



**computational mathematics**

Prof. Dr. S. Sauter  
Institut für Mathematik  
Universität Zürich

# Numerische Methoden für elliptische und parabolische Differentialgleichungen

S. Sauter

Herbstsemester 2017

Version 13.12. 2017

## Literatur

- [1] S. Brenner and L. Scott. *The Mathematical Theory of Finite Element Methods*. Springer-Verlag, New York, 1994.
- [2] W. Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen*. Teubner, Stuttgart, 1996.
- [3] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems*. Springer, Berlin, 1997.

# 1 Partielle Differentialgleichungen

Für viele physikalische Problemstellungen ist eine Grösse (Lösungsfunktion) gesucht, die im allgemeinen von mehr als einer Variablen abhängt – typischerweise von den drei Raumkoordinaten, von der Zeit oder auch von der Temperatur.

Physikalische Problemstellungen lassen sich sehr häufig als Differentialgleichungen formulieren, deren Lösung entweder direkt die gesuchte physikalische Grösse ergibt oder daraus abgeleitet werden kann.

Hängt die unbekannte Funktion lediglich von einer Variablen ab, spricht man von gewöhnlichen Differentialgleichungen, andernfalls von partiellen Differentialgleichungen. Die Analysis und numerischen Lösungsmethoden für partielle Differentialgleichungen unterscheiden sich ganz wesentlich von denjenigen für gewöhnliche Differentialgleichungen.

Man unterscheidet drei Typen partieller Differentialgleichungen, für die sehr unterschiedliche Methoden zur Analyse und Numerik erforderlich sind.

Die unbekannte Funktion wird immer mit  $u$  bezeichnet und deren freie Variablen mit  $x_1, x_2, \dots, x_d$ . Die partiellen Ableitungen sind durch

$$\partial_i u := \partial_{x_i} u(x_1, \dots, x_d) := \frac{\partial u(x_1, \dots, x_d)}{\partial x_i} \quad 1 \leq i \leq d.$$

gegeben. Die Variablen  $(x_1, x_2, \dots, x_d)^\top$  werden zum Vektor  $\mathbf{x} \in \mathbb{R}^d$  zusammengefasst.

Wir motivieren die Typeneinteilung der partiellen Differentialgleichungen an Hand einiger Beispiele.

## 1.1 Beispiele

Die ersten beiden Beispiele sind partielle Differentialgleichungen *erster* Ordnung, da sie nur partielle Ableitungen erster Ordnung enthalten.

**Beispiel 1.1** *Gesucht ist die Lösung  $u(x_1, x_2)$  von*

$$\frac{\partial u(x_1, x_2)}{\partial x_2} = 0. \tag{1.1}$$

*Offensichtlich muss die Lösung  $u$  unabhängig von  $x_2$  sein und besitzt daher die Form*

$$u(x_1, x_2) = \varphi(x_1). \tag{1.2}$$

*Andererseits erfüllt jede Funktion der Bauart (1.2) die Gleichung (1.1).*

Die Differentialgleichung (1.1) ist ein Spezialfall des folgenden Beispiels.

**Beispiel 1.2** *Sei  $c \in \mathbb{R}$  gegeben. Gesucht ist eine Funktion  $u(x_1, x_2)$  mit*

$$c \frac{\partial u(x_1, x_2)}{\partial x_1} - \frac{\partial u(x_1, x_2)}{\partial x_2} = 0. \tag{1.3}$$

*Die Koordinatentransformation*

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \chi(\xi, \eta) := \begin{pmatrix} \xi - c\eta \\ \eta \end{pmatrix}$$

führt mit der Bezeichnung  $\hat{u} = u \circ \chi$  auf die folgende Differentialgleichung für  $\hat{u}$  :

$$\frac{\partial \hat{u}(\xi, \eta)}{\partial \eta} = 0.$$

Beispiel 1.1 zeigt, dass die Lösung die Form  $\hat{u}(\xi, \eta) = \varphi(\xi)$  besitzt. Daraus ergibt sich die allgemeine Lösung von (1.3)

$$u(x_1, x_2) = \varphi(x_1 + cx_2), \quad (1.4)$$

vorausgesetzt die Funktion  $\varphi$  ist stetig differenzierbar.

Wie für die gewöhnlichen Differentialgleichungen benötigen wir noch Zusatzbedingungen, um die Lösung der partiellen Differentialgleichung (1.3) eindeutig zu charakterisieren. Dies kann durch die Anfangsbedingung

$$u(x_1, t_0) = u_0(x_1) \quad (1.5)$$

mit einem gegebenen  $t_0$  und zugehöriger Funktion  $u_0$  geschehen. Vergleich von (1.4) und (1.5) führt auf

$$u(x_1, t_0) = \varphi(x_1 + ct_0) = u_0(x_1).$$

Durch Translation der Variablen erhält man schliesslich

$$\varphi(x_1) = u_0(x_1 - ct_0) \quad \text{und} \quad u(x_1, x_2) = \varphi(x_1 + cx_2) = u_0(x_1 + c(x_2 - t_0)).$$

Die folgenden Beispiele sind Gleichungen zweiter Ordnung. Sei im folgenden  $\Omega$  immer eine offene Teilmenge des  $\mathbb{R}^d$ . Für eine zweimal stetig differenzierbare Funktion  $u : \Omega \rightarrow \mathbb{R}$  ist der Laplace-Operator durch

$$\Delta u(x_1, x_2, \dots, x_d) = \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2} u(x_1, x_2, \dots, x_d)$$

gegeben.

**Beispiel 1.3 (Potential- oder Laplace-Gleichung)** *Gesucht ist eine Lösung von*

$$\Delta u = 0 \quad \text{in } \Omega. \quad (1.6)$$

In zwei Dimensionen lassen sich die Lösungen dieser Gleichung direkt angeben. Identifiziert man  $(x_1, x_2)$  mit der komplexen Variablen  $x_1 + ix_2$ , so sind die Real- und Imaginärteile aller in  $\Omega$  holomorphen Funktionen Lösungen von (1.6). (Wir erinnern hier an die Cauchy-Riemannschen Differentialgleichungen: Sei  $u$  eine komplexe Funktion mit Realteil  $u_R := \operatorname{Re} u$  und Imaginärteil  $u_I := \operatorname{Im} u$ . Dann ist  $u$  komplex differenzierbar, genau dann wenn

$$\begin{aligned} \frac{\partial_1 u_R(x_1, x_2)}{\partial x_1} &= \frac{\partial_2 u_I(x_1, x_2)}{\partial x_2} \\ \frac{\partial u_R(x_1, x_2)}{\partial x_2} &= -\frac{\partial u_I(x_1, x_2)}{\partial x_1} \end{aligned}$$

gilt. Differenziert man die erste Gleichung nach  $x_1$  und die zweite nach  $x_2$  erhält man

$$\frac{\partial_1^2 u_R}{\partial x_1^2} = -\frac{\partial^2 u_R(x_1, x_2)}{\partial x_2^2},$$

d.h.  $u_R$  erfüllt die Laplace-Gleichung. Analog erhält man die Eigenschaft für den Imaginärteil. Beispiele für Lösungen von (1.6) sind  $\operatorname{Re} z^0 = 1$ ,  $\operatorname{Im} z^3 = y(3x^2 - y^2)$ . ) Zur eindeutigen Lösbarkeit benötigt man Randwerte

$$u(x_1, x_2, \dots, x_d) = \varphi(x_1, x_2, \dots, x_d) \quad \text{für } (x_1, x_2, \dots, x_d) \in \partial\Omega$$

mit einer vorgegebenen Funktion  $\varphi$ .

**Beispiel 1.4 (Wellengleichung)** Alle Lösungen der Gleichung

$$\frac{\partial^2 u}{\partial x_1^2} - \frac{\partial^2 u}{\partial x_2^2} = 0 \tag{1.7}$$

besitzen die Form

$$u(x_1, x_2) = \varphi(x_1 + x_2) + \psi(x_1 - x_2),$$

wobei  $\varphi, \psi$  beliebige, zweifach stetig differenzierbare Funktionen sind. Geeignete Anfangswerte sind

$$u(x_1, 0) = u_0(x_1), \quad \frac{\partial u(x_1, 0)}{\partial x_2} = u_1(x_1)$$

mit gegebenen Funktionen  $u_0$  und  $u_1$ .

**Beispiel 1.5 (Wärmeleitungsgleichung)** Gesucht ist die Lösung von

$$\Delta u(x_1, x_2, \dots, x_d, t) - \partial_t u(x_1, x_2, \dots, x_d, t) = 0, \tag{1.8}$$

wobei sich der Laplace-Operator auf die  $x_1, \dots, x_d$ -Variablen bezieht. Verwendet man den Ansatz  $u(\mathbf{x}, t) = v(\mathbf{x})w(t)$  erhält man

$$w(t) \Delta v(\mathbf{x}) = v(\mathbf{x}) \partial_t w(t).$$

Division durch  $v$  und  $w$  ergibt

$$\frac{\Delta v(\mathbf{x})}{v(\mathbf{x})} = \frac{\partial_t w(t)}{w(t)}.$$

Die linke Seite hängt nicht von  $t$  und die rechte nicht  $\mathbf{x}$  ab und daher stimmen beide Seite mit einer freien Konstanten  $c \in \mathbb{R}$  überein:

$$\Delta v = cv \quad \text{und} \quad \partial_t w = cw.$$

Beschränken wir uns auf den Fall  $d = 1$  sind dies zwei gewöhnliche Differentialgleichungen ( $\Delta = \partial_x^2$ ), die beispielsweise die Lösungen

$$u(x, t) = e^{\sqrt{c}x + ct}$$

besitzen. Bezüglich der Zeit ist dies eine Gleichung erster Ordnung, und es genügt (im Gegensatz zur Wellengleichung) eine Anfangsbedingung vorzugeben:

$$u(x_1, \dots, x_d, 0) = u_0(x_1, \dots, x_d).$$

Genauso wie für gewöhnliche Differentialgleichungen gibt es auch physikalische Anwendungen, welche auf Systeme partieller Differentialgleichungen. Ein Beispiel hierfür ist die Stokes-Gleichung, welche laminare (wirbelfreie) Strömungen beschreibt. Die Komponenten der Lösungsfunktion ist das Geschwindigkeitsfeld der Strömung  $u_1, \dots, u_d$  und der Druck  $p$ .

**Beispiel 1.6 (Stokes-Gleichung)** Die Stokes-Gleichungen sind durch

$$\frac{\partial^2 u_1}{\partial^2 x_1} + \frac{\partial^2 u_1}{\partial^2 x_2} + \frac{\partial p}{\partial x_1} = 0$$

$$\frac{\partial^2 u_2}{\partial^2 x_1} + \frac{\partial^2 u_2}{\partial^2 x_2} + \frac{\partial p}{\partial x_2} = 0$$

$$\frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} = 0$$

gegeben.

## 1.2 Typeneinteilung bei Gleichungen zweiter Ordnung

Eine Differentialgleichung heisst linear, falls die Gleichung linear von der unbekanntem Funktion abhängt. Sei  $Lu$  eine Abkürzung für die linke Seite der Differentialgleichung. Für lineare Differentialgleichungen gilt das Superpositionsprinzip. Löst die Funktion  $u_1$  die Gleichung  $Lu_1 = f_1$  und die Funktion  $u_2$  die Gleichung  $Lu_2 = f_2$ , so löst die Linearkombination  $u = \alpha u_1 + \beta u_2$  die Gleichung  $Lu = \alpha f_1 + \beta f_2$ .

Die allgemeine Differentialgleichung zweiter Ordnung lautet

$$-\operatorname{div}(\mathbf{A} \operatorname{grad} u) + \langle \mathbf{b}, \operatorname{grad} u \rangle + cu = g. \quad (1.9)$$

Für eine stetig differenzierbare Funktion  $u : \mathbb{R}^d \rightarrow \mathbb{R}$  ist der **Gradient** durch

$$\operatorname{grad} u(\mathbf{x}) = \left( \frac{\partial u(x_1, \dots, x_d)}{\partial x_i} \right)_{i=1}^d$$

gegeben und für ein differenzierbares Vektorfeld  $\mathbf{v} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  ist die **Divergenz** durch

$$\operatorname{div} \mathbf{v}(\mathbf{x}) = \sum_{i=1}^d \frac{\partial v_i(x_1, \dots, x_d)}{\partial x_i}$$

gegeben. Die Koeffizienten in (1.9) sind Matrix- (bzw. Vektor-) wertige (bzw. skalare) Funktionen  $\mathbf{A} : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$  (bzw.  $\mathbf{b} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  bzw.  $c : \mathbb{R}^d \rightarrow \mathbb{R}$ ). Die Matrix  $\mathbf{A}$  ist hierbei differenzierbar und symmetrisch.

**Übungsaufgabe 1.7** Zeigen Sie, dass der allgemeinere Ansatz

$$-\sum_{i,j=1}^d \tilde{\mathbf{A}}_{i,j}(\mathbf{x}) \partial_j \partial_i u(\mathbf{x}) + \sum_{i=1}^d \tilde{\mathbf{b}}_i(\mathbf{x}) \partial_i u(\mathbf{x}) + \tilde{c}(\mathbf{x})$$

für zweimal stetig differenzierbare Funktionen  $u$  immer auf die Form der linken Seite in (1.9) transformiert werden kann.

**Bemerkung 1.8** Aus der Symmetrie der Matrix  $\mathbf{A}(\mathbf{x})$  folgt, dass alle Eigenwerte reell sind.

**Definition 1.9** (a) Die Gleichung (1.9) heisst **elliptisch** in  $\mathbf{x} \in \mathbb{R}^d$ , falls  $\mathbf{A}(\mathbf{x})$  positiv definit oder negativ definit ist.

(b) Die Gleichung (1.9) heisst **hyperbolisch** in  $\mathbf{x} \in \mathbb{R}^d$ , falls  $d - 1$  Eigenwerte von  $\mathbf{A}(\mathbf{x})$  gleiches Vorzeichen besitzen und ein Eigenwert entgegengesetztes Vorzeichen besitzt.

(c) Die Gleichung (1.9) heisst **parabolisch** in  $\mathbf{x} \in \mathbb{R}^d$ , falls  $d - 1$  Eigenwerte von  $\mathbf{A}(\mathbf{x})$  gleiches Vorzeichen besitzen, ein Eigenwert gleich Null ist und  $\text{Rang}(\mathbf{A}(\mathbf{x}), \mathbf{b}(\mathbf{x})) = d$  gilt.

(d) Die Gleichung (1.9) heisst elliptisch in  $\Omega \subset \mathbb{R}^d$ , falls  $\mathbf{A}(\mathbf{x})$  für alle  $\mathbf{x} \in \Omega$  positive definit oder für alle  $\mathbf{x} \in \Omega$  negativ definit ist.

Man beachte, dass diese Typeneinteilung nicht alle möglichen Koeffizientenfälle abdeckt, die in (1.9) auftreten.

**Beispiel 1.10** (a) Die Potentialgleichung (1.6) ist elliptisch. Es gilt  $\mathbf{A} = \text{Einheitsmatrix}$ .

**Example 1.11** (b) Die Wärmeleitungsgleichung (1.8) ist parabolisch. Es gilt

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & & 0 & \\ 0 & & & 0 & 1 \\ 0 & \dots & & & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -1 \end{pmatrix}.$$

(c) Die Wellengleichung (1.7) ist hyperbolisch. Es gilt

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & & 0 & \\ 0 & & & 0 & 1 \\ 0 & \dots & & 0 & -1 \end{bmatrix}.$$

In dieser Vorlesung werden wir uns in erster Linie mit der Lösung von elliptischen Differentialgleichungen beschäftigen. Es sei hier angemerkt, dass typischerweise die Diskretisierung parabolischer und hyperbolischer Gleichungen bzgl. der Ortsvariablen  $x_1, \dots, x_d$  die partielle Differentialgleichung in ein *System* gewöhnlicher Differentialgleichungen umwandelt. Dieses lässt sich dann mit geeigneten numerischen Methoden für gewöhnliche Differentialgleichung vollständig diskretisieren.

## 2 Sobolev-Räume

Heutzutage werden zur Lösung elliptischer Differentialgleichungen vor allem sogenannte *Finite-Elemente-Methoden* eingesetzt. Diese werden seit etwa 1965 entwickelt und haben die bis dahin verwendeten *Differenzenverfahren* weitestgehend verdrängt. Letztere beruhten auf der punktweisen Sichtweise der Differentialgleichung: Der Ableitungsoperator wurde in geeigneten Punkten durch einen Differenzenquotienten ersetzt.

Die punktweise Betrachtung einer partiellen Differentialgleichung besitzt jedoch einen entscheidenden Nachteil: Existenz und Eindeutigkeitssätze lassen sich nicht in befriedigender

Weise formulieren. Dies liegt u.a. daran, dass der Raum aller  $k$ -mal stetig differenzierbaren Funktionen kein Hilbert-Raum ist.

Auf der anderen Seite wissen wir aus der Analysis, dass der Lebesgue-Raum  $L^2$  ein Hilbert-Raum ist. Nun sind Funktionen  $f \in L^2$  nicht punktweise definiert. Dies macht die Definition einer Ableitung -die klassisch als Limes eines geeigneten Differenzquotienten gegeben sind- auf den ersten Blick unmöglich.

Wir werden in diesem Unterkapitel zeigen, wie sich der Ableitungsbegriff erweitern lässt auf gewisse Teilmengen von  $L^2$ , den sogenannten Sobolev-Räumen.

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  $p' \in [1, \infty[$ ,  $\frac{1}{p} + \frac{1}{p'} = 1$ . Mehrdimensionale Ableitungen lassen sich mittels Multiindizes kompakt schreiben. Für eine Multiindex  $\alpha \in \mathbb{N}_0^d$  setzen wir  $|\alpha| = \alpha_1 + \dots + \alpha_d$  und

$$D^\alpha \varphi := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \dots \partial x_d^{\alpha_d}} \varphi \quad \forall \varphi \in C^{|\alpha|}(\Omega).$$

Da  $\Omega$  beschränkt ist, gilt  $L^p(\Omega) \subset L^1(\Omega)$  und die triviale Injektion  $\iota : L^p(\Omega) \rightarrow L^1(\Omega)$  ist stetig. Wir nehmen darüber hinaus immer an, dass  $\Omega$  ein *Normalgebiet* ist, für das der Gaussische Integralsatz angewendet werden kann. Daraus folgt für alle  $\varphi, \psi \in C_0^\infty(\Omega)$  mit<sup>1</sup>

$$C_0^\infty(\Omega) := \{v \in C^\infty(\Omega) : \text{supp } v \subset \Omega\}.$$

und alle  $\alpha \in \mathbb{N}_0^d$

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

**Definition 2.1** Seien  $\varphi, \psi \in L^1(\Omega)$  und  $\alpha \in \mathbb{N}_0^d$ . Dann heisst  $\psi$  die  $\alpha$ te **schwache Ableitung** von  $\varphi$ , kurz  $\psi = D^\alpha \varphi$ , wenn für all  $\rho \in C_0^\infty(\Omega)$  gilt

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

**Bemerkung 2.2** (1) Die  $\alpha$ -te schwache Ableitung ist, sofern sie existiert, eindeutig bestimmt (im Sinne von  $L^1$ -Funktionen).

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

**Beweis.** zu (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

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

Proposition 2.3 zeigt, dass daraus  $\psi_1 = \psi_2$  im  $L^1$ -Sinne folgt.

zu (2): Folgt aus dem Gaussischen Integralsatz.

---

<sup>1</sup>Der Träger einer Funktion  $v$  ist durch

$$\text{supp } v := \overline{\{x \in \Omega \mid v(x) \neq 0\}}.$$



**Proposition 2.3** Sei  $g \in L^1(\Omega)$ ,  $\Omega \subset \mathbb{R}^d$  offen. Dann gilt

$$\int_{\Omega} f \rho = 0 \quad \forall \rho \in C_0^\infty(\Omega) \iff f = 0 \text{ fast überall in } \Omega.$$

**Beweis.** Die Richtung „ $\Leftarrow$ “ folgt offensichtlich aus der Tatsache, dass das Integral über Nullmengen verschwindet, und wir beweisen im Folgenden „ $\Rightarrow$ “.

(a) Konstruktion einer Dirac-Folge. Sei  $\varphi \in C_0^\infty(B_1(0))$  mit

$$\varphi \geq 0 \quad \text{und} \quad \int_{\mathbb{R}^d} \varphi = 1.$$

Dann ist für  $\varepsilon > 0$  die **Dirac-Folge** durch

$$\varphi_\varepsilon(x) = \varepsilon^{-d} \varphi(x/\varepsilon)$$

gegeben. Es gilt für jedes  $\delta > 0$

$$\int_{\mathbb{R}^d} \varphi_\varepsilon = 1 \quad \text{und} \quad \int_{\mathbb{R}^d \setminus B(0,\delta)} \varphi_\varepsilon \xrightarrow{\varepsilon \rightarrow 0} 0.$$

Sei  $E$  messbar und beschränkt mit  $\bar{E} \subset \Omega$ . Betrachte Funktionen

$$\rho_\varepsilon(x) := \int_E \varphi_\varepsilon(x-y) dy$$

mit der Dirac-Folge  $\varphi_\varepsilon$ . Wie aus der Analysis (Differentiation nach Parametern) bekannt ist, gilt  $\rho_\varepsilon \in C_0^\infty(\Omega)$  für hinreichend kleines  $\varepsilon$  und ausserdem  $0 \leq \rho_\varepsilon \leq 1$ .

(b)  $\rho_\varepsilon \rightarrow \chi_E$ :

Es gilt

$$(\rho_\varepsilon - \chi_E)(x) = \int_{\mathbb{R}^d} (\chi_E(y) - \chi_E(x)) \varphi_\varepsilon(x-y) dy.$$

Für eine beliebige Funktion messbare Funktion  $\psi$  auf  $\mathbb{R}^d$  gilt für

$$F(x) := \int_{\mathbb{R}^d} (\chi_E(y) - \chi_E(x)) \psi(x-y) dy$$

die Abschätzung

$$\begin{aligned} \|F\|_{L^p(\mathbb{R}^d)}^p &\leq \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} |\psi(x-y)|^{\frac{1}{p'}} |\psi(x-y)|^{\frac{1}{p}} |\chi_E(y) - \chi_E(x)| dy \right|^p dx \\ &\stackrel{\text{Hölder}}{\leq} \int_{\mathbb{R}^d} \left( \int_{\mathbb{R}^d} |\psi(x-y)| dy \right)^{\frac{p}{p'}} \left( \int_{\mathbb{R}^d} |\psi(x-y)| |\chi_E(y) - \chi_E(x)|^p dy \right) dx \\ &\leq \|\psi\|_{L^1(\mathbb{R}^d)}^{p/p'} \int_{\mathbb{R}^d} |\psi(x)| \int_{\mathbb{R}^d} |\chi_E(y) - \chi_E(y+x)|^p dy dx \\ &\leq \|\psi\|_{L^1(\mathbb{R}^d)}^p \sup_{x \in \text{supp } \psi} \int_{\mathbb{R}^d} |\chi_E(y) - \chi_E(y+x)|^p dy, \end{aligned}$$

d.h.

$$\|F\|_{L^p(\mathbb{R}^d)} \leq \|\psi\|_{L^1(\mathbb{R}^d)} \sup_{x \in \text{supp } \psi} \|\chi_E(\cdot) - \chi_E(\cdot+x)\|_{L^p(\mathbb{R}^d)}.$$

Wir wenden diese Aussage auf eine additive Zerlegung von  $\varphi_\varepsilon$  an und setzen für  $\delta > 0$  in  $\varphi_{\varepsilon,\delta} := \chi_{B(0,\delta)}\varphi_\varepsilon$  und  $\psi_{\varepsilon,\delta} := \chi_{\mathbb{R}^d \setminus B(0,\delta)}\varphi_\varepsilon$ . Wir erhalten

$$\left\| \int_{\mathbb{R}^d} (\chi_E(y) - \chi_E(\cdot)) \varphi_{\varepsilon,\delta}(\cdot - y) dy \right\|_{L^p(\mathbb{R}^d)} \leq \underbrace{\left( \int_{\mathbb{R}^d} \varphi_{\varepsilon,\delta} \right)}_{\leq 1} \sup_{x \in \text{supp } \varphi_{\varepsilon,\delta}} \|\chi_E(\cdot) - \chi_E(\cdot + x)\|_{L^p(\mathbb{R}^d)} \quad (2.1)$$

und

$$\left\| \int_{\mathbb{R}^d} (\chi_E(y) - \chi_E(\cdot)) \psi_{\varepsilon,\delta}(\cdot - y) dy \right\|_{L^p(\mathbb{R}^d)} \leq \underbrace{\left( \int_{\mathbb{R}^d} \psi_{\varepsilon,\delta} \right)}_{\xrightarrow{\varepsilon \rightarrow 0} 0} \underbrace{\sup_{x \in \text{supp } \psi} \|\chi_E(\cdot) - \chi_E(\cdot + x)\|_{L^p(\mathbb{R}^d)}}_{\leq 2\|\chi_E\|_{L^p(\mathbb{R}^d)}}.$$

Es bleibt das Supremum auf der rechten Seite von (2.1) abzuschätzen:

$$\sup_{x \in B(0,\delta)} \|\chi_E(\cdot) - \chi_E(\cdot + x)\|_{L^p(\mathbb{R}^d)}$$

Wir wählen eine Folge von Funktionen  $\chi_j \in C_0^\infty(\mathbb{R}^d)$  mit  $\|\chi_j - \chi_E\|_{L^p(\mathbb{R}^d)} \rightarrow 0$  für  $j \rightarrow \infty$ . Damit gilt

$$\begin{aligned} \|\chi_E(\cdot + x) - \chi_E(\cdot)\|_{L^p(\mathbb{R}^d)} &\leq \|\chi_E(\cdot + x) - \chi_j(\cdot + x)\|_{L^p(\mathbb{R}^d)} + \|\chi_E - \chi_j\|_{L^p(\mathbb{R}^d)} \\ &\quad + \|\chi_j(\cdot + x) - \chi_j(\cdot)\|_{L^p(\mathbb{R}^d)} \\ &\leq 2\|\chi_E - \chi_j\|_{L^p(\mathbb{R}^d)} + |\text{Tr } \chi_j|^{1/p} \|\chi_j(\cdot + x) - \chi_j(\cdot)\|_\infty. \end{aligned}$$

Der erste Term konvergiert per Konstruktion gegen Null, der zweite wegen der gleichmässigen Stetigkeit von  $\chi_j$  für  $\delta \rightarrow 0$ .

