist wegen der Definition von Sprung und Mittelwert

$$\int_E \mathbf{n}_E \cdot \nabla u w = \int_E \mathbb{A}_E(\mathbf{n}_E \cdot \nabla u)\mathbb{J}_E(w).$$

Insgesamt erhalten wir somit

$$\sum_{K\in\mathcal{T}} \int_K fw = \sum_{K\in\mathcal{T}} \int_K \nabla u \cdot \nabla w - \sum_{E\in\mathcal{E}} \int_E \mathbb{A}_E(\mathbf{n}_E \cdot \nabla u)\mathbb{J}_E(w).$$

Daher sind

$$\begin{aligned} w &\mapsto \sum_{K\in\mathcal{T}} \int_K fw, \\ v, w &\mapsto \sum_{K\in\mathcal{T}} \int_K \nabla v \cdot \nabla w - \sum_{E\in\mathcal{E}} \int_E \mathbb{A}_E(\mathbf{n}_E \cdot \nabla v)\mathbb{J}_E(w)\end{aligned}$$

Kandidaten für $\ell_\mathcal{T}$ und $B_\mathcal{T}$, die die Konsistenzbedingung erfüllen. Offensichtlich ist dieser Kandidat für $B_\mathcal{T}$ aber nicht symmetrisch. Die Symmetrie erhalten wir, indem wir $-\sum_{E\in\mathcal{E}} \int_E \mathbb{A}_E(\mathbf{n}_E \cdot \nabla w)\mathbb{J}_E(v)$ zu unserem Kandidaten hinzufügen. Wegen $\mathbb{J}_E(u) = 0$ für alle Kanten zerstört dies nicht die Konsistenz.

Als nächstes untersuchen wir die Beschränktheit dieses Kandidaten für $B_{\mathcal{T}}$, da dies die Wahl von $\|\cdot\|_{X_{\mathcal{T}}}$ beeinflusst. Dazu setzen wir von nun an voraus, dass v und w in $S^{k,-1}(\mathcal{T})$ sind und überlegen uns mit einem Skalierungsargument (d.h. Transformation auf das Referenzelement, Äquivalenz von Normen auf endlich dimensionalen Räumen dort und Rücktransformation), dass es eine Konstante c_{tr} gibt, die nur von dem Polynomgrad k und dem Formparameter $\mathrm{C}_{\mathcal{T}}$ abhängt, so dass für alle Elemente K, alle Kanten E von K und alle Polynome $\varphi \in R_k(K)$ gilt

$$\|\varphi\|_E \leq c_{tr} h_E^{-\frac{1}{2}} \|\varphi\|_K .$$

Dann erhalten wir mit der Cauchy-Schwarzschen Ungleichung

$$\begin{aligned}
&\left| \sum_{K\in\mathcal{T}} \int_K \nabla v \cdot \nabla w - \sum_{E\in\mathcal{E}} \int_E \mathbb{A}_E(\mathbf{n}_E \cdot \nabla v)\mathbb{J}_E(w) \right. \\
&\qquad \left. - \sum_{E\in\mathcal{E}} \int_E \mathbb{A}_E(\mathbf{n}_E \cdot \nabla w)\mathbb{J}_E(v) \right| \\
&\leq \sum_{K\in\mathcal{T}} |v|_{1;K} |w|_{1;K} + \sum_{E\in\mathcal{E}} c_{tr} h_E^{-\frac{1}{2}} |v|_{1;K} \|\mathbb{J}_E(w)\|_E \\
&\qquad + \sum_{E\in\mathcal{E}} c_{tr} h_E^{-\frac{1}{2}} |w|_{1;K} \|\mathbb{J}_E(v)\|_E \\
&\leq c_{tr} \left\{ \sum_{K\in\mathcal{T}} |v|_{1;K}^2 + \sum_{E\in\mathcal{E}} h_E^{-1} \|\mathbb{J}_E(v)\|_E^2 \right\}^{\frac{1}{2}} \cdot \\
&\qquad \cdot \left\{ \sum_{K\in\mathcal{T}} |w|_{1;K}^2 + \sum_{E\in\mathcal{E}} h_E^{-1} \|\mathbb{J}_E(w)\|_E^2 \right\}^{\frac{1}{2}} .
\end{aligned}$$

Also ist unser Kandidat für $B_{\mathcal{T}}$ bezüglich

$$\|v\|_{X_{\mathcal{T}}} = \left\{ \sum_{K\in\mathcal{T}} |v|_{1;K}^2 + \sum_{E\in\mathcal{E}} h_E^{-1} \|\mathbb{J}_E(v)\|_E^2 \right\}^{\frac{1}{2}}$$

gleichmäßig beschränkt.

Bevor wir uns der Koerzivität zuwenden, müssen wir noch nachprüfen, dass $\|\cdot\|_{X_{\mathcal{T}}}$ tatsächlich eine Norm ist. Die Homogenität und die Dreiecksungleichung sind klar. Wir müssen nur nachprüfen, dass aus $\|v\|_{X_{\mathcal{T}}} = 0$ auch $v = 0$ folgt. Aus $\|v\|_{X_{\mathcal{T}}} = 0$ folgt $|v|_{1;K} = 0$ für jedes Element K. Also ist v elementweise konstant. Aus $\|v\|_{X_{\mathcal{T}}} = 0$ folgt $\|\mathbb{J}_E(v)\|_E = 0$ für jede innere Kante E. Also ist v global konstant. Da aber $\|v\|_{X_{\mathcal{T}}}$ mindestens einen Beitrag $\|\mathbb{J}_E(v)\|_E = \|v\|_E$ einer Randkante enthält, folgt wie gewünscht $v = 0$.

Setzen wir in unseren Kandidaten für $B_{\mathcal{T}}$ zweimal das gleiche Argument $v \in S^{k,-1}(\mathcal{T})$ ein, erhalten wir mit obiger Spurungleichung

$$\begin{aligned}
&\sum_{K\in\mathcal{T}} |v|^2_{1;K} - 2\sum_{E\in\mathcal{E}} \int_E \mathbb{A}_E(\mathbf{n}_E\cdot v)\mathbb{J}_E(v) \\
&\quad\geq \sum_{K\in\mathcal{T}} |v|^2_{1;K} - 2\sum_{K\in\mathcal{T}}\sum_{E\in\mathcal{E}_K} c_{tr}\,|v|_{1;K}\, h_E^{-\frac{1}{2}}\,\|\mathbb{J}_E(v)\|_E \\
&\quad\geq \frac{1}{2}\sum_{K\in\mathcal{T}} |v|^2_{1;K} - 4c_{tr}^2 \sum_{E\in\mathcal{E}} h_E^{-1}\,\|\mathbb{J}_E(v)\|_E^2\,.
\end{aligned}$$

Dies zeigt, dass wir Koerzivität erhalten, wenn wir zu unserem Kandidaten für $B_{\mathcal{T}}$ noch die Terme $\delta\sum_{E\in\mathcal{E}} h_E^{-1}\int_E \mathbb{J}_E(v)\mathbb{J}_E(w)$ mit $\delta > 4c_{tr}^2$ hinzufügen. Solange $\delta \approx 4c_{tr}^2$ ist, bleibt die Beschränktheit bestehen. Die Symmetrie wird offensichtlich auch nicht zerstört. Da $\mathbb{J}_E(u) = 0$ ist für jede Kante E, bleibt auch die Konsistenz erhalten.

Diese Überlegungen führen auf die folgenden Daten für die *discontinuous Galerkin Methode*:

$$\begin{aligned}
X_{\mathcal{T}} &= S^{k,-1}(\mathcal{T}),\\
\|v\|_{X_{\mathcal{T}}} &= \left\{\sum_{K\in\mathcal{T}} |v|^2_{1;K} + \sum_{E\in\mathcal{E}} h_E^{-1}\,\|\mathbb{J}_E(v)\|_E^2\right\}^{\frac{1}{2}},\\
\ell_{\mathcal{T}}(w) &= \sum_{K\in\mathcal{T}}\int_K fw,\\
B_{\mathcal{T}}(v,w) &= \sum_{K\in\mathcal{T}}\int_K \nabla v\cdot\nabla w - \sum_{E\in\mathcal{E}}\int_E \mathbb{A}_E(\mathbf{n}_E\cdot\nabla v)\mathbb{J}_E(w)\\
&\quad - \sum_{E\in\mathcal{E}}\int_E \mathbb{A}_E(\mathbf{n}_E\cdot\nabla w)\mathbb{J}_E(v) + \delta\sum_{E\in\mathcal{E}} h_E^{-1}\int_E \mathbb{J}_E(v)\mathbb{J}_E(w).
\end{aligned}$$

Die discontinuous Galerkin Methode passt in den abstrakten Rahmen von §IV.1. Daher liefern die Sätze IV.1.1 (S. 110) und IV.1.2 (S. 110) Fehlerabschätzungen für $\|u-u_{\mathcal{T}}\|_{X_{\mathcal{T}}}$ und $\|u-u_{\mathcal{T}}\|$. Wegen der Konsistenz fallen in diesen Sätzen die Konsistenzfehler weg. Insbesondere folgt

$$\|u-u_{\mathcal{T}}\|_{X_{\mathcal{T}}} \leq \left(1+\frac{\mathcal{B}_{\mathcal{T}}}{\beta_{\mathcal{T}}}\right)\inf_{v_{\mathcal{T}}\in S^{k,-1}(\mathcal{T})}\|u-v_{\mathcal{T}}\|_{X_{\mathcal{T}}}\,.$$

Für $k\geq 1$ ist $S_0^{1,0}(\mathcal{T}) \subset S_0^{k,0}(\mathcal{T}) \subset S^{k,-1}(\mathcal{T})$. Da zudem $\|\cdot\|_{X_{\mathcal{T}}}$ auf $H^1(\Omega)$ mit $|\cdot|_1$ übereinstimmt, folgt für $u\in H^2(\Omega)$

$$\|u-u_{\mathcal{T}}\|_{X_{\mathcal{T}}} \leq c\inf_{v_{\mathcal{T}}\in S_0^{1,0}(\mathcal{T})}|u-v_{\mathcal{T}}|_1 \leq ch_{\mathcal{T}}\,\|u\|_2$$

und für $u\in H^{k+1}(\Omega)$

$$\|u-u_{\mathcal{T}}\|_{X_{\mathcal{T}}} \leq c\inf_{v_{\mathcal{T}}\in S_0^{k,0}(\mathcal{T})}|u-v_{\mathcal{T}}|_1 \leq ch_{\mathcal{T}}^k\,\|u\|_{k+1}\,.$$

Damit folgt aus den Sätzen IV.1.1 und IV.1.2:

SATZ IV.2.1 (A priori Fehlerabschätzung). *Seien* $k \geq 1$, $u_{\mathcal{T}} \in S^{k,-1}(\mathcal{T})$ *die Lösung der discontinuous Galerkin Diskretisierung und* $u \in H_0^1(\Omega) \cap H^{k+1}(\Omega)$ *die schwache Lösung der Poissongleichung. Dann ist*

$$\|u - u_{\mathcal{T}}\|_{X_{\mathcal{T}}} \leq c_1 h_{\mathcal{T}}^k \, |u|_{k+1} .$$

Ist zusätzlich Ω *konvex, gilt*

$$\|u - u_{\mathcal{T}}\| \leq c_2 h_{\mathcal{T}}^{k+1} \, |u|_{k+1} .$$

Die Konstanten c_1, c_2 *hängen von* Ω, k *und* $\mathrm{C}_{\mathcal{T}} = \max_K \frac{h_K}{\rho_K}$ *ab.*

BEMERKUNG IV.2.2. (1) Man kann auch $k = 0$ für die discontinuous Galerkin Diskretisierung wählen. Dann enthält $B_{\mathcal{T}}$ nur die Sprungterme $\delta \sum_{E \in \mathcal{E}} h_E^{-1} \int_E \mathbb{J}_E(v) \mathbb{J}_E(w)$. Der obige Beweis der a priori Fehlerabschätzung funktioniert dann nicht mehr, da wir nicht mehr auf die Inklusion $S_0^{1,0}(\mathcal{T}) \subset S^{k,-1}(\mathcal{T})$ zurückgreifen können. Technisch aufwändigere, verfeinerte Methoden liefern aber noch eine Fehlerabschätzung der Form $\|u - u_{\mathcal{T}}\|_{X_{\mathcal{T}}} \leq c h_{\mathcal{T}}^{\frac{1}{2}} \, |u|_2$.
(2) Die discontinuous Galerkin Methode ist bei Problemen mit dominanter Konvektion besonders beliebt, da sie besonders flexible upwind- oder SDFEM-artige Approximationen des Konvektionstermes erlaubt [8].
(3) Wie in Bemerkung IV.1.5(4) (S. 115) folgt $\dim S^{1,-1}(\mathcal{T}) = 3\sharp\mathcal{T} \approx 6\sharp\mathcal{N} \approx 6 \dim S_0^{1,0}(\mathcal{T})$ und $\dim S^{2,-1}(\mathcal{T}) = 6\sharp\mathcal{T} \approx 12\sharp\mathcal{N} \approx 3(\sharp\mathcal{N} + \sharp\mathcal{E}) \approx 3 \dim S_0^{2,0}(\mathcal{T})$. Für $k \geq 3$ können bei der konformen Finite Element Methode und bei der discontinuous Galerkin Methode die inneren Freiheitsgrade eliminiert werden, sog. statische Kondensation, so dass der Aufwand der discontinuous Galerkin Methode asymptotisch etwa doppelt so groß ist wie derjenige der konformen Methode gleichen Polynomgrads.