Zusammen haben wir

$$\|\rho_\varepsilon - \chi_E\|_{L^p(\Omega)} \xrightarrow{\varepsilon \rightarrow 0} 0 \quad (2.2)$$

gezeigt für beliebiges  $1 \leq p < \infty$ .

c)  $f = 0$

Daher gibt es eine Funktionenfolge  $(\rho_{\varepsilon_n})_{n \in \mathbb{N}}$ , die fast überall gegen  $\chi_E$  konvergiert. Mit dem Satz von Lebesgue folgt dann  $f\rho_{\varepsilon_n} \rightarrow f\chi_E$  in  $L^1(\Omega)$  und somit

$$0 \stackrel{\text{Voraussetzung}}{=} \int_{\Omega} f\rho_{\varepsilon_n} \rightarrow \int_{\Omega} f\chi_E = \int_E f.$$

Nun müssen wir  $\Omega$  noch geeignet *ausschöpfen*. Für  $\delta > 0$  setzen wir

$$\Omega_\delta := \{x \in \Omega : \text{dist}(x, \partial\Omega) > \delta\} \quad \text{und} \quad D_\delta := \Omega_\delta \cap B\left(0, \frac{1}{\delta}\right).$$

Offensichtlich schöpfen die Gebiete  $D_\delta$  für  $\delta \rightarrow 0$  das Gebiet  $\Omega$  aus. Wir setzen

$$E_\pm = \{x \in D_\delta : \pm f(x) > 0\}$$

und erhalten

$$\int_{D_\delta} |f| = \int_{E_+} f - \int_{E_-} f = 0.$$

Daraus folgt  $f = 0$  fast überall. ■

**Beispiel 2.4** Sei  $\Omega = ]-1, 1[$  und  $\varphi(x) = |x|$ . Dann ist  $\varphi$  im Sinne von Definition 2.1 differenzierbar und es gilt

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

**Beweis.** Für alle  $\rho \in C_0^\infty(\Omega)$  gilt

$$\begin{aligned} \int_{-1}^1 \varphi \rho' &= \int_{-1}^0 \varphi \rho' + \int_0^1 \varphi \rho' \\ &= - \int_{-1}^0 \psi \rho + \varphi(0-) \rho(0-) - \varphi(-1) \rho(-1) - \int_0^1 \psi \rho + \varphi(1) \rho(1) - \varphi(0+) \rho(0+). \end{aligned}$$

Die Stetigkeit von  $\rho$  über 0 impliziert  $\rho(0+) = \rho(0-)$ , und wir erhalten

$$\int_{-1}^1 \varphi \rho' = - \int_{-1}^1 \psi \rho + [\varphi]_0 \rho(0) - \varphi(-1) \rho(-1) + \varphi(1) \rho(1)$$

mit dem Sprung  $[\varphi]_0 = \varphi(0-) - \varphi(0+)$ . Wegen  $[\varphi]_0 = 0$  und  $\rho(-1) = \rho(1) = 0$  gilt daher

$$\int_{-1}^1 \varphi \rho' = - \int_{-1}^1 \psi \rho \quad \forall \rho \in C_0^\infty(\Omega).$$

**Übungsaufgabe 2.5** Zeigen sie, dass  $\psi(x)$  auf  $]-1, 1[$  nicht schwach differenzierbar.

Für die folgende Definition benötigen wir den Begriff der Seminorm, d.h. die Bedingung:

$$\|\varphi\| \geq 0 \quad \text{und} \quad \varphi = 0 \iff \|\varphi\| = 0$$

für eine Norm ist für eine Seminorm durch die schwächere Bedingung  $\|\varphi\| \geq 0$  ersetzt.

**Definition 2.6** (1) Für  $k \in \mathbb{N}$  und  $p \in [1, \infty[$  ist der Sobolev-Raum  $W^{k,p}(\Omega)$  und seine Norm durch

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

(2) Durch

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

ist eine Seminorm auf  $W^{k,p}(\Omega)$  gegeben.

In der Vorlesung wird der Fall  $p = 2$  am häufigsten auftreten, und wir verwenden die Abkürzung  $H^k(\Omega)$  für  $W^{k,2}(\Omega)$  und lassen den Index  $p$  bei der Norm und Seminorm weg.

**Beispiel 2.7** (1) Seien  $\varphi$  und  $\Omega$  wie in Bsp. 2.4. Dann gilt  $\varphi \in W^{1,p}$  für alle  $p \in [1, \infty[$ .

(2) Seien  $\Omega = B(0, \frac{1}{2})$  die offene  $d$ -dimensionale Kreisscheibe um den Ursprung mit Radius  $\frac{1}{2}$  und  $\varphi(x) = \|x\|^s$  mit  $s \in \mathbb{R}$ . Die Euklidische Norm im  $\mathbb{R}^d$  wird mit  $\|\cdot\|$  bezeichnet. Dann gilt

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

und

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

Der Integrand im letzten Integral ist integrierbar, genau dann wenn

$$p(s - |\alpha|) + d - 1 > -1, \quad \text{d.h.} \quad s > |\alpha| - \frac{d}{p}$$

gilt.

**Übungsaufgabe 2.8** (a) Sei  $\Omega = ]0, 1/2[$ . Finden Sie eine Funktion  $f : \Omega \rightarrow \mathbb{R}$  mit  $f \in L^p(\Omega)$  und  $f \notin L^{p+\varepsilon}(\Omega)$  für alle  $\varepsilon > 0$ .

(b) Sei  $\Omega$  wie in Beispiel 2.7(2). Verwenden Sie Teil (a), um eine unbeschränkte Funktion in  $H^1(\Omega)$  zu konstruieren.

**Satz 2.9** (1)  $W^{k,p}(\Omega)$  versehen mit der Norm  $\|\cdot\|_{k,p}$  ist ein Banach-Raum.

(2)  $C^\infty(\Omega) \cap W^{k,p}(\Omega)$  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.** Zu (1): Sei  $n_{k,d} := \#\{\alpha \in \mathbb{N}_0^d : |\alpha| \leq k\}$ . Dann können wir  $W^{k,p}(\Omega)$  vermöge der Abbildung  $\iota : \varphi \rightarrow (D^\alpha \varphi)_{|\alpha| \leq k}$  mit  $L^p(\Omega, \mathbb{R}^{n_{k,d}})$  identifizieren. Insbesondere gilt dann  $\|\varphi\|_{k,p} := \|\iota(\varphi)\|_{L^p(\Omega, \mathbb{R}^{n_{k,d}})}$ . 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  $(\iota(\varphi_n))_{n \in \mathbb{N}} \subset L^p(\Omega, \mathbb{R}^{n_{k,d}})$  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  $\psi_\alpha$  konvergiert. Insbesondere konvergiert (nach Übergang zu einer geeigneten Teilfolge, die gleich bezeichnet wird)  $D^\alpha \varphi_n$  punktweise f.ü. gegen  $\psi_\alpha$ . Für jede  $\rho \in C_0^\infty(\Omega)$  gilt andererseits

$$\int_{\Omega} \varphi_n D^\alpha \rho = (-1)^{|\alpha|} \int_{\Omega} (D^\alpha \varphi_n) \rho. \quad (2.3)$$

Wegen des Lebesgueschen Konvergenzsatzes können wir in (2.3) den Grenzübergang  $n \rightarrow \infty$  durchführen und erhalten  $\int_{\Omega} \psi_0 D^\alpha \rho = \lim_{n \rightarrow \infty} \int_{\Omega} \varphi_n D^\alpha \rho = \lim_{n \rightarrow \infty} (-1)^{|\alpha|} \int_{\Omega} (D^\alpha \varphi_n) \rho = (-1)^{|\alpha|} \int_{\Omega} \psi_\alpha \rho$ . Also ist  $\psi_\alpha$  die  $\alpha$ -te schwache Ableitung von  $\psi_0$ , und  $(\varphi_n)_{n \in \mathbb{N}_0}$  konvergiert in  $W^{k,p}(\Omega)$  gegen  $\psi_0$ .

Zu (2): Der Beweis von „ $C^\infty(\Omega) \cap L^p(\Omega)$  ist dicht in  $L^p(\Omega)$ “ aus der Analysis III kann wortwörtlich wiederholt werden.

Zu (3): Offensichtlich ist  $(\cdot, \cdot)_k$  bilinear und  $\|\varphi\|_k^2 = (\varphi, \varphi)_k$ . ■

Die Betragsfunktion war ein Beispiel einer Funktion, die stückweise glatt ist. Der folgende Satz überträgt dieses Differenzierbarkeitsresultat auf allgemeinere Funktionen.

**Satz 2.10** Sei  $\Omega$  ein Gebiet und  $\Omega_1, \Omega_2$  zwei nichtleere, offene, beschränkte und disjunkte Teilmengen von  $\Omega$  mit stückweise glattem Rand auf  $\bar{\Omega} = \bar{\Omega}_1 \cup \bar{\Omega}_2$ . Weiter sei  $\varphi \in L^p(\Omega)$  so, dass  $\varphi|_{\Omega_i} \in C^k(\Omega_i) \cap W^{k,p}(\Omega_i)$ ,  $i \in \{1, 2\}$ ,  $k \geq 1$  gilt. Dann ist  $\varphi \in W^{k,p}(\Omega)$  genau dann, wenn  $\varphi \in C^{k-1}(\Omega)$  ist.

**Beweis.** Es genügt den Fall  $k = 1$  zu betrachten. Der allgemeine Fall folgt dann durch Induktion. Der innere Rand wird mit  $\gamma$  bezeichnet:  $\Omega = \Omega_1 \cup \Omega_2 \cup \gamma$ . Sei  $n$  die äussere Normale an  $\Omega_1$  und für  $x \in \gamma$  wird der Sprung von  $\varphi$  über  $\gamma$  mit

$$[\varphi](x) := \lim_{\varepsilon \rightarrow 0^+} \{\varphi(x + \varepsilon n(x)) - \varphi(x - \varepsilon n(x))\}$$

bezeichnet. Seien  $\rho \in C_0^\infty(\Omega)$  und  $i \in \{1, 2, \dots, d\}$  beliebig. Dann folgt aus dem Gaussischen Integralsatz

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

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

$$\int_{\gamma} [\varphi] \rho n_i = 0 \quad \forall \rho \in C_0^\infty(\Omega), \quad i \in \{1, 2, \dots, d\}.$$

Also ist  $[\varphi] = 0$  f.ü. auf  $\gamma$ , d.h. aber  $\varphi \in C(\Omega)$ .

Ist umgekehrt  $\varphi \in C(\Omega)$ , so verschwindet  $[\varphi]$  auf  $\gamma$ , und aus obiger Identität folgt  $\varphi \in W^{1,p}(\Omega)$ . ■

Man beachte, dass  $C_0^\infty(\Omega)$  im allgemeinen nicht dicht in  $W^{k,p}(\Omega)$  sind. Der Abschluss von  $C_0^\infty(\Omega)$  unter der  $\|\cdot\|_{k,p}$ -Norm definiert Sobolev-Räume mit Nullrandbedingungen in einem „schwachen“ Sinne.

**Definition 2.11**  $W_0^{k,p}(\Omega)$  ist die Vervollständigung von  $C_0^\infty(\Omega)$  unter der  $W^{k,p}(\Omega)$ -Norm und  $H_0^k(\Omega) = W_0^{k,2}(\Omega)$ .

**Definition 2.12** Das Gebiet  $\Omega$  besitzt einen **Lipschitz-Rand** bzw.  $\Omega$  ist ein **Lipschitz-Gebiet**, wenn es ein  $N \in \mathbb{N}$  und offene Mengen  $U_1, \dots, U_N \subset \mathbb{R}^d$  mit folgenden Eigenschaften gibt:

- (1)  $\partial\Omega \subset \bigcup_{i=1}^N U_i$ ,
- (2) Für jedes  $1 \leq i \leq N$  ist  $\partial\Omega \cap U_i$  darstellbar als Graph einer Lipschitz-stetigen Funktion.

**Bemerkung 2.13** (1)  $\Omega$  sei ein Lipschitz-Gebiet. Dann existiert fast überall auf  $\partial\Omega$  das äussere Einheitsnormalenfeld  $n$  zu  $\Omega$ .

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

Eine wichtige Eigenschaft für Funktionen aus Sobolev-Räumen ist die Existenz von Spuren, d.h., die Einschränkung von Sobolev-Funktionen auf niederdimensionale Teilmengen  $\gamma \subset \Omega$  lässt sich sinnvoll definieren. Das ist aus dem Blickwinkel überraschend, da niederdimensionale Teilmengen  $\gamma \subset \Omega$  Nullmengen in  $\Omega$  sind und Funktionen, beispielsweise in  $L^p$ , beliebig auf Nullmengen abgeändert werden können, ohne die Restklasse zu ändern. Der Beweis des Spursatzes benötigt tiefere Hilfsmittel aus der Funktionalanalysis, und wir geben hier nur ein illustrierendes Beispiel an.

**Beispiel 2.14** Sei  $\Omega$  die offene Einheitskreisscheibe in  $\mathbb{R}^2$ . Für  $u \in C^1(\overline{\Omega})$  betrachten wir die Restriktion auf den Rand  $\partial\Omega$  in Polarkoordinaten  $\hat{u}(r, \varphi) := u \circ \psi(r, \varphi)$  mit  $\psi(r, \varphi) := (r \cos \varphi, r \sin \varphi)^\top$  und erhalten mit  $x = (x_1, x_2)^\top$

$$\begin{aligned} \hat{u}(1, \varphi)^2 &= \int_0^1 \frac{\partial}{\partial r} (r^2 \hat{u}(r, \varphi)^2) dr = \int_0^1 2 (r^2 \hat{u} \hat{u}_r + r \hat{u}^2) (r, \varphi) dr \\ &= \int_0^1 2 (r \hat{u} \langle \nabla u, x \rangle + r \hat{u}^2) \Big|_{x=\psi(r, \varphi)} dr \\ &\leq \int_0^1 2 (|\hat{u}| \|\nabla u\| r^2 + r \hat{u}^2) \Big|_{x=\psi(r, \varphi)} dr \\ &\leq \int_0^1 2 (|u| \|\nabla u\| + u^2) \Big|_{x=\psi(r, \varphi)} r dr. \end{aligned}$$

Integration bezüglich der Winkelvariable liefert

$$\|u\|_{L^2(\partial\Omega)}^2 := \int_{\partial\Omega} \hat{u}^2(1, \varphi) d\varphi \leq 2 \int_{\Omega} (|u| \|\nabla u\| + u^2) dx dy.$$

Die Cauchy-Schwarzsche-Ungleichung liefert

$$\|u\|_{L^2(\partial\Omega)}^2 \leq 2 \|u\|_{L^2(\Omega)} \|u\|_{H^1(\Omega)} + 2 \|u\|_{L^2(\Omega)}^2.$$

Die Binomische Formel liefert  $a + b \leq \sqrt{2(a^2 + b^2)}$  und daraus folgt

$$\|u\|_{L^2(\partial\Omega)}^2 \leq 2\sqrt{2} \|u\|_{L^2(\Omega)} \|u\|_{H^1(\Omega)} \leq 2\sqrt{2} \|u\|_{H^1(\Omega)}^2$$

und schliesslich

$$\|u\|_{L^2(\partial\Omega)} \leq 8^{1/4} \|u\|_{H^1(\Omega)}. \quad (2.4)$$

Das vorige Beispiel ist ein Spursatz für  $C^1(\Omega)$ -Funktionen. Da sich  $H^1(\Omega)$ -Funktionen durch  $C^1(\Omega)$ -Funktionen beliebig gut approximieren lassen und die Norm stetig ist, überträgt sich die Spurgleichung (2.4) auf  $H^1(\Omega)$ -Funktionen. Die Interpretation ist wie folgt. Sei  $u_n \in C^\infty(\overline{\Omega})$  eine beliebige Cauchy-Folge in  $H^1(\Omega)$ , die gegen  $u$  konvergiert. Wegen  $u_n \in C^\infty(\overline{\Omega})$  ist die Spurbildung  $\varphi_n := u_n|_{\partial\Omega}$  wohldefiniert. Abschätzung (2.4) besagt, dass  $\varphi_n$  gegen eine  $L^2(\partial\Omega)$ -Funktion  $\varphi$  konvergiert, die wir als „Einschränkung“ von  $u$  auf  $\partial\Omega$  interpretieren. Aussage (2.4) drückt die Stetigkeit des Spuroperators  $\gamma_0 : H^1(\Omega) \rightarrow L^2(\Omega)$ ,  $\gamma_0(u) := \varphi$  aus.

Wir kommen nun zum allgemeinen Spursatz.

**Satz 2.15 (Spursatz)** Seien  $\Omega$  ein Lipschitz-Gebiet und  $k \in \mathbb{N}$ ,  $\ell \in \{0, 1, \dots, k-1\}$ . Dann gibt es eine stetige lineare Abbildung  $\gamma_\ell : W^{k,p}(\Omega) \rightarrow L^p(\partial\Omega)$  mit der Eigenschaft

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

Der Beweis dieses Satzes übersteigt den Rahmen dieser Vorlesung. Wir verweisen stattdessen für diesen Beweis und auch für die übrigen Sätze dieses Kapitels auf [H.-W. Alt, Funktionalanalysis].

Der Spursatz ist essentiell für die Formulierung von elliptischen Differentialgleichungen auf Gebieten  $\Omega$ . Wie bereits in der Einleitung gesagt, benötigen wir für die Existenz und Eindeutigkeit Randbedingungen -beispielsweise Dirichlet-Randbedingungen, bei denen die Werte der unbekanntes Funktion auf dem Rand vorgegeben sind. Daher ist es wesentlich, präzise Kenntnis zu haben, welche Räume zur Beschreibung der vorgegebenen Randbedingungen geeignet sind. Der Spursatz in der Form von Satz 2.15 ist hierfür nicht geeignet, da man beweisen kann:  $\gamma_\ell(W^{k,p}(\Omega)) \subsetneq L^p(\partial\Omega)$ . Man kann zeigen, dass das Bild von  $W^{k,p}(\Omega)$  einen abgeschlossenen Unterraum von  $L^p(\partial\Omega)$  definiert, und wir führen die Bezeichnung

$$\gamma_\ell(W^{k,p}(\Omega)) =: W^{k-\ell-\frac{1}{p},p}(\partial\Omega)$$

ein. Für unsere Anwendungen wird der Fall  $\ell = 0, 1, k = 1$  und  $p = 2$  am häufigsten auftreten und wir setzen:

$$\gamma_0(H^1(\Omega)) =: H^{1/2}(\partial\Omega) \quad \text{und} \quad \gamma_1(H^1(\Omega)) =: H^{-1/2}(\partial\Omega).$$

**Satz 2.16**  $W_0^{k,p}(\Omega) = \{\varphi \in W^{k,p}(\Omega) \mid 0 \leq \ell \leq k-1 : \gamma_\ell(\varphi) = 0\}$ .

Ein für viele Anwendungen ganz wesentliches Hilfsmittel für Randwertprobleme mit Dirichlet-Randbedingungen ist die Friedrichssche Ungleichung.

**Satz 2.17 (Friedrichssche Ungleichung)**  $\|\cdot\|_{k,p}$  und  $|\cdot|_{k,p}$  definieren äquivalente Normen auf  $W_0^{k,p}(\Omega)$ .

**Beweis.** Offensichtlich gilt  $|v|_{k,p} \leq \|v\|_{k,p}$  für alle  $v \in W^{k,p}(\Omega)$ .

Für die umgekehrte Abschätzung wähle  $R > 0$ , so dass  $\Omega \subset B_{|\cdot|_\infty}(0, R)$  gilt. Dabei bezeichne  $|\cdot|_\infty$  die Maximumsnorm auf  $\mathbb{R}^d$ . Sei  $\alpha \in \mathbb{N}_0^d$  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

$$\begin{aligned} |\psi(x)|^p &= \left| \int_{-R}^{x_1} \frac{\partial}{\partial x_1} \psi(t, x_2, x_3, \dots, x_d) dt \right|^p \\ &\leq \|1\|_{L^{p'}(-R,R)}^p \int_{-R}^{x_1} \left| \frac{\partial}{\partial x_1} \psi(t, x_2, x_3, \dots, x_d) \right|^p dt \\ &= (2R)^{p-1} \int_{-R}^R \left| \frac{\partial}{\partial x_1} \psi(t, x_2, x_3, \dots, x_d) \right|^p dt. \end{aligned}$$

Integration über  $\Omega$  liefert mit dem ersten Einheitsvektor  $e_1$  in  $\mathbb{R}^d$ :

$$\begin{aligned} \|D^\alpha \varphi\|_{L^p(\Omega)}^p &= \|\psi\|_{L^p(\Omega)}^p \leq \|\psi\|_{L^p(B_{|\cdot|_\infty}(0,R))}^p \leq (2R)^{p-1} \int_{-R}^R \int_{B_{|\cdot|_\infty}(0,r)} \left| \frac{\partial}{\partial x_1} \psi(t, x_2, x_3, \dots, x_d) \right|^p dx dt \\ &= (2R)^p \|\partial_1 \psi\|_{L^p(\Omega)}^p = (2R)^p \|D^{\alpha+e_1} \varphi\|_{L^p(\Omega)}^p. \end{aligned}$$

Summation über alle Multiindizes mit  $|\alpha| = k-1$  ergibt

$$|\varphi|_{W^{k-1,p}(\Omega)}^p \leq (2R)^p |\varphi|_{W^{k,p}(\Omega)}^p.$$

Summation über alle  $k$  ergibt:

$$\begin{aligned} \|\varphi\|_{W^{k,p}(\Omega)}^p &= |\varphi|_{W^{k,p}(\Omega)}^p + \sum_{i=0}^{k-1} |\varphi|_{W^{i,p}(\Omega)}^p \\ &\leq \left(1 + (2R)^p + (2R)^p (2R)^p + \dots (2R)^{pk}\right) |\varphi|_{W^{k,p}(\Omega)}^p. \end{aligned}$$

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

**Bemerkung 2.18** Die Friedrichssche Ungleichung lässt sich verallgemeinern für alle Funktionen, welche auf einem Teil  $\Gamma_0 \subset \Gamma$  verschwinden, welcher positives  $d - 1$ -dimensionales Mass besitzt.

**Definition 2.19** Seien  $(X, \|\cdot\|_X)$  und  $(Y, \|\cdot\|_Y)$  zwei normierte Vektorräume.

1. Eine lineare Abbildung  $A : X \rightarrow Y$  heisst **kompakt**, wenn der Abschluss des Bilds  $A\left(\overline{B_X(0,1)}\right)$  in  $Y$  kompakt ist.
2.  $X$  ist **stetig eingebettet** in  $Y$ , kurz  $X \hookrightarrow Y$ , wenn  $X \subset Y$  und die kanonische Injektion  $\iota : X \rightarrow Y$  stetig ist.
3.  $X$  ist **kompakt eingebettet** in  $Y$ , kurz  $X \xhookrightarrow{c} Y$ , wenn  $X \subset Y$  und die kanonische Injektion  $\iota : X \rightarrow Y$  kompakt ist.

**Bemerkung 2.20**

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

**Beweis.** zu (1): Folgt aus der Definition der Stetigkeit für lineare Operatoren (Beschränktheit=Stetigkeit).

zu (2): Sei  $A : X \rightarrow Y$  ein kompakter, linearer Operator. Dann ist nach Voraussetzung  $A\left(\overline{B_X(0,1)}\right)$  kompakt und somit insbesondere beschränkt. Also gibt es ein  $C > 0$  mit  $\|A\varphi\|_Y \leq C$  für alle  $\varphi \in X$  mit  $\|\varphi\|_X = 1$ . Also ist  $A$  stetig.

zu (3):  $(\iota(\varphi_n))_{n \in \mathbb{N}} \subset Y$  ist in der kompakten Menge  $\iota\left(\overline{B_{\|\cdot\|_X}(0,R)}\right)$  mit  $R := \max_{n \in \mathbb{N}} \|\varphi_n\|_X$  enthalten.

**Satz 2.21 (Sobolevscher Einbettungssatz)**

1. Sei  $p < d$ . Dann gilt  $W^{k,p}(\Omega) \hookrightarrow W^{k-1,q}(\Omega)$  für alle  $q \in \left[1, \frac{pd}{d-p}\right]$  und  $W^{k,p}(\Omega) \xhookrightarrow{c} W^{k-1,q}(\Omega)$  für alle  $q \in \left[1, \frac{pd}{d-p}\right[$ .



2. Sei  $p = d$ . Dann gilt  $W^{k,p}(\Omega) \xrightarrow{c} W^{k-1,q}(\Omega)$  für alle  $q \in [1, \infty[$ .

3. Sei  $k > d/p$ . Dann gilt  $W^{k,p}(\Omega) \xrightarrow{c} C^\ell(\overline{\Omega})$  für alle  $\ell \in \mathbb{N}$  mit  $0 \leq \ell < k - d/p$ .

**Bemerkung 2.22** Sei  $p = 2$  und  $d \in \{2, 3\}$ . Dann ist  $H^2(\Omega) \hookrightarrow C^0(\overline{\Omega})$  aber  $H^1(\Omega)$  ist nicht in  $C^0(\overline{\Omega})$  enthalten, d.h., Punktauswertungen sind im allgemeinen für Funktionen in  $H^1(\Omega)$  nicht definiert.

**Satz 2.23 (Poincarésche Ungleichung)** Für die Raumdimension gelte  $d \geq 2$ . Dann sind  $|\cdot|_1$  und  $\|\cdot\|_1$  äquivalent auf  $V := \{\varphi \in H^1(\Omega) : \int_\Omega \varphi = 0\}$ .

**Beweis.** Wie im Beweis von Satz 2.17 müssen wir nur zeigen, dass es eine Konstante  $C > 0$  gibt mit

$$\|\varphi\|_1 \leq C |\varphi|_1 \quad \forall \varphi \in V. \quad (2.5)$$

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

$$\|\varphi_n\|_1 = 1 \quad \forall n \in \mathbb{N} \quad (2.6)$$

und

$$\lim_{n \rightarrow \infty} |\varphi_n|_1 = 0. \quad (2.7)$$

Wegen Satz 2.21 und Bemerkung 2.20(3) gibt es eine Teilfolge  $(\varphi_{n_k})_{k \in \mathbb{N}}$  von  $(\varphi_n)_{n \in \mathbb{N}}$  und eine Funktion  $\varphi \in L^2(\Omega)$  mit

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

Wegen (2.7) konvergiert  $(\varphi_{n_k})_{k \in \mathbb{N}}$  sogar in  $H^1(\Omega)$ . Mithin ist  $\varphi \in H^1(\Omega)$  und  $|\varphi|_1 = 0$ . Daher ist  $\varphi$  konstant. Aus  $\int_\Omega \varphi = 0$  folgt daher  $\varphi = 0$  im Widerspruch zu (2.6). ■

**Bemerkung 2.24** Satz 2.23 kann für  $H^1(\Omega)$  nicht gelten, da die rechte Seite von (2.5) für die konstante Funktion  $v = 1$  verschwindet.

### 3 Abstrakte Variationsprobleme

Wir werden die elliptischen Differentialgleichungen -als Ausgangspunkt für deren Diskretisierung- in (fast äquivalente) *Variationsprobleme* umwandeln. Wir beginnen, die notwendigen funktionalanalytischen Hilfsmittel zur Verfügung zu stellen.

**Satz 3.1 (Lax-Milgram)** Seien  $(X, \|\cdot\|_X)$  ein Banach-Raum,  $\ell \in \mathcal{L}(X, \mathbb{R})$  ein stetiges lineares Funktional und  $a \in \mathcal{L}^2(X, \mathbb{R})$  eine stetige Bilinearform. Zusätzlich sei  $a$  symmetrisch, d.h.

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

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

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

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

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

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

$$a(u^*, v) = \ell(v) \quad \forall v \in X. \quad (3.1)$$

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

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

Also ist jeder kritische Punkt von  $J$  eine Lösung von (3.1).

**2. Schritt:** Seien  $u_1, u_2 \in X$  zwei Lösungen von (3.1). Dann folgt

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

und

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

Also besitzt (3.1) höchstens eine Lösung.

**3. Schritt:** Für alle  $u \in X$  gilt

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

Also ist  $J$  nach unten beschränkt. Sei

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

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

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

Dann folgt für  $n, m \in \mathbb{N}$

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

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

**Satz 3.2** Die Voraussetzungen und Bezeichnungen seien wie in Satz 3.1. Setze zur Abkürzung

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

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

$$\|u - u_S\|_X \leq \frac{A}{\alpha} \inf_{v \in S} \|u - v\|_X.$$

Sei zusätzlich  $H$  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

$$a(v, u_\varphi) = (\varphi, v)_H \quad \forall v \in X. \quad (3.2)$$

Dann gilt

$$\|u - u_S\|_H \leq A \|u - u_S\|_X \sup_{\varphi \in H \setminus \{0\}} \inf_{v \in S} \frac{\|u_\varphi - v\|_X}{\|\varphi\|_H}.$$

**Beweis.** Wegen Satz 3.1 besitzt  $J$  ein eindeutiges Minimum  $u_S \in S$ . Dieses ist eindeutig charakterisiert durch

$$a(u_S, v) = \ell(v) \quad \forall v \in S. \quad (3.3)$$

Aus (3.1) und (3.3) folgt

$$a(u - u_S, v) = 0 \quad \forall v \in S. \quad (3.4)$$

Hieraus ergibt sich für jedes  $v \in S$

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

Da  $v \in S$  beliebig war, folgt hieraus die erste Fehlerabschätzung.

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

$$v \rightarrow (\varphi, v)_H \quad \forall v \in X$$

ein stetiges lineare Funktional auf  $X$ . Wegen Satz 3.1 besitzt somit (3.2) eine eindeutige Lösung  $u_\varphi \in X$ . Aus (3.2) und (3.4) folgt für beliebiges  $\varphi \in H$  und beliebiges  $v \in S$

$$(u - u_S, \varphi)_H = a(u - u_S, u_\varphi) = a(u - u_S, u_\varphi - v) \leq A \|u - u_S\|_X \|u_\varphi - v\|_X.$$

Da

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

ist, folgt hieraus die zweite Fehlerabschätzung. ■

**Bemerkung 3.3** *Der erste Teil von Satz 3.2 ist bekannt unter dem Namen „Céa’s Lemma“, der zweite Teil unter dem Namen „Dualitätsargument von Aubin-Nitsche“.*

Die Voraussetzungen der Sätze 3.1 und 3.2 lassen sich im wesentlichen nur für positiv definite Bilinearformen (Skalarprodukte) anwenden und nicht für unsymmetrische Probleme. Daher schwächen wir sie in den folgenden beiden Sätzen entsprechend ab.

**Satz 3.4** *Seien  $(X, \|\cdot\|_X)$  und  $(Y, \|\cdot\|_Y)$  Banach-Räume und  $X \xrightarrow{c} Y$  und  $a_0, a_1 \in \mathcal{L}^2(X, \mathbb{R})$  zwei stetige Bilinearformen. Die Bilinearform  $a_0$  sei symmetrisch und koerziv. Für die Bilinearform  $a_1$  gebe es eine Konstante  $\bar{A} \in \mathbb{R}_{>0}$  mit*

$$a_1(u, v) \leq \bar{A} \|u\|_X \|v\|_Y \quad \forall u, v \in X. \quad (3.5)$$

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

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

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

$$a(u, u) > 0. \quad (3.6)$$

Dann besitzen die Probleme

$$a(u, v) = \ell(v) \quad \forall v \in X \quad (3.7)$$

und

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

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

**Beweis.** Wir beweisen die Behauptung nur für Problem (3.7). Der Beweis für das andere Problem ist völlig analog. Sei  $\ell \in \mathcal{L}(X, \mathbb{R})$  beliebig. Wegen Satz 3.1 gibt es genau ein  $u_\ell \in X$  mit

$$a_0(u_\ell, v) = \ell(v) \quad \forall v \in X.$$

Dann ist (3.7) äquivalent zu

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

Wiederum wegen Satz 3.1 gibt es zu jedem  $w \in X$  ein eindeutiges  $u_w \in X$  mit

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

Die Zuordnung  $w \rightarrow u_w$  definiert eine lineare Abbildung  $K \in \mathcal{L}(X, X)$ , und (3.7) ist damit äquivalent zu

$$(I + K)u = u_\ell. \quad (3.8)$$

Wegen  $X \xrightarrow{c} Y$  und (3.5) ist  $K$  kompakt (vgl. Übungsaufgabe 3.5). Daher erfüllt  $I + K$  die Fredholmsche Alternative: Entweder besitzt (3.8) für jede rechte Seite eine eindeutige Lösung oder das zugehörige homogene Problem besitzt eine nichttriviale Lösung  $u \neq 0$ . Wegen (3.6) besitzt (3.7) mit  $\ell = 0$  und damit (3.8) mit  $u_\ell = 0$  aber nur die triviale Lösung. ■

**Übungsaufgabe 3.5** Zeigen Sie, dass der Operator  $K : X \rightarrow X$  in (3.8) kompakt ist.

**Satz 3.6** Die Bezeichnungen und Voraussetzungen seien wie in Satz 3.4. Zusätzlich sei  $S \subset X$  ein endlichdimensionaler Unterraum. Dann besitzt das Problem

$$a(u_S, v) = \ell(v) \quad \forall v \in S \quad (3.9)$$

für jedes  $\ell \in \mathcal{L}(X, \mathbb{R})$  eine eindeutige Lösung  $u_S \in S$ . Die Bilinearform  $a$  sei zusätzlich koerziv, d.h., es gibt ein  $\beta > 0$  mit  $a(u, u) \geq \beta \|u\|_X^2$  für alle  $u \in X$ . Dann gilt für die eindeutigen Lösungen  $u$  und  $u_S$  der Probleme (3.7) und (3.9) die Fehlerabschätzung

$$\|u - u_S\|_X \leq \frac{A}{\beta} \inf_{v \in S} \|u - v\|_X. \quad (3.10)$$

Dabei ist  $A := \|a\|_{\mathcal{L}^2(X, \mathbb{R})}$ . Seien schliesslich  $H$ ,  $\varphi$  und  $u_\varphi$  wie in Satz 3.2. Dann gilt die Fehlerabschätzung

$$\|u - u_S\|_H \leq A \|u - u_S\|_X \sup_{\varphi \in H \setminus \{0\}} \inf_{v \in S \setminus \{0\}} \frac{\|u_\varphi - v\|_X}{\|\varphi\|_H}. \quad (3.11)$$

**Beweis.** Die eindeutige Lösbarkeit von (3.9) folgt aus Satz 3.4. Die Fehlerabschätzung (3.10) folgt wie im Beweis von Satz 3.2. Man beachte, dass wir dort nur die zu (3.7) und (3.9) analogen Eigenschaften (3.1) und (3.3) ausgenützt haben. Wegen  $X \hookrightarrow H$  definiert jedes  $\varphi \in H$  durch  $v \rightarrow (\varphi, v)_H$  ein stetiges lineares Funktional auf  $X$ . Daher folgt die Existenz und Eindeutigkeit von  $u_\varphi$  aus Satz 3.4. Die Fehlerabschätzung (3.11) folgt dann wie im Beweis von Satz 3.2. ■

## 4 Schwache Lösungen

Wir setzen im Folgenden generell voraus, dass  $\Omega \subset \mathbb{R}^d$ ,  $d \geq 2$ , eine offene beschränkte Menge mit Lipschitz-Rand  $\Gamma := \partial\Omega$  und äusserem Einheitsnormalenfeld  $n$  ist. Wir betrachten skalare, elliptische Differentialgleichungen zweiter Ordnung. Ihre allgemeine Form lautet (vgl. (1.9))

$$-\sum_{1 \leq i, j \leq d} \frac{\partial}{\partial x_i} \left( A_{i,j} \frac{\partial u}{\partial x_j} \right) + \sum_{i=1}^d b_i \frac{\partial u}{\partial x_i} + cu = f \quad \text{in } \Omega. \quad (4.1)$$

Dabei ist  $f \in L^2(\Omega)$ ,  $c \in C(\Omega, \mathbb{R}_{\geq 0})$ ,  $\mathbf{b} = (b_1, b_2, \dots, b_d)^\top \in C^1(\Omega, \mathbb{R}^d)$  und  $\mathbf{A} = (A_{i,j})_{i,j=1}^d \in C^1(\Omega, \mathbb{R}^{d \times d})$  mit  $\mathbf{A}(\mathbf{x}) = \mathbf{A}^\top(\mathbf{x})$  für alle  $\mathbf{x} \in \Omega$  und

$$0 < \lambda := \inf_{\mathbf{x} \in \Omega} \inf_{\mathbf{z} \in \mathbb{R}^d \setminus \{0\}} \frac{\mathbf{z}^\top \mathbf{A}(\mathbf{x}) \mathbf{z}}{\mathbf{z}^\top \mathbf{z}} \leq \sup_{\mathbf{x} \in \Omega} \sup_{\mathbf{z} \in \mathbb{R}^d \setminus \{0\}} \frac{\mathbf{z}^\top \mathbf{A}(\mathbf{x}) \mathbf{z}}{\mathbf{z}^\top \mathbf{z}} =: \Lambda < \infty. \quad (4.2)$$

Die Differentialgleichung muss mit Randbedingungen versehen werden. Wir betrachten drei Typen von Randbedingungen

- (homogene) **Dirichlet-Randbedingungen:**  $u = 0$  auf  $\Gamma$ ,
- (inhomogene) **Neumann-Randbedingungen:**  $\langle \mathbf{A}n, \text{grad } u \rangle = g$  auf  $\Gamma$ ,
- **gemischte Dirichlet-Neumann-Randbedingungen:**  $u = 0$  auf  $\Gamma_D$  und  $\langle \mathbf{A}n, \text{grad } u \rangle = g$  auf  $\Gamma_N$ .

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

Sei nun  $u \in C^2(\Omega)$  eine Lösung von (4.1) mit homogenen Dirichlet-Randbedingungen und  $v \in C_0^\infty(\Omega)$ . Multiplikation von (4.1) mit  $v$ , Integration über  $\Omega$  und Anwenden des Gaußschen Integralsatzes liefert

$$\begin{aligned} \int_{\Omega} f v &= - \int_{\Omega} v \operatorname{div}(\mathbf{A} \operatorname{grad} u) + \int_{\Omega} \langle \mathbf{b}, \operatorname{grad} u \rangle v + \int_{\Omega} c u v \\ &= \int_{\Omega} (\langle \operatorname{grad} v, \mathbf{A} \operatorname{grad} u \rangle + \langle \mathbf{b}, \operatorname{grad} u \rangle v + c u v). \end{aligned} \quad (4.3)$$

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

$$\int_{\Omega} (\langle \operatorname{grad} v, \mathbf{A} \operatorname{grad} u \rangle + \langle \mathbf{b}, \operatorname{grad} u \rangle v + c u v) = \int_{\Omega} f v \quad \forall v \in H_0^1(\Omega) \quad (4.4)$$

erfüllt. Umgekehrt folgt aus (4.3), dass eine Lösung von (4.4) die Differentialgleichung (4.1) erfüllt, sofern sie glatt genug, d.h. in  $C^2(\Omega)$  ist. In diesem Sinne ist (4.4) zur Differentialgleichung (4.1) mit homogenen Dirichlet-Randbedingungen äquivalent.

Betrachten wir in obigem Argument Funktionen  $v \in C^\infty(\overline{\Omega})$ , so treten in (4.3) zusätzliche Randterme  $\int_{\Gamma} \langle \mathbf{A}n, \operatorname{grad} u \rangle v$  auf. Erfüllt  $u$  Neumann-Randbedingungen, so lassen sich diese substituieren:

$$\int_{\Gamma} \langle \mathbf{A}n, \operatorname{grad} u \rangle v = \int_{\Gamma} g v.$$

Wir werden daher in diesem Fall die Gleichung (4.4) durch den zusätzlichen Term  $\int_{\Gamma} gv$  auf der rechten Seite modifizieren. Diese Überlegungen führen auf folgende Definition.

**Definition 4.1**

1.  $u \in H_0^1(\Omega)$  heisst **schwache Lösung** der Differentialgleichung (4.1) mit homogenen Dirichlet-Randbedingungen, wenn gilt

$$\int_{\Omega} \langle \mathbf{A} \operatorname{grad} u, \operatorname{grad} v \rangle + \langle \mathbf{b}, \operatorname{grad} u \rangle v + cuv = \int_{\Omega} fv \quad \forall v \in H_0^1(\Omega).$$

2.  $u \in H_D^1(\Omega) := \{\varphi \in H^1(\Omega) : \varphi|_{\Gamma_D} = 0\}$  heisst **schwache Lösung** der Differentialgleichungen (4.1) mit gemischten Randbedingungen, wenn gilt

$$\int_{\Omega} \langle \mathbf{A} \operatorname{grad} u, \operatorname{grad} v \rangle + \langle \mathbf{b}, \operatorname{grad} u \rangle v + cuv = \int_{\Omega} fv + \int_{\Gamma_N} gv \quad \forall v \in H_D^1(\Omega).$$

3.  $u \in H^1(\Omega)$  heisst **schwache Lösung** der Differentialgleichung (4.1) mit Neumann-Randbedingungen, wenn gilt

$$\int_{\Omega} \langle \mathbf{A} \operatorname{grad} u, \operatorname{grad} v \rangle + \langle \mathbf{b}, \operatorname{grad} u \rangle v + cuv = \int_{\Omega} fv + \int_{\Gamma} gv \quad \forall v \in H^1(\Omega).$$

**Bemerkung 4.2**

1. Jede klassische Lösung von (4.1) ist auch eine schwache Lösung. Jede schwache Lösung, die zweimal stetig differenzierbar ist, ist eine klassische Lösung von (4.1).
2. Für schwache Lösungen benötigen wir für die Koeffizienten nur die Regularitätsvoraussetzungen  $c \in L^\infty(\Omega)$ ,  $\mathbf{b} \in L^\infty(\Omega, \mathbb{R}^d)$ ,  $\mathbf{A} \in L^\infty(\Omega, \mathbb{R}^{d \times d})$ .
3. Bei inhomogenen Dirichlet-Randbedingungen  $u = u_D$  auf  $\Gamma$  bzw.  $\Gamma_D$  muss in Definition 4.1 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 4.3 (Existenz- und Eindeutigkeitsatz für schwache Lösungen)**

1. Ist  $-\frac{1}{2} \operatorname{div} \mathbf{b} + c \geq 0$ , so besitzt die Differentialgleichung (4.1) mit homogenen Dirichlet-Randbedingungen eine eindeutige schwache Lösung.
2. Ist  $-\frac{1}{2} \operatorname{div} \mathbf{b} + c \geq 0$  und  $\langle \mathbf{b}, \mathbf{n} \rangle \geq 0$  auf  $\Gamma_N$ , so besitzt die Differentialgleichung (4.1) mit gemischten Randbedingungen eine eindeutige schwache Lösung.
3. Ist  $c \geq c_0 > 0$ ,  $-\frac{1}{2} \operatorname{div} \mathbf{b} + c \geq 0$  und  $\langle \mathbf{b}, \mathbf{n} \rangle \geq 0$  auf  $\Gamma$ , so besitzt die Differentialgleichung (4.1) mit Neumann-Randbedingungen eine eindeutige schwache Lösung.
4. Ist  $c = 0$ ,  $-\frac{1}{2} \operatorname{div} \mathbf{b} \geq 0$  und  $\langle \mathbf{b}, \mathbf{n} \rangle = 0$  auf  $\Gamma$  sowie  $\int_{\Omega} f + \int_{\Gamma} g = 0$ , so besitzt die Differentialgleichung (4.1) mit Neumann-Randbedingungen eine eindeutige schwache Lösung  $u$  mit  $\int_{\Omega} u = 0$ .

**Beweis.** Wir wenden Satz 3.4 an.

**zu 1:** Setze  $X := H_0^1(\Omega)$ ,  $Y := L^2(\Omega)$  und

$$\ell(v) := \int_{\Omega} f v, \quad a_0(u, v) = \int_{\Omega} \langle \mathbf{A} \operatorname{grad} u, \operatorname{grad} v \rangle + c u v, \quad a_1(u, v) := \int_{\Gamma} v \langle \mathbf{b}, \operatorname{grad} u \rangle.$$

Aus der Cauchy-Schwarzschen-Ungleichung folgt

$$|\ell(v)| \leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)} \|v\|_{H^1(\Omega)}$$

und somit ist  $\ell$  ein stetige lineares Funktional auf  $H^1(\Omega)$ . Die Bilinearformen ist stetig auf  $H^1(\Omega) \times H^1(\Omega)$

$$\begin{aligned} |a_0(u, v)| &\leq \|\mathbf{A}\|_{\infty} |u|_{H^1(\Omega)} |v|_{H^1(\Omega)} + \|c\|_{\infty} \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\ &\leq \max\{\|\mathbf{A}\|_{\infty}, \|c\|_{\infty}\} \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}. \end{aligned}$$

und die Bilinearform  $a_1$  ist stetig auf  $H^1(\Omega) \times L^2(\Omega)$ :

$$|a_1(u, v)| \leq \|\mathbf{b}\|_{\infty} |u|_{H^1(\Omega)} \|v\|_{L^2(\Omega)} \leq \|\mathbf{b}\|_{\infty} \|u\|_{H^1(\Omega)} \|v\|_{L^2(\Omega)}.$$

Die Bilinearform  $a_0$  ist wegen  $c \geq 0$  koerziv auf  $H_0^1(\Omega)$ :

$$a_0(u, u) = \int_{\Omega} \langle \mathbf{A} \operatorname{grad} u, \operatorname{grad} u \rangle + c u^2 \stackrel{\text{vgl. (4.2)}}{\geq} \lambda |u|_{H^1(\Omega)}^2 \geq \lambda c \|u\|_{H^1(\Omega)}^2$$

mit der Konstante  $c$  aus der Friedrichsschen Ungleichung.

Aus dem Gaussischen Integralsatz und (4.2) folgt schliesslich

$$\begin{aligned} a(u, u) &= \int_{\Omega} \langle \mathbf{A} \operatorname{grad} u, \operatorname{grad} u \rangle + u \langle \mathbf{b}, \operatorname{grad} u \rangle + c u^2 \\ &\geq \lambda |u|_{H^1(\Omega)}^2 + \int_{\Omega} \frac{1}{2} \langle \mathbf{b}, \operatorname{grad} (u^2) \rangle + c u^2 \\ &= \lambda |u|_{H^1(\Omega)}^2 + \int_{\Omega} \left( -\frac{1}{2} \operatorname{div} \mathbf{b} + c \right) u^2. \end{aligned}$$

Die Bilinearform  $a$  ist wegen  $-\frac{1}{2} \operatorname{div} \mathbf{b} + c \geq 0$  koerziv auf  $H_0^1(\Omega)$ .

Die Behauptung folgt daher aus Satz 3.4.

**zu 2:** In diesem Fall setzen wir  $X = H_D^1(\Omega)$  und  $\ell(v) = \int_{\Omega} f v + \int_{\Gamma_N} g v$  und definieren die anderen Grössen wie im ersten Fall. Aus der Cauchy-Schwarzschen-Ungleichung und dem Spursatz 2.15 folgt die Stetigkeit von  $\ell$

$$\begin{aligned} |\ell(v)| &\leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma_N)} \|v\|_{L^2(\Gamma_N)} \\ &\leq \left( \|f\|_{L^2(\Omega)} + c \|g\|_{L^2(\Gamma_N)} \right) \|v\|_{H^1(\Omega)}. \end{aligned}$$

Die Koerzivitat von  $a_0$  bleibt wegen Bemerkung 2.18 erhalten. Bei der Anwendung des Gaussischen Integralsatzes in der Abschatzung von  $a(u, u)$  tritt der zusatzliche Randterm  $\int_{\Gamma_N} \langle \mathbf{n}, \mathbf{b} \rangle u^2$  auf. Wegen  $\langle \mathbf{n}, \mathbf{b} \rangle \geq 0$  auf  $\Gamma_N$  ist er nicht negativ, und die Koerzivitat von  $a$  bleibt erhalten.

**zu 3:** In diesem Fall gilt  $X := H^1(\Omega)$ . Die anderen Grössen sind wie im Fall (2) mit  $\Gamma$  statt  $\Gamma_N$ . Wegen  $c \geq c_0 > 0$  erhalten wir

$$a_0(u, u) \geq \lambda |u|_{H^1(\Omega)}^2 + c_0 \|u\|_{L^2(\Omega)}^2 \geq \min\{\lambda, c_0\} \|u\|_{H^1(\Omega)}^2$$

und somit die Koerzivitat von  $a_0$ . Die anderen Abschatzungen andern sich nicht, insbesondere gilt  $a(u, u) \geq 0$  fur alle  $u \in X$ .

**zu 4:** Alle Grossen sind wie in Fall 3. Wegen  $c = 0$  gilt

$$a_0(u, u) \geq \lambda |u|_{H^1(\Omega)}^2.$$

Hieraus und aus der Poincareschen Ungleichung (Satz 2.23) folgt die Koerzivitat von  $a_0$  auf  $V := \{u \in H^1(\Omega) : \int_{\Omega} \varphi = 0\}$ . Die Abschatzung  $a(u, u) \geq 0$  fur alle  $u \in X$  bleibt gultig, ebenso die Stetigkeit von  $a_0$  und  $a_1$ . Lediglich bei der Stetigkeit von  $\ell$  ist Sorgfalt geboten. Da  $V \cong H^1(\Omega)/\mathbb{R}$  ist, muss fur die Stetigkeit von  $\ell$  die Inklusion  $\mathbb{R} \subset \text{Kern}(\ell)$  gelten. Dies ist aber wegen  $\int_{\Omega} f + \int_{\Gamma} g = 0$  der Fall. ■

Das folgende Beispiel zeigt, dass wir eine Regularitatsaussage der Form  $u \in H^2(\Omega)$  fur schwache Losungen nur unter zusatzlichen Annahmen an den Rand  $\Gamma$  erwarten konnen.

**Beispiel 4.4** Sei  $0 < \alpha < 2\pi$  und  $\Omega_{\alpha}$  das Kreissegment

$$\Omega_{\alpha} := \left\{ \mathbf{x} = r \begin{pmatrix} \cos \varphi \\ \sin \varphi \end{pmatrix} \in \mathbb{R}^2 : 0 < r < 1, 0 < \varphi < \alpha \right\}.$$

Definiere die Funktion  $v \in \Omega_{\alpha} \rightarrow \mathbb{R}$  durch

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

Dann gilt fur jedes  $\mathbf{x} \in \Omega_{\alpha}$

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

Sei  $w \in C_0^{\infty}(\mathbb{R}^2, \mathbb{R})$  mit  $\text{Tr } w \subset B(0, \frac{2}{3})$  und  $w = 1$  auf  $\overline{B(0, \frac{1}{3})}$ .

Definiere

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

Dann gilt

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

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

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

Wie man leicht nachrechnet gilt

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

und

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

Wir konnen also bei gegebenem  $\alpha$  i.a. **keine** Abschatzung der Form

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

und **keine** Abschatzung der Form

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

erwarten, wie sie fur gewohnliche Differentialgleichungen gelten wurde.



**Satz 4.5 (Regularitätssatz)** Sei  $\Gamma$  ein  $C^1$ -Mannigfaltigkeit oder  $\Omega$  konvex und  $f \in L^2(\Omega)$ . Bei gemischten Neumann-Randbedingungen gebe es eine Funktion  $u_g \in H^2(\Omega)$  mit  $g = u_g|_{\Gamma_N}$ . Dann gilt für die schwache Lösung  $u$  der elliptischen Differentialgleichung mit homogenen oder gemischten oder Neumann-Randbedingungen die Regularitätsaussage  $u \in H^2(\Omega)$  und die a-priori-Abschätzung

$$\|u\|_{H^2(\Omega)} \leq c \left\{ \|f\|_{L^2(\Omega)} + \|u_g\|_{H^2(\Omega)} \right\}.$$

Die Konstante  $c$  hängt nur von  $\Omega$  und den Koeffizienten  $c, \mathbf{b}, \mathbf{A}$  in der Differentialgleichung ab.

## 5 Eindimensionale lineare finite Elemente

Elliptische Differentialgleichungen werden am günstigsten mit *finiten Elementen* diskretisiert. Zur Einführung geben wir das Verfahren zunächst im eindimensionalen Fall an.

Sei  $\Omega = (0, 1)$  und

$$X = H_0^1(0, 1), \quad \ell(v) = \int_0^1 f v, \quad a(u, v) = \int_0^1 \{Au'v' + cuv\}.$$

**Kontinuierliches Problem:**

Sei  $f \in L^2(\Omega)$  gegeben. Finde  $u \in H_0^1(\Omega)$  mit

$$a(u, v) = \ell(v) \quad \forall v \in H_0^1(\Omega).$$

Das Ziel dieses Kapitels ist es, dieses Problem mit finite Elementen zu diskretisieren. Die Diskretisierung wird im Rahmen von Satz 3.2 und Satz 3.6 geschehen. Daher müssen wir im ersten Schritt endlich-dimensionale Teilräume des Sobolev-Raums  $H_0^1(\Omega)$  definieren. Als Teilräume verwenden wir die Spline-Räume, welche bereits in der Vorlesung Numerik I zur Approximation von Funktionen verwendet wurden.