IV.3. Gemischte Finite Elemente

Um technische Schwierigkeiten zu vermeiden und die wesentlichen Aspekte besser herauszuarbeiten, beschränken wir uns im Folgenden auf ein einfaches Modellbeispiel: die Poissongleichung in einem zweidimensionalen, beschränkten, *konvexen* Polygon Ω mit homogenen Dirichlet-Randbedingungen

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \Gamma. \end{aligned} \tag{IV.3.1}$$

Dies ist ein extrem einfaches Modell für die vertikale Auslenkung u einer ebenen, dünnen, eingespannten Membran unter Einfluss einer vertikalen Last f. Die bisher betrachteten Methoden liefern Approximationen der Auslenkung. Unter mechanischen Gesichtspunkten sind aber

häufig die aus der Belastung resultierenden inneren Spannungskräfte viel wichtiger. Im Rahmen dieses einfachen Modells ist dies der Gradient ∇u der Auslenkung. Diese Größe muss bei den bisher betrachteten Methoden durch Differentiation aus der Auslenkung u berechnet werden und wird i.a. um eine h-Potenz schlechter approximiert als die Auslenkung. Wegen der Bedeutung dieser Größe sucht man nach Verfahren, die sie direkt und genauer approximieren.

Dies leisten sog. *gemischte Finite Element Methoden*. Dazu führen wir $\sigma = \nabla u$ als zusätzliche Variable ein. Damit geht (IV.3.1) über in das folgende Differentialgleichungssystem 1. Ordnung

$$\begin{aligned} \sigma - \nabla u &= 0 \quad \text{in } \Omega \\ -\operatorname{div}\sigma &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \Gamma. \end{aligned} \tag{IV.3.2}$$

Multiplizieren wir die erste Gleichung von (IV.3.2) mit einem hinreichend glatten Vektorfeld τ, integrieren über Ω, wenden den Gaußschen Integralsatz an und nutzen die Randbedingungen für u aus, erhalten wir

$$\begin{aligned} 0 &= \int_\Omega \sigma\cdot\tau - \int_\Omega \tau\cdot\nabla u = \int_\Omega \sigma\cdot\tau + \int_\Omega u \operatorname{div}\tau - \int_\Gamma u\tau\cdot n \\ &= \int_\Omega \sigma\cdot\tau + \int_\Omega u\operatorname{div}\tau. \end{aligned}$$

Offensichtlich ist die zweite Gleichung von (IV.3.2) äquivalent zu

$$-\int_\Omega v\operatorname{div}\sigma = \int_\Omega fv \quad \forall v\in L^2(\Omega).$$

Diese Beobachtung führt auf folgende schwache Formulierung von Problem (IV.3.1): Finde $[\sigma,u]\in X = M\times Q$, so dass

$$\begin{aligned} \int_\Omega \sigma\cdot\tau + \int_\Omega u\operatorname{div}\tau &= 0 \qquad \forall\tau\in M \\ -\int_\Omega v\operatorname{div}\sigma &= \int_\Omega fv \quad \forall v\in Q. \end{aligned} \tag{IV.3.3}$$

Dabei ist

$$\begin{aligned} M &= H(\operatorname{div},\Omega) = \{\sigma\in L^2(\Omega;\mathbb{R}^2) : \operatorname{div}\sigma\in L^2(\Omega;\mathbb{R})\}, \\ Q &= L^2(\Omega). \end{aligned}$$

M wird versehen mit der Norm

$$\|\sigma\|_{H(\operatorname{div};\Omega)} = \left\{\|\sigma\|^2 + \|\operatorname{div}\sigma\|^2\right\}^{\frac{1}{2}}.$$

Wie in Satz I.2.6 (S. 23) kann man zeigen, dass M mit dieser Norm ein Banach-Raum ist.

Die Herleitung von (IV.3.3) zeigt, dass dieses Problem im üblichen Sinne zu (IV.3.2) äquivalent ist: Jede klassische Lösung von (IV.3.2)

löst auch (IV.3.3) und jede hinreichend glatte Lösung von (IV.3.3) ist auch eine klassische Lösung von (IV.3.2).

Problem (IV.3.3) ist ein sogenanntes *gemischtes* oder *Sattelpunktsproblem*. Es stellt die Euler-Lagrange Gleichungen des folgenden Optimierungsproblems mit Nebenbedingungen dar:

$$J(\sigma) = \frac{1}{2}\int_\Omega \sigma\cdot\sigma \to \min \quad \text{in } M_f = \{\sigma \in M : -\operatorname{div}\sigma = f\}.$$

Für die Analyse von (IV.3.3) führen wir folgende Abkürzungen ein:

$$\begin{aligned} \|[\sigma,u]\|_X &= \left\{\|\sigma\|^2 + \|\operatorname{div}\sigma\|^2 + \|u\|^2\right\}^{\frac{1}{2}}, \\ B([\sigma,u],[\tau,v]) &= \int_\Omega \sigma\cdot\tau + \int_\Omega u\operatorname{div}\tau - \int_\Omega v\operatorname{div}\sigma, \\ \ell([\tau,v]) &= \int_\Omega fv. \end{aligned}$$

Dann ist (IV.3.3) offensichtlich äquivalent zu

$$B([\sigma,u],[\tau,v]) = \ell([\tau,v]) \quad \forall [\tau,v]\in X. \tag{IV.3.4}$$

Auf den ersten Blick scheint Problem (IV.3.4) in den abstrakten Rahmen von §I.1 (S. 15) zu passen. Aber die Bilinearform B ist nicht koerziv. Dies spiegelt die Sattelpunktsstruktur von (IV.3.3) wider. Statt der Koerzivität erfüllt B allerdings eine sog. *inf-sup Bedingung*.

Lemma IV.3.1 (inf-sup Bedingung). *Es gibt eine Konstante $\beta > 0$, die nur von Ω abhängt, mit*

$$\inf_{[\sigma,u]\in X\setminus\{0\}} \sup_{[\tau,v]\in X\setminus\{0\}} \frac{B([\sigma,u],[\tau,v])}{\|[\sigma,u]\|_X \, \|[\tau,v]\|_X} \geq \beta.$$

Beweis. Sei $[\sigma,u]\in X\setminus\{0\}$ beliebig, aber fest. Dann gilt

$$B([\sigma,u],[\sigma,u]) = \|\sigma\|^2$$

und wegen $\operatorname{div}\sigma \in L^2(\Omega)$

$$B([\sigma,u],[0,-\operatorname{div}\sigma]) = \|\operatorname{div}\sigma\|^2.$$

Da Ω beschränkt ist, gibt es ein $R>0$ mit $\Omega\subset(-R,R)^2$. Setze u durch 0 auf ganz $\mathbb{R}^2$ fort und definiere

$$\tau_u(x) = e_1\int_{-R}^{x_1} u(s,x_2)\mathrm{d}s \quad \forall x=(x_1,x_2)\in\Omega.$$

Dabei ist e_1 der erste Einheitsvektor in $\mathbb{R}^2$. Offensichtlich ist $\tau_u\in M$ und erfüllt

$$\operatorname{div}\tau_u = u, \quad \|\tau_u\| \leq c_0\|u\|.$$

Dabei hängt die Konstante c_0 nur vom Durchmesser von Ω ab. O.E. ist $c_0 \geq 1$. Hieraus folgt

$$\begin{aligned} &B([\sigma,u],[\tau_u,0]) \\ &\quad = \int_\Omega \sigma\cdot\tau_u + \|u\|^2 \geq -\|\sigma\|\,\|\tau_u\| + \|u\|^2 \geq -\|\sigma\|\,c_0\,\|u\| + \|u\|^2 \\ &\quad \geq \frac{1}{2}\|u\|^2 - \frac{1}{2}c_0^2\|\sigma\|^2 . \end{aligned}$$

Setze

$$\rho = c_0^2\sigma + \tau_u, \quad w = c_0^2 u - \frac{1}{2}\operatorname{div}\sigma.$$

Dann folgt

$$\begin{aligned} &B([\sigma,u],[\rho,w]) \\ &\quad = c_0^2 B([\sigma,u],[\sigma,u]) + B([\sigma,u],[\tau_u,0]) + \frac{1}{2}B([\sigma,u],[0,-\operatorname{div}\sigma]) \\ &\quad \geq c_0^2\|\sigma\|^2 + \frac{1}{2}\|u\|^2 - \frac{1}{2}c_0^2\|\sigma\|^2 + \frac{1}{2}\|\operatorname{div}\sigma\|^2 \\ &\quad \geq \frac{1}{2}\|[\sigma,u]\|_X^2 \end{aligned}$$

und wegen $\|\tau_u\|^2 \leq c_0^2\|u\|^2$

$$\begin{aligned} \|[\rho,w]\|_X &\leq c_0^2\|[\sigma,u]\|_X + \|[\tau_u,0]\|_X + \frac{1}{2}\|[0,-\operatorname{div}\sigma]\|_X \\ &= c_0^2\|[\sigma,u]\|_X + \left\{\|u\|^2 + \|\tau_u\|^2\right\}^{\frac{1}{2}} + \frac{1}{2}\|\operatorname{div}\sigma\| \\ &\leq \left\{c_0^4 + c_0^2 + 1 + \frac{1}{4}\right\}^{\frac{1}{2}}\|[\sigma,u]\|_X \\ &\leq 2c_0^2\|[\sigma,u]\|_X . \end{aligned}$$

Aus diesen Abschätzungen folgt die Behauptung mit $\beta = \frac{1}{4c_0^2}$. □

Aus Lemma IV.3.1 folgt:

Satz IV.3.2 (Eindeutige Lösbarkeit des Sattelpunktproblems). *Das Problem* (IV.3.3) *besitzt eine eindeutige Lösung.*

Beweis. Sei $u \in H_0^1(\Omega)$ die schwache Lösung von (IV.3.1) im Sinne von Definition I.3.1 (S. 30). Da Ω konvex ist, folgt aus Satz I.3.6 (S. 33) $u \in H^2(\Omega)$. Also ist $\sigma = \nabla u \in H(\operatorname{div},\Omega)$. Aus der Herleitung von (IV.3.3) folgt, dass $[\sigma,u]$ eine Lösung von (IV.3.3) ist. Wir müssen also noch die Eindeutigkeit zeigen. Dazu reicht es, zu zeigen, dass das homogene Problem, d.h. (IV.3.3) mit $f = 0$ bzw. (IV.3.4) mit $\ell = 0$, nur die triviale Lösung hat. Ist aber $[\sigma,u]$ eine Lösung des homogenen Problems (IV.3.4), so folgt aus Lemma IV.3.1

$$\beta\,\|[\sigma,u]\|_X \leq \sup_{[\tau,v]\in X\setminus\{0\}} \frac{B([\sigma,u],[\tau,v])}{\|[\tau,v]\|_X} = 0.$$

Also ist $\sigma = u = 0$. □

Für die Diskretisierung von (IV.3.3) betrachten wir nur das einfachste Beispiel, das sog. *Raviart-Thomas Element* niedrigster Ordnung. Dazu bezeichnet $\mathcal{T}$ eine Familie zulässiger und regulärer Triangulierungen von Ω und $\mathcal{E}$ die Menge der Dreieckskanten in $\mathcal{T}$. Für ein Dreieck K sei

$$RT(K) = \left\{ \begin{pmatrix} p \\ q \end{pmatrix} + r \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : p, q, r \in \mathbb{R} \right\}.$$

LEMMA IV.3.3 (Eigenschaften der Raviart-Thomas Elemente). *Sei K ein Dreieck und $\mathbf{n}_K$ das äußere Einheitsnormalenfeld zu ∂K. Dann gilt für jedes $\sigma \in RT(K)$:*

(1) *$\sigma \cdot \mathbf{n}_K$ ist konstant auf den Kanten von K.*
(2) *σ ist eindeutig bestimmt durch die Werte von $\sigma \cdot \mathbf{n}_K$ auf den Kanten von K.*

BEWEIS. *ad (1):* Die Funktion $x \mapsto x \cdot \mathbf{n}_K$ 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 $\sigma \cdot \mathbf{n}_K$ auf allen Kanten von K verschwindet. Sei $\sigma = \left(\begin{smallmatrix} p \\ q \end{smallmatrix}\right) + r \left(\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right)$ eine solche Funktion. Dann folgt aus dem Gaußschen Integralsatz

$$0 = \int_{\partial K} \sigma \cdot \mathbf{n}_K = \int_K \operatorname{div} \sigma = \int_K 2r \quad \Longrightarrow \quad r = 0.$$

Also steht der Vektor $\left(\begin{smallmatrix} p \\ q \end{smallmatrix}\right)$ senkrecht auf dem zweidimensionalen Raum, der von den Richtungsvektoren der drei Kanten von K aufgespannt wird. Also ist auch $p = q = 0$. □

Wegen Lemma IV.3.3 ist folgende Definition sinnvoll:

$$RT^{-1}(\mathcal{T}) = \left\{ \sigma : \Omega \longrightarrow \mathbb{R}^2 : \sigma|_K \in RT(K) \, \forall K \in \mathcal{T} \right\},$$
$$RT(\mathcal{T}) = RT^{-1}(\mathcal{T}) \cap H(\operatorname{div}, \Omega).$$