Sei  $\mathcal{G} := \{\tau_i : 0 \leq i \leq n\}$  mit  $\tau_i := [x_i, x_{i+1}]$  und  $0 = x_0 < x_1 < \dots < x_{n+1} = 1$  eine Unterteilung von  $[0, 1]$  in  $n + 1$  Teilintervalle. Setze  $h_i := x_{i+1} - x_i$  für  $0 \leq i \leq n$ . Die maximale Schrittweite ist durch

$$h := \max_{0 \leq i \leq n} h_i$$

gegeben. Wir führen den Spline-Raum  $S_{\mathcal{G}}^{1,0}$  ein:

$$S := S_{\mathcal{G}}^{1,0} := \{\varphi \in C^0([0, 1]) \mid \forall \tau \in \mathcal{G} : \varphi|_{\tau} \in \mathbb{P}_1\} \cap H_0^1(\Omega).$$

Eine Basis von  $S$  ist durch die Funktionen  $b_i, 1 \leq i \leq n_{\mathcal{G}}$ , gegeben:

$$b_i(x) := \begin{cases} \frac{x-x_{i-1}}{h_{i-1}} & \text{in } \tau_{i-1}, \\ \frac{x_{i+1}-x}{h_i} & \text{in } \tau_i, \\ 0 & \text{sonst.} \end{cases} \quad (5.1)$$

Das diskrete Problem

$$a(u_S, v) = \ell(v) \quad \forall v \in S \quad (5.2)$$

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

Aus Satz 3.2 folgt, dass der Fehler  $u - u_S$  zwischen der schwachen Lösung  $u$  und der Finite-Elemente-Lösung  $u_S$  (bis auf einen Faktor) durch den Approximationsfehler

$$\inf_{v \in S} \|u - v\|_{H^1(\Omega)}$$

kontrolliert wird. Im nächsten Schritt werden wir diesen Term abschätzen. Wir betrachten ein Intervall  $\tau = [A, B] \in \mathcal{G}$  der Länge  $h_\tau = B - A$  und setzen  $u \in H^2(\Omega)$  voraus. Analog wie für die Friedrichssche Ungleichung genügt es, die Fehlerabschätzung für Funktionen aus  $C^\infty(\overline{\Omega})$  zu beweisen.

Sei  $\tilde{u}$  die lineare Interpolation von  $u$  auf  $\tau$ . Dann gilt die Fehlerdarstellung

$$u(x) - \tilde{u}(x) = (x - A)(x - B) \frac{u''(\xi)}{2}$$

für  $x \in [A, B]$  und einer Zwischenstelle  $\xi = \xi(x) \in \tau$ .

Die Ableitung  $u'$  wird durch die (konstante) Funktion  $\tilde{u}'$  approximiert, welche nach dem Satz von Rolle mit der Ableitung  $u'$  an einer Zwischenstelle  $\xi \in \tau$  übereinstimmt. Daraus folgt

$$e'(x) = u'(x) - \tilde{u}'(x) = \int_\xi^x u''(t) dt \quad \forall x \in \tau.$$

Quadrieren und die Hölder-Ungleichung liefert:

$$|e'(x)|^2 \leq \left( \int_A^B |u''(t)| dt \right)^2 \leq (B - A) \int_A^B |u''(t)|^2 dt = h_\tau \|u''\|_{L^2(\tau)}^2.$$

Integration über  $\tau$  ergibt

$$\|u' - \tilde{u}'\|_{L^2(\tau)}^2 \leq h_\tau^2 \|u''\|_{L^2(\tau)}^2.$$

Durch Summation über alle Intervalle erhalten wir

$$\|u - \tilde{u}\|_{H^1(\Omega)}^2 \leq \sum_{\tau \in \mathcal{G}} h_\tau^2 \|u''\|_{L^2(\tau)}^2 = h^2 \|u''\|_{H^2(\Omega)}^2$$

und haben insgesamt die Approximationseigenschaft:

$$\inf_{v \in S} \|u - v\|_{H^1(\Omega)} \leq h \|u''\|_{H^2(\Omega)}$$

gezeigt. Wegen  $u, v \in H_0^1(\Omega)$  lässt sich die Friedrichssche Ungleichung anwenden:

$$\inf_{v \in S} \|u - v\|_{H^1(\Omega)} \leq \sqrt{2} \inf_{v \in S} \|u - v\|_{H^1(\Omega)} \leq \sqrt{2} h \|u''\|_{H^2(\Omega)}. \quad (5.3)$$

Der Satz von Lax-Milgram impliziert, dass das Infimum auf der linken Seite in (5.3) angenommen wird und eindeutig ist. Wir verwenden hierfür die Bezeichnung  $v_{\text{opt}} \in S$ :

$$\|u - v_{\text{opt}}\|_{H^1(\Omega)} = \inf_{v \in S} \|u - v\|_{H^1(\Omega)}.$$

Mit Satz 3.2 folgt daraus die Fehlerabschätzung

$$\|u - v_{\text{opt}}\|_{L^2(\Omega)} \leq \|u - v_{\text{opt}}\|_{H^1(\Omega)} \sup_{\varphi \in L^2(\Omega) \setminus \{0\}} \inf_{v \in S} \frac{\|u_\varphi - v\|_{H^1(\Omega)}}{\|\varphi\|_{L^2(\Omega)}}.$$

Die Funktion  $u_\varphi \in H^2(\Omega)$  ist dabei die Lösung von

$$\begin{aligned} -u_\varphi'' + u_\varphi &= \varphi && \text{in } \Omega, \\ u_\varphi(0) &= u_\varphi(1) = 0 \end{aligned} \tag{5.4}$$

Nochmaliges Anwenden des vorigen Approximationsresultats liefert:

$$\sup_{\varphi \in L^2(\Omega) \setminus \{0\}} \inf_{v \in S} \frac{\|u_\varphi - v\|_{H^1(\Omega)}}{\|\varphi\|_{L^2(\Omega)}} \leq \sqrt{2}h \sup_{\varphi \in L^2(\Omega) \setminus \{0\}} \frac{|u_\varphi|_{H^2(\Omega)}}{\|\varphi\|_{L^2(\Omega)}}.$$

In Übungsaufgabe 5.1 ist zu zeigen, dass eine Konstante  $C$  existiert mit

$$|u_\varphi|_{H^2(\Omega)} \leq C \|\varphi\|_{L^2(\Omega)} \quad \forall \varphi \in L^2(\Omega). \tag{5.5}$$

Insgesamt haben wir

$$\|u - v_{\text{opt}}\|_{L^2(\Omega)} \leq Ch^2 \|u''\|_{L^2(\Omega)}$$

gezeigt. Die Quasi-Optimalität des Céa's Lemmas liefert daher die Fehlerabschätzung für die Lösung von (5.2):

$$\|u - u_S\|_{L^2(\Omega)} + h |u - u_S|_{H^1(\Omega)} \leq ch^2 |u|_{H^2(\Omega)}.$$

**Übungsaufgabe 5.1** Zeigen Sie Abschätzung (5.5) für Problem (5.4).

Um diese Vorgehensweise zu verallgemeinern, fassen wir die wesentlichen Diskretisierungsschritte im folgenden nochmals zusammen.

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

## 6 Bilineare Rechteckselemente

Im folgenden bezeichnet  $\Omega \subset \mathbb{R}^2$  ein zusammenhängendes Polygon mit achsenparallelen Kanten. Wir betrachten zunächst die allgemeine elliptische Differentialgleichung mit

$$\mathbf{b} \equiv \mathbf{0}, \quad \inf_{\mathbf{x} \in \Omega} c(\mathbf{x}) \geq 0 \quad \text{und homogenen Dirichlet-Randbedingungen.} \tag{6.1}$$

Wir setzen im Rahmen der abstrakten Theorie von Bilinearformen

$$X = H_0^1(\Omega), \quad \ell(v) = \int_{\Omega} f v \quad a(u, v) = \int_{\Omega} \langle \nabla u, A \nabla v \rangle + c u v.$$

Sei  $\mathcal{G} := \{\tau_i : 1 \leq i \leq N\}$  eine Zerlegung von  $\Omega$  in achsenparallele Rechtecke, so dass je zwei Rechtecke entweder disjunktes Inneres besitzen oder einen Eckpunkt oder eine Kante

gemeinsam haben (**Zulässigkeitsbedingung**). Für  $\tau \in \mathcal{G}$  bezeichnet  $h_\tau$  die längste Kante. Setze

$$h := \max_{\tau \in \mathcal{G}} h_\tau.$$

Die Eckpunkte der Rechtecke bezeichnen wir mit  $\Theta_{\mathcal{G}}$  und  $\overset{\circ}{\Theta}_{\mathcal{G}}$  die Teilmenge von  $\Theta_{\mathcal{G}}$ , welche keine Punkte auf dem Rand  $\partial\Omega$  enthält. Wir numerieren zunächst die Punkte in  $\overset{\circ}{\Theta}_{\mathcal{G}}$  von 1 bis  $n_{\mathcal{G}}$  und anschliessend die Randpunkte  $\Theta_{\mathcal{G}} \setminus \overset{\circ}{\Theta}_{\mathcal{G}}$  von  $n_{\mathcal{G}} + 1$  bis  $\overline{n_{\mathcal{G}}}$ .

Sei  $Q_1 = \text{span}\{1, x_1, x_2, x_1x_2\}$  der Raum aller Polynome in zwei Variablen, die in jeder Variablen höchstens den Grad 1 haben. Als endlichdimensionalen Teilraum von  $H_0^1(\Omega)$  setzen wir

$$S := S_{\mathcal{G}} := \{\varphi \in C(\overline{\Omega}) \mid \forall \tau \in \mathcal{G} : \varphi|_{\tau} \in Q_1\} \cap H_0^1(\Omega).$$

Der Schnitt mit  $H_0^1(\Omega)$  ist äquivalent zur Zusatzbedingung  $\varphi|_{\partial\Omega} = 0$ . Man überlegt sich leicht, dass für jedes achsenparallele Rechteck  $\tau$  und jedes Polynom  $p \in Q_1$  das Polynom  $p$  eindeutig bestimmt wird durch seine Werte in den vier Eckpunkten von  $\tau$ . Da die Einschränkung von  $p$  auf eine beliebige Kante von  $\tau$  ein lineares Polynom einer Veränderlichen ist, folgt ausserdem, dass für je zwei achsenparallele Rechtecke  $\tau_1, \tau_2$ , die eine Kante  $E$  gemeinsam haben, und je zwei Polynome  $p_1, p_2 \in Q_1$  die Funktion

$$\varphi := \begin{cases} p_1 & \text{auf } \tau_1 \\ p_2 & \text{auf } \tau_2 \end{cases}$$

genau dann stetig ist, wenn  $p_1$  und  $p_2$  in den Eckpunkten von  $E$  übereinstimmen. Daher sind die Funktionen in  $S$  eindeutig bestimmt durch ihre Werte in den Punkten von  $\overset{\circ}{\Theta}_{\mathcal{G}}$ . Insbesondere gilt  $\dim S = n_{\mathcal{G}}$ .

Wir wollen nun eine Basis für  $S$  konstruieren. Sei dazu  $\mathbf{x}_i = (x_{i,1}, x_{i,2}) \in \overset{\circ}{\Theta}_{\mathcal{G}}$  beliebig. Seien  $h_W, h_E, h_N, h_S$  die Längen der Kanten, die von  $\mathbf{x}_i$  in die vier Himmelsrichtungen ausgehen. Damit lassen sich zwei Funktionen einer Veränderlichen analog wie in (5.1) durch

$$\varphi_{i,1}(t_1) := \begin{cases} 0 & \text{falls } t_1 \leq x_{i,1} - h_W \text{ oder } t_1 \geq x_{i,1} + h_E, \\ \frac{t_1 - x_{i,1} + h_W}{h_W} & \text{falls } x_{i,1} - h_W \leq t_1 \leq x_{i,1}, \\ \frac{x_{i,1} + h_E - t_1}{h_E} & \text{falls } x_{i,1} \leq t_1 \leq x_{i,1} + h_E, \end{cases}$$

und

$$\varphi_{i,2}(t_2) := \begin{cases} 0 & \text{falls } t_2 \leq x_{i,2} - h_S \text{ oder } t_2 \geq x_{i,2} + h_N, \\ \frac{t_2 - x_{i,2} + h_S}{h_S} & \text{falls } x_{i,2} - h_S \leq t_2 \leq x_{i,2}, \\ \frac{x_{i,2} + h_N - t_2}{h_N} & \text{falls } x_{i,2} \leq t_2 \leq x_{i,2} + h_N \end{cases}$$

definieren. Offensichtlich sind die Funktionen  $\varphi_{i,1}$  und  $\varphi_{i,2}$  stetig und stückweise linear. Daher ist eine Basis auf  $S$  durch

$$b_i(t_1, t_2) := \varphi_{i,1}(t_1) \varphi_{i,2}(t_2) \in S$$

gegeben und somit  $S = \text{span}\{b_i : 1 \leq i \leq n_{\mathcal{G}}\}$ .

Man beachte, dass der Träger von  $b_i$  genau aus den vier Rechtecken besteht, die den Knoten  $\mathbf{x}_i$  gemeinsam haben.

Jedes  $u \in S$  lässt sich eindeutig darstellen als

$$u = \sum_{i=1}^{n_{\mathcal{G}}} \mathbf{u}_i b_i, \tag{6.2}$$

und es gilt

$$\mathbf{u}_i = u(\mathbf{x}_i) \quad 1 \leq i \leq n_G.$$

Daher ist das das diskrete Problem: Finde  $u_S \in S$  mit

$$a(u_S, v) = \ell(v) \quad \forall v \in S \quad (6.3)$$

äquivalent zu einem linearen Gleichungssystem mit  $n_G$  Gleichungen und Unbekannten. Die Unbekannten sind genau die Werte von  $u_S$  in den Gitterpunkten. Die Matrix des linearen Gleichungssystems (LGS) heisst **Systemmatrix**. Wir setzen

$$\mathbf{A} := (a(b_i, b_j))_{i,j=1}^{n_G} \quad \mathbf{r} := (\ell(b_j))_{j=1}^{n_G}.$$

Damit ist der Koeffizientenvektor  $\mathbf{u} := (\mathbf{u}_i)_{i=1}^{n_G}$  in der Basisdarstellung (6.2) die Lösung von

$$\mathbf{A}\mathbf{u} = \mathbf{r}. \quad (6.4)$$

- Wegen der Symmetrie von  $a$  ist die Systemmatrix symmetrisch.
- Wegen der Koerzivität von  $a$  ist die Systemmatrix positiv definit.
- Da die Träger der Basisfunktionen  $b_i$  aus jeweils vier Rechtecken bestehen, hat die Systemmatrix höchstens 9 von Null verschiedene Einträge pro Zeile, d.h., sie ist schwachbesetzt.

Da die Unbekannten des zu (6.3) äquivalenten linearen Gleichungssystem (6.4) die Werte von  $u_S$  in den Gitterpunkten sind, lässt sich (6.4) auch Differenzenverfahren interpretieren. Dies gibt dann auch einen Einblick in die Struktur der Systemmatrix. Wir betrachten hierzu den einfachsten Spezialfall der Poisson-Gleichung auf einem äquidistanten Gitter, d.h.,  $c \equiv 0$  und  $\mathbf{A} \equiv \mathbf{I}$  und alle Kanten haben die Länge  $h$ . Für jedes  $p \in Q_1$  sind die partiellen Ableitungen  $\partial p / \partial x_1$  bzw.  $\partial p / \partial x_2$  offensichtlich lineare Funktionen allein der Variablen  $x_2$  bzw.  $x_1$ . Daher ist für alle  $u, v \in S$  der Ausdruck  $\langle \nabla u, \nabla v \rangle$  stückweise (auf jedem Rechteck) von der Form  $\alpha \psi_{i,1}(x_1) + \beta \psi_{i,2}(x_2)$  mit  $\alpha, \beta \in \mathbb{R}$  und  $\psi_{i,1}, \psi_{i,2} \in \mathbb{P}_2$ . Da die eindimensionale Simpson-Regel die Ordnung 3 hat, folgt aus dem Satz von Fubini, dass ein Ausdruck der Form  $\int_\tau \langle \nabla u, \nabla v \rangle$  durch die zweidimensionale Simpsonregel

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

mit

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

exakt integriert wird. Hierbei ist  $\tau$  ein achsenparalleles Rechteck mit Kantenlängen  $h_1$  und  $h_2$  und unterer linken Ecke  $\mathbf{z} = (z_1, z_2)^\top$ .

Seien  $p \in Q_1$  und  $p_{\mu,\nu} := p(z_1 + \mu h_1, z_2 + \nu h_2)$ ,  $\mu, \nu \in \{0, 1\}$ , die Werte von  $p$  in den Eckpunkten von  $\tau$ . Dann sind die Werte von  $\nabla p$  in den Eckpunkten von  $\tau$  gegeben durch

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

Die Werte in den restlichen Knoten obiger Quadraturformel ergeben sich durch lineare Interpolation. Wenden wir diese Überlegungen auf  $\int_{\Omega} \langle \nabla u, \nabla b_i \rangle$  für alle  $1 \leq i \leq n_{\mathcal{G}}$  an, so erhalten wir für eine Unterteilung in Quadrate, d.h.,  $h_1 = h_2 = h$ ,

$$\int_{\Omega} \langle \nabla u^{\top}, \nabla b_i \rangle = \sum_{j=-1}^1 \sum_{k=-1}^1 \beta_{j,k} u(\mathbf{x}_i + (jh, kh)^{\top})$$

mit

$$\beta_{0,0} = 8/3 \quad \text{und} \quad \beta_{j,k} = -1/3 \quad \forall (j, k) \neq (0, 0).$$

Dies liefert die gewünschte Darstellung der linken Seite von (6.3) als Differenzenverfahren.

Sei nun  $u \in H_0^1(\Omega)$  die eindeutige schwache Lösung der allgemeinen elliptischen Differentialgleichung mit (6.1) und  $u_S \in S$  die eindeutige Lösung von (6.3). Gemäss Satz 3.2 kann der Fehler  $|u - u_S|_{H^1(\Omega)}$  durch den Approximationsfehler  $\inf_{v \in S} |u - v|_{H^1(\Omega)}$  abgeschätzt werden. Um diesen Approximationsfehler zu kontrollieren, konstruieren wir ein spezielles  $v$  und zwar den *bilinearen Interpolanten*. Der bilineare Interpolant  $I : C(\overline{\Omega}) \rightarrow S$  ist definiert durch

$$I_{\mathcal{G}}u \in S \quad \text{und} \quad (I_{\mathcal{G}}u)(\mathbf{x}_i) = u(\mathbf{x}_i) \quad \forall 1 \leq i \leq n_{\mathcal{G}}$$

oder äquivalent

$$(I_{\mathcal{G}}u) = \sum_{i=1}^{n_{\mathcal{G}}} u(\mathbf{x}_i) b_i.$$

**Satz 6.1** Für alle  $v \in H_0^1(\Omega) \cap H^2(\Omega)$  gilt die Interpolationsfehlerabschätzung:

$$|v - I_{\mathcal{G}}v|_{H^1(\Omega)} \leq \frac{h}{\sqrt{3}} |v|_{H^2(\Omega)}.$$

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

$$\varphi(t) = \varphi(0) + \int_0^t \varphi'(s) ds = \varphi(0) + t\varphi'(t) - \int_0^t s\varphi''(s) ds$$

und

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

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

$$\varphi(t) = \iota_h \varphi(t) + \frac{h-t}{h} \int_0^t \varphi'(s) ds - \frac{t}{h} \int_t^h \varphi'(s) ds \quad (6.5)$$

$$= \iota_h \varphi(t) - \frac{h-t}{h} \int_0^t s\varphi''(s) ds - \frac{t}{h} \int_t^h (h-s)\varphi''(s) ds. \quad (6.6)$$

Sei nun  $v \in C_0^\infty(\Omega)$  und  $\tau$  ein achsenparalleles Rechteck mit Kantenlängen  $h_1$  und  $h_2$ . Durch Translation des Koordinatensystems können wir erreichen, dass o.E.  $\tau = [0, h_1] \times [0, h_2]$  gilt. Anwenden der Gleichung (6.6) auf die Variable  $x$  liefert für alle  $(x, y) \in \tau$

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

Wenden wir Gleichung (6.5) für festes  $x$  auf die Variable  $y$  und

$$(\iota_h v(\cdot, y))(x) = v(0, y) \frac{h_1 - x}{h_1} + v(h_1, y) \frac{x}{h_1}$$

an, erhalten wir

$$\begin{aligned} (\iota_h v(\cdot, y))(x) &= \frac{h_1 - x}{h_1} \left\{ \frac{h_2 - y}{h_2} v(0, 0) + \frac{y}{h_2} v(0, h_2) \right. \\ &\quad \left. + \frac{h_2 - y}{h_2} \int_0^y \partial_2 v(0, t) dt - \frac{y}{h_2} \int_y^{h_2} \partial_2 v(0, t) dt \right\} \\ &\quad + \frac{x}{h_1} \left\{ \frac{h_2 - y}{h_2} v(h_1, 0) + \frac{y}{h_2} v(h_1, h_2) \right. \\ &\quad \left. + \frac{h_2 - y}{h_2} \int_0^y \partial_2 v(h_1, t) dt - \frac{y}{h_2} \int_y^{h_2} \partial_2 v(h_1, t) dt \right\}. \end{aligned}$$

Daraus folgt weiter

$$\begin{aligned} (\iota_h v(\cdot, y))(x) &= I_G v(x, y) + \frac{h_2 - y}{h_2} \int_0^y \partial_2 v(0, t) dt - \frac{y}{h_2} \int_y^{h_2} \partial_2 v(0, t) dt \\ &\quad + \frac{x}{h_1} \frac{h_2 - y}{h_2} \int_0^y \{ \partial_2 v(h_1, t) - \partial_2 v(0, t) \} dt \\ &\quad - \frac{x}{h_1} \frac{y}{h_2} \int_y^{h_2} \{ \partial_2 v(h_1, t) - \partial_2 v(0, t) \} dt. \end{aligned}$$

Da für alle  $t \in [0, h_2]$

$$\partial_2 v(h_1, t) - \partial_2 v(0, t) = \int_0^{h_1} \partial_1 \partial_2 v(s, t) ds$$

gilt, erhalten wir insgesamt die Darstellung

$$\begin{aligned} v(x, y) &= (I_G v)(x, y) + \frac{h_2 - y}{h_2} \int_0^y \partial_2 v(0, t) dt - \frac{y}{h_2} \int_y^{h_2} \partial_2 v(0, t) dt \\ &\quad + \frac{x}{h_1} \frac{h_2 - y}{h_2} \int_0^y \int_0^{h_1} \partial_1 \partial_2 v(s, t) ds dt \\ &\quad - \frac{x}{h_1} \frac{y}{h_2} \int_y^{h_2} \int_0^{h_1} \partial_1 \partial_2 v(s, t) ds dt \\ &\quad - \frac{h_1 - x}{h_1} \int_0^x s \partial_1^2 v(s, y) ds - \frac{x}{h_1} \int_x^{h_1} (h_1 - s) \partial_1^2 v(s, y) ds. \end{aligned}$$

Differentiation bzgl.  $x$  liefert

$$\begin{aligned}\frac{\partial}{\partial x}(v - I_{\mathcal{G}}v)(x, y) &= \frac{1}{h_1} \int_0^{h_1} \int_0^y \frac{h_2 - y}{h_2} \partial_1 \partial_2 v(s, t) dt ds - \frac{1}{h_1} \int_0^{h_1} \int_y^{h_2} \frac{y}{h_2} \partial_1 \partial_2 v(s, t) dt ds \\ &\quad + \frac{1}{h_1} \int_0^x s \partial_1^2 v(s, y) ds - \frac{1}{h_1} \int_x^{h_1} (h_1 - s) \partial_1^2 v(s, y) ds \\ &= \frac{1}{h_1} \int_0^{h_1} \int_0^{h_2} K_1(t, y) \partial_1 \partial_2 v(s, t) dt ds + \frac{1}{h_1} \int_0^{h_1} K_2(s, x) \partial_1^2 v(s, y) ds\end{aligned}$$

mit den *Integralkernen*

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

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

Quadrieren dieser Identität und Anwenden der Cauchy-Schwarzschen-Ungleichung ergibt

$$\begin{aligned}\left| \frac{\partial}{\partial x}(v - I_{\mathcal{G}}v)(x, y) \right|^2 &\leq \frac{2}{h_1^2} \int_0^{h_1} \int_0^{h_2} |K_1(t, y)|^2 dt ds \int_0^{h_1} \int_0^{h_2} |\partial_1 \partial_2 v(s, t)|^2 dt ds \\ &\quad + \frac{2}{h_1^2} \int_0^{h_1} |K_2(s, x)|^2 ds \int_0^{h_1} |\partial_1^2 v(s, y)|^2 ds.\end{aligned}$$

Integration über das Rechteck  $\tau$  liefert:

$$\begin{aligned}\int_{\tau} \left| \frac{\partial}{\partial x}(v - I_{\mathcal{G}}v) \right|^2 &\leq \frac{2}{h_1^2} \int_0^{h_1} \int_0^{h_2} \int_0^{h_1} \int_0^{h_2} |K_1(t, y)|^2 dt ds dy dx \|\partial_1 \partial_2 v\|_{L^2(\tau)}^2 \\ &\quad + \frac{2}{h_1^2} \int_0^{h_1} \int_0^{h_2} |K_2(s, x)|^2 dx ds \|\partial_1^2 v\|_{L^2(\tau)}^2.\end{aligned}$$

Offensichtlich gilt

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

und

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

Insgesamt haben wir damit die Abschätzung

$$\begin{aligned}\left\| \frac{\partial}{\partial x}(v - I_{\mathcal{G}}v) \right\|_{L^2(\tau)}^2 &\leq \frac{h_2^2}{3} \|\partial_1 \partial_2 v\|_{L^2(\tau)}^2 + \frac{h_1^2}{3} \|\partial_1^2 v\|_{L^2(\tau)}^2 \\ &\leq \frac{h^2}{3} \left\{ \|\partial_1 \partial_2 v\|_{L^2(\tau)}^2 + \|\partial_1^2 v\|_{L^2(\tau)}^2 \right\}.\end{aligned}$$



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

$$\left\| \frac{\partial}{\partial y} (v - I_{\mathcal{G}}v) \right\|_{L^2(\tau)}^2 \leq \frac{h^2}{3} \left\{ \|\partial_1 \partial_2 v\|_{L^2(\tau)}^2 + \|\partial_2^2 v\|_{L^2(\tau)}^2 \right\}.$$

Addition dieser beiden Abschätzungen und Summation über alle Rechtecke beweist mit

$$|v|_{H^2(\tau)}^2 = \|\partial_1^2 v\|_{L^2(\tau)}^2 + 2\|\partial_1 \partial_2 v\|_{L^2(\tau)}^2 + \|\partial_2^2 v\|_{L^2(\tau)}^2$$

die Behauptung. ■

Aus Satz 3.2, Satz 4.5 und Satz 6.1 folgt unmittelbar die folgende Fehlerabschätzung.

**Satz 6.2** Sei  $u \in H_0^1(\Omega)$  die schwache Lösung der allgemeinen elliptischen Differentialgleichung mit (6.1) und homogenen Dirichlet-Randbedingungen. Die Lösung des diskreten Problems (6.3) wird mit  $u_S$  bezeichnet. Es sei  $u \in H^2(\Omega)$ . Dann gilt

$$|u - u_S|_{H^1(\Omega)} \leq c_1 h |u|_{H^2(\Omega)}.$$

Ist zusätzlich  $\Omega$  konvex, so gilt

$$|u - u_S|_{L^2(\Omega)} \leq c_2 h^2 |u|_{H^2(\Omega)}.$$

Die Konstanten  $c_1$  und  $c_2$  hängen nur von  $\Omega$  und den Koeffizienten in der Differentialgleichung  $c$  und  $\mathbf{A}$  ab.

### Bemerkung 6.3

1. Bei inhomogenen Dirichlet-Randbedingungen  $u = u_D$  auf  $\Gamma$  sucht man die Lösung  $u_S$  von (6.3) statt in  $S_{\mathcal{G}}$  in  $I_{\mathcal{G}}u_D + S_{\mathcal{G}}$ , d.h., man setzt  $u_S$  in der Form

$$u_S = \sum_{i=1}^{n_{\mathcal{G}}} \mathbf{u}_i b_i + \sum_{i=n_{\mathcal{G}}+1}^{\overline{n_{\mathcal{G}}}} u_D(\mathbf{x}_i) b_i$$

mit unbekanntem Koeffizienten  $\mathbf{u} \in \mathbb{R}^{n_{\mathcal{G}}}$  an. Satz 6.2 bleibt gültig.

2. Bei gemischten oder Neumann-Randbedingungen muss man analog zu Definition 4.1 die rechte Seite von (6.3) durch  $\int_{\Gamma_N} gv$  ergänzen. Satz 6.2 bleibt gültig.
3. Unter den Voraussetzungen von Satz 4.3 gilt Satz 6.2 auch für die allgemeine elliptische Gleichung. Die Konstanten  $c_1, c_2$  aus Satz 6.2 verhalten sich dann im wesentlichen wie  $\lambda^{-1} \max \{\|\mathbf{A}\|_{\infty}, \|\mathbf{b}\|_{\infty}, \|c\|_{\infty}\}$ , wobei  $\lambda$  wie in Gleichung (4.2) ist. Für diffusionsdominierte Probleme - das sind Probleme, bei denen das Lösungsverhalten in erster Linie durch den Hauptteil  $\mathbf{A}$  der elliptischen Differentialgleichung bestimmt wird, d.h.,  $\|\mathbf{b}\|_{\infty} \sim \|\mathbf{A}\|_{\infty} \sim \lambda$  - ist diese Abschätzung gut. Für Konvektionsdominante Probleme, d.h.  $\|\mathbf{A}\|_{\infty} \sim \lambda \ll \|\mathbf{b}\|_{\infty}$ , ist sie dagegen unbrauchbar. Ausserdem entspricht die Diskretisierung dann einer zentralen Differenzendiskretisierung des Konvektionsterms  $\langle \mathbf{b}, \text{grad } u \rangle$ . Daher treten bei der numerischen Lösung  $u_S$  unphysikalische Oszillationen auf, wenn die **Peclet-Zahl**  $\lambda^{-1} \|\mathbf{b}\|_{\infty} h$  gross ist. Im nächsten Paragraphen beschreiben wir für diesen Fall eine Modifikation von (6.3), welche die genannten Schwierigkeiten vermeidet und die auch auf Rechteckselemente angewendet werden kann. ■

## 7 Finite Elemente in zwei Dimensionen

Sei  $\Omega \subset \mathbb{R}^2$  ein offenes, beschränktes, zusammenhängendes Gebiet mit polygonalem Rand  $\Gamma$ . Wir betrachten die allgemeine, *symmetrische*, elliptische Differentialgleichung mit Dirichlet-Randbedingungen:

$$\text{Finde } u \in H_0^1(\Omega) \quad a(u, v) = \ell(v) \quad \forall v \in H_0^1(\Omega) \quad (7.1)$$

mit

$$a(u, v) = \int_{\Omega} (\mathbf{A} \nabla u, \nabla v + cuv) \quad \text{und} \quad \ell(v) = \int_{\Omega} f v. \quad (7.2)$$

Die Voraussetzungen an die Koeffizienten  $\mathbf{A}$ ,  $c$  und die rechte Seite  $f$  sind

$$\mathbf{A} \in L^\infty(\Omega, \mathbb{R}_{\text{symm}}^{2 \times 2}), \quad c \in L^\infty(\Omega, \mathbb{R}_{\geq 0}), \quad f \in L^2(\Omega) \quad (7.3)$$

und

$$0 < \lambda := \inf_{\mathbf{x} \in \Omega} \inf_{\mathbf{z} \in \mathbb{R}^d \setminus \{0\}} \frac{\mathbf{z}^\top \mathbf{A}(\mathbf{x}) \mathbf{z}}{\mathbf{z}^\top \mathbf{z}} \leq \sup_{\mathbf{x} \in \Omega} \sup_{\mathbf{z} \in \mathbb{R}^d \setminus \{0\}} \frac{\mathbf{z}^\top \mathbf{A}(\mathbf{x}) \mathbf{z}}{\mathbf{z}^\top \mathbf{z}} =: \Lambda < \infty. \quad (7.4)$$

Sei  $\mathcal{G} = \{\tau_i : 1 \leq i \leq n_{\mathcal{G}}\}$  eine Unterteilung von  $\Omega$  in Dreiecke, derart, dass je zwei Dreiecke entweder disjunkt sind oder einen Eckpunkt oder eine Kante gemeinsam haben (**Zulässigkeitsbedingung**). Für jedes  $\tau \in \mathcal{G}$  bezeichnet  $h_\tau$  den Durchmesser von  $\tau$  und  $\rho_\tau$  den Durchmesser des grössten, eingeschriebenen Kreises. Die Triangulierung sei **form-regulär**, d.h., es existiert ein (moderate) Konstante mit

$$h_\tau / \rho_\tau \leq c_{\mathcal{G}}. \quad (7.5)$$

Dies ist äquivalent zur Bedingung, dass der kleinste Winkel aller Dreiecke gleichmässig von Null weg beschränkt ist (**Winkelbedingung**). Setze

$$h := \max_{\tau \in \mathcal{G}} h_\tau.$$

Für  $\tau \in \mathcal{G}$  bezeichnen wir die Eckpunkte mit  $\mathbf{x}_i^\tau$ ,  $1 \leq i \leq 3$ . Die Numerierung sei immer entgegen des Uhrzeigersinns gewählt. Die Menge aller Eckpunkte der Dreiecke werden wieder mit  $\Theta_{\mathcal{G}}$  bezeichnet und die inneren mit  $\overset{\circ}{\Theta}_{\mathcal{G}}$ . Wir numerieren zunächst die inneren Punkte in  $\overset{\circ}{\Theta}_{\mathcal{G}}$  von 1 bis  $n_{\mathcal{G}}$  und danach die Randpunkte von  $n_{\mathcal{G}} + 1$  bis  $\overline{n_{\mathcal{G}}}$ .

Im Rahmen von abstrakten Finite-Elemente-Diskretisierungen von Bilinearformen müssen wir zunächst wieder einen endlichdimensionalen Teilraum  $S$  von  $H_0^1(\Omega)$  definieren. Wir setzen

$$S := S_{\mathcal{G}} := \{\varphi \in C^0(\overline{\Omega}) \mid \forall \tau \in \mathcal{G} : \varphi|_{\tau} \in \mathbb{P}_1\} \cap H_0^1(\Omega). \quad (*)$$

Da eine Ebene im  $\mathbb{R}^3$  durch drei nicht kollineare Punkte eindeutig bestimmt ist, ist für jedes  $\tau \in \mathcal{G}$  und jedes  $p \in \mathbb{P}_1$  das Polynom  $p$  eindeutig durch seine Werte in der drei Eckpunkten von  $\tau$  bestimmt. Da die Einschränkung von  $p$  auf eine beliebige Kante eine lineare Funktion einer Veränderlichen ist, folgt ausserdem, dass für je zwei Dreiecke  $\tau_1, \tau_2$ , die eine Kante  $E$  gemeinsam haben und für je zwei Polynome  $p_1, p_2 \in \mathbb{P}_1$  die Funktion

$$\varphi := \begin{cases} p_1 & \text{auf } \tau_1 \\ p_2 & \text{auf } \tau_2 \end{cases}$$

genau dann stetig ist, wenn  $p_1$  und  $p_2$  in den Eckpunkten von  $E$  übereinstimmen. Daher sind die Funktionen in  $S$  eindeutig bestimmt durch ihre Werte in den Punkten von  $\overset{\circ}{\Theta}_{\mathcal{G}}$ . Insbesondere ist

$$\dim S = n_{\mathcal{G}}.$$

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

$$\lambda_{\tau,i}(\mathbf{x}) := \frac{\det(\mathbf{x}_{i+1}^{\tau} - \mathbf{x}, \mathbf{x}_{i-1}^{\tau} - \mathbf{x}_{i+1}^{\tau})}{\det(\mathbf{x}_{i+1}^{\tau} - \mathbf{x}_i^{\tau}, \mathbf{x}_{i-1}^{\tau} - \mathbf{x}_{i+1}^{\tau})}.$$

Dabei sind die Indizes immer Modulo 3 zu interpretieren  $4 \longleftrightarrow 1$ ,  $5 \longleftrightarrow 2$  usw. Die Funktionen  $\lambda_{\tau,i}$  heißen **Schwerpunktskoordinaten** von  $\tau$ . Sie haben offensichtlich die Eigenschaft

$$\lambda_{\tau,i} \in \mathbb{P}_1, \quad \lambda_{\tau,i}(\mathbf{x}_j^{\tau}) = \begin{cases} 1 & i = j, \\ 0 & \text{sonst,} \end{cases} \quad \sum_{i=1}^3 \lambda_{\tau,i} = 1.$$

Man beachte, dass  $\frac{1}{2} \det(\mathbf{x}_{i+1}^{\tau} - \mathbf{x}_i^{\tau}, \mathbf{x}_{i-1}^{\tau} - \mathbf{x}_{i+1}^{\tau})$  die Fläche  $|\tau|$  von  $\tau$  angibt. Für beliebiges  $\mathbf{x}_i \in \Theta_{\mathcal{G}}$  definieren wir

$$\omega_i := \bigcup_{\substack{\tau \in \mathcal{G} \\ \mathbf{x}_i \in \tau}} \tau$$

und

$$b_i(\mathbf{x}) := \begin{cases} 0 & \text{falls } \mathbf{x} \notin \omega_i, \\ \lambda_{\tau,j}(\mathbf{x}) & \text{falls } \mathbf{x} \in \tau \subset \omega_i \text{ und } \mathbf{x}_i = \mathbf{x}_j^{\tau}. \end{cases}$$

Offensichtlich sind die Funktionen  $b_i$  stetig, stückweise linear, haben den Wert 1 im Punkt  $\mathbf{x}_i$  und verschwinden in allen anderen Punkten von  $\Theta_{\mathcal{G}}$ . Daher gilt

$$S = \text{span} \{b_i : 1 \leq i \leq n_{\mathcal{G}}\}.$$

Man beachte, dass  $\omega_i$  genau aus den Dreiecken besteht, die den Punkt  $\mathbf{x}_i$  als Eckpunkt besitzen. Die Funktionen  $b_i$  heißen **Knotenbasis** von  $S$ .

Jedes  $u \in S$  lässt sich eindeutig schreiben in der Form

$$u = \sum_{i=1}^{n_{\mathcal{G}}} u_i b_i \quad \text{mit} \quad u_i = u(\mathbf{x}_i). \quad (7.6)$$

Daher ist das diskrete Problem: Finde  $u_S \in S$  mit

$$a(u_S, v) = \ell(v) \quad \forall v \in S \quad (7.7)$$

äquivalent zu einem linearen Gleichungssystem mit  $n_{\mathcal{G}}$  Gleichungen und Unbekannten. Wir definieren dazu die Systemmatrix  $\mathbf{A} = (a_{i,j})_{i,j=1}^{n_{\mathcal{G}}} \in \mathbb{R}^{n_{\mathcal{G}} \times n_{\mathcal{G}}}$  und den rechte-Seite-Vektor  $\mathbf{r} = (r_i)_{i=1}^{n_{\mathcal{G}}} \in \mathbb{R}^{n_{\mathcal{G}}}$  durch

$$a_{i,j} = a(b_j, b_i) \quad \text{und} \quad r_i = \ell(b_i) \quad \forall 1 \leq i, j \leq n_{\mathcal{G}}.$$

**Proposition 7.1**

a. Die Systemmatrix  $\mathbf{A}$  für Problem (7.2) ist symmetrisch und positiv definit.

b. Das lineare Gleichungssystem

$$\mathbf{A}\mathbf{u} = \mathbf{r} \quad (7.8)$$

besitzt eine eindeutige Lösung. Die zugehörige Finite-Elemente-Funktion  $u_S = \sum_{i=1}^{n_G} u_i b_i$  ist die eindeutige Lösung von (7.7). Ist umgekehrt  $u_S$  die eindeutige Lösung von (7.7), so erfüllt der durch (7.6) definierte Koeffizientenvektor  $\mathbf{u}$  das Gleichungssystem (7.8).

c. Die Matrix ist schwach besetzt, d.h. die Anzahl der Nicht-Nullelemente pro Zeile ist beschränkt durch „1+ maximale Anzahl von Dreiecksanten, die von einem Knotenpunkt im Gitter  $\mathcal{G}$  ausgehen“.

**Beweis. @ a:** Die Symmetrie folgt aus der Symmetrie der Bilinearform  $a(\cdot, \cdot)$  (wegen  $\mathbf{b} = \mathbf{0}$  und der Symmetrie des Diffusionskoeffizienten  $\mathbf{A}(\mathbf{x}) \in \mathbb{R}^{d \times d}$ ). Für einen Koeffizientenvektor  $\mathbf{u} = (u_i)_{i=1}^{n_G}$  assoziieren wir die Finite-Elemente-Funktion  $u$  wie in (7.6). Da  $(b_i)_{i=1}^{n_G}$  eine Basis von  $S$  ist, gilt die Äquivalenz  $\mathbf{u} = \mathbf{0} \iff u = 0$ .

Daraus folgt mit dem Beweis von Satz 4.3(Teil 1) die Abschätzung

$$\langle \mathbf{u}, \mathbf{A}\mathbf{u} \rangle = \sum_{i,j=1}^{n_G} u_i u_j a_{i,j} = \sum_{i,j=1}^{n_G} u_i u_j a(b_i, b_j) = a\left(\sum_{i=1}^{n_G} u_i b_i, \sum_{j=1}^{n_G} u_j b_j\right) = a(u, u) \geq C \|u\|_{H^1(\Omega)}^2$$

und wegen der oben erklärten Äquivalenz die positive Definitheit

$$\langle \mathbf{u}, \mathbf{A}\mathbf{u} \rangle > 0 \quad \forall \mathbf{u} \in \mathbb{R}^{n_G} \setminus \{0\}.$$

**@ b:** Das lineare Gleichungssystem (7.8) ist die Basisdarstellung der Variationsaufgabe (7.7). Daher folgt die Aussage aus elementaren Sätzen der linearen Algebra.

**@ c:** Sei  $x_i$  der innere Gitterpunkt zum Index  $1 \leq i \leq n_G$ . Der Träger von  $b_i$  (und damit auch von  $\nabla b_i$ ) ist durch  $\omega_i$  gegeben. Das Matrixelement  $a_{i,j}$  ist also genau dann von Null verschieden, falls die Fläche  $\omega_i \cap \omega_j$  positives zweidimensionales Mass besitzt. Dies ist aber genau dann der Fall, falls  $x_i$  und  $x_j$  durch eine Kante verbunden sind. ■

Jedes Dreieck  $\tau \in \mathcal{G}$  lässt sich mittels einer (nicht-eindeutigen) affinen Abbildung  $F_\tau : \hat{\tau} \rightarrow \tau$  auf das Referenzelement

$$\hat{\tau} := \text{conv} \{(0, 0)^\top, (1, 0)^\top, (1, 1)^\top\}$$

abbilden.

Mit  $\Theta_{\mathcal{G}}$  und  $\overset{\circ}{\Theta}_{\mathcal{G}}$  bezeichnen wir die Menge aller Eckpunkte aller  $\tau \in \mathcal{G}$  bzw. der Eckpunkte im Innern von  $\Omega$ .

**Definition 7.2** Sei  $k \in \mathbb{N}$ . Die Menge aller äquidistanten kartesischen Gitterpunkte im zwei-dimensionalen Einheitswürfel zur Maschenweite  $1/k$  ist durch

$$K_k := \left\{ \left( \frac{\mu_1}{k}, \frac{\mu_2}{k} \right) : \mu_1, \mu_2 \in \{0, 1, \dots, k\} \right\}$$

gegeben.

1. Für  $k \in \mathbb{N}$  ist die Menge aller **Referenz-Knotenpunkte** durch

$$\hat{\Sigma}_k := \bar{\hat{\tau}} \cap K_k.$$

2. Für  $\tau = F_\tau(\hat{\tau})$  und  $k \in \mathbb{N}$  ist die Menge aller **lokalen Knotenpunkte** durch

$$\Sigma_k := \Sigma_k(\tau) := F_\tau(\hat{\Sigma}_k)$$

gegeben.

3. Die Menge aller **globalen Knotenpunkte** bzw. aller **globalen inneren Knotenpunkte** ist durch

$$K := K_{\mathcal{G}} := K_{\mathcal{G}}^k := \bigcup_{\tau \in \mathcal{G}} \Sigma_k(\tau) \quad \text{bzw.} \quad \mathring{K} := \mathring{K}_{\mathcal{G}} := \mathring{K}_{\mathcal{G}}^k := \left( \bigcup_{\tau \in \mathcal{G}} \Sigma_k(\tau) \right) \cap \Omega$$

gegeben.

4. Setze  $\mathbb{P}_0 = \mathbb{R}$  und definiere für  $k \in \mathbb{N}_{\geq 0}$

$$\mathbb{P}_k := \text{span} \{x_1^{\alpha_1} x_2^{\alpha_2} : \alpha_1, \alpha_2 \in \mathbb{N}_0 \text{ mit } \alpha_1 + \alpha_2 \leq k\}.$$

**Bemerkung 7.3** Die Polynomräume  $\mathbb{P}_k$  sind invariant bezüglich affiner Koordinaten-Transformation.

**Satz 7.4** Sei  $k \in \mathbb{N}$ . Dann ist jedes  $p \in \mathbb{P}_k$  eindeutig bestimmt durch seine Werte auf  $\hat{\Sigma}_k$  bzw. auf  $\Sigma_k(\tau)$ ,  $\tau \in \mathcal{G}$ .

**Beweis.** Wegen der Definition von  $\Sigma_k$  und Bemerkung 7.3 reicht es, die Behauptung für  $\hat{\Sigma}_k$  zu zeigen. Eine leichte Rechnung zeigt, dass

$$\dim \mathbb{P}_k = \#\hat{\Sigma}_k$$

gilt. Daher genügt es, eine der Aussagen (a) und (b) zu beweisen.

a. Zu jedem Vektor  $(b_{\mathbf{z}})_{\mathbf{z} \in \hat{\Sigma}_k}$  existiert ein  $p \in \mathbb{P}_k$  mit  $p(\mathbf{z}) = b_{\mathbf{z}}$  für alle  $\mathbf{z} \in \hat{\Sigma}_k$ .

b. Ist  $p \in \mathbb{P}_k$  und  $p(\mathbf{z}) = 0$  für alle  $\mathbf{z} \in \hat{\Sigma}_k$ , so ist  $p \equiv 0$ .

Die Schwerpunktskoordinaten für das Referenzelement sind gegeben durch

$$\hat{\lambda}_1(\mathbf{x}) := 1 - \hat{x}_1, \quad \hat{\lambda}_2(\mathbf{x}) := \hat{x}_1 - \hat{x}_2, \quad \hat{\lambda}_3(\mathbf{x}) := \hat{x}_2$$

gegeben. Offensichtlich sind diese Funktionen in  $\mathbb{P}_1$  und besitzen die Lagrange-Eigenschaft

$$\forall 1 \leq i, j \leq 3 : \hat{\lambda}_i(A_j) = \delta_{i,j} \quad \text{mit} \quad A_1 = (0, 0)^\top, \quad A_2 = (1, 0)^\top, \quad A_3 = (1, 1)^\top.$$

1.  $k = 1$  : Zu gegebenem  $(b_i)_{i=1}^3 \in \mathbb{R}^3$  besitzt offensichtlich  $p \in \mathbb{P}_1$  :

$$p(\mathbf{x}) = \sum_{i=1}^3 b_i \hat{\lambda}_i(\mathbf{x})$$

die Eigenschaft (a).

2.  $k = 2$  : Für  $1 \leq i < j \leq 3$  bezeichnen  $A_{i,j} := (A_i + A_j) / 2$  die Kantenmittelpunkte von  $\hat{\tau}$ . Definiere

$$\begin{aligned}\widehat{N}_i &:= \widehat{\lambda}_i (2\widehat{\lambda}_i - 1) & 1 \leq i \leq 3, \\ \widehat{N}_{i,j} &:= 4\widehat{\lambda}_i \widehat{\lambda}_j & 1 \leq i < j \leq 3.\end{aligned}$$

Dann gilt offensichtlich  $\widehat{N}_k, \widehat{N}_{i,j} \in \mathbb{P}_2$  und

$$\begin{aligned}\widehat{N}_i(A_j) &= \delta_{i,j} & \widehat{N}_i(A_{k,\ell}) &= 0 & \forall i, k, \ell, \\ \widehat{N}_{i,j}(A_k) &= 0 & \widehat{N}_{i,j}(A_{k,\ell}) &= \delta_{i,k} \delta_{j,\ell} & \forall i, j, k, \ell.\end{aligned}$$

Für gegebenes  $\{b_{\mathbf{z}} : \mathbf{z} \in \widehat{\Sigma}_2\} = \{b_i, b_{k,\ell}\}$  besitzt  $p \in \mathbb{P}_2$  :

$$p(\mathbf{x}) := \sum_{i=1}^3 b_i \widehat{N}_i(\mathbf{x}) + \sum_{1 \leq k < \ell \leq 3} b_{k,\ell} \widehat{N}_{k,\ell}(\mathbf{x})$$

die Eigenschaft (a).

3.  $k = 3$  : Dieser Fall wird in Übungsaufgabe 7.5 behandelt.  
 4.  $k \geq 4$  : Sei  $p \in \mathbb{P}_k$  mit  $p(\mathbf{z}) = 0$  für alle  $\mathbf{z} \in \widehat{\Sigma}_k$ . Dann verschwindet  $p$  auf allen Kanten von  $\hat{\tau}$ . Daher gibt es ein  $\psi \in \mathbb{P}_{k-3}$  mit

$$p = \widehat{\lambda}_1 \widehat{\lambda}_2 \widehat{\lambda}_3 \psi \quad \text{und} \quad \forall \mathbf{z} \in \widehat{\Sigma}_k \cap \hat{\tau} : \psi(\mathbf{z}) = 0.$$

(Man beachte hierbei, daß  $\hat{\tau}$  offen ist.) Die Aufgabe ist daher auf das Problem zurückgeführt

$$(\psi \in \mathbb{P}_{k-3}) \wedge \left( \forall \mathbf{z} \in \widehat{\Sigma}_k \cap \hat{\tau} : \psi(\mathbf{z}) = 0 \right) \implies \psi \equiv 0. \quad (7.9)$$

Eigenschaft (b) folgt durch Induktion über  $k$ :

Sei  $\hat{\tau}'$  das Dreieck mit Ecken  $A = \left(\frac{2}{k+1}, \frac{1}{k+1}\right)^\top$ ,  $B = \left(\frac{k}{k+1}, \frac{1}{k+1}\right)^\top$ ,  $C = \left(\frac{k}{k+1}, \frac{k-1}{k+1}\right)^\top$ . Dann gilt  $\widehat{\Sigma}_k \cap \hat{\tau} =: \widehat{\Sigma}'_k \subset \hat{\tau}'$ . Die Transformation

$$T : \hat{\tau} \rightarrow \hat{\tau}' : T\xi = A + \left(1 - \frac{3}{k+1}\right) \xi$$

ist affin und daher  $\tilde{\psi} = \psi \circ T \in \mathbb{P}_{k-3}$ . Weiter gilt  $T^{-1}\widehat{\Sigma}'_k = \widehat{\Sigma}_{k-3}$ . Daher ist (7.9) äquivalent zu

$$\left( \tilde{\psi} \in \mathbb{P}_{k-3} \right) \wedge \left( \forall \mathbf{z} \in \widehat{\Sigma}_{k-3} : \tilde{\psi}(\mathbf{z}) = 0 \right) \implies \tilde{\psi} \equiv 0.$$

Das ist aber Aussage (b) für  $k \leftarrow k - 3$ . Da der Induktionsanfang für  $k = 1, 2, 3$  durch die Beweisschritte 1-3 wegen der Äquivalenz der Aussagen (a) und (b) gegeben ist, folgt die Behauptung. ■

**Übungsaufgabe 7.5** Sei  $\hat{\tau}$  das Einheitsdreieck. Konstruieren Sie für  $\mathbb{P}_3$  eine Lagrange-Basis zur Stützstellenmenge  $\widehat{\Sigma}_k$ .

**Satz 7.6** Seien  $k \in \mathbb{N}$ ,  $\tau_1, \tau_2 \in \mathcal{G}$  mit  $\tau_1 \cap \tau_2 \neq \emptyset$  und  $p_1, p_2 \in \mathbb{P}_k$ . Definiere die Funktion  $\varphi$  auf  $\tau_1 \cup \tau_2$  durch

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

Dann ist  $\varphi \in C(\overline{\tau_1} \cup \overline{\tau_2})$  genau dann, wenn für alle  $\mathbf{x} \in F_{\tau_1}(\hat{\Sigma}_k) \cap F_{\tau_2}(\hat{\Sigma}_k)$  gilt

$$p_1(\mathbf{x}) = p_2(\mathbf{x}).$$

**Beweis.** Die Richtung „ $\implies$ “ ist offensichtlich.

„ $\impliedby$ “: Offensichtlich ist nur etwas zu zeigen, wenn  $\tau_1$  und  $\tau_2$  eine Kante besitzen.

Wir betrachten zunächst den Fall, dass  $\tau_1$  und  $\tau_2$  eine Kante  $E$  gemeinsam haben. Setze  $\hat{E} := \hat{\tau} \cap \{x_2 = 0\}$ . Dann ist  $\hat{E}$  das Einheitsintervall, d.h. der Standard 1-Simplex. Die Transformationen  $F_{\tau_1}$  und  $F_{\tau_2}$  können so gewählt werden, dass  $E = F_{\tau_1}(\hat{E}) = F_{\tau_2}(\hat{E})$  ist. Definiere  $q := p_1 \circ F_{\tau_1}|_{\hat{E}} - p_2 \circ F_{\tau_2}|_{\hat{E}}$ . Wegen Bemerkung 7.3 ist  $q \in \mathbb{P}_k$  ein Polynom einer Veränderlichen vom Maximalgrad  $k$ . Nach Voraussetzung verschwindet  $q$  in allen Punkten von  $\hat{\Sigma}_k \cap \hat{E}$ . Also ist  $q = 0$  damit  $\varphi$  stetig.