Wie im Beweis von Satz I.2.7 (S. 24) folgt, dass $\sigma \in RT^{-1}(\mathcal{T})$ genau dann in $RT(\mathcal{T})$ liegt, wenn $\sigma \cdot \mathbf{n}_K$ stetig ist über alle Dreieckskanten, die in Ω liegen. Wegen Lemma IV.3.3 ist daher $RT(\mathcal{T}) \neq \{0\}$. Die Freiheitsgrade der Funktionen $\sigma_\mathcal{T} \in RT(\mathcal{T})$ sind genau die Werte von $\sigma_\mathcal{T} \cdot \mathbf{n}_K$ auf den Kanten in $\mathcal{E}$. Insbesondere ist $\dim RT(\mathcal{T}) = \#\mathcal{E}$. Wir setzen nun

$$M_\mathcal{T} = RT(\mathcal{T}), \quad Q_\mathcal{T} = S^{0,-1}(\mathcal{T}), \quad X_\mathcal{T} = M_\mathcal{T} \times Q_\mathcal{T}$$

und approximieren Problem (IV.3.4) durch: Finde $[\sigma_\mathcal{T}, u_\mathcal{T}] \in X_\mathcal{T}$, so dass

$$B([\sigma_\mathcal{T}, u_\mathcal{T}], [\tau_\mathcal{T}, v_\mathcal{T}]) = \ell([\tau_\mathcal{T}, v_\mathcal{T}]) \quad \forall [\tau_\mathcal{T}, v_\mathcal{T}] \in X_\mathcal{T} \tag{IV.3.5}$$

oder in anderer Schreibweise: Finde $[\sigma_\mathcal{T}, u_\mathcal{T}] \in X_\mathcal{T}$, so dass

$$\begin{aligned} \int_\Omega \sigma_\mathcal{T} \cdot \tau_\mathcal{T} + \int_\Omega u_\mathcal{T} \operatorname{div} \tau_\mathcal{T} &= 0 \qquad \forall \tau_\mathcal{T} \in M_\mathcal{T} \\ -\int_\Omega v_\mathcal{T} \operatorname{div} \sigma_\mathcal{T} &= \int_\Omega f v_\mathcal{T} \quad \forall v_\mathcal{T} \in Q_\mathcal{T}. \end{aligned} \tag{IV.3.6}$$

Satz IV.3.4 (Eindeutige Lösbarkeit des diskreten Problems). *(1) Es ist* $\operatorname{div} M_\mathcal{T} = Q_\mathcal{T}$. *Zu jedem* $u_\mathcal{T} \in Q_\mathcal{T}$ *gibt es ein* $\tau_{u_\mathcal{T},\mathcal{T}} \in M_\mathcal{T}$ *mit*

$$\operatorname{div} \tau_{u_\mathcal{T},\mathcal{T}} = u_\mathcal{T}, \quad \|\tau_{u_\mathcal{T},\mathcal{T}}\| \le c_1 \|u_\mathcal{T}\|.$$

Die Konstante c_1 *hängt nur vom Durchmesser von* Ω *und dem Formparameter* $\mathrm{C}_\mathcal{T} = \max_K \frac{h_K}{\rho_K}$ *ab.*
(2) Es gibt eine Konstante $\beta > 0$, *die nur von der Konstanten* c_1 *aus Teil (1) abhängt, so dass die* diskrete inf-sup Bedingung *gilt*

$$\inf_{[\sigma_\mathcal{T},u_\mathcal{T}]\in X_\mathcal{T}\setminus\{0\}} \sup_{[\tau_\mathcal{T},v_\mathcal{T}]\in X_\mathcal{T}\setminus\{0\}} \frac{B([\sigma_\mathcal{T},u_\mathcal{T}],[\tau_\mathcal{T},v_\mathcal{T}])}{\|[\sigma_\mathcal{T},u_\mathcal{T}]\|_X \, \|[\tau_\mathcal{T},v_\mathcal{T}]\|_X} \ge \beta.$$

(3) Problem (IV.3.5) *bzw.* (IV.3.6) *besitzt eine eindeutige Lösung.*

Beweis. *ad (1):* Aus der Definition von $RT(\mathcal{T})$ folgt sofort $\operatorname{div} M_\mathcal{T} \subset Q_\mathcal{T}$. Sei nun $u_\mathcal{T} \in Q_\mathcal{T}$ beliebig und

$$\tau_{u_\mathcal{T}} = e_1 \int_{-R}^{x_1} u_\mathcal{T}(s, x_2) \mathrm{d}s \quad \forall x = (x_1, x_2) \in \Omega$$

wie im Beweis von Lemma IV.3.1. Wir definieren einen Operator $J_\mathcal{T} : H(\operatorname{div}, \Omega) \to M_\mathcal{T}$ durch

$$(J_\mathcal{T}\tau) \cdot \mathbf{n}_K = h_E^{-1} \int_E \tau \cdot \mathbf{n}_K \quad \forall K \in \mathcal{T}, E \in \mathcal{E}, E \subset \partial K. \tag{IV.3.7}$$

Dabei ist h_E die Länge von E. Wegen Lemma IV.3.3 ist diese Definition sinnvoll. Für jedes $K \in \mathcal{T}$ folgt mit dem Gaußschen Integralsatz

$$\begin{aligned} \int_K \operatorname{div}(J_\mathcal{T}\tau_{u_\mathcal{T}}) &= \sum_{E\subset\partial K} \int_E (J_\mathcal{T}\tau_{u_\mathcal{T}}) \cdot \mathbf{n}_K = \sum_{E\subset\partial K} \int_E \tau_{u_\mathcal{T}} \cdot \mathbf{n}_K \\ &= \int_K \operatorname{div} \tau_{u_\mathcal{T}} = \int_K u_\mathcal{T}. \end{aligned}$$

Da $u_\mathcal{T}$ und $\operatorname{div}(J_\mathcal{T}\tau_{u_\mathcal{T}})$ auf K konstant sind, bedeutet dies

$$\operatorname{div}(J_\mathcal{T}\tau_{u_\mathcal{T}}) = u_\mathcal{T}.$$

Mit dem üblichen Skalierungsargument, d.h. Transformation auf das Referenzelement und Äquivalenz von Normen auf endlich dimensionalen Räumen, zeigt man, dass

$$\|J_\mathcal{T}\tau_{u_\mathcal{T}}\| \le c_2 \|\tau_{u_\mathcal{T}}\|_{H(\operatorname{div};\Omega)}$$

ist mit einer Konstanten c_2, die nur von $\mathrm{C}_\mathcal{T}$ abhängt. Also leistet $\tau_{u_\mathcal{T},\mathcal{T}} = J_\mathcal{T}\tau_{u_\mathcal{T}}$ das Gewünschte mit $c_1 = c_2(1 + c_0^2)^{\frac{1}{2}}$ und c_0 wie im Beweis von Lemma IV.3.1.

ad (2): Wegen Teil (1) können wir den Beweis von Lemma IV.3.1 kopieren. Dabei übernimmt $\tau_{u_\mathcal{T},\mathcal{T}}$ aus Teil (1) die Rolle von τ_u aus dem Beweis von Lemma IV.3.1.
ad (3): Wie im Beweis von Satz IV.3.2 folgt aus Teil (2), dass das homogene Problem (IV.3.5), d.h. (IV.3.5) mit $\ell = 0$, nur die triviale Lösung besitzt. Da (IV.3.5) ein lineares Gleichungssystem mit der gleichen Anzahl von Gleichungen und Unbekannten ist, folgt hieraus die Behauptung. □

Satz IV.3.5 (A priori Fehlerabschätzung). *Sei $[\sigma, u] \in X$ die eindeutige Lösung von Problem (IV.3.4) und $[\sigma_\mathcal{T}, u_\mathcal{T}] \in X_\mathcal{T}$ die eindeutige Lösung von (IV.3.5). 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*

$$\begin{aligned}\|\sigma - \sigma_\mathcal{T}\| + \|\operatorname{div}(\sigma - \sigma_\mathcal{T})\| + \|u - u_\mathcal{T}\| \\ \leq c h_\mathcal{T} \left\{ |\sigma|_1 + |\operatorname{div}\sigma|_1 + |u|_1 \right\}.\end{aligned}$$

Die Konstante c hängt nur von Ω und $\mathrm{C}_\mathcal{T} = \max_K \frac{h_K}{\rho_K}$ ab.

Beweis. Sei $[\tau_\mathcal{T}, v_\mathcal{T}] \in X_\mathcal{T}$ beliebig. Mit der Dreiecksungleichung folgt

$$\|[\sigma - \sigma_\mathcal{T}, u - u_\mathcal{T}]\|_X \leq \|[\sigma - \tau_\mathcal{T}, u - v_\mathcal{T}]\|_X + \|[\tau_\mathcal{T} - \sigma_\mathcal{T}, v_\mathcal{T} - u_\mathcal{T}]\|_X .$$

Wegen $X_\mathcal{T} \subset X$ folgt aus (IV.3.4) und (IV.3.5) die Galerkin-Orthogonalität

$$B([\sigma - \sigma_\mathcal{T}, u - u_\mathcal{T}], [\rho_\mathcal{T}, w_\mathcal{T}]) = 0 \quad \forall [\rho_\mathcal{T}, w_\mathcal{T}] \in X_\mathcal{T}.$$

Hieraus und aus Satz IV.3.4 (3) folgt

$$\begin{aligned}&\beta \|[\tau_\mathcal{T} - \sigma_\mathcal{T}, v_\mathcal{T} - u_\mathcal{T}]\|_X \\ &\quad \leq \sup_{[\rho_\mathcal{T}, w_\mathcal{T}] \in X_\mathcal{T} \setminus \{0\}} \frac{B([\tau_\mathcal{T} - \sigma_\mathcal{T}, v_\mathcal{T} - u_\mathcal{T}], [\rho_\mathcal{T}, w_\mathcal{T}])}{\|[\rho_\mathcal{T}, w_\mathcal{T}]\|_X} \\ &\quad = \sup_{[\rho_\mathcal{T}, w_\mathcal{T}] \in X_\mathcal{T} \setminus \{0\}} \frac{B([\tau_\mathcal{T} - \sigma, v_\mathcal{T} - u], [\rho_\mathcal{T}, w_\mathcal{T}])}{\|[\rho_\mathcal{T}, w_\mathcal{T}]\|_X} \\ &\quad \leq \|[\tau_\mathcal{T} - \sigma, v_\mathcal{T} - u]\|_X .\end{aligned}$$

Hierbei haben wir im letzten Schritt die Cauchy-Schwarzsche Ungleichung für Integrale und Summen und die Definition von $\|\cdot\|_X$ ausgenutzt. Da $[\tau_\mathcal{T}, v_\mathcal{T}] \in X_\mathcal{T}$ beliebig war, beweisen diese Abschätzungen das folgende Analogon zum Céa-Lemma, Satz I.1.2 (S. 17),

$$\|[\sigma - \sigma_\mathcal{T}, u - u_\mathcal{T}]\|_X \leq \left(1 + \frac{1}{\beta}\right) \inf_{[\tau_\mathcal{T}, v_\mathcal{T}] \in X_\mathcal{T}} \|[\sigma - \tau_\mathcal{T}, u - v_\mathcal{T}]\|_X .$$

Bezeichne mit $\pi_{\mathcal{T}} : L^2(\Omega) \longrightarrow Q_{\mathcal{T}}$ die L^2-Projektion. Dann folgt aus obiger Abschätzung mit dem Interpolationsoperator $J_{\mathcal{T}}$ aus (IV.3.7)

$$\begin{aligned}
&\left\{\|\sigma-\sigma_{\mathcal{T}}\|^2+\|\operatorname{div}(\sigma-\sigma_{\mathcal{T}})\|^2+\|u-u_{\mathcal{T}}\|^2\right\}^{\frac{1}{2}}\\
&\quad\le\left(1+\frac{1}{\beta}\right)\|[\sigma-J_{\mathcal{T}}\sigma,u-\pi_{\mathcal{T}}u]\|_X\\
&\quad=\left(1+\frac{1}{\beta}\right)\left\{\|\sigma-J_{\mathcal{T}}\sigma\|^2+\|\operatorname{div}(\sigma-J_{\mathcal{T}}\sigma)\|^2+\|u-\pi_{\mathcal{T}}u\|^2\right\}^{\frac{1}{2}}.
\end{aligned}$$

Aus der Poincaréschen Ungleichung, Satz I.2.21 (S. 28), folgt

$$\begin{aligned}
\|u-\pi_{\mathcal{T}}u\|&=\left\{\sum_{K\in\mathcal{T}}\|u-\pi_{\mathcal{T}}u\|_K^2\right\}^{\frac{1}{2}}=\left\{\sum_{K\in\mathcal{T}}\left\|u-\frac{1}{|K|}\int_K u\right\|_K^2\right\}^{\frac{1}{2}}\\
&\le\left\{\sum_{K\in\mathcal{T}}c^2h_K^2\,|u|_{1;K}^2\right\}^{\frac{1}{2}}\\
&\le ch_{\mathcal{T}}\,|u|_1\,.
\end{aligned}$$

Durch Transformation auf das Referenzdreieck folgt wie im Beweis von Satz II.2.7 (S. 51) für jedes $K\in\mathcal{T}$

$$\begin{aligned}
\|\sigma-J_{\mathcal{T}}\sigma\|_K&\le c\inf_{\tau_K\in RT(K)}\|\sigma-\tau_K\|_K\le c\inf_{\rho_K\in\mathbb{R}^2}\|\sigma-\rho_K\|_K\\
&\le c'h_K\,|\sigma|_{1;K}\,.
\end{aligned}$$

Dabei haben wir wieder im letzten Schritt die Poincarésche Ungleichung, Satz I.2.21 (S. 28), ausgenutzt. Quadrieren dieser Abschätzung und Summieren über alle Dreiecke liefert

$$\|\sigma-J_{\mathcal{T}}\sigma\|\le ch_{\mathcal{T}}\,|\sigma|_1\,.$$

Sei nun $v\in L^2(\Omega)$ beliebig. Dann ist

$$\begin{aligned}
\int_\Omega\operatorname{div}(\sigma-J_{\mathcal{T}}\sigma)v&=\int_\Omega\operatorname{div}(\sigma-J_{\mathcal{T}}\sigma)(v-\pi_{\mathcal{T}}v)\\
&\quad+\int_\Omega\operatorname{div}(\sigma-J_{\mathcal{T}}\sigma)\pi_{\mathcal{T}}v.
\end{aligned}$$

Wegen (IV.3.7) erhalten wir für den zweiten Summanden

$$\begin{aligned}
\int_\Omega\operatorname{div}(\sigma-J_{\mathcal{T}}\sigma)\pi_{\mathcal{T}}v&=\sum_{K\in\mathcal{T}}\int_K\operatorname{div}(\sigma-J_{\mathcal{T}}\sigma)\pi_{\mathcal{T}}v\\
&=\sum_{K\in\mathcal{T}}\int_{\partial K}\mathbf{n}_K\cdot(\sigma-J_{\mathcal{T}}\sigma)\pi_{\mathcal{T}}v\\
&=\sum_{K\in\mathcal{T}}\sum_{E\subset\partial K}\left\{\int_E\mathbf{n}_K\cdot\sigma-\int_E\mathbf{n}_K\cdot(J_{\mathcal{T}}\sigma)\right\}\pi_{\mathcal{T}}v\\
&=0.
\end{aligned}$$

Für den ersten Summanden folgt aus der Cauchy-Schwarzschen und der Poincaréschen Ungleichung

$$\int_\Omega \operatorname{div}(\sigma - J_\mathcal{T}\sigma)(v - \pi_\mathcal{T} v) \le \|\operatorname{div}(\sigma - J_\mathcal{T}\sigma)\| \left\{ \sum_{K\in\mathcal{T}} \|v - \pi_\mathcal{T} v\|_K^2 \right\}^{\frac{1}{2}}$$
$$\le c\, \|\operatorname{div}(\sigma - J_\mathcal{T}\sigma)\| \left\{ \sum_{K\in\mathcal{T}} h_K^2\, |v|_{1;K}^2 \right\}^{\frac{1}{2}}.$$

Wählen wir speziell $v = \operatorname{div}(\sigma - J_\mathcal{T}\sigma)$ und beachten, dass $\operatorname{div} J_\mathcal{T}\sigma$ elementweise konstant ist, erhalten wir aus obiger Abschätzung

$$\|\operatorname{div}(\sigma - \mathcal{J}_\mathcal{T}\sigma)\|^2 \le c\, \|\operatorname{div}(\sigma - J_\mathcal{T}\sigma)\| \left\{ \sum_{K\in\mathcal{T}} h_K^2\, |\operatorname{div}\sigma|_{1;K}^2 \right\}^{\frac{1}{2}}$$
$$\le c h_\mathcal{T}\, \|\operatorname{div}(\sigma - J_\mathcal{T}\sigma)\|\, |\operatorname{div}\sigma|_1$$

und damit

$$\|\operatorname{div}(\sigma - J_\mathcal{T}\sigma)\| \le c h_\mathcal{T}\, |\operatorname{div}\sigma|_1 .$$

Hieraus folgt die Fehlerabschätzung des Satzes. □

Bemerkung IV.3.6 (Regularitätsannahmen). Da Ω konvex ist, ist $u \in H^2(\Omega)$ und $\sigma = \nabla u \in H^1(\Omega)$. Wegen $\operatorname{div}\sigma = -f$ sind daher die Regularitätsannahmen von Satz IV.3.5 erfüllt, wenn $f \in H^1(\Omega)$ ist.

Das lineare Gleichungssystem (IV.3.6) hat folgende Struktur

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \sigma_\mathcal{T} \\ u_\mathcal{T} \end{pmatrix} = \begin{pmatrix} 0 \\ -b_\mathcal{T} \end{pmatrix}.$$

Die Koeffizientenmatrix dieses LGS ist symmetrisch, aber indefinit. Dies spiegelt die Sattelpunktsstruktur von (IV.3.3) wider. Die Matrix A ist symmetrisch, positiv definit, und die Matrix B hat wegen Satz IV.3.4 (1) maximalen Rang. Die Größe des LGS (IV.3.6) ist $\#\mathcal{E} + \#\mathcal{T}$. 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 $\mathcal{E}_\Omega$ die Menge aller Kanten im Innern von Ω. Jedem $E \in \mathcal{E}_\Omega$ ordnen wir wieder einen dazu orthogonalen Einheitsvektor $\mathbf{n}_E$ zu und bezeichnen mit $\mathbb{J}_E(\varphi)$ den Sprung von φ über E in Richtung $\mathbf{n}_E$. Setze

$$\Sigma = \bigcup_{E\in\mathcal{E}_\Omega} E$$

und bezeichne mit

$$S^{0,-1}(\Sigma) = \{\lambda : \Sigma \longrightarrow \mathbb{R} : \lambda|_E \in \mathbb{R}\ \forall E \in \mathcal{E}_\Omega\}$$

die stückweise konstanten Funktionen auf Σ.

LEMMA IV.3.7 (Weitere Eigenschaften der Raviart-Thomas Elemente). *(1) Es ist*

$$RT(\mathcal{T}) = \left\{ \sigma \in RT^{-1}(\mathcal{T}) : \sum_{E\in\mathcal{E}_\Omega} \int_E \mathbb{J}_E(\sigma\cdot\mathbf{n}_E)\lambda = 0 \ \forall \lambda \in S^{0,-1}(\Sigma) \right\}.$$

(2) Sei $\varphi \in \mathcal{L}(RT^{-1}(\mathcal{T}),\mathbb{R})$ mit $\varphi(\sigma) = 0$ für alle $\sigma \in RT(\mathcal{T})$. Dann gibt es genau ein $\lambda_\varphi \in S^{0,-1}(\Sigma)$ mit

$$\varphi(\sigma) = \sum_{E\in\mathcal{E}_\Omega} \int_E \mathbb{J}_E(\sigma\cdot\mathbf{n}_E)\lambda_\varphi \quad \forall \sigma \in RT^{-1}(\mathcal{T}).$$

BEWEIS. *ad (1):* Wie im Beweis von Satz I.2.7 (S. 24) folgt, dass $\sigma \in RT^{-1}(\mathcal{T})$ genau dann in $RT(\mathcal{T})$ liegt, wenn $\sigma\cdot\mathbf{n}_K$ stetig ist über alle inneren Kanten, d.h., wenn $\mathbb{J}_E(\sigma\cdot\mathbf{n}_E)$ für alle $E \in \mathcal{E}_\Omega$ verschwindet. Wegen Lemma IV.3.3 (1) ist dies genau dann der Fall, wenn gilt

$$\sum_{E\in\mathcal{E}_\Omega} \int_E \mathbb{J}_E(\sigma\cdot\mathbf{n}_E)\lambda = 0 \quad \forall \lambda \in S^{0,-1}(\Sigma).$$

ad (2): Sei $\varphi \in \mathcal{L}(RT^{-1}(\mathcal{T}),\mathbb{R})$ eine lineare Abbildung, die auf $RT(\mathcal{T})$ verschwindet. Wegen des Rangsatzes gibt es ein $\lambda_\varphi \in S^{0,-1}(\Sigma)$ mit der gewünschten Eigenschaft. Wir müssen also nur noch die Eindeutigkeit von λ_φ zeigen. Sei dazu $\mu \in S^{0,-1}(\Sigma)$ mit

$$\sum_{E\in\mathcal{E}_\Omega} \int_E \mathbb{J}_E(\sigma\cdot\mathbf{n}_E)\mu = 0 \quad \forall \sigma \in RT^{-1}(\mathcal{T}).$$

Dann müssen wir $\mu = 0$ zeigen. Sei dazu $E^* \in \mathcal{E}_\Omega$ beliebig und $K^* \in \mathcal{T}$ ein Dreieck, das E^* als Kante hat. Wegen Lemma IV.3.3 gibt es ein $\sigma^* \in RT^{-1}(\mathcal{T})$ mit

$$\begin{aligned} \sigma^*|_K &= 0 \quad \text{für alle } K \in \mathcal{T}\setminus\{K^*\} \\ \sigma^*\cdot\mathbf{n}_E &= 0 \quad \text{für alle Kanten } E \text{ von } K^* \text{ mit } E \neq E^* \\ \sigma^*\cdot\mathbf{n}_{E^*} &= 1. \end{aligned}$$

Damit folgt

$$0 = \sum_{E\in\mathcal{E}_\Omega} \int_E \mathbb{J}_E(\sigma^*\cdot\mathbf{n}_E)\mu = \int_{E^*} \mathbb{J}_E(\sigma^*\cdot\mathbf{n}_{E^*})\mu = \pm\,|E^*|\,\mu|_{E^*}.$$

Da E^* beliebig war, folgt $\mu = 0$. □

Wir betrachten nun das folgende Problem: Finde $\widetilde{\sigma}_{\mathcal{T}} \in RT^{-1}(\mathcal{T})$, $\widetilde{u}_{\mathcal{T}} \in S^{0,-1}(\mathcal{T})$, $\widetilde{\mu}_{\mathcal{T}} \in S^{0,-1}(\Sigma)$, so dass

$$\begin{aligned}
&\int_\Omega \widetilde{\sigma}_{\mathcal{T}} \cdot \tau_{\mathcal{T}} + \int_\Omega \widetilde{u}_{\mathcal{T}} \operatorname{div} \tau_{\mathcal{T}} \\
&+ \sum_{E\in\mathcal{E}_\Omega} \int_E \mathbb{J}_E(\tau_{\mathcal{T}} \cdot \mathbf{n}_E)\widetilde{\mu}_{\mathcal{T}} = 0 \qquad\qquad \forall \tau_{\mathcal{T}} \in RT^{-1}(\mathcal{T}) \\
&- \int_\Omega v_{\mathcal{T}} \operatorname{div} \widetilde{\sigma}_{\mathcal{T}} = \int_\Omega f v_{\mathcal{T}} \quad \forall v_{\mathcal{T}} \in S^{0,-1}(\mathcal{T}) \\
&\sum_{E\in\mathcal{E}_\Omega} \int_E \mathbb{J}_E(\widetilde{\sigma}_{\mathcal{T}} \cdot \mathbf{n}_E)\lambda_{\mathcal{T}} = 0 \qquad\qquad \forall \lambda_{\mathcal{T}} \in S^{0,-1}(\Sigma).
\end{aligned} \tag{IV.3.8}$$

Satz IV.3.8 (Äquivalenz der diskreten Probleme). *Die diskreten Probleme* (IV.3.6) *und* (IV.3.8) *sind äquivalent.*

Beweis. Sei $\widetilde{\sigma}_{\mathcal{T}}$, $\widetilde{u}_{\mathcal{T}}$, $\widetilde{\mu}_{\mathcal{T}}$ eine beliebige Lösung von (IV.3.8). Aus der dritten Gleichung von (IV.3.8) und Lemma IV.3.7 (1) folgt $\widetilde{\sigma}_{\mathcal{T}} \in RT(\mathcal{T}) = M_{\mathcal{T}}$. Indem wir in der ersten Gleichung von (IV.3.8) nur Vektorfelder $\tau_{\mathcal{T}} \in RT(\mathcal{T})$ als Testfunktionen betrachten, sehen wir, dass $\widetilde{\sigma}_{\mathcal{T}}$, $\widetilde{u}_{\mathcal{T}}$ eine Lösung von (IV.3.6) ist.
Betrachte nun umgekehrt die Lösung $\sigma_{\mathcal{T}}$, $u_{\mathcal{T}}$ von (IV.3.6). Diese erfüllt wegen (IV.3.6) und Lemma IV.3.7 (1) die zweite und dritte Gleichung von (IV.3.8). Wegen (IV.3.6) verschwindet zudem die lineare Abbildung $\tau \mapsto \int_\Omega \sigma_{\mathcal{T}} \cdot \tau + \int_\Omega u_{\mathcal{T}} \operatorname{div} \tau$ auf $RT(\mathcal{T})$. Wegen Lemma IV.3.7 (2) gibt es daher genau ein $\mu_{\mathcal{T}} \in S^{0,-1}(\Sigma)$ mit

$$\int_\Omega \sigma_{\mathcal{T}} \cdot \tau_{\mathcal{T}} + \int_\Omega u_{\mathcal{T}} \operatorname{div} \tau_{\mathcal{T}} + \sum_{E\in\mathcal{E}_\Omega} \int_E \mathbb{J}_E(\tau_{\mathcal{T}} \cdot \mathbf{n}_E)\mu_{\mathcal{T}} = 0 \quad \forall \tau_{\mathcal{T}} \in RT^{-1}(\mathcal{T}).$$

Also ist $\sigma_{\mathcal{T}}$, $u_{\mathcal{T}}$, $\mu_{\mathcal{T}}$ eine Lösung von (IV.3.8). □

Problem (IV.3.8) ist ein LGS der Form

$$\begin{pmatrix} \widetilde{A} & B^T & C^T \\ B & 0 & 0 \\ C & 0 & 0 \end{pmatrix} \begin{pmatrix} \sigma_{\mathcal{T}} \\ u_{\mathcal{T}} \\ \mu_{\mathcal{T}} \end{pmatrix} = \begin{pmatrix} 0 \\ -b_{\mathcal{T}} \\ 0 \end{pmatrix}.$$

Da es die Größe $\#\mathcal{E} + \#\mathcal{T} + \#\mathcal{E}_\Omega$ hat, haben wir auf den ersten Blick im Vergleich zu (IV.3.6) nichts gewonnen. Dieser Eindruck täuscht aber.