■

Nach diesen Vereinbarungen können wir nun die Finite-Elemente-Räume höherer Ordnung zu  $\mathcal{G}$  wie folgt definieren:

$$\begin{aligned} S^{k,-1} &:= \{\varphi \in L^1(\Omega) \mid \forall \tau \in \mathcal{G} : \varphi|_{\tau} \in \mathbb{P}_k\}, \\ S^{k,0} &:= S^{k,-1} \cap C(\overline{\Omega}), \\ S_0^{k,0} &:= \{\varphi \in S^{k,0} : \varphi = 0 \text{ auf } \Gamma\}. \end{aligned}$$

Wegen Satz 2.10 ist  $S^{k,0} \subset W^{1,p}(\Omega)$  und  $S_0^{k,0} \subset W_0^{1,p}(\Omega)$  für jedes  $1 \leq p < \infty$ . Wir können daher im Rahmen von konformen Finite-Elemente-Diskretisierungen  $S = S_0^{k,0}$  mit beliebigem  $k \in \mathbb{N}$  wählen und dadurch die Diskretisierung (7.6) verallgemeinern. Aus den Sätzen 7.4, 7.6 folgt, dass die Funktionen in  $S^{k,0}$  und  $S_0^{k,0}$  eindeutig bestimmt sind durch ihre Werte in den Gitterpunkten  $K_{\mathcal{G}}^k$  bzw.  $K_{\mathcal{G}}^{k,0}$ . Insbesondere gibt es zu jedem  $\mathbf{z}_i \in K_{\mathcal{G}}^k$  eine eindeutige Funktion  $b_i \in S^{k,0}$  mit  $b_i(\mathbf{z}_j) = \delta_{i,j}$  für alle  $1 \leq i, j \leq N$ , wobei  $N := \#K_{\mathcal{G}}^k$  gesetzt wurde. Die Funktionen  $b_i$  heissen **Knotenbasis** von  $S^{k,0}$ . Jedes  $u \in S^{k,0}$  lässt sich eindeutig darstellen als

$$u = \sum_{i=1}^N \mathbf{u}_i b_i,$$

und die  $\mathbf{u}_i$  sind genau die Werte  $u(\mathbf{z}_i)$

Für die Fehlerabschätzung – genauer die Abschätzung des Bestapproximationsfehlers

$$\inf_{v \in S} \|u - v\|_{H^1(\Omega)}$$

definieren wir einen Interpolationsoperator  $I_{\mathcal{G}}^k : H^2(\Omega) \rightarrow S^{k,0}$  durch

$$(I_{\mathcal{G}}^k u)(\mathbf{z}) = u(\mathbf{z}) \quad \forall \mathbf{z} \in K_{\mathcal{G}}^k$$

oder äquivalent

$$(I_{\mathcal{G}}^k u)(\mathbf{x}) = \sum_{i=1}^N u(\mathbf{z}_i) b_i.$$

Man beachte, dass  $I_{\mathcal{G}}^k(H_0^1(\Omega) \cap H^2(\Omega)) \subset S_0^{k,0}$  gilt.

Die Abschätzung des Interpolationsfehlers geschieht zunächst auf dem Referenzelement. Dazu definieren wir  $\hat{\Pi}_k : H^2(\hat{\tau}) \rightarrow \mathbb{P}_k$  durch

$$\left(\hat{\Pi}_k u\right)(\mathbf{z}) = u(\mathbf{z}) \quad \forall \mathbf{z} \in \hat{\Sigma}_k. \quad (7.10)$$

Man beachte, dass für jedes  $u \in H^2(\Omega)$  und jedes  $\tau \in \mathcal{G}$  die Identität

$$(I_{\mathcal{G}}^k u)|_{\tau} = \left(\hat{\Pi}_k \hat{u}\right) \circ F_{\tau}^{-1} \quad \text{mit} \quad \hat{u} = u|_{\tau} \circ F_{\tau}.$$

**Satz 7.7** Sei  $k \in \mathbb{N}$ . Dann wird durch

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

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

**Beweis.** Der Sobolevsche Einbettungssatz impliziert  $H^{k+1}(\hat{\tau}) \hookrightarrow C(\bar{\hat{\tau}})$  und daher ist  $[u]_{k+1}$  wohldefiniert, und es existiert eine Konstante  $c_1 > 0$  mit

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

Wir müssen also noch zeigen, dass es eine Konstante  $c_2 > 0$  gibt mit

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

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

$$\|u_n\|_{H^{k+1}(\hat{\tau})} = 1 \quad \forall n \in \mathbb{N} \quad (7.11)$$

und

$$\lim_{n \rightarrow \infty} [u_n]_{k+1} = 0. \quad (7.12)$$

Wegen Bemerkung 2.20 und Satz 2.21 gibt es eine Teilfolge  $(u_{n_m})_{m \in \mathbb{N}}$  und ein  $u \in H^k(\hat{\tau})$  mit

$$\lim_{m \rightarrow \infty} \|u_{n_m} - u\|_{H^k(\hat{\tau})} = 0.$$

Diese Folge ist eine Cauchy-Folge in  $H^{k+1}(\hat{\tau})$ , da auch

$$|u_{n_m} - u_{n_k}|_{H^{k+1}(\hat{\tau})} \leq |u_{n_m}|_{H^{k+1}(\hat{\tau})} + |u_{n_k}|_{H^{k+1}(\hat{\tau})} \leq [u_{n_m}]_{H^{k+1}(\hat{\tau})} + [u_{n_k}]_{H^{k+1}(\hat{\tau})} \rightarrow 0$$

gilt, so dass  $u \in H^{k+1}(\Omega)$  folgt. Wegen (7.12) ist insbesondere

$$\lim_{m \rightarrow \infty} |u_{n_m} - u|_{H^{k+1}(\hat{\tau})} = 0.$$

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

$$\lim_{m \rightarrow \infty} \|u_{n_m} - u\|_{H^{k+1}(\hat{\tau})} = 0.$$



Wegen  $|u|_{H^{k+1}(\hat{\tau})} = 0$  ist  $u \in \mathbb{P}_k$ . Wegen Satz 2.21 gilt

$$u(\mathbf{z}) = \lim_{m \rightarrow \infty} u_{n_m}(\mathbf{z}) \quad \forall \mathbf{z} \in \hat{\Sigma}_k,$$

wobei  $\hat{\Sigma}_k$  hier die Knotenmenge für den Einheitssimplex bezeichnet. Hieraus und aus (7.12) folgt aber

$$u(\mathbf{z}) = 0 \quad \forall \mathbf{z} \in \hat{\Sigma}_k.$$

Wegen Satz 7.4 ist also  $u = 0$  im Widerspruch zu (7.11). ■

**Satz 7.8** Sei  $k \in \mathbb{N}$ . Dann gibt es eine Konstante  $c$ , die von  $k$  abhängt, mit

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

**Beweis.** Aus Satz 7.7 folgt für beliebiges  $u \in H^{k+1}(\hat{\tau})$

$$\left\| u - \hat{\Pi}_k u \right\|_{H^{k+1}(\hat{\tau})} \leq c_2 \left[ u - \hat{\Pi}_k u \right]_{H^{k+1}(\hat{\tau})} = c_2 |u|_{H^{k+1}(\hat{\tau})},$$

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

Für  $\tau \in \mathcal{G}$  bezeichnet

$$F_\tau(\hat{\mathbf{x}}) = b_\tau + \mathbf{B}_\tau \hat{\mathbf{x}} \tag{7.13}$$

eine affine Transformation des Referenzelements  $\hat{\tau}$  auf  $\tau$ . Das folgende Lemma schätzt die Jacobi-Matrix  $\mathbf{B}_\tau$  dieser Transformation ab.

**Lemma 7.9** Für jedes  $\tau \in \mathcal{G}$  gilt für die Matrix  $\mathbf{B}_\tau$  aus (7.13)

$$\|\mathbf{B}_\tau\| \leq \frac{2 + \sqrt{2}}{2} h_\tau, \quad \|\mathbf{B}_\tau^{-1}\| \leq \frac{\sqrt{2}}{\rho_\tau}.$$

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

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

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

$$\|\mathbf{B}_\tau \hat{\mathbf{z}}\| = \|F_\tau(\hat{\mathbf{x}}) - F_\tau(\hat{\mathbf{y}})\| \leq h_\tau.$$

Also ist

$$\|\mathbf{B}_\tau\| = \rho_{\hat{\tau}}^{-1} \sup_{\substack{\hat{\mathbf{z}} \in \mathbb{R}^2 \\ \|\hat{\mathbf{z}}\| = \rho_{\hat{\tau}}}} \|\mathbf{B}_\tau \hat{\mathbf{z}}\| \leq \frac{h_\tau}{\rho_{\hat{\tau}}} = \frac{2 + \sqrt{2}}{2} h_\tau.$$

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

$$\|\mathbf{B}_\tau^{-1}\| \leq \frac{h_{\hat{\tau}}}{\rho_\tau} = \frac{\sqrt{2}}{\rho_\tau}.$$

■

**Satz 7.10** Sei  $k \in \mathbb{N}$ . Dann gelten für alle  $u \in H^{k+1}(\Omega)$  die Interpolationsfehlerabschätzungen

$$\begin{aligned} \|u - I_{\mathcal{G}}^k u\|_{L^2(\Omega)} &\leq c_1 h^{k+1} |u|_{H^{k+1}(\Omega)} \\ |u - I_{\mathcal{G}}^k u|_{H^1(\Omega)} &\leq c_2 h^k |u|_{H^{k+1}(\Omega)}. \end{aligned}$$

Dabei ist  $h := \max_{\tau \in \mathcal{G}} h_\tau$ . Die Konstanten  $c_1$  und  $c_2$  hängen von  $\Omega$ ,  $k$  und der Konstanten  $c_{\mathcal{G}}$  in der Regularitätsbedingung an  $\mathcal{G}$  ab.

**Beweis.** Sei  $\tau \in \mathcal{G}$  und  $F_\tau(\hat{\mathbf{x}}) = b_\tau + \mathbf{B}_\tau \hat{\mathbf{x}}$  eine affine Transformation des Referenzelements  $\hat{\tau}$  auf  $\tau$ . Die Transformationsformel für Integrale liefert<sup>2</sup>

$$\begin{aligned} \|u - I_{\mathcal{G}}^k u\|_{L^2(\tau)} &= |\det \mathbf{B}_\tau|^{1/2} \|(u - I_{\mathcal{G}}^k u) \circ F_\tau\|_{L^2(\hat{\tau})}, \\ |u - I_{\mathcal{G}}^k u|_{H^1(\tau)} &\leq |\det \mathbf{B}_\tau|^{1/2} \|\mathbf{B}_\tau^{-1}\| \|(u - I_{\mathcal{G}}^k u) \circ F_\tau\|_{H^1(\hat{\tau})}, \\ |u \circ F_\tau|_{H^{k+1}(\hat{\tau})} &\leq C_k |\det \mathbf{B}_\tau|^{-1/2} \|\mathbf{B}_\tau\|^{k+1} |u|_{H^{k+1}(\tau)}. \end{aligned} \quad (7.15)$$

Aus Lemma 7.9 folgt

$$\|\mathbf{B}_\tau\| \leq h_\tau / \rho_{\hat{\tau}}, \quad \|\mathbf{B}_\tau^{-1}\| \leq h_{\hat{\tau}} / \rho_\tau.$$

Aus diesen Abschätzungen und Satz 7.8 ergibt sich wegen  $I_{\mathcal{G}}^k u \circ F_\tau = \hat{\Pi}_\tau(u \circ F_\tau)$

$$\|u - I_{\mathcal{G}}^k u\|_{L^2(\tau)} \leq c (h_\tau / \rho_{\hat{\tau}})^{k+1} |u|_{H^{k+1}(\tau)}, \quad (7.16a)$$

$$|u - I_{\mathcal{G}}^k u|_{H^1(\tau)} \leq c h_{\hat{\tau}} / \rho_\tau (h_\tau / \rho_{\hat{\tau}})^{k+1} |u|_{H^{k+1}(\tau)} = c h_{\hat{\tau}} \rho_{\hat{\tau}}^{-k-1} \left(\frac{h_\tau}{\rho_\tau}\right) h_\tau^k |u|_{H^{k+1}(\tau)}. \quad (7.16b)$$

Hieraus folgt die Behauptung durch Quadrieren und Summieren über alle Elemente  $\tau \in \mathcal{G}$ . ■

<sup>2</sup>Für (7.15), setzen wir  $\hat{v} = v \circ F_\tau$  und verwenden wir die Kettenregel

$$((\partial^\alpha \hat{v}) \circ F_\tau^{-1}) = (\mathbf{B}_\tau^\top \nabla)^\alpha v.$$

Wir bezeichnen die Spaltenvektoren von  $\mathbf{B}^\top$  mit  $\mathbf{b}_1, \mathbf{b}_2$  und verwenden die Konvention

$$\sum_{\mu \leq \alpha} \dots := \sum_{\mu_1=0}^{\alpha_1} \sum_{\mu_2=0}^{\alpha_2} \dots, \quad (\alpha) = \binom{\alpha_1}{\mu_1} \binom{\alpha_2}{\mu_2}$$

für Multiindizes  $\mu, \alpha \in \mathbb{N}_0^2$ . Daraus folgt

$$(\mathbf{B}_\tau^\top \nabla)^\alpha v = \sum_{\mu \leq \alpha} \binom{\alpha}{\mu} \mathbf{b}_1^\mu \mathbf{b}_2^{\alpha-\mu} \partial_1^{|\mu|} \partial_2^{|\alpha-\mu|} v.$$

Um den Betrag punktweise abzuschätzen verwenden wir

$$|\mathbf{b}_{i,j}| \leq \|\mathbf{B}_\tau\|$$

und erhalten mit  $|\alpha| = k+1$

$$|((\partial^\alpha \hat{v}) \circ F_\tau^{-1})(\mathbf{x})|^2 = |(\mathbf{B}_\tau^\top \nabla)^\alpha v(\mathbf{x})|^2 \leq C \|\mathbf{B}_\tau\|^{2(k+1)} \sum_{\mu \leq \alpha} \left| \partial_1^{|\mu|} \partial_2^{|\alpha-\mu|} v(\mathbf{x}) \right|^2. \quad (7.14)$$

Summation über alle  $|\alpha| = k+1$  und Integration über  $\tau$  liefert (7.15).

**Theorem 7.11** *Wir betrachten das Problem (7.1) mit der Bilinearform  $a$  und Linearform  $\ell$  aus (7.2). Die Daten erfüllen die Voraussetzungen (7.3) und (7.4). Dann besitzt das Finite-Elemente-Galerkin-Verfahren:*

$$\text{Finde } u_S \in S_{\mathcal{G},0}^{k,0} \quad a(u_S, v) = \ell(v) \quad \forall v \in S_{\mathcal{G},0}^{k,0}$$

eine eindeutige Lösung, die die Fehlerabschätzungen

$$\begin{aligned} \|u - u_S\|_{H^1(\Omega)} &\leq \frac{A}{\alpha} \inf_{v \in S_{\mathcal{G},0}^{k,0}} \|u - v\|_{H^1(\Omega)} \\ \|u - u_S\|_{L^2(\Omega)} &\leq \frac{A^2}{\alpha} \inf_{v \in S_{\mathcal{G},0}^{k,0}} \|u - v\|_{H^1(\Omega)} \sup_{\varphi \in L^2 \setminus \{0\}} \inf_{v \in S_{\mathcal{G},0}^{k,0}} \frac{\|u_\varphi - v\|_{H^1(\Omega)}}{\|\varphi\|_{L^2(\Omega)}} \end{aligned}$$

erfüllt. Hier bezeichnet  $u_\varphi \in H_0^1(\Omega)$  die eindeutige Lösung des Problems

$$a(v, u_\varphi) = (\varphi, v)_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega). \quad (7.17)$$

Falls  $u \in H^{k+1}(\Omega)$  gilt, folgt die Fehlerabschätzung

$$\|u - u_S\|_{H^1(\Omega)} \leq Ch^k |u|_{H^{k+1}(\Omega)}.$$

Ist das Problem darüber hinaus  $H^2(\Omega)$ -regulär<sup>3</sup>, erhalten wir

$$\|u - u_S\|_{L^2(\Omega)} \leq Ch^{k+1} |u|_{H^{k+1}(\Omega)}.$$

Für die Bewertung von Verfahren spielen verschiedene Kriterien eine Rolle. Dazu zählen die Konvergenzgeschwindigkeit -genauer der Rechen- und Speicheraufwand eine Näherungslösung mit vorgeschriebener Genauigkeit zu berechnen, die Grösse der Problemklasse, für die das Verfahren anwendbar ist und die Robustheit bei parameterabhängigen Problemen.

Exemplarisch wollen wir im folgenden die Situation betrachten, dass der Koeffizientenvektor  $\mathbf{b}$  im Konvektionsterm  $\langle \mathbf{b}, \nabla u \rangle$  „gross“ ist verglichen zur Koeffizientenmatrix  $\mathbf{A}$  des Diffusionsterms  $\text{div}(\mathbf{A} \text{grad } u)$ . In diesem Fall beobachtet man in der Finite-Elemente-Lösung unphysikalische Oszillationen, d.h., das Verfahren ist nicht robust bezüglich dieser Parameterabhängigkeit. Wir werden im folgenden eine Modifikation erklären, bei dem diese Schwierigkeit eliminiert wird.

Dieses Verfahren wird in der Literatur als **Stromlinien-Diffusions-Methode** (engl. **Streamline-diffusion finite element method**, kurz **SDFEM**) bzw. **streamline upwind Petrov-Galerkin** Verfahren, kurz **SUPG**, bezeichnet. Für die Darstellung des zugrundeliegenden Prinzips beschränken wir uns auf den einfachsten Spezialfall konstanter Koeffizienten  $\mathbf{A} = \varepsilon \mathbf{I}$ ,  $\mathbf{b} \in \mathbb{R}^2 \setminus \{0\}$  und  $c \equiv 0$  und homogenen Dirichlet-Randbedingungen. Dabei ist die Skalierung so gewählt, dass

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

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

---

<sup>3</sup>  $H^2(\Omega)$ -regulär bedeutet, dass die Lösung von (7.17) für jede rechte Seite  $\varphi \in L^2(\Omega)$  in  $H^2(\Omega)$  ist und  $\|u\|_{H^2(\Omega)} \leq C \|\varphi\|_{L^2(\Omega)}$  gilt (vgl. Satz 4.5).

Im Rahmen von Satz 3.1 ist jetzt

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

Der diskrete Raum  $S$  ist wieder durch

$$S := S_{\mathcal{G}} := \{ \varphi \in C^0(\overline{\Omega}) \mid \forall \tau \in \mathcal{G} : \varphi|_{\tau} \in \mathbb{P}_1 \} \cap H_0^1(\Omega)$$

definiert. Wir definieren auf  $S$  eine gitterabhängige Bilinearform  $a_S$ , eine Linearform  $\ell_S$  und eine Norm  $\|\cdot\|_S$  gemäss

$$a_S(u, v) := a(u, v) + \sum_{\tau \in \mathcal{G}} \delta_{\tau} \int_{\tau} \langle \mathbf{b}, \nabla u \rangle \langle \mathbf{b}, \nabla v \rangle \quad (7.18a)$$

$$\ell_S(v) := \ell(v) + \sum_{\tau \in \mathcal{G}} \delta_{\tau} \int_{\tau} f \langle \mathbf{b}, \nabla v \rangle \quad (7.18b)$$

$$\|u\|_{1,S} := \left\{ \varepsilon |u|_{H^1(\Omega)}^2 + \sum_{\tau \in \mathcal{G}} \delta_{\tau} \|\langle \mathbf{b}, \nabla u \rangle\|_{L^2(\tau)}^2 \right\}^{1/2}. \quad (7.18c)$$

Das neue diskrete Problem lautet dann: Finde  $u_S \in S$  mit

$$a_S(u_S, v) = \ell_S(v) \quad \forall v \in S. \quad (7.19)$$

Dabei sind die  $\delta_{\tau}$  nicht-negative Parameter, die wir später fixieren werden. Die  $\delta_{\tau}$  werden zur numerischen Stabilisierung verwendet. Der zusätzliche Term in  $a_S$  entspricht einer Diskretisierung von  $-\partial^2 u / \partial \mathbf{b}^2$ , d.h. der zweiten Ableitung in Richtung von  $\mathbf{b}$ . Daher wird eine zusätzlich Diffusion in Stromrichtung eingeführt. Insbesondere kann man erhoffen, dass senkrecht zur Stromrichtung keine künstliche Diffusion, d.h. kein Verschmieren auftritt.

Wir illustrieren das Lösungsverhalten an Hand des Eigenwertproblems

$$\begin{aligned} -\varepsilon \Delta u + \partial_1 u &= \lambda u & \text{in } \Omega = (0, 1)^2, \\ u &= 0 & \text{auf } \Gamma = \partial\Omega. \end{aligned} \quad (7.20)$$

Ein Separationsansatz  $u(x, y) = v(x)w(y)$  führt auf das System gewöhnlicher Differentialgleichungen

$$-\varepsilon v'' + v' - \delta^2 v = 0 \quad \text{und} \quad -\varepsilon w'' - (\delta^2 + \lambda) w = 0$$

Die Lösungen sind durch

$$u(x, y) := e^{x/(2\varepsilon)} \left( e^{x \frac{\sqrt{1-4\delta^2\varepsilon}}{2\varepsilon}} - e^{-x \frac{\sqrt{1-4\delta^2\varepsilon}}{2\varepsilon}} \right) \left( C_8 e^{iy \sqrt{\frac{\lambda+\delta^2}{\varepsilon}}} + C_7 e^{-iy \sqrt{\frac{\lambda+\delta^2}{\varepsilon}}} \right)$$

gegeben. Die Randbedingungen führen auf

$$u(x, y) = e^{\frac{x}{2\varepsilon}} \sin(k_1 \pi x) \sin(k_2 \pi y)$$

zum Eigenwert  $\lambda = \frac{1}{4\varepsilon} + \varepsilon(k_1^2 + k_2^2)\pi^2$ . Für moderates  $k_1$  und  $k_2$  ist die Ableitung in Strömungsrichtung um die Grössenordnung  $(2\varepsilon)^{-1}$  grösser als die Ableitung in Richtung von  $y$ . Die Stromlinien verlaufen also in  $x$ -Richtung. Für die numerische Lösung beobachtet man unphysikalische Oszillationen in  $y$ -Richtung.

Formal lässt sich die Modifikation (7.18) so interpretieren, dass die Konvektions-Diffusions-Gleichung statt mit  $v$  mit  $v + \sum_{\tau \in \mathcal{G}} \delta_\tau \langle \mathbf{b}, \nabla v \rangle$  getestet wird. Daher sind im allgemeinen Fall noch zusätzliche Terme der Form

$$\sum_{\tau \in \mathcal{G}} \delta_\tau \int_\tau \{-\operatorname{div}(\mathbf{A} \operatorname{grad} u_S) + cu_S\} \langle \mathbf{b}, \nabla v \rangle$$

in  $a_S$  aufzunehmen.

**Satz 7.12** Die Bilinearform  $a_S$  ist koerziv bzgl. der  $|\cdot|_{1,S}$ -Norm, d.h.

$$a_S(u, u) \geq |u|_{1,S}^2 \quad \forall u \in S.$$

**Beweis.** Aus dem Beweis von Satz 4.3(1) ergibt sich für den aktuellen Spezialfall die Abschätzung

$$a(v, v) \geq \varepsilon |v|_{H^1(\Omega)}^2 \quad \forall v \in H_0^1(\Omega).$$

Hieraus und aus der Definition von  $a_S$  und  $|\cdot|_{1,S}$  folgt sofort die Behauptung. ■

**Bemerkung 7.13**

1. Aus Satz 7.12 folgt, dass das diskrete Problem (7.19) eindeutig lösbar ist.
2. Satz 7.12 benötigt ausser der Annahme  $\delta_\tau \geq 0$  keine weiteren Voraussetzungen an die  $\delta_\tau$ .
3. Im allgemeinen Fall bleibt Satz 7.12 gültig, wenn wie in Satz 4.3  $-\frac{1}{2} \operatorname{div} \mathbf{b} + c \geq 0$  und  $h^{-1} \|\mathbf{A}\|_{L^\infty(\tau)} \delta_\tau^{1/2} \leq 1$  ist.