Da die Funktionen in $RT^{-1}(\mathcal{T})$ keine globalen Stetigkeitsbedingungen erfüllen müssen, gibt es eine Basis von $RT^{-1}(\mathcal{T})$ aus Funktionen, deren Träger jeweils auf ein einziges Dreieck konzentriert ist. Daher ist $\widetilde{A}$ eine blockdiagonale Matrix; die Zahl der Blöcke ist $\#\mathcal{T}$. Wegen Lemma IV.3.3 (2) ist jeder Block eine 3×3 Matrix. Wir können daher die Unbekannte $\sigma_{\mathcal{T}}$ elementweise durch $u_{\mathcal{T}}$ und $\mu_{\mathcal{T}}$ ausdrücken und erhalten

$$\sigma_{\mathcal{T}} = -\widetilde{A}^{-1}\{B^T u_{\mathcal{T}} + C^T \mu_{\mathcal{T}}\}.$$

Einsetzen in die Gleichung für $u_\mathcal{T}$ liefert

$$-b_\mathcal{T} = B\sigma_\mathcal{T} = -B\widetilde{A}^{-1}B^T u_\mathcal{T} - B\widetilde{A}^{-1}C^T\mu_\mathcal{T}.$$

Da die Funktionen $u_\mathcal{T}$ elementweise konstant sind, ist die Matrix $B\widetilde{A}^{-1}B^T$ diagonal. Wir können daher $u_\mathcal{T}$ durch $\mu_\mathcal{T}$ ausdrücken und erhalten

$$u_\mathcal{T} = \{B\widetilde{A}^{-1}B^T\}^{-1}\{b_\mathcal{T} - B\widetilde{A}^{-1}C^T\mu_\mathcal{T}\}.$$

Setzen wir dies in die Gleichung für $\mu_\mathcal{T}$ ein, erhalten wir

$$\begin{aligned}
0 &= C\sigma_\mathcal{T} \\
&= -C\widetilde{A}^{-1}B^T u_\mathcal{T} - C\widetilde{A}^{-1}C^T\mu_\mathcal{T} \\
&= -C\widetilde{A}^{-1}B^T\left\{B\widetilde{A}^{-1}B^T\right\}^{-1}\left\{b_\mathcal{T} - B\widetilde{A}^{-1}C^T\mu_\mathcal{T}\right\} - C\widetilde{A}^{-1}C^T\mu_\mathcal{T} \\
&= \left[C\widetilde{A}^{-1}B^T\left\{B\widetilde{A}^{-1}B^T\right\}^{-1}B\widetilde{A}^{-1}C^T - C\widetilde{A}^{-1}C^T\right]\mu_\mathcal{T} \\
&\quad - C\widetilde{A}^{-1}B^T\left\{B\widetilde{A}^{-1}B^T\right\}^{-1}b_\mathcal{T}.
\end{aligned}$$

Wir können also die Unbekannten $\sigma_\mathcal{T}$ und $u_\mathcal{T}$ elementweise eliminieren und erhalten ein LGS der Form

$$H\mu_\mathcal{T} = g_\mathcal{T} \tag{IV.3.9}$$

für $\mu_\mathcal{T}$. Es hat nur noch die Größe $\#\mathcal{E}_\Omega$. Die Matrix

$$H = C\widetilde{A}^{-1}B^T\left\{B\widetilde{A}^{-1}B^T\right\}^{-1}B\widetilde{A}^{-1}C^T - C\widetilde{A}^{-1}C^T$$

ist offensichtlich symmetrisch. Da die Matrix des LGS (IV.3.8) wegen Satz IV.3.4 (3) und Satz IV.3.8 regulär ist, ist H auch regulär. Man kann zeigen, dass H sogar positiv definit ist. Daher kann Problem (IV.3.9) z.B. mit einem CG- oder PCG-Verfahren gelöst werden.

Die Freiheitsgrade von $\mu_\mathcal{T}$ 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.3.9) und die Crouzeix-Raviart Diskretisierung eng verwandt. Aufgrund dieser Verwandschaft kann man zudem die Unbekannte $\mu_\mathcal{T}$ benutzen, um eine Approximation der Ordnung $\mathrm{O}(h_\mathcal{T}^2)$ für die Verschiebung zu erhalten.

Satz IV.3.9 (Verbesserte a priori Fehlerabschätzung). *Seien* $[\sigma, u] \in X$ *und* $[\sigma_\mathcal{T}, u_\mathcal{T}, \mu_\mathcal{T}] \in RT^{-1}(\mathcal{T}) \times S^{0,-1}(\mathcal{T}) \times S^{0,-1}(\Sigma)$ *die eindeutigen Lösungen der Probleme* (IV.3.3) *und* (IV.3.8). *Bezeichne mit* $CR(\mathcal{T})$ *den Raum der Crouzeix-Raviart Elemente aus* §*IV.1 und definiere* $\widehat{u}_\mathcal{T} \in CR(\mathcal{T})$ *durch*

$$\widehat{u}_\mathcal{T}(m_E) = -\mu_\mathcal{T}|_E \quad \forall E \in \mathcal{E}_\Omega.$$

Dabei ist m_E *der Mittelpunkt der Kante* E. *Dann gilt*

$$\|u - \widehat{u}_\mathcal{T}\| \leq ch_\mathcal{T}^2\left\{|f|_1 + \|u\|_2\right\}.$$

BEWEIS. Da die Mittelpunktsregel für lineare Funktionen exakt ist, folgt aus der Definition von $\widehat{u}_{\mathcal{T}}$

$$\int_E (\widehat{u}_{\mathcal{T}} + \mu_{\mathcal{T}}) = 0 \quad \forall E \in \mathcal{E}_\Omega.$$

Definiere analog $u^*_{\mathcal{T}} \in CR(\mathcal{T})$ durch

$$\int_E (u^*_{\mathcal{T}} - u) = 0 \quad \forall E \in \mathcal{E}_\Omega.$$

Dann ist

$$\|u - \widehat{u}_{\mathcal{T}}\| \le \|u - u^*_{\mathcal{T}}\| + \|u^*_{\mathcal{T}} - \widehat{u}_{\mathcal{T}}\|$$

und gemäß §IV.1

$$\|u - u^*_{\mathcal{T}}\| \le c h^2_{\mathcal{T}} \|u\|_2 .$$

Daher müssen wir nur noch $\|u^*_{\mathcal{T}} - \widehat{u}_{\mathcal{T}}\|$ passend abschätzen. Dazu benötigen wir die L^2-Projektion $\overline{u}_{\mathcal{T}} = \pi_{\mathcal{T}} u$ von u auf $S^{0,-1}(\mathcal{T})$.
Sei $\tau_{\mathcal{T}} \in RT^{-1}(\mathcal{T})$ beliebig. Da $\operatorname{div} \tau_{\mathcal{T}} \in S^{0,-1}(\mathcal{T})$ ist, folgt

$$\int_\Omega \overline{u}_{\mathcal{T}} \operatorname{div} \tau_{\mathcal{T}} = \int_\Omega u \operatorname{div} \tau_{\mathcal{T}}.$$

Elementweise partielle Integration ergibt

$$\int_\Omega u \operatorname{div} \tau_{\mathcal{T}} = \sum_{K \in \mathcal{T}} \left\{ - \int_K \nabla u \cdot \tau_{\mathcal{T}} + \int_{\partial K} u \mathbf{n}_K \cdot \tau_{\mathcal{T}} \right\}.$$

Wegen $\sigma = \nabla u$ folgt aus diesen beiden Gleichungen

$$\begin{aligned} \int_\Omega \sigma \cdot \tau_{\mathcal{T}} + \int_\Omega \overline{u}_{\mathcal{T}} \operatorname{div} \tau_{\mathcal{T}} &= \sum_{K \in \mathcal{T}} \int_{\partial K} u \mathbf{n}_K \cdot \tau_{\mathcal{T}} \\ &= \sum_{E \in \mathcal{E}_\Omega} \int_E u \mathbb{J}_E(\mathbf{n}_E \cdot \tau_{\mathcal{T}}) \\ &= \sum_{E \in \mathcal{E}_\Omega} \int_E u^*_{\mathcal{T}} \mathbb{J}_E(\mathbf{n}_E \cdot \tau_{\mathcal{T}}). \end{aligned}$$

Da $\mathbb{J}_E(\mathbf{n}_E \cdot \tau_{\mathcal{T}})$ auf den inneren Kanten konstant ist, gilt

$$\int_E \widehat{u}_{\mathcal{T}} \mathbb{J}_E(\mathbf{n}_E \cdot \tau_{\mathcal{T}}) = - \int_E \mu_{\mathcal{T}} \mathbb{J}_E(\mathbf{n}_E \cdot \tau_{\mathcal{T}}) \quad \forall E \in \mathcal{E}_\Omega.$$

Aus diesen beiden Beziehungen und der ersten Gleichung von (IV.3.8) erhalten wir

$$\begin{aligned} &\sum_{E \in \mathcal{E}_\Omega} \int_E (\widehat{u}_{\mathcal{T}} - u^*_{\mathcal{T}}) \mathbb{J}_E(\mathbf{n}_E \cdot \tau_{\mathcal{T}}) \\ &\quad = \sum_{E \in \mathcal{E}_\Omega} \left\{ - \int_E \mu_{\mathcal{T}} \mathbb{J}_E(\mathbf{n}_E \cdot \tau_{\mathcal{T}}) - \int_E u^*_{\mathcal{T}} \mathbb{J}_E(\mathbf{n}_E \cdot \tau_{\mathcal{T}}) \right\} \\ &\quad = \int_\Omega (\sigma_{\mathcal{T}} - \sigma) \cdot \tau_{\mathcal{T}} + \int_\Omega (u_{\mathcal{T}} - \overline{u}_{\mathcal{T}}) \operatorname{div} \tau_{\mathcal{T}}. \end{aligned}$$

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 $\mathrm{C}_{\mathcal{T}}$ abhängt, mit

$$\|\widehat{u}_{\mathcal{T}} - u^*_{\mathcal{T}}\| \le c \sup_{\tau_{\mathcal{T}} \in RT^{-1}(\mathcal{T})} \frac{\left\{ \sum_{E \in \mathcal{E}_\Omega} \int_E (\widehat{u}_{\mathcal{T}} - u^*_{\mathcal{T}}) \mathbb{J}_E (\mathbf{n}_E \cdot \tau_{\mathcal{T}}) \right\}}{\left\{ \sum_{K \in \mathcal{T}} \left[h_K^{-2} \|\tau_{\mathcal{T}}\|_K^2 + \|\operatorname{div} \tau_{\mathcal{T}}\|_K^2 \right] \right\}^{\frac{1}{2}}}.$$

Damit folgt aus obiger Identität

$$\begin{aligned} \|\widehat{u}_{\mathcal{T}} - u^*_{\mathcal{T}}\| &\le c \left\{ \sum_{K \in \mathcal{T}} \left[h_K^2 \|\sigma - \sigma_{\mathcal{T}}\|_K^2 + \|u_{\mathcal{T}} - \overline{u}_{\mathcal{T}}\|_K^2 \right] \right\}^{\frac{1}{2}} \\ &\le c' \left\{ h_{\mathcal{T}} \|\sigma - \sigma_{\mathcal{T}}\| + \|u_{\mathcal{T}} - \overline{u}_{\mathcal{T}}\| \right\}. \end{aligned}$$

Gemäß Satz IV.3.5 und Bemerkung IV.3.6 ist

$$h_{\mathcal{T}} \|\sigma - \sigma_{\mathcal{T}}\| \le c h_{\mathcal{T}}^2 \left\{ |f|_1 + \|u\|_2 \right\}.$$

Wir müssen also noch zeigen, dass gleiches für $\|u_{\mathcal{T}} - \overline{u}_{\mathcal{T}}\|$ gilt. Dazu benutzen wir ein Dualitätsargument und bezeichnen mit z die schwache Lösung von

$$\begin{aligned} \Delta z &= \overline{u}_{\mathcal{T}} - u_{\mathcal{T}} \quad \text{in } \Omega \\ z &= 0 \qquad\qquad\quad \text{auf } \Gamma \end{aligned}$$

und setzen $\varphi = \nabla z$. Mit dem Interpolationsoperator $J_{\mathcal{T}}$ aus (IV.3.7) folgt dann wegen $\operatorname{div} \varphi = \Delta z = \overline{u}_{\mathcal{T}} - u_{\mathcal{T}}$ und $\operatorname{div}(J_{\mathcal{T}}\varphi) \in S^{0,-1}(\mathcal{T})$

$$\begin{aligned} \|\overline{u}_{\mathcal{T}} - u_{\mathcal{T}}\|^2 &= \int_\Omega (\overline{u}_{\mathcal{T}} - u_{\mathcal{T}}) \operatorname{div} \varphi = \int_\Omega (\overline{u}_{\mathcal{T}} - u_{\mathcal{T}}) \operatorname{div}(J_{\mathcal{T}}\varphi) \\ &= \int_\Omega (u - u_{\mathcal{T}}) \operatorname{div}(J_{\mathcal{T}}\varphi) \\ &= \int_\Omega (\sigma_{\mathcal{T}} - \sigma) J_{\mathcal{T}}\varphi. \end{aligned}$$

Weiter ist

$$\int_\Omega (\sigma_{\mathcal{T}} - \sigma) J_{\mathcal{T}}\varphi = \int_\Omega (\sigma_{\mathcal{T}} - \sigma)(J_{\mathcal{T}}\varphi - \varphi) + \int_\Omega (\sigma_{\mathcal{T}} - \sigma)\varphi$$

und

$$\begin{aligned} \int_\Omega (\sigma_{\mathcal{T}} - \sigma)\varphi &= \int_\Omega (\sigma_{\mathcal{T}} - \sigma) \cdot \nabla z = -\int_\Omega \operatorname{div}(\sigma_{\mathcal{T}} - \sigma) z \\ &= -\int_\Omega \operatorname{div}(\sigma_{\mathcal{T}} - \sigma)(z - \pi_{\mathcal{T}} z). \end{aligned}$$

Hieraus folgt mit der Cauchy-Schwarzschen Ungleichung

$$\begin{aligned}\int_\Omega &(\sigma_\mathcal{T} - \sigma) J_\mathcal{T}\varphi \\ &\leq \|\sigma - \sigma_\mathcal{T}\| \, \|\varphi - J_\mathcal{T}\varphi\| + \|\operatorname{div}(\sigma - \sigma_\mathcal{T})\| \, \|z - \pi_\mathcal{T} z\| \\ &\leq c h_\mathcal{T} \left\{ \|\sigma - \sigma_\mathcal{T}\| \, |\varphi|_1 + \|\operatorname{div}(\sigma - \sigma_\mathcal{T})\| \, |z|_1 \right\}.\end{aligned}$$

Da Ω konvex ist, folgt aus dem Regularitätssatz, Satz I.3.6 (S. 33)

$$|z|_1 + |\varphi|_1 \leq c \, \|z\|_2 \leq c' \, \|\overline{u}_\mathcal{T} - u_\mathcal{T}\| .$$

Aus diesen beiden Abschätzungen und Satz IV.3.5 folgt schließlich

$$\begin{aligned}\|\overline{u}_\mathcal{T} - u_\mathcal{T}\| &\leq c h_\mathcal{T} \left\{ \|\sigma - \sigma_\mathcal{T}\| + \|\operatorname{div}(\sigma - \sigma_\mathcal{T})\| \right\} \\ &\leq c' h_\mathcal{T}^2 \left\{ |f|_1 + \|u\|_2 \right\}.\end{aligned}$$

□

IV.4. Finite Volumen Methoden

In diesem Abschnitt betrachten wir sogenannte *Systeme in Divergenzform*. Gegeben sind dabei ein beschränktes Gebiet $\Omega \subset \mathbb{R}^n$ mit üblicherweise $n = 2$ oder $n = 3$, eine Zahl $m \geq 1$ und Funktionen $\mathbf{g} : \mathbb{R}^m \times \Omega \times (0, \infty) \to \mathbb{R}^m$, $\mathbf{M} : \mathbb{R}^m \to \mathbb{R}^m$, $\underline{\mathbf{F}} : \mathbb{R}^m \to \mathbb{R}^{m \times n}$ und $\mathbf{U}_0 : \Omega \to \mathbb{R}^m$, gesucht wird eine Funktion $\mathbf{U} : \Omega \times (0, \infty) \to \mathbb{R}^m$ mit

$$\begin{aligned}\frac{\partial \mathbf{M}(\mathbf{U})}{\partial t} + \operatorname{div} \underline{\mathbf{F}}(\mathbf{U}) &= \mathbf{g}(\mathbf{U}, x, t) \quad \text{in } \Omega \times (0, \infty) \\ \mathbf{U}(\cdot, 0) &= \mathbf{U}_0 \qquad\qquad \text{in } \Omega.\end{aligned} \tag{IV.4.1}$$

Für ein wohlgestelltes Problem müssen zu dieser Gleichung noch geeignete Randbedingungen gestellt werden. Diese vernachlässigen wir aber im Folgenden durchweg, um die Darstellung möglichst einfach zu gestalten.

Man beachte, dass die Divergenz in Gleichung (IV.4.1) zeilenweise zu verstehen ist, d.h.

$$\operatorname{div} \underline{\mathbf{F}}(\mathbf{U}) = \left(\sum_{j=1}^{n} \frac{\partial \underline{\mathbf{F}}(\mathbf{U})_{i,j}}{\partial x_j} \right)_{1 \leq i \leq m}.$$

Die Funktion $\underline{\mathbf{F}}$ heißt der *Fluss* des Systems. Er wird im allgemeinen additiv in einen sogenannten *advektiven Fluss* $\underline{\mathbf{F}}_{\text{adv}}$, der keine Ableitungen enthält, und einen sogenannten *viskosen Fluss* $\underline{\mathbf{F}}_{\text{visc}}$, der Ortsableitungen enthält, zerlegt d.h. $\underline{\mathbf{F}} = \underline{\mathbf{F}}_{\text{adv}} + \underline{\mathbf{F}}_{\text{visc}}$.

Beispiel IV.4.1 (Zeitabhängige Diffusions-Konvektions-Reaktions Gleichung). Die zeitabhängige Diffusions-Konvektions-Reaktions Gleichung $\frac{\partial u}{\partial t} - \operatorname{div}(A \nabla u) + \mathbf{a} \cdot \nabla u + \alpha u = f$ ist wegen $\operatorname{div}(\mathbf{a} u) = u \operatorname{div} \mathbf{a} + \mathbf{a} \cdot \nabla u$ ein System in Divergenzform mit $m = 1$, $\mathbf{U} = u$, $M(\mathbf{U}) = u$, $\mathbf{g}(\mathbf{U}, x, t) = f - (\alpha - \operatorname{div} \mathbf{a}) u$, $\underline{\mathbf{F}}_{\text{visc}}(\mathbf{U}) = -A \nabla u$ und $\underline{\mathbf{F}}_{\text{adv}}(\mathbf{U}) = \mathbf{a} u$.

BEISPIEL IV.4.2 (Burgers Gleichung). Die Burgers Gleichung $\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0$ ist wegen $u\frac{\partial u}{\partial x} = \frac{\partial}{\partial x}\left(\frac{1}{2}u^2\right)$ ein System in Divergenzform mit $n = m = 1$, $\mathbf{U} = u$, $M(\mathbf{U}) = u$, $\mathbf{g}(\mathbf{U}, x, t) = 0$, $\underline{\mathbf{F}}_{\text{visc}}(\mathbf{U}) = 0$ und $\underline{\mathbf{F}}_{\text{adv}}(\mathbf{U}) = \frac{1}{2}u^2$.

BEISPIEL IV.4.3 (Euler und Navier-Stokes Gleichungen). Die reibungsfreie bzw. reibungsbehaftete Strömung einer kompressiblen Flüssigkeit mit Dichte ρ, Geschwindigkeit $\mathbf{v}$ und interner Energie e, die häufig mit der Temperatur identifiziert wird, wird durch die Euler bzw. Navier-Stokes Gleichungen beschrieben. Beide sind Systeme in Divergenzform mit $m = n + 2$, $\mathbf{U} = \begin{pmatrix} \rho \\ \mathbf{v} \\ e \end{pmatrix}$, $M(\mathbf{U}) = \begin{pmatrix} \rho \\ \rho\mathbf{v} \\ e \end{pmatrix}$, $\mathbf{g} = \begin{pmatrix} 0 \\ \rho\mathbf{f} \\ \mathbf{f}\cdot\mathbf{v} \end{pmatrix}$ und $\underline{\mathbf{F}}_{\text{adv}}(\mathbf{U}) = \begin{pmatrix} \rho\mathbf{v} \\ \rho\mathbf{v}\otimes\mathbf{v} + p\underline{\mathbf{I}} \\ e\mathbf{v} + p\mathbf{v} \end{pmatrix}$. Für die Euler Gleichungen ist $\underline{\mathbf{F}}_{\text{visc}}(\mathbf{U}) = 0$ und für die Navier-Stokes Gleichungen $\underline{\mathbf{F}}_{\text{visc}}(\mathbf{U}) = \begin{pmatrix} 0 \\ \underline{\mathbf{T}} + p\underline{\mathbf{I}} \\ (\underline{\mathbf{T}} + p\underline{\mathbf{I}})\cdot\mathbf{v} + \sigma \end{pmatrix}$ mit $\underline{\mathbf{T}} = \frac{1}{2}\left(\nabla u + \nabla u^T\right)$, $p = p(\rho, e)$ und $\sigma = \alpha\nabla e$.

Zur Beschreibung der grundlegenden Ideen vom Finite Volumen Verfahren wählen wir eine Zeitschrittweite $\tau > 0$ und eine feste Unterteilung $\mathcal{T}$ von Ω. Man beachte, dass $\mathcal{T}$ aus beliebigen Polyedern bestehen darf, die aber nicht einander überlappen dürfen.

In einem ersten Schritt wählen wir eine Zahl i und ein Element $K \in \mathcal{T}$. Dann integrieren wir das System (IV.4.1) über den Raum-Zeit-Zylinder $K \times [(i-1)\tau, i\tau]$. Dies liefert

$$\int_{(i-1)\tau}^{i\tau}\int_K \frac{\partial \mathbf{M}(\mathbf{U})}{\partial t} dx dt + \int_{(i-1)\tau}^{i\tau}\int_K \operatorname{div}\underline{\mathbf{F}}(\mathbf{U}) \mathrm{d}x\mathrm{d}t = \int_{(i-1)\tau}^{i\tau}\int_K \mathbf{g}(\mathbf{U}, x, t)\mathrm{d}x\mathrm{d}t.$$

Partielle Integration liefert mit der äußeren Normalen $\mathbf{n}_K$ von K für die linke Seite

$$\int_{(i-1)\tau}^{i\tau}\int_K \frac{\partial \mathbf{M}(\mathbf{U})}{\partial t}\mathrm{d}x\mathrm{d}t = \int_K \mathbf{M}(\mathbf{U}(x, i\tau))\mathrm{d}x - \int_K \mathbf{M}(\mathbf{U}(x, (i-1)\tau))\mathrm{d}x$$

$$\int_{(i-1)\tau}^{i\tau}\int_K \operatorname{div}\underline{\mathbf{F}}(\mathbf{U})\mathrm{d}x\mathrm{d}t = \int_{(i-1)\tau}^{i\tau}\int_{\partial K} \underline{\mathbf{F}}(\mathbf{U})\cdot\mathbf{n}_K \mathrm{d}S\mathrm{d}t.$$

Im Folgenden nehmen wir an, dass $\mathbf{U}$ bezüglich Ort und Zeit stückweise konstant ist und bezeichnen mit $\mathbf{U}_K^i$ und $\mathbf{U}_K^{i-1}$ die Werte von $\mathbf{U}$ auf K zu den Zeiten $i\tau$ und $(i-1)\tau$. Dann ist

$$\int_K \mathbf{M}(\mathbf{U}(x, i\tau))\mathrm{d}x \approx |K|\,\mathbf{M}(\mathbf{U}_K^i)$$

$$\int_K \mathbf{M}(\mathbf{U}(x, (i-1)\tau))\mathrm{d}x \approx |K|\,\mathbf{M}(\mathbf{U}_K^{i-1}),$$

wobei $|K|$ das n-dimensionale Lebesgue-Maß von K bezeichnet.
Als nächstes approximieren wir den Fluss-Term durch

$$\int_{(i-1)\tau}^{i\tau}\int_{\partial K}\underline{\mathbf{F}}(\mathbf{U})\cdot\mathbf{n}_K\mathrm{d}S\mathrm{d}t \approx \tau\int_{\partial K}\underline{\mathbf{F}}(\mathbf{U}_K^{i-1})\cdot\mathbf{n}_K\mathrm{d}S.$$

Die rechte Seite dagegen approximieren wir durch

$$\int_{(i-1)\tau}^{i\tau}\int_{K}\mathbf{g}(\mathbf{U},x,t)\mathrm{d}x\mathrm{d}t \approx \tau\,|K|\,\mathbf{g}(\mathbf{U}_K^{i-1},x_K,(i-1)\tau),$$

wobei x_K ein fester Punkt in K ist, z.B. der Schwerpunkt.
Im letzten Schritt ersetzen wir das Randintegral des Flusses durch einen *numerischen Fluss*

$$\tau\int_{\partial K}\underline{\mathbf{F}}(\mathbf{U}_K^{i-1})\cdot\mathbf{n}_K\mathrm{d}S \approx \tau\sum_{\substack{K'\in\mathcal{T}\\ \partial K\cap\partial K'\in\mathcal{E}}}|\partial K\cap\partial K'|\,\mathbf{F}_{\mathcal{T}}(\mathbf{U}_K^{i-1},\mathbf{U}_{K'}^{i-1}).$$

Dabei bezeichnet wie üblich $\mathcal{E}$ die Menge aller $n-1$-dimensionalen Elementflächen.

Damit lautet die einfachste Finite Volumen Diskretisierung von Gleichung (IV.4.1).