Für die Fehlerabschätzung der Diskretisierung (7.19) benötigen wir eine Abschätzung der  $L^2$ -Norm des Interpolationsfehlers, wie wir sie bereits in (7.16) bewiesen haben:

1. Für alle  $v \in H^2(\hat{\tau})$  gilt

$$\|v - \hat{\Pi}_1 v\|_{L^2(\hat{\tau})} \leq C |v|_{H^2(\hat{\tau})}.$$

2. Für alle  $\tau \in \mathcal{G}$  und alle  $v \in H^2(\tau)$  gilt

$$\|v - I_{\mathcal{G}} v\|_{L^2(\tau)} \leq Ch_\tau^2 |v|_{H^2(\tau)}.$$

**Satz 7.14** Sei  $u \in H_0^1(\Omega)$  die Lösung der Konvektions-Diffusions-Gleichung mit  $\mathbf{A} = \varepsilon \mathbf{I}$ ,  $\mathbf{b} \in \mathbb{R}^2$  mit  $\|\mathbf{b}\| = 1$ ,  $c \equiv 0$  und homogenen Dirichlet-Randbedingungen und  $u_S \in S$  die Lösung der SDFEM-Diskretisierung (7.19). Es sei  $u \in H^2(\Omega)$ . Dann gilt die Fehlerabschätzung

$$|u - u_S|_{1,S} \leq C \left\{ \sum_{\tau \in \mathcal{G}} (\varepsilon^2 \delta_\tau + \varepsilon h_\tau^2 + \delta_\tau h_\tau^2 + \delta_\tau^{-1} h_\tau^4) |u|_{H^2(\tau)}^2 \right\}^{1/2}.$$

Dabei ist  $c_{\mathcal{G}}$  wie in (7.5) das Mass für die Formregularität der Triangulierung. Die Fehlerabschätzung ist optimal für die Wahl

$$\delta_\tau = \frac{h_\tau^2}{\sqrt{\varepsilon^2 + h_\tau^2}}.$$

**Beweis.** Die Dreiecksungleichung führt zur Aufspaltung

$$|u - u_S|_{1,S} \leq |u - I_{\mathcal{G}}u|_{1,S} + |I_{\mathcal{G}}u - u_S|_{1,S}.$$

Aus (7.16) folgt

$$|u - I_{\mathcal{G}}u|_{1,S} \leq C \left\{ \sum_{\tau \in \mathcal{G}} (\varepsilon + \delta_\tau) h_\tau^2 |u|_{H^2(\tau)}^2 \right\}^{1/2}.$$

Setze zur Abkürzung  $w_S := u_S - I_{\mathcal{G}}u$ . Aus Satz 7.12 folgt

$$|w_S|_{1,S}^2 \leq a_S(w_S, w_S) = a_S(u - I_{\mathcal{G}}u, w_S) + a_S(u_S - u, w_S).$$

Da  $u \in H^2(\Omega)$  ist, folgt aus der Definition von  $a_S$ ,  $\ell_S$  und (7.19)

$$\begin{aligned} a_S(u_S - u, w_S) &= \ell_S(w_S) - a_S(u, w_S) \\ &= \ell(w_S) + \sum_{\tau \in \mathcal{G}} \delta_\tau \int_\tau f \langle \mathbf{b}, \nabla w_S \rangle - a(u, w_S) - \sum_{\tau \in \mathcal{G}} \delta_\tau \int_\tau \langle \mathbf{b}, \nabla u \rangle \langle \mathbf{b}, \nabla w_S \rangle \\ &= \sum_{\tau \in \mathcal{G}} \delta_\tau \int_\tau \{f - \langle \mathbf{b}, \nabla u \rangle\} \langle \mathbf{b}, \nabla w_S \rangle = \sum_{\tau \in \mathcal{G}} \delta_\tau \int_\tau \{-\varepsilon \Delta u\} \langle \mathbf{b}, \nabla w_S \rangle \\ &\leq \sum_{\tau \in \mathcal{G}} \varepsilon \delta_\tau |u|_{H^2(\tau)} \|\langle \mathbf{b}, \nabla w_S \rangle\|_{L^2(\tau)} \leq \left\{ \sum_{\tau \in \mathcal{G}} \varepsilon^2 \delta_\tau |u|_{H^2(\tau)}^2 \right\}^{1/2} |w_S|_{1,S}. \end{aligned}$$

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

$$\begin{aligned} a_S(u - I_{\mathcal{G}}u, w_S) &= \varepsilon \int_\Omega \langle \nabla(u - I_{\mathcal{G}}u), \nabla w_S \rangle + \int_\Omega \langle \mathbf{b}, \nabla(u - I_{\mathcal{G}}u) \rangle w_S \\ &\quad + \sum_{\tau \in \mathcal{G}} \delta_\tau \int_\tau \langle \mathbf{b}, \nabla(u - I_{\mathcal{G}}u) \rangle \langle \mathbf{b}, \nabla w_S \rangle \\ &= \varepsilon \int_\Omega \langle \nabla(u - I_{\mathcal{G}}u), \nabla w_S \rangle - \int_\Omega \langle \mathbf{b}, \nabla w_S \rangle (u - I_{\mathcal{G}}u) \\ &\quad + \sum_{\tau \in \mathcal{G}} \delta_\tau \int_\tau \langle \mathbf{b}, \nabla(u - I_{\mathcal{G}}u) \rangle \langle \mathbf{b}, \nabla w_S \rangle \\ &\leq \varepsilon |u - I_{\mathcal{G}}u|_{H^1(\Omega)} |w_S|_{H^1(\Omega)} + \sum_{\tau \in \mathcal{G}} \|u - I_{\mathcal{G}}u\|_{L^2(\tau)} \|\langle \mathbf{b}, \nabla w_S \rangle\|_{L^2(\tau)} \\ &\quad + \sum_{\tau \in \mathcal{G}} \delta_\tau \|\langle \mathbf{b}, \nabla(u - I_{\mathcal{G}}u) \rangle\|_{L^2(\tau)} \|\langle \mathbf{b}, \nabla w_S \rangle\|_{L^2(\tau)} \\ &\leq 2 |w_S|_{1,S} \left\{ \sum_{\tau \in \mathcal{G}} (\varepsilon + \delta_\tau) |u - I_{\mathcal{G}}u|_{H^1(\tau)}^2 + \delta_\tau^{-1} \|u - I_{\mathcal{G}}u\|_{L^2(\tau)}^2 \right\}^{1/2} \\ &\leq 2 |w_S|_{1,S} \left\{ \sum_{\tau \in \mathcal{G}} (\varepsilon + \delta_\tau) C^2 h_\tau^2 |u|_{H^2(\tau)}^2 + \sum_{\tau \in \mathcal{G}} \delta_\tau^{-1} C h_\tau^4 |u|_{H^2(\tau)}^2 \right\}^{1/2} \\ &\leq 2 \cdot 10.5 c_{\mathcal{G}} |w|_{1,S} \left\{ \sum_{\tau \in \mathcal{G}} h_\tau^2 (\varepsilon + \delta_\tau + \delta_\tau^{-1} h_\tau^2) |u|_{H^2(\tau)}^2 \right\}^{1/2}. \end{aligned}$$

Aus den letzten beiden Abschätzungen folgt

$$|u_S - I_{\mathcal{G}}u|_{1,S} = |w|_{1,S} \leq C \left\{ \sum_{\tau \in \mathcal{G}} (\varepsilon h_{\tau}^2 + h_{\tau}^2 \delta_{\tau} + \delta_{\tau}^{-1} h_{\tau}^4 + \varepsilon^2 \delta_{\tau}) |u|_{H^2(\tau)}^2 \right\}^{1/2}.$$

Zusammen mit der bewiesenen Abschätzung für  $|u - I_{\mathcal{G}}u|_{1,S}$  folgt hieraus die Fehlerabschätzung für  $|u - u_S|_{1,S}$ . Offensichtlich ist sie optimal, wenn die  $\delta$ -Terme und  $\delta^{-1}$ -Terme gleich sind. Hieraus folgt die Behauptung über die optimale Wahl von  $\delta$ . ■

**Bemerkung 7.15** *Es gilt  $\delta_{\tau} \sim h_{\tau}$ , wenn  $\varepsilon \ll h_{\tau}$  ist. In diesem Fall liefert Satz 7.14 eine optimale Fehlerabschätzung der Form*

$$\|\langle \mathbf{b}, \nabla(u - u_S) \rangle\|_{L^2(\Omega)} \leq ch |u|_{H^2(\Omega)}$$

*in Stromrichtung. Im Fall  $h_{\tau} \ll \varepsilon$ , in dem auch die Diskretisierung (7.7) gut funktioniert, ist  $\delta_{\tau} \sim h_{\tau}^2 \varepsilon^{-1}$ , und Satz 7.14 liefert eine Fehlerabschätzung der Form*

$$|u - u_S|_{H^1(\Omega)} \leq ch |u|_{H^2(\Omega)},$$

*die vergleichbar ist zu derjenigen von Satz 7.11.*

Die folgenden beiden Beispiele illustrieren das verbesserte Konvergenzverhalten von SUPG für konvektionsdominierte Probleme verglichen zur Standard-Finite-Elemente-Methode.<sup>4</sup>

**Beispiel 7.16** *Wir betrachten zunächst ein eindimensionales Modellproblem. Sei  $\Omega = (0, 1)$  und  $0 < \varepsilon \ll 1$  ein Parameter. Aufgabe:*

$$\text{Finde } u \in H_0^1(\Omega), \text{ so dass } \int_0^1 \varepsilon u' v' + u' = \int_0^1 v \quad \forall v \in H_0^1(\Omega).$$

*Die exakte Lösung*

$$\eta(x) := x - \frac{e^{x/\varepsilon} - 1}{e^{1/\varepsilon} - 1}$$

*ist in Abb. 1 dargestellt. Die numerische Lösung mit linearen finiten Elementen auf einer äquidistanten Zerlegung des Gebiets  $\Omega = (0, 1)$  zeigt unphysikalische Oszillationen und ist unbrauchbar (siehe Abbildung 2). Abbildung 3 zeigt den Fehlerverlauf bei kleiner werdender Schrittweite. Man sieht, dass die asymptotische Konvergenzrate für kleiner werdende Diffusion immer später einstellt.*

**Beispiel 7.17** *Wir betrachten nun ein zweidimensionales Modellproblem. Sei  $\Omega = (0, 1)^2$  und  $0 < \varepsilon \ll 1$  wieder ein Parameter. Aufgabe:*

$$\text{Finde } u \in H_0^1(\Omega) \text{ mit } \int_{\Omega} \varepsilon \langle \nabla u, \nabla v \rangle + \langle \mathbf{b}, \nabla u \rangle v = \int_{\Omega} f v \quad \forall v \in H_0^1(\Omega).$$

*Wir setzen  $\mathbf{b} = (1, 1)^{\top}$  und verwenden die Funktion  $\eta$  aus Beispiel 7.16. Die exakte Lösung für die rechte Seite  $f(x, y) = \eta(x) + \eta(y)$  ist gegeben durch*

$$u(x, y) := \eta(x) \eta(y).$$

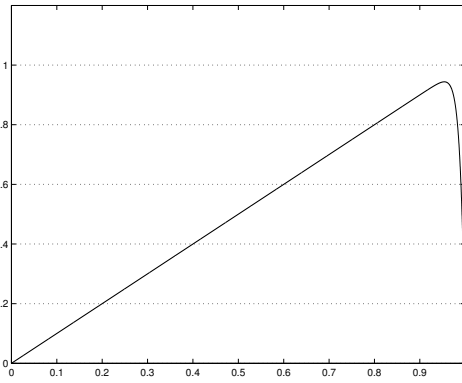


Abbildung 1: Die exakte Lösung für die eindimensionale Konvektions-Diffusionsgleichung. Der kritische Bereich ist die Rand-Grenzschicht der Breite  $O(\varepsilon)$  bei  $x = 1$ .

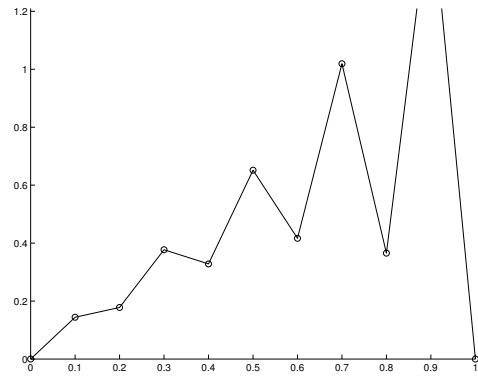


Abbildung 2: Numerische Lösung mit linearen finiten Elementen auf einem äquidistanten Gitter.

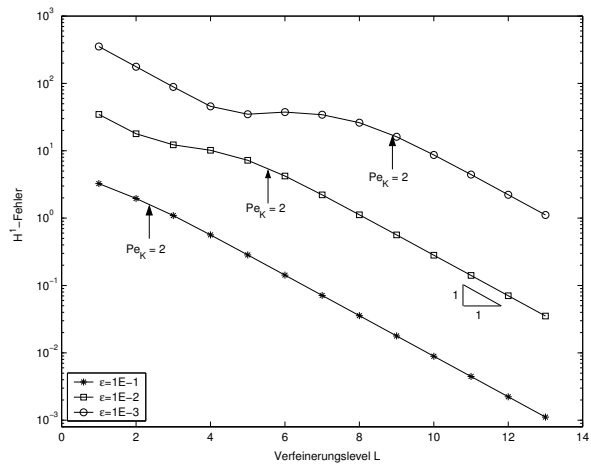


Abbildung 3:  $H^1$ -Fehler für verschiedene, moderate Parameter.



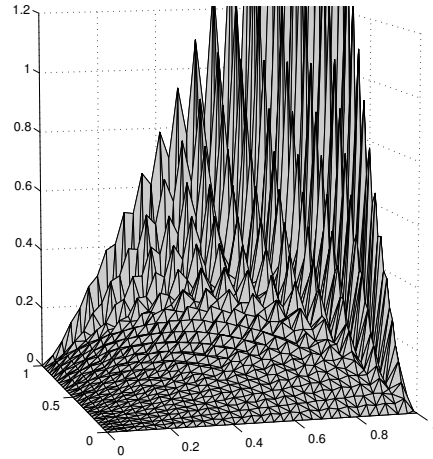


Abbildung 4: Numerische Lösung mit linearen finiten Elementen auf einem regelmässigen Dreiecksgitter.

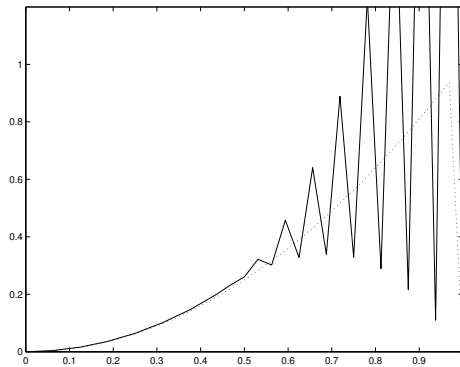


Abbildung 5: Vergleich der exakten und der numerischen Lösung mit linearen finiten Elementen auf einem regelmässigen Dreiecksgitter.

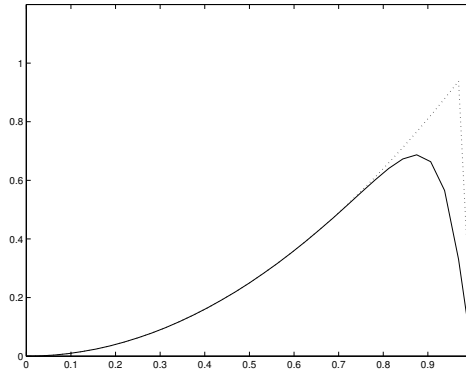
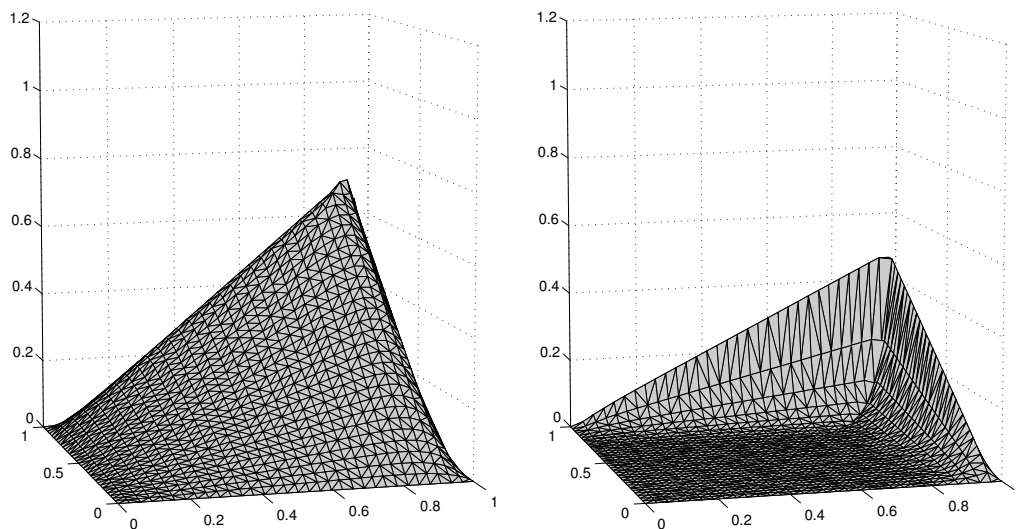


Abbildung 6: SUPG-Lösung und exakte Lösung über der Diagonalen. Es treten keine unphysikalischen Oszillationen auf und das qualitative Verhalten wird korrekt wiedergegeben.

*Diskretisiert man dieses Problem mit linearen finiten Elementen auf einem regelmässigen Dreiecksgitter zeigen sich wiederum unphysikalische Oszillationen (siehe Abbildung 4) Die Oszillationen werden noch deutlicher, wenn man die Funktion entlang der Diagonalen  $\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}$  zeichnet (siehe Abbildung 5). Qualitativ zeigt der Fehler das gleiche Verhalten wie im eindimensionalen Fall. Für kleiner werdende Diffusion stellt sich die asymptotische Konvergenzrate immer später ein.*

*Die folgende Abbildung zeigt die SUPG-Lösung links und den zugehörigen Fehler rechts.*



*Man sieht, dass die SUPG-Lösung das Verhalten der exakten Lösung wesentlich besser widerspiegelt und keine unphysikalischen Oszillationen auftreten. Der Fehler ist in der Randschicht konzentriert, wie die Graphik über der Diagonalen  $\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}$  deutlich zeigt (siehe Abb. 6)*

<sup>4</sup>Die folgenden beiden Beispiele sind der Diplomarbeit von C. Wüst entnommen, die auf <http://www.math.unizh.ch/compmath/index.php?id=dipl> heruntergeladen werden kann.

## 8 Implementierung

In diesem Kapitel gehen wir auf die Implementierung der Methoden der letzten Kapitel und die dafür benötigten Datenstrukturen ein.

Sei  $\mathcal{G}$  eine Triangulierung des polygonalen Gebietes  $\Omega \subset \mathbb{R}^2$ . Bezeichne mit  $NT$  die Anzahl aller Dreiecke und mit  $NV$  die Anzahl aller Dreieckseckpunkte. Die Dreieckseckpunkte werden beliebig numeriert. Zweidimensionale **real**-Zahlen bzw. **integer**-Zahlen definieren die Typen

$$\text{real2d} : \text{array}[1..2] \text{ of } \mathbf{real} \quad \text{und} \quad \text{int3D} : \text{array}[1..3] \text{ of } \mathbf{integer}$$

und die Koordinaten der Ecken werden im Feld

$$\mathbf{x} : \text{array}[1..NV] \text{ of } \text{real2d}$$

abgespeichert. In jedem Dreieck  $\tau \in \mathcal{G}$  werden die Eckpunkte lokal gegen den Uhrzeigersinn nummeriert. Wir definieren den Typ Dreieck durch

$$\begin{aligned} \Delta = \text{record} \\ & \text{loc2glob} : \text{array}[1..3] \text{ of } \mathbf{integer} \\ & \text{neighbor} : \text{array}[1..3] \text{ of } \mathbf{integer} \\ & \text{mark} : \text{array}[1..2] \text{ of } \text{int3D} \\ & p : \mathbf{integer}; \\ & \text{loc2globdof} : \text{array}\left[1..\frac{(p+1)(p+2)}{2}\right] \text{ of } \mathbf{integer} \\ \text{end,} \end{aligned} \tag{8.1}$$

so dass die Dreiecke im Feld

$$\tau : \text{array}[1..NT] \text{ of } \Delta$$

abgespeichert werden. Die einzelnen Grössen sind wie folgt definiert:

1.  $\tau(i).loc2glob(j) = k$ : Die  $j$ -te Ecke des Dreiecks  $i$  hat die globale Nummer  $k$  und die Koordinaten  $\mathbf{x}(k)$
2.  $\tau(i).neighbor(j) = k$ : Das Nachbardreieck von  $\tau(i)$ , welches die  $j$ -te Kante<sup>5</sup> mit  $\tau(i)$  gemeinsam hat, ist  $\tau(k)$ . Falls die  $j$ -te Kante von  $\tau(i)$  eine Randkante ist wird  $\tau(i).neighbor(j) = 0$  gesetzt<sup>6</sup>.

Moderne numerische Verfahren zur Lösung partieller Differentialgleichungen sind adaptiv; das heisst – ausgehend von einer sehr groben Triangulierung des Gebiets – wird das Gitter sukzessive verfeinert (oder auch der lokale Polynomgrad variiert). Das hat zwei positive Implikationen: 1) Das Gitter kann problemangepasst verfeinert werden, falls ein „Fehlerschätzer“ lokale Informationen über den Fehler der numerischen Lösung liefert und 2) zur Lösung des Problems auf der feinsten Stufe steht die ganze Diskretisierungshierarchie zur Verfügung, und es lassen sich Mehrgitterverfahren zur numerischen Lösung einsetzen. Dazu müssen aber mehrere Gitter  $(\mathcal{G}_\ell)_{\ell=0}^L$  und geometrische Abhängigkeiten zwischen diesen Gittern verwaltet werden. Für Dreiecke existieren verschiedene Verfeinerungsmuster, die in Abbildung 7 dargestellt sind.

Wir definieren die Grösse  $mark$  in (8.1) wie folgt

<sup>5</sup>Die  $j$ -te Kante eines Dreiecks ist diejenige, welche der  $j$ -ten Ecke gegenüber liegt.

<sup>6</sup>Für Kanten auf dem Rand von  $\Gamma$  kann  $neighbor(j)$  noch verfeinert genutzt werden, um den Typ des Randes, (Dirichlet/Neumann-Kante oder/und die gegebenen Randdaten) zu beschreiben.

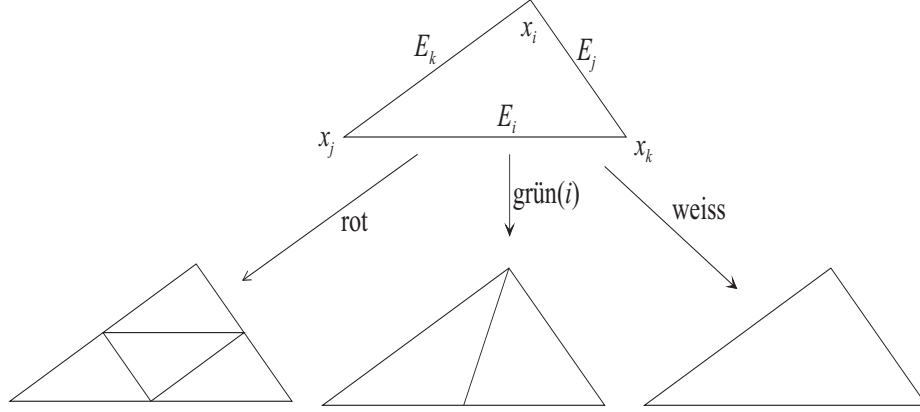


Abbildung 7: Verschiedene Verfeinerungsmuster für Dreiecke:  $(4, 4, 4)$ : Zerlegung in vier kongruente Dreiecke,  $(i, j, k)$ : Verbindung des  $i$ -ten Eckpunkts mit der gegenüberliegenden Kantenmitte  $\mathbf{m}_i$  und (für  $j \neq 0$ ) Verbindung  $\overline{\mathbf{m}_i, \mathbf{m}_j}$  und (für  $k \neq 0$ ) Verbindung  $\overline{\mathbf{m}_i, \mathbf{m}_k}$ . Für  $(0, 0, 0)$ : keine Verfeinerung.

1.  $\tau(k) \cdot \text{mark}(1) = (j, \ell)$ . Das Dreieck  $\tau(k)$  ist durch Unterteilung des Dreiecks  $\tau_j$  der vorigen Stufe entstanden und das Dreieck  $\tau(\ell)$  ist ein „Kind“ von  $\tau(k)$  auf der feineren Stufe.
2.  $\tau(k) \cdot \text{mark}(2) = (\ell, m, n)$ . Wir unterscheiden die Fälle<sup>7</sup>
  - (a)  $(\ell, m, n) = (0, 0, 0)$ . Das Dreieck  $\tau(k)$  wird nicht unterteilt.
  - (b)  $(\ell, 0, 0)$ ,  $1 \leq \ell \leq 3$ . Wir setzen  $\ell_{\text{glob}} := \tau(k) \cdot \text{loc2glob}(\ell)$ . Das Dreieck  $\tau(k)$  wird durch die Strecke  $\mathbf{x}(\ell_{\text{glob}}) \mathbf{m}_{k,\ell}$  unterteilt.
  - (c)  $(\ell, m, 0)$ ,  $1 \leq \ell, m \leq 3$ ,  $\ell \neq m$ . Wir setzen  $\ell_{\text{glob}} := \tau(k) \cdot \text{loc2glob}(\ell)$ . Das Dreieck  $\tau(k)$  wird unterteilt, indem die Strecken  $\mathbf{x}(\ell_{\text{glob}}) \mathbf{m}_{k,\ell}$  und  $\overline{\mathbf{m}_{k,\ell} \mathbf{m}_{k,m}}$  eingefügt werden.
  - (d)  $(\ell, m, n)$ ,  $1 \leq \ell, m, n \leq 3$ , mit paarweise verschiedenen  $\ell, m, n$ . Wir setzen  $\ell_{\text{glob}} := \tau(k) \cdot \text{loc2glob}(\ell)$  und zerteilen  $\tau(k)$  durch Einfügen der Strecken  $\mathbf{x}(\ell_{\text{glob}}) \mathbf{m}_{k,\ell}$  und  $\overline{\mathbf{m}_{k,\ell} \mathbf{m}_{k,m}}$  und  $\overline{\mathbf{m}_{k,\ell} \mathbf{m}_{k,n}}$ .
  - (e)  $(\ell, m, n) = (4, 4, 4)$ . Wir zerteilen  $\tau(k)$  in vier kongruente Dreiecke durch Verbinden der Seitenmitten, durch Einfügen der Strecken  $\overline{\mathbf{m}_{k,1} \mathbf{m}_{k,2}}$ ,  $\overline{\mathbf{m}_{k,2} \mathbf{m}_{k,3}}$ ,  $\overline{\mathbf{m}_{k,3} \mathbf{m}_{k,1}}$ .

Die grösste Triangulierung kann auf zwei Weisen erzeugt werden. Bei der ersten Möglichkeit gibt man die Daten für die Dreiecke und Eckpunkte des grössten Gitters von Hand ein. Nach Eingabe der grössten Triangulierung werden weitere Triangulierungen durch gleichmässige oder adaptive Verfeinerung erzeugt.

Bei der zweiten Variante erzeugt man die grösste Triangulierung automatisch aus einer orientierten Liste von Eckpunkten des Randpolygons  $\Gamma = \partial\Omega$ . Weitere Gitterpunkte und die Triangulierung werden aufgrund einer der folgenden beiden Strategien, die auch kombiniert werden können, erzeugt:

<sup>7</sup>Konvention: Die Kante, welchem dem  $j$ -ten Eckpunkt des Dreiecks  $\tau(k)$  gegenüberliegt, wird mit  $E_{k,j}$  bezeichnet und dessen Mittelpunkt mit  $\mathbf{m}_{k,j}$ .

- a. Schlage eine Ecke ab d.h. wähle ein  $i$ , verbinde die Punkte  $i - 1$  und  $i + 1$  der Liste und erzeuge so ein Dreieck, dessen Eckpunkte die Punkte  $i - 1, i$  und  $i + 1$  der Liste sind. Streiche aus der Liste der Punkte auf dem Randpolygon den Punkt  $i$ .
- b. Verbinde alle Punkte der aktuellen Liste mit ihrem Schwerpunkt.

Beide Varianten sind natürlich nur zulässig, wenn die so erzeugten Dreiecke in  $\Omega$  liegen. Zusätzlich sollte man die Qualität der erzeugten Dreiecke kontrollieren, um zu spitze oder zu stumpfe Dreiecke zu vermeiden (um die Konstante  $c_{\mathcal{G}}$  zu kontrollieren, welche in den Fehlerabschätzungen eine wesentliche Rolle spielt). In einer Implementierung ist die Bestimmung des Inkreisdurchmessers etwas unhandlich, und man sollte das (äquivalente) Mass für die Güte eines Dreiecks verwenden:

$$c_{\tau} = \frac{\sum_{i=1}^3 \|x_{\tau}^i - x_{\tau}^{i+1}\|^2}{\text{area}(\tau)}.$$

Man beachte, dass zur Berechnung von  $c_{\tau}$  lediglich elementare arithmetische Operationen ( $+$ ,  $-$ ,  $\times$ ,  $\div$ ) erforderlich sind. Der Verfeinerungsalgorithmus kann dann wie folgt durchgeführt werden. Wir nehmen an, dass eine Teilmenge von  $\mathcal{G}$  zum Verfeinern (gemäß Strategie 2d oder 2e) markiert ist. Das folgende Unterprogramm berechnet eine globale Markierung ohne hängende Knoten, d.h. einen „grünen Abschluss“.