Berechne für jedes Element $K\in\mathcal{T}$

$$\mathbf{U}_K^0 = \frac{1}{|K|}\int_K \mathbf{U}_0(x)\mathrm{d}x.$$

Und für $i=1,2,\ldots$ berechne sukzessive für jedes Element $k\in\mathcal{T}$

$$\begin{aligned}\mathbf{M}(\mathbf{U}_K^i) = \mathbf{M}(\mathbf{U}_K^{i-1})&\\ -\tau\sum_{\substack{K'\in\mathcal{T}\\ \partial K\cap\partial K'\in\mathcal{E}_{\mathcal{T}}}}&\frac{|\partial K\cap\partial K'|}{|K|}\mathbf{F}_{\mathcal{T}}(\mathbf{U}_K^{i-1},\mathbf{U}_{K'}^{i-1})\\ +\tau\mathbf{g}(\mathbf{U}_K^{i-1},x_K,(i-1)\tau)&.\end{aligned}$$

Bemerkung IV.4.4. Wegen der Flussterme hängt $\mathbf{U}_K^i$ von $\mathbf{U}_K^{i-1}$ und allen $\mathbf{U}_{K'}^{i-1}$ der an K grenzenden Elemente K' ab. Der Zeitschritt $i-1\mapsto i$ ist explizit. Daher ist eine CFL-artige Koppelung der Zeitschrittweite τ an eine geeignete Gitterweite von $\mathcal{T}$ zu erwarten. In der Praxis arbeitet man mit variablen Zeitschrittweiten und Unterteilungen. Dazu wählt man eine streng monoton wachsende Folge $0=t_0<t_1<t_2<\ldots$ von Zeiten und ordnet jeder Zeit t_i eine Zerlegung $\mathcal{T}_i$ von Ω zu. Dann muss man natürlich τ durch $\tau_i=t_i-t_{i-1}$ ersetzen, und K und K' sind Elemente in $\mathcal{T}_{i-1}$ oder $\mathcal{T}_i$. Zusätzlich benötigt man einen Interpolationsoperator, der stückweise konstante Funktionen bezüglich einer Unterteilung in stückweise konstante Funktionen bezüglich einer anderen Unterteilung abbildet.

Für eine konkrete Realisierung des Finite Volumen Verfahrens müssen wir im Folgenden noch die Zerlegung $\mathcal{T}$ und den numerischen Fluss $\mathbf{F}_{\mathcal{T}}$ bestimmen.

Zuerst beschreiben wir die Konstruktion der Unterteilung $\mathcal{T}$. Selbstverständlich kann man hierzu wie bei Finite Element Unterteilungen vorgehen. In der Praxis aber bevorzugt man sogenannte *duale Gitter*. Zur Beschreibung der grundlegenden Idee betrachten wir den Fall $n = 2$. Wir beginnen mit einer sogenannten primalen Finite Element Unterteilung $\widetilde{\mathcal{T}}$, die die Voraussetzungen von §II.1 (S. 35) erfüllt. Dann unterteilen wir jedes Element $\widetilde{K} \in \widetilde{\mathcal{T}}$ in kleinere Elemente, indem wir entweder

- die Mittelsenkrechten von $\widetilde{K}$ ziehen (siehe Abbildung IV.4.1) oder
- die Kantenmittelpunkte von $\widetilde{K}$ mit seinem Schwerpunkt verbinden (siehe Abbildung IV.4.2).

Die Element von $\mathcal{T}$ bestehen dann aus der Vereinigung der kleinen Elemente, die einen Eckpunkt von $\widetilde{\mathcal{T}}$ gemeinsam haben.

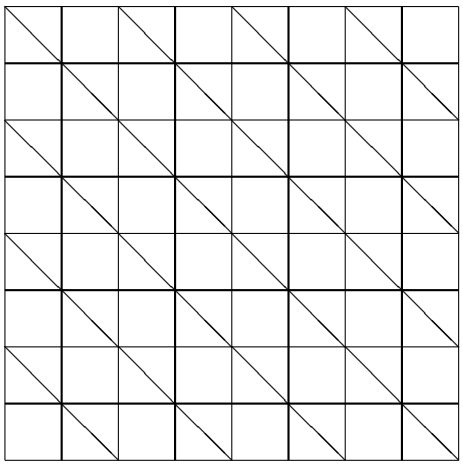

ABBILDUNG IV.4.1. Duales Gitter (fette Linien) mittels Mittelsenkrechten eines primalen Gitters (dünne Linien)

Bei beiden Konstruktionen kann man jedem Element in $\mathcal{T}$ eindeutig einen Eckpunkt eines Elementes in $\widetilde{\mathcal{T}}$ zuordnen und zu jeder Kante eines Elementes in $\mathcal{T}$ gibt es genau zwei Eckpunkte von Elementen in $\widetilde{\mathcal{T}}$, so dass deren Verbindungslinie die gegebene Kante schneidet (siehe Abbildung IV.4.3). Die erste Konstruktion hat den Vorteil, dass dieser Schnitt orthogonal ist. Allerdings hat sie einige gravierende Nachteile, die die zweite Konstruktion nicht hat:

- Der Schnittpunkt der Mittelsenkrechten eines Dreieckes kann außerhalb des Dreieckes liegen. Er liegt genau dann innerhalb des Dreieckes oder auf seinem Rand, wenn der größte Winkel höchstens ein rechter ist.

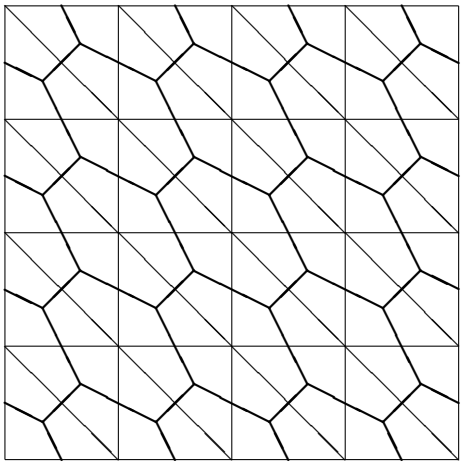

ABBILDUNG IV.4.2. Duales Gitter (fette Linien) mittels der Schwerpunkte eines primalen Gitters (dünne Linien)

- Die Mittelsenkrechten eines Vierecks schneiden sich genau dann, wenn das Viereck ein Rechteck ist.
- Es gibt kein dreidimensionales Gegenstück der ersten Konstruktion.

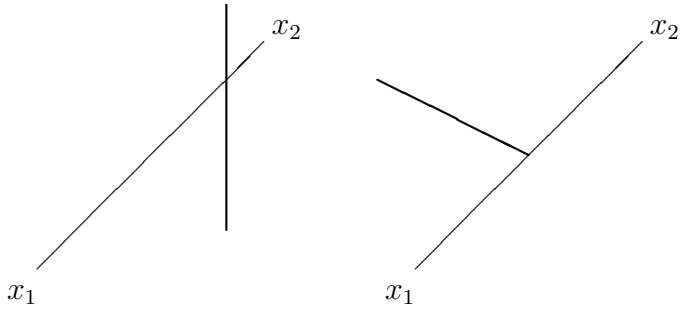

ABBILDUNG IV.4.3. Beispiele für gemeinsame Kanten E (fette Linien) zweier Elemente eines dualen Gitters und zugehörige Elementeckpunkte x_1 und x_2 des entsprechenden primalen Gitters mit ihrer Verbindungslinie (dünne Linien)

Als nächstes beschreiben wir die Konstruktion des numerischen Flusses $\mathbf{F}_{\mathcal{T}}(\mathbf{U}_K^{i-1}, \mathbf{U}_{K'}^{i-1})$. Dazu betrachten wir zwei aneinandergrenzende Elemente K und K' und spalten $\partial K \cap \partial K'$ in gerade Kanten, falls $n = 2$ ist, oder Flächen, falls $n = 3$ ist, auf. Betrachte eine derartige Kante oder Fläche E. Bezeichne die angrenzenden Elemente mit K_1 und K_2 und schreibe $\mathbf{U}_1$ und $\mathbf{U}_2$ statt $\mathbf{U}_{K_1}^{i-1}$ und $\mathbf{U}_{K_2}^{i-1}$. Wir nehmen zudem an, dass $\mathcal{T}$ das duale Gitter zu einer primalen Finite Element Unterteilung $\widetilde{\mathcal{T}}$ ist und bezeichnen mit x_1 und x_2 die Elementeckpunkte in $\widetilde{\mathcal{T}}$, deren Verbindungslinie E schneidet (siehe Abbildung IV.4.3).

Für die Konstruktion des zu E gehörenden Flusses $\mathbf{F}_{\mathcal{T}}(\mathbf{U}_1, \mathbf{U}_2)$ spalten wir diesen wie den analytischen Fluss $\underline{\mathbf{F}}$ additiv in einen advektiven Anteil $\mathbf{F}_{\mathcal{T},\text{adv}}(\mathbf{U}_1, \mathbf{U}_2)$ und einen viskosen Anteil $\mathbf{F}_{\mathcal{T},\text{visc}}(\mathbf{U}_1, \mathbf{U}_2)$ auf.

Für die Konstruktion des viskosen Flusses führen wir ein lokales Koordinatensystem $\eta_1, \ldots, \eta_n$ derart ein, dass die Richtung η_1 parallel ist zu der Richtung $\overline{x_1 x_2}$ und die anderen Richtungen tangential sind zu E. Im allgemeinen sind diese Koordinatenrichtungen nicht paarweise orthogonal. Wir drücken nun alle Ableitungen in $\underline{\mathbf{F}}_{\text{visc}}$in dem neuen Koordinatensystem aus und vernachlässigen alle Ableitungen, die nicht η_1 enthalten. Die Ableitungen bezüglich η_1 dagegen approximieren wir durch Differenzenquotienten der Form $\frac{\varphi_1 - \varphi_2}{|x_1 - x_2|}$.

Beispiel IV.4.5 (Zeitabhängige Diffusions-Konvektions-Reaktions Gleichung). Mit $\nabla_x u = B \nabla_\eta u$, dem äußeren Einheitsnormalenvektor $\mathbf{n}_{K_1}$ zu K_1 und dem ersten Einheitsvektor $\mathbf{e}_1$ ergibt sich für die zeitabhängige Diffusions-Konvektions-Reaktions Gleichung aus Beispiel IV.4.1 $\mathbf{F}_{\mathcal{T},\text{visc}}(\mathbf{U}_1, \mathbf{U}_2) = \mathbf{n}_{K_1} \cdot AB \cdot \mathbf{e}_1 \frac{u_1 - u_2}{|x_1 - x_2|}$.

Für den advektiven Teil des numerischen Flusses bezeichnen wir für einen beliebigen Vektor $\mathbf{V} \in \mathbb{R}^m$ mit $D(\underline{\mathbf{F}}_{\text{adv}}(\mathbf{V}) \cdot \mathbf{n}_{K_1}) \in \mathbb{R}^{m \times m}$ die Ableitung von $\underline{\mathbf{F}}_{\text{adv}}(\mathbf{V}) \cdot \mathbf{n}_{K_1}$ nach $\mathbf{V}$. Für die zeitabhängige Diffusions-Konvektions-Reaktions Gleichung aus Beispiel IV.4.1 und die Burgers Gleichung aus Beispiel IV.4.2 ist dies wegen $m = 1$ eine Zahl. Man kann zeigen, dass für viele Systeme mit $m > 1$, insbesondere die Euler- und Navier-Stokes Gleichungen aus Beispiel IV.4.3, die Matrix $D(\underline{\mathbf{F}}_{\text{adv}}(\mathbf{V}) \cdot \mathbf{n}_{K_1})$ diagonalisierbar ist, d.h. es gibt eine invertierbare Matrix $Q(\mathbf{V}) \in \mathbb{R}^{m \times m}$ und eine Diagonalmatrix $\Delta(\mathbf{V}) \in \mathbb{R}^{m \times m}$ mit

$$Q(\mathbf{V})^{-1} D(\underline{\mathbf{F}}_{\text{adv}}(\mathbf{V}) \cdot \mathbf{n}_{K_1}) Q(\mathbf{V}) = \Delta(\mathbf{V}).$$

Die Diagonalelemente von $\Delta(\mathbf{V})$ sind naürlich die Eigenwerte von $D(\underline{\mathbf{F}}_{\text{adv}}(\mathbf{V}) \cdot \mathbf{n}_{K_1})$. Für reelle Zahlen z definieren wir wie üblich

$$z^+ = \max\{z, 0\} \ , \ z^- = \min\{z, 0\}$$

und setzen

$$\Delta(\mathbf{V})^{\pm} = \begin{pmatrix} \Delta(\mathbf{V})^{\pm}_{11} & 0 & \ldots & 0 \\ 0 & \Delta(\mathbf{V})^{\pm}_{22} & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & \Delta(\mathbf{V})^{\pm}_{mm} \end{pmatrix}$$

sowie

$$C(\mathbf{V})^{\pm} = Q(\mathbf{V}) \Delta(\mathbf{V})^{\pm} Q(\mathbf{V})^{-1}.$$

Mit diesen Notationen lautet die *Steger-Warming Approximation* des advektiven Flusses

$$\mathbf{F}_{\mathcal{T},\text{adv}}(\mathbf{U}_1, \mathbf{U}_2) = C(\mathbf{U}_1)^+ \mathbf{U}_1 + C(\mathbf{U}_2)^- \mathbf{U}_2.$$

Ein anderer weit verbreiteter advektiver numerischer Fluss ist die *van Leer Approximation*

$$\begin{aligned}\mathbf{F}_{\mathcal{T},\mathrm{adv}}&(\mathbf{U}_1,\mathbf{U}_2)\\ &= \left[\frac{1}{2}C(\mathbf{U}_1) + C\left(\frac{1}{2}(\mathbf{U}_1+\mathbf{U}_2)\right)^+ - C\left(\frac{1}{2}(\mathbf{U}_1+\mathbf{U}_2)\right)^-\right]\mathbf{U}_1\\ &\quad + \left[\frac{1}{2}C(\mathbf{U}_2) - C\left(\frac{1}{2}(\mathbf{U}_1+\mathbf{U}_2)\right)^+ + C\left(\frac{1}{2}(\mathbf{U}_1+\mathbf{U}_2)\right)^-\right]\mathbf{U}_2.\end{aligned}$$

Bei beiden Approximationen muss man für jede gemeinsame Kante oder Seitenfläche von zwei benachbarten Elementen die Ableitung $D\underline{\mathbf{F}}_{\mathrm{adv}}(\mathbf{V})\cdot\mathbf{n}_{K_1}$ und ihre Eigenwerte und Eigenvektoren bestimmen.

Beispiel IV.4.6 (Zeitabhängige Diffusions-Konvektions-Reaktions Gleichung). Für die zeitabhängige Diffusions-Konvektions-Reaktions Gleichung aus Beispiel IV.4.1 ist $D(\underline{\mathbf{F}}_{\mathrm{adv}}(\mathbf{V})\cdot\mathbf{n}_{K_1}) = \mathbf{n}_{K_1}\cdot\mathbf{a}$. Damit ergibt sich für die Steger-Warming Approximation

$$\mathbf{F}_{\mathcal{T},\mathrm{adv}}(\mathbf{U}_1,\mathbf{U}_2) = \begin{cases}\mathbf{n}_{K_1}\cdot\mathbf{a}u_1 & \text{falls } \mathbf{n}_{K_1}\cdot\mathbf{a}\geq 0\\ \mathbf{n}_{K_1}\cdot\mathbf{a}u_2 & \text{falls } \mathbf{n}_{K_1}\cdot\mathbf{a}\leq 0\end{cases}$$

und für die van Leer Approximation

$$\begin{aligned}\mathbf{F}_{\mathcal{T},\mathrm{adv}}(\mathbf{U}_1,\mathbf{U}_2) &= \begin{cases}\frac{3}{2}\mathbf{n}_{K_1}\cdot\mathbf{a}u_1 - \frac{1}{2}\mathbf{n}_{K_1}\cdot\mathbf{a}u_2 & \text{falls } \mathbf{n}_{K_1}\cdot\mathbf{a}\geq 0\\ \frac{3}{2}\mathbf{n}_{K_1}\cdot\mathbf{a}u_2 - \frac{1}{2}\mathbf{n}_{K_1}\cdot\mathbf{a}u_1 & \text{falls } \mathbf{n}_{K_1}\cdot\mathbf{a}\leq 0\end{cases}\\ &= \mathbf{F}^{\mathrm{SW}}_{\mathcal{T},\mathrm{adv}}(\mathbf{U}_1,\mathbf{U}_2) + \frac{1}{2}\left|\mathbf{n}_{K_1}\cdot\mathbf{a}\right|(u_1-u_2),\end{aligned}$$

wenn $\mathbf{F}^{\mathrm{SW}}_{\mathcal{T},\mathrm{adv}}(\mathbf{U}_1,\mathbf{U}_2)$ die Steger-Warming Approximation bezeichnet.

Beispiel IV.4.7 (Burgers Gleichung). Für die Burgers Gleichung aus Beispiel IV.4.2 ist $D(\underline{\mathbf{F}}_{\mathrm{adv}}(\mathbf{V})\cdot\mathbf{n}_{K_1}) = \mathbf{V}$. Damit ergibt sich für die Steger-Warming Approximation

$$\underline{\mathbf{F}}_{\mathcal{T},\mathrm{adv}}(\mathbf{U}_1,\mathbf{U}_2) = \begin{cases}u_1^2 & \text{falls } u_1\geq 0, u_2\geq 0\\ u_1^2+u_2^2 & \text{falls } u_1\geq 0, u_2\leq 0\\ u_2^2 & \text{falls } u_1\leq 0, u_2\leq 0\\ 0 & \text{falls } u_1\leq 0, u_2\geq 0\end{cases}$$

und für die van Leer Approximation

$$\underline{\mathbf{F}}_{\mathcal{T},\mathrm{adv}}(\mathbf{U}_1,\mathbf{U}_2) = \begin{cases}u_1^2 & \text{falls } u_1\geq -u_2\\ u_2^2 & \text{falls } u_1\leq -u_2.\end{cases}$$

Die Tatsache, dass die Elemente eines dualen Gitters mit den Elementeckpunkten eines primalen Finite Element Gitters assoziiert werden können, gibt eine einfache und sehr nützliche Brücke zwischen Finite Volumen und Finite Element Verfahren.

Betrachte dazu eine stückweise konstante Funktion $\varphi \in S^{0,-1}(\mathcal{T})$ auf einem dualen Gitter $\mathcal{T}$ zu einem primalen Gitter $\widetilde{\mathcal{T}}$. Der Funktion φ können wir dann ein-eindeutig die stetige stückweise lineare Funktion $\Phi \in S^{1,0}(\widetilde{\mathcal{T}})$ zuordnen, die im Elementeckpunkt x_K von $\widetilde{\mathcal{T}}$, der in $K \in \mathcal{T}$ liegt, den Wert φ_K annimmt, d.h. $\Phi(x_K) = \varphi_K$.
Diese Beziehung erleichtert manchmal die Analyse von Finite Volumen Methoden ganz erheblich. So kann man mit ihrer Hilfe ganz einfach a posteriori Fehlerschätzer und adaptive Gitterverfeinerungen für Finite Volumen Methoden wie folgt erzeugen:

- Gegeben sei die Lösung φ der Finite Volumen Diskretisierung. Berechne die Finite Element Funktion Φ.
- Wende einen üblichen a posteriori Fehlerschätzer auf Φ und die Differentialgleichung an.
- Basierend auf diesem Fehlerschätzer führe eine übliche adaptive Verfeinerung von $\widetilde{\mathcal{T}}$ aus und bestimme so eine verfeinerte primale Finite Element Unterteilung $\widehat{\mathcal{T}}$.
- Bestimme das zu $\widehat{\mathcal{T}}$ gehörige duale Gitter $\mathcal{T}'$. Dies ist die adaptive Verfeinerung von $\mathcal{T}$.

Eine weitere Brücke zwischen Finite Element und Finite Volumen Verfahren bilden die *discontinuous Galerkin Methoden*. Im einfachsten Fall geht man ähnlich wie in §IV.2 vor:

- Approximiere $\mathbf{U}$ durch unstetige Funktionen, die bezüglich Ort und Zeit Polynome auf Orts-Zeit-Zylindern $K \times [(n-1)\tau, n\tau]$ mit $K \in \mathcal{T}$ sind.
- Für jeden Orts-Zeit-Zylinder multipliziere die Differentialgleichung mit einem Testpolynom und integriere das Ergebnis über den Zylinder.
- Führe partielle Integration für die Flussterme aus.
- Akkumuliere die Beiträge aller Elemente in $\mathcal{T}$.
- Kompensiere die unzulässige partielle Integration durch geeignete Sprungterme über die Elementgrenzen.
- Stabilisiere die Diskretisierung durch geeignete Elementresiduen wie bei den Petrov-Galerkin Methoden.

In der einfachsten Form führt dies auf das folgende diskrete Problem:

Berechne die L^2-Projektion $\mathbf{U}_{\mathcal{T}}^0$ von $\mathbf{U}_0$ auf $S^{k,-1}(\mathcal{T})$.
Berechne für $n = 1, 2, \ldots$ sukzessive $\mathbf{U}_{\mathcal{T}}^n \in S^{k,-1}(\mathcal{T})$ so,

dass für alle $\mathbf{V}_\mathcal{T} \in S^{k,-1}(\mathcal{T})$ gilt

$$\begin{aligned}
&\sum_{K\in\mathcal{T}} \frac{1}{\tau}\int_K M(\mathbf{U}_\mathcal{T}^n)\cdot\mathbf{V}_\mathcal{T} - \sum_{K\in\mathcal{T}}\int_K \underline{\mathbf{F}}(\mathbf{U}_\mathcal{T}^n) : \nabla\mathbf{V}_\mathcal{T} \\
&+ \sum_{E\in\mathcal{E}} \delta_E \int_E \mathbb{J}_E(\mathbf{n}_E\cdot\underline{\mathbf{F}}(\mathbf{U}_\mathcal{T}^n)\mathbf{V}_\mathcal{T}) \\
&+ \sum_{K\in\mathcal{T}} \delta_K \int_K \operatorname{div}\underline{\mathbf{F}}(\mathbf{U}_\mathcal{T}^n)\cdot\operatorname{div}\underline{\mathbf{F}}(\mathbf{V}_\mathcal{T}) \\
&+ \sum_{E\in\mathcal{E}} \vartheta_E \int_E \mathbb{J}_E(\mathbf{U}_\mathcal{T}^n)\mathbb{J}_E(\mathbf{V}_\mathcal{T}) \\
&= \sum_{K\in\mathcal{T}} \frac{1}{\tau}\int_K M(\mathbf{U}_\mathcal{T}^{n-1})\cdot\mathbf{V}_\mathcal{T} + \sum_{K\in\mathcal{T}}\int_K \mathbf{g}(\cdot,n\tau)\cdot\mathbf{V}_\mathcal{T} \\
&+ \sum_{K\in\mathcal{T}} \delta_K \int_K \mathbf{g}(\cdot,n\tau)\cdot\operatorname{div}\underline{\mathbf{F}}(\mathbf{V}_\mathcal{T})
\end{aligned}$$

Dieser Ansatz kann wie folgt ausgebaut werden:

- Die Sprung- und Stabilisierungsterme können verfeinert werden.
- Der Zeitschritt kann variabel sein.
- Das Ortsgitter kann für jede Zwischenzeit unterschiedlich gewählt werden.
- Die Funktionen $\mathbf{U}_\mathcal{T}$ und $\mathbf{V}_\mathcal{T}$ können auch bezüglich der Zeit Polynome höheren Grades sein. Dann müssen Terme der Form

$$\sum_{K\in\mathcal{T}} \int_{(n-1)\tau}^{n\tau}\int_K \frac{\partial M(\mathbf{U}_\mathcal{T})}{\partial t}\cdot\mathbf{V}_\mathcal{T}$$

auf der linken Seite obiger Gleichung und Terme der Form

$$\frac{\partial M(\mathbf{U}_\mathcal{T})}{\partial t}\cdot\mathbf{V}_\mathcal{T}$$

zu den Elementresiduen hinzugefügt werden.

Literaturverzeichnis

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, 1975.

[2] H. W. Alt, *Lineare Funktionalanalysis. Eine anwendungsorientierte Einführung.*, Hochschultext. Berlin etc.: Springer-Verlag. IX, 292 S., 1985.

[3] D. Braess, *Finite Elements: Theory, Fast Solvers and Applications in Solid Mechanics*, 3 rd ed., Cambridge University Press, 2007.

[4] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, Berlin - Heidelberg - New York, 1994.

[5] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, Berlin, 2011.

[6] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Series in Computational Mathematics, vol. 15, Springer, Berlin, 1991.

[7] P. G. Ciarlet, *Basic Error Estimates for Elliptic Problems*, Handbook of Numerical Analysis (P. G. Ciarlet and J. L. Lions, eds.), vol. 2, North-Holland, 1991, pp. 17 – 351.

[8] D. A. Di Pietro and A. Ern, *Mathematical aspects of discontinuous Galerkin methods*, Mathématiques & Applications (Berlin) [Mathematics & Applications], vol. 69, Springer, Heidelberg, 2012.

[9] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Classics in Mathematics, Springer-Verlag, Berlin, 2001, Reprint of the 1998 edition.

[10] H. Goering, H.-G. Roos, and L. Tobiska, *Die Finite Element Methode für Anfänger*, Wiley-VCH, Weinheim, 2010.

[11] J. Nečas, *Les Méthodes Directes en Théorie des Equations Elliptiques*, Masson et Cie, Éditeurs, Paris, 1967.

[12] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical Mathematics*, second ed., Texts in Applied Mathematics, vol. 37, Springer-Verlag, Berlin, 2007.

[13] R. Verfürth, *A Posteriori Error Estimation Techniques for Finite Element Methods*, Oxford University Press, Oxford, 2013.

[14] ———, *Adaptive Finite Element Methods*, Skriptum, Ruhr-Universität Bochum, Bochum, August 2018, 129 Seiten, www.rub.de/num1/files/lectures/AdaptiveFEM.pdf.

[15] ———, *Einführung in die Numerische Mathematik*, Skriptum, Ruhr-Universität Bochum, Bochum, Januar 2018, 137 Seiten, www.rub.de/num1/files/lectures/EinfNumerik.pdf.

[16] ———, *Numerik I - Gewöhnliche Differentialgleichungen und Differenzenverfahren für partielle Differentialgleichungen*, Skriptum, Ruhr-Universität Bochum, Bochum, August 2018, 145 Seiten, www.rub.de/num1/files/lectures/NumDgl1.pdf.

[17] D. Werner, *Funktionalanalysis*, extended ed., Springer-Verlag, Berlin, 2000.

Index