**procedure mark;**

**begin**

1: **for all**  $\tau(k) \in \mathcal{G}$  **do**

**for**  $j := 1$  **to** 3 **do** (\*Schleife über alle Nachbarn von  $\tau(k)$ \*)

**if** “ $\tau(k)$ . mark (2) ist derart, dass  $\mathbf{m}_{k,j}$  beim Verfeinern nicht generiert werden würde” **then begin**

$\ell := \tau(k)$ . neighbor ( $j$ ); (\* $\tau(\ell)$  ist der Nachbar von  $\tau(k)$  über  $E_{k,j}$ \*)

berechne  $m \in \{1, 2, 3\}$  mit  $k = \tau(\ell)$ . neighbor ( $m$ );

**if** “ $\tau(\ell)$ . mark (2) ist derart, dass  $\mathbf{m}_{\ell,m}$  beim Verfeinern generiert werden würde” **then**

ändere  $\tau(k)$ . mark (2) in eine “nächst höhere” Strategie,  
die  $\mathbf{m}_{k,j}$  als Verfeinerungspunkt enthält;

**end**

**if** “Markierung von  $\mathcal{G}$  enthält noch hängende Knoten” **then goto 1**

**end**

Für das Aufstellen des linearen Gleichungssystems muss man die Einträge der Rechte-Seite-Vektors  $\mathbf{r} \in \mathbb{R}^N$  und der Systemmatrix  $\mathbf{A} = (a_{i,j})_{i,j=1}^N$  berechnen.

$$r_i = \ell(b_i) \quad \text{und} \quad a_{i,j} := a(b_j, b_i).$$

Dabei bezeichnet  $b_i$  wieder die Basisfunktion zum  $i$ -ten Knoten. Bezeichne dazu für  $\tau \in \mathcal{G}$  mit  $\ell_{\tau}$  und  $a_{\tau}$  die Einschränkung des Funktionals  $\ell$  bzw. der Bilinearform  $a$  auf das Element  $\tau$ :

$$\ell_{\tau}(v) := \int_{\tau} f v dx \quad \text{und} \quad a_{\tau}(u, v) := \int_{\tau} \langle \nabla v, \mathbf{A} \nabla u \rangle + \langle \mathbf{b}, \nabla u \rangle v + c v.$$

Hier sind wir von Dirichlet-Randbedingungen ausgegangen; im Fall inhomogener Neumann-Randbedingungen muss das Funktional  $\ell$  durch ein Randintegral ergänzt werden. Es muss

nun zwischen Geometrie und Algebra unterschieden werden: Die *Geometrie* eines Dreiecks ist durch die drei Eckpunkte definiert und durch das Feld `loc2glob` realisiert. Die algebraischen Freiheitsgrade sind die Knotenpunkte auf dem Dreieck für die Polynominterpolation. Diese werden lokal auf dem Dreieck durch das Feld `loc2globdof` verwaltet, welches wiederum auf den globalen Eintrag im Koeffizientenvektor für Finite-Elemente-Funktionen verweist. Wir definieren den Datentyp

```
dof _ type = record
    r : real
    u : real
    sys : matrix_type
end
```

und das Feld

```
dof : array [1..NDOF] of dof _ type;
```

Der Datentyp `matrix_type` dient zur Abspeicherung der Matrix in einem schwachbesetzten Format. `dof(i).sys` speichert die  $i$ -te Zeile der Systemmatrix. Sei  $Z_{\max}$  die maximale Anzahl der Nicht-Null-Elemente in einer Matrixzeile der Systemmatrix. Dann definieren wir

```
element_type = record
    pos : integer
    value : real
end
```

und

```
matrix_type = record
    count : integer
    el : array [1..Z_max] of element_type
end
```

Wir nehmen an, dass die algebraischen Strukturen zu Beginn auf Null initialisiert sind. Die Berechnung der rechten Seite erfolgt dann nach folgendem Schema:

```
for all  $\tau(k) \in \mathcal{G}$  do begin
     $p_{\text{loc}} := \tau(k).p$ ;
    for  $i = 1$  to  $\binom{p_{\text{loc}}+2}{2}$  do begin (*Schleife über alle Freiheitsgrade von  $\tau(k)$ *)
         $j := \tau(k).loc2globdof(i)$ ;
         $dof(j).r := dof(j).r + \ell_{\tau_k}(b_{k,i})$ ; (* $b_{k,i}$  ist die  $i$ -te lokale Basisfunktion auf  $\tau(k)$ *)
    end end.
```

Dabei wird  $\ell_{\tau_k}(b_{k,i})$  mit einer Quadraturformel auf dem Dreieck  $\tau_k$  berechnet. Eine Möglichkeit ist beispielsweise die Schwerpunktsregel

$$\int_{\tau} g(x) \approx \text{area}(\tau) \times g(M_{\tau}), \quad M_{\tau} : \text{Schwerpunkt von } \tau$$

oder die Kantenmittelpunktsregel

$$\int_{\tau} g(x) = \frac{\text{area}(\tau)}{3} \sum_{i=1}^3 g(m_i), \quad m_i : \text{Mittelpunkt der Kante } E_i.$$

**Übungsaufgabe 8.1** Berechnen Sie den Exaktheitsgrad der Schwerpunktsregel und der Kantenmittelpunktsregel.

**Bemerkung 8.2** Das Ersetzen der exakten Integrale durch Quadraturverfahren lässt sich als Störung der Finite-Elemente-Methode interpretieren. Die Genauigkeit der Quadraturverfahren muss daher so gewählt werden, dass die Konvergenzordnung der Gesamtdiskretisierung erhalten bleibt. Für lineare Elemente ist dies durch die Kantenmittelpunktsregel gesichert, vorausgesetzt, die rechte Seite  $f$ , eingeschränkt auf ein beliebiges Dreieck  $\tau \in \mathcal{G}$ , ist in  $C^2(\bar{\tau})$ .

Ganz analog berechnet man die Systemmatrix gemäss folgendem Schema:

```

for all  $\tau(k) \in \mathcal{G}$  do begin
   $p_{\text{loc}} := \tau(k).p$ ;
  for  $i = 1$  to  $\binom{p_{\text{loc}}+2}{2}$  do begin
     $n := \tau(k).loc2globdof(i)$ ;
     $c := dof(n).sys.count$ ;
    for  $j = 1$  to  $\binom{p_{\text{loc}}+2}{2}$  do begin
       $m := \tau(k).loc2globdof(j)$ ;
      if "es gibt  $1 \leq \ell \leq c$  mit  $m = dof(n).sys.el(\ell).pos$ " then
         $dof(n).sys.el(\ell).value := dof(n).sys.el(\ell).value + a_{\tau_k}(b_{k,j}, b_{k,i})$ 
      else begin
         $c := c + 1$ ;
         $dof(n).sys.el(c).value := a_{\tau_k}(b_{k,j}, b_{k,i})$ ;
         $dof(n).sys.el(c).pos := m$ ;
         $dof(n).sys.count := c$ ;
      end
    end
  end
end

```

Dabei werden die Grössen  $a_{\tau_k}(b_{k,j}, b_{k,i})$  wieder mittels numerischer Integration berechnet.

Mit Hilfe dieser Datenstrukturen lässt sich eine Matrix-Vektor-Multiplikation  $\mathbf{y} = \mathbf{A}\mathbf{u}$ , die man beispielsweise für die iterative Lösung linearer Gleichungssysteme benötigt, wie folgt berechnen (Der Typ `dof _ type` wird noch um die Komponente  $y$  ergänzt.).

```

procedure mat _ vec _ mult;
begin
  for  $i := 1$  to  $NDOF$  do begin
     $dof(i).y = 0$ ;
    for  $j := 1$  to  $dof(i).sys.count$  do
       $dof(i).y := dof(i).y + dof(i).sys.el(j).value * dof(dof(i).sys.el(j).pos).u$ 
    end;
  end;
end;

```

## 9 A posteriori Fehlerschätzung und adaptive Gitterverfeinerung

Bisher haben wir in den beiden Finite-Elemente-Vorlesungen immer a-priori-Fehlerabschätzungen hergeleitet, ohne dass Kenntnis der berechneten Lösung verwendet wurde. Die Fehlerabschätzungen sind alle **asymptotisch**, d.h., sie sagen etwas über die Konvergenzgeschwindigkeit des Fehlers aus, wenn die Elementgrößen gegen Null streben. Für ein gegebenes Problem und eine gegebene Unterteilung sind sie aber unbrauchbar, da sie u.a. von Normen der unbekanntenen Lösung der Differentialgleichung abhängen. Für die Praxis stellt sich aber natürlich die Frage nach dem tatsächlichen Fehler der berechneten Finite-Elemente-Lösung. Zudem will man i.a. eine bestimmte Genauigkeit mit minimalem Aufwand, d.h. möglichst wenigen Elementen erreichen. Diese Fragen werden durch **a posteriori Fehlerabschätzungen** und **adaptive Gitterverfeinerungstechniken** gelöst. Hiermit wollen wir uns in diesem Abschnitt beschäftigen.

Um die wesentlichen Prinzipien herauszuarbeiten und um technische Schwierigkeiten zu vermeiden, beschränken wir uns auf die Reaktions-Diffusions-Gleichung in  $\Omega \subset \mathbb{R}^2$  mit homogenen Dirichlet-Randbedingungen und Dreieckselemente der Ordnung  $k \geq 1$ , d.h.  $\mathcal{G}$  besteht aus Dreiecken und  $S = S_{\mathcal{G},0}^{k,0}$ . Zudem nehmen wir im ganzen Kapitel an, dass die Diffusionsmatrix  $\mathbf{A}$  und der Reaktionskoeffizient konstant sind.

Zunächst müssen wir einige zusätzliche Notationen einführen. Die Menge aller Dreieckskanten, die im Innern von  $\Omega$  liegen, bezeichnen wir mit  $\mathcal{E}_\Omega$ . Jedes  $E$  ist dann die gemeinsame Kante von genau zwei Dreiecken, die wir mit  $\tau_{E,1}$  und  $\tau_{E,2}$  bezeichnen. Jedem  $E \in \mathcal{E}_\Omega$  ordnen wir einen Einheitsvektor  $n_E$ , der senkrecht auf  $E$  steht, zu und bezeichnen für stückweise stetige Funktionen  $\varphi$  mit  $[\varphi]_E$  den Sprung von  $\varphi$  über  $E$  in Richtung  $n_E$ , d.h.

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

Der Sprung  $[\varphi]_E$  hängt von der Orientierung von  $n_E$  ab. Größen der Form  $[\langle n_E, \nabla \varphi \rangle]_E$  sind aber von der Orientierung von  $n_E$  unabhängig. Für  $E \in \mathcal{E}_\Omega$  bezeichnet  $h_E$  die Länge von  $E$ . Wegen der Regularitätsbedingungen (Formregularität) an das Gitter, können Größen der Form  $h_\tau/h_{\tau'}$  und  $h_\tau/h_E$  mit  $\bar{\tau} \cap \bar{\tau}' \neq \emptyset$  und  $\bar{E} \cap \bar{\tau} \neq \emptyset$  nach oben und unten durch die Konstanten  $c_{\mathcal{G}}$  abgeschätzt werden. Schliesslich benötigen wir verschiedene Umgebungen von Punkten  $z \in \Theta_\Omega$ , Kanten  $E \in \mathcal{E}_\Omega$  und Dreiecken  $\tau \in \mathcal{G}$ :

$$\begin{aligned} \omega_z &:= \text{Tr } b_z = \overline{\bigcup_{\substack{\tau' \in \mathcal{G} \\ z \in \tau'}} \tau'}, & \omega_\tau &:= \overline{\bigcup_{\bar{\tau} \cap \bar{\tau}' \in \mathcal{E}_\Omega} \tau'}, & \tilde{\omega}_\tau &:= \overline{\bigcup_{\bar{\tau} \cap \bar{\tau}' \neq \emptyset} \tau'}, \\ \omega_E &:= \overline{\bigcup_{E \subset \partial \tau'} \tau'}, & \tilde{\omega}_E &:= \overline{\bigcup_{\bar{E} \cap \bar{\tau}' \neq \emptyset} \tau'}. \end{aligned}$$

Dabei ist  $b_z \in S_{\mathcal{G},0}^{1,0}$  die Basisfunktion zu  $z$ , und  $\bar{\tau} \cap \bar{\tau}' \in \mathcal{E}_\Omega$  bedeutet, dass  $\bar{\tau}$  und  $\bar{\tau}'$  eine Kante gemeinsam haben.

Im folgenden bezeichnen  $u \in H_0^1(\Omega)$  und  $u_S \in S$  die schwache Lösung der Reaktions-Diffusions-Gleichung und ihre Finite-Elemente-Approximation. Die Koerzivität von  $a(\cdot, \cdot)$  impliziert die Existenz von  $\alpha > 0$ , so dass

$$\alpha \|u - u_S\|_{H^1(\Omega)}^2 \leq a(u - u_S, u - u_S)$$



gilt.  $\alpha$  hängt von  $\Omega$  und den Koeffizienten  $\mathbf{A}$  und  $c$  des Differentialoperators ab. Aus dieser Abschätzung folgt insbesondere die **Stabilität**, d.h.

$$\|u - u_S\|_{H^1(\Omega)} \leq \frac{1}{\alpha} \sup_{\substack{v \in H_0^1(\Omega) \\ \|v\|_{H^1(\Omega)}=1}} a(u - u_S, v). \quad (9.1)$$

Da  $S \subset H_0^1(\Omega)$  ist, haben wir **Galerkin-Orthogonalität** des Fehlers, d.h.

$$a(u - u_S, v) = 0 \quad \forall v \in S. \quad (9.2)$$

Die Abbildung  $H_0^1(\Omega) \ni v \rightarrow a(u - u_S, v) = \ell(v) - a(u_S, v) \in \mathbb{R}$  definiert ein stetiges lineares Funktional auf  $H_0^1(\Omega)$ . Dies ist das **Residuum**. Als nächstes wollen wir eine Darstellung des Residuums angeben, die praktisch handhabbar ist. Sei dazu  $v \in H_0^1(\Omega)$  mit  $\|v\|_{H^1(\Omega)} = 1$  beliebig, aber fest. Anwenden des Gaussischen Integralsatzes auf jedem Element  $\tau \in \mathcal{G}$  liefert

$$\begin{aligned} a(u - u_S, v) &= \ell(v) - a(u_S, v) = \int_{\Omega} f v - \int_{\Omega} \{ \langle \nabla v_S, \mathbf{A} \nabla u_S \rangle + c u_S v \} \\ &= \sum_{\tau \in \mathcal{G}} \left\{ \int_{\tau} f v - \int_{\tau} \langle \nabla u_S, \mathbf{A} \nabla v \rangle - \int_{\tau} c u_S v \right\} \\ &= \sum_{\tau \in \mathcal{G}} \left\{ \int_{\tau} f v + \int_{\tau} \operatorname{div}(\mathbf{A} \nabla u_S) v - \int_{\partial\tau} \langle n_{\tau}, \mathbf{A} \nabla u_S \rangle v - \int_{\tau} c u_S v \right\} \\ &= \sum_{\tau \in \mathcal{G}} \int_{\tau} \{ f + \operatorname{div}(\mathbf{A} \nabla u_S) - c u_S \} v + \sum_{E \in \mathcal{E}_{\Omega}} \int_E [\langle n_E, \mathbf{A} \nabla u_S \rangle]_E v. \end{aligned} \quad (9.3)$$

Dabei bezeichnet  $n_{\tau}$  die äussere Normale zu  $\tau$ . Wegen der Galerkin-Orthogonalität (9.2) können wir auf der rechten Seite von (9.3) ein beliebiges Element  $v_S \in S$  von  $v$  subtrahieren. Aus der Cauchy-Schwarzschen Ungleichung für Integrale folgt dann

$$\begin{aligned} a(u - u_S, v) &\leq \sum_{\tau \in \mathcal{G}} \|f + \operatorname{div}(\mathbf{A} \nabla u_S) - c u_S\|_{L^2(\tau)} \|v - v_S\|_{L^2(\tau)} \\ &\quad + \sum_{E \in \mathcal{E}_{\Omega}} \|[\langle n_E, \mathbf{A} \nabla u_S \rangle]_E\|_{L^2(E)} \|v - v_S\|_{L^2(E)}. \end{aligned} \quad (9.4)$$

Als nächstes müssen wir  $v_S \in S$  geschickt wählen, so dass die Normen  $\|v - v_S\|_{L^2(\tau)}$  und  $\|v - v_S\|_{L^2(E)}$  möglichst klein werden. Eine erste Idee wäre für  $v_S$  die Interpolierende  $I_S v$  zu wählen. Wegen  $v \in H_0^1(\Omega) \not\subset C^0(\overline{\Omega})$  ist dies jedoch im allgemeinen nicht möglich. Statt dessen verwenden wir den **Quasi-Interpolationsoperator**  $R_S : H_0^1(\Omega) \rightarrow S_{\mathcal{G},0}^{1,0}$  wie folgt. Für jedes  $z \in \Theta_{\Omega}$  bezeichnen wir mit  $\pi_z : L^2(\omega_z) \rightarrow \mathbb{R}$  die orthogonale Projektion, d.h.

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

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

$$R_S \varphi := \sum_{z \in \Theta_{\Omega}} (\pi_z \varphi) b_z. \quad (9.5)$$

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

**Satz 9.1** *Es gibt zwei Konstanten  $c_1, c_2$ , die nur von der Konstanten  $c_{\mathcal{G}}$  in der Formregularitätsbedingung von  $\mathcal{G}$  abhängen, so dass für alle  $v \in H_0^1(\Omega)$ , alle Dreiecke  $\tau \in \mathcal{G}$  und alle Kanten  $E \in \mathcal{E}_\Omega$  gilt*

$$\begin{aligned}\|v - R_S v\|_{L^2(\tau)} &\leq c_1 h_\tau |v|_{H^1(\tilde{\omega}_\tau)}, \\ \|v - R_S v\|_{L^2(E)} &\leq c_2 h_E^{1/2} |v|_{H^1(\tilde{\omega}_E)}.\end{aligned}$$

**Beweis. 1. Schritt:** Aus der Poincaréschen Ungleichung folgt für jedes  $z \in \Theta_\Omega$  und jedes  $\varphi \in H^1(\omega_z)$

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

Wegen der Regularitätsannahme an  $\mathcal{G}$  gibt es eine Konstante  $c'$  mit  $\operatorname{diam} \omega_z \leq c' h_\tau$  für jedes Dreieck  $\tau$ , das  $z$  als Eckpunkt hat.

**2. Schritt:** Bezeichne mit  $\hat{E}$  die horizontale Kante des Referenzdreiecks  $\hat{\tau}$ . Wegen des Spursatzes gibt es eine Konstante  $\hat{c}$ , so dass für alle  $\varphi \in H^1(\hat{\tau})$  gilt

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

Seien nun  $\tau \in \mathcal{G}$  ein beliebiges Dreieck,  $E$  eine Kante von  $\tau$  und  $\varphi \in H^1(\tau)$ . Wähle die affine Transformation  $\chi_\tau : \hat{\tau} \rightarrow \tau$  so, dass  $\hat{E}$  auf  $E$  abgebildet wird. Setze  $\hat{\varphi} := \varphi \circ \chi_\tau \in H^1(\hat{\tau})$ . Wegen der Regularitätsannahme an  $\mathcal{G}$  gilt

$$h_E^{1/2} |\det D\chi_\tau|^{-1/2} \leq c h_\tau^{-1/2}, \quad h_E^{1/2} |\det D\chi_\tau|^{-1/2} \|D\chi_\tau\| \leq c h_\tau^{1/2}.$$

Damit folgt durch Transformation auf  $\hat{\tau}$ :

$$\begin{aligned}\|\varphi\|_{L^2(E)} &= h_E^{1/2} \|\hat{\varphi}\|_{L^2(\hat{E})} \leq \hat{c} h_E^{1/2} \|\hat{\varphi}\|_{H^1(\hat{\tau})} = \hat{c} h_E^{1/2} \left\{ \|\hat{\varphi}\|_{L^2(\hat{\tau})}^2 + |\hat{\varphi}|_{H^1(\hat{\tau})}^2 \right\}^{1/2} \\ &\leq \hat{c} h_E^{1/2} \left\{ |\det D\chi_\tau|^{-1} \|\varphi\|_{L^2(\tau)}^2 + |\det D\chi_\tau|^{-1} \|D\chi_\tau\|^2 |\varphi|_{H^1(\tau)}^2 \right\}^{1/2} \\ &\leq \hat{c} c \left\{ h_\tau^{-1/2} \|\varphi\|_{L^2(\tau)} + h_\tau^{1/2} |\varphi|_{H^1(\tau)} \right\}.\end{aligned}$$

**3. Schritt:** Sei  $\tau \in \mathcal{G}$  ein Dreieck, das keinen Eckpunkt auf dem Rand  $\Gamma$  hat. Bezeichne die Menge der Eckpunkte von  $\tau$  mit  $\Theta_\tau$ . Dann ist

$$\sum_{z \in \Theta_\tau} b_z = 1 \quad \text{auf } \tau.$$

Dann folgt aus Schritt 1

$$\begin{aligned}\|v - R_S v\|_{L^2(\tau)} &= \left\| \sum_{z \in \Theta_\tau} b_z (v - \pi_z v) \right\|_{L^2(\tau)} \leq \sum_{z \in \Theta_\tau} \|b_z (v - \pi_z v)\|_{L^2(\tau)} \\ &\leq \sum_{z \in \Theta_\tau} \|v - \pi_z v\|_{L^2(\tau)} \leq \sum_{z \in \Theta_\tau} \|v - \pi_z v\|_{L^2(\omega_z)} \leq \sum_{z \in \Theta_\tau} c h_\tau |v|_{H^1(\omega_z)} \\ &\leq c' h_\tau |v|_{H^1(\tilde{\omega}_\tau)}.\end{aligned}$$

**4. Schritt:** Betrachte nun ein Dreieck  $\tau$ , das mindestens einen Eckpunkt auf dem Rand  $\Gamma$  hat. Dann gilt auf  $\tau$

$$v - R_S v = \sum_{z \in \Theta_\tau} b_z v - \sum_{z \in \Theta_\tau \cap \Theta_\Omega} b_z \pi_z v = \sum_{z \in \Theta_\tau} b_z (v - \pi_z v) + \sum_{z \in \Theta_\tau \setminus \Theta_\Omega} b_z \pi_z v$$

und somit

$$\|v - R_S v\|_{L^2(\tau)} \leq \sum_{z \in \Theta_\tau} \|b_z(v - \pi_z v)\|_{L^2(\tau)} + \sum_{z \in \Theta_\tau \setminus \Theta_\Omega} \|b_z \pi_z v\|_{L^2(\tau)}.$$

Der erste Summand wurde bereits in Schritt 3 abgeschätzt. Sei also  $z \in \Theta_\tau \setminus \Theta_\Omega$  ein Eckpunkt von  $\tau$ , der auf  $\Gamma$  liegt. Dann ist

$$\|b_z \pi_z v\|_{L^2(\tau)} \leq ch_\tau |\pi_z v|.$$

Da  $z \in \Gamma$  ist, gibt es ein Dreieck  $\tau' \in \mathcal{G}$  und eine Kante  $E'$  von  $\tau'$ , so dass  $z$  ein Endpunkt von  $E'$  und  $E' \subset \Gamma$  ist. Da  $v$  auf  $E'$  verschwindet, folgt mit Schritt 2

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

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

**5 Schritt:** Sei  $E \in \mathcal{E}_\Omega$  eine Kante die keinen Eckpunkt auf dem Rand  $\Gamma$  hat. Bezeichne mit  $\Theta_E$  die Menge der Eckpunkte von  $E$ . Dann ist

$$\sum_{z \in \Theta_E} b_z = 1 \quad \text{auf } E.$$

Hieraus und aus Schritt 2 folgt

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

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

**6. Schritt:** Betrachte nun eine Kante  $E$ , die einen Eckpunkt auf dem Rand  $\Gamma$  hat. Dann ist auf  $E$

$$v - R_S v = \sum_{z \in \Theta_E} b_z (v - \pi_z v) + \sum_{z \in \Theta_E \setminus \Theta_\Omega} b_z \pi_z v.$$

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

Wir greifen nun die Abschätzung (9.4) wieder auf. Wir setzen  $v_S = R_S v$ , benutzen Satz

9.1 und wenden die Cauchy-Schwarzsche Ungleichung für endliche Summen an:

$$\begin{aligned}
a(u - u_S, v) &\leq c_1 \sum_{\tau \in \mathcal{G}} h_\tau \|f + \operatorname{div}(\mathbf{A} \nabla u_S) - cu_S\|_{L^2(\tau)} |v|_{H^1(\tilde{\omega}_\tau)} \\
&\quad + c_2 \sum_{E \in \mathcal{E}_\Omega} h_E^{1/2} \|[\langle n_E, \mathbf{A} \nabla u_S \rangle]_E\|_{L^2(E)} |v|_{H^1(\tilde{\omega}_E)} \\
&\leq c_1 \left\{ \sum_{\tau \in \mathcal{G}} h_\tau^2 \|f + \operatorname{div}(\mathbf{A} \nabla u_S) - cu_S\|_{L^2(\tau)}^2 \right\}^{1/2} \left\{ \sum_{\tau \in \mathcal{G}} |v|_{H^1(\tilde{\omega}_\tau)}^2 \right\}^{1/2} \\
&\quad + c_2 \left\{ \sum_{E \in \mathcal{E}_\Omega} h_E \|[\langle n_E, \mathbf{A} \nabla u_S \rangle]_E\|_{L^2(E)}^2 \right\}^{1/2} \left\{ \sum_{E \in \mathcal{E}_\Omega} |v|_{H^1(\tilde{\omega}_E)}^2 \right\}^{1/2} \\
&\leq c' |v|_{H^1(\Omega)} \left\{ \sum_{\tau \in \mathcal{G}} h_\tau^2 \|f + \operatorname{div}(\mathbf{A} \nabla u_S) - cu_S\|_{L^2(\tau)}^2 + \sum_{E \in \mathcal{E}_\Omega} h_E \|[\langle n_E, \mathbf{A} \nabla u_S \rangle]_E\|_{L^2(E)}^2 \right\}^{1/2}.
\end{aligned}$$

Dabei haben wir im letzten Schritt die Formregularität von  $\mathcal{G}$  ausgenutzt. Hieraus und aus (9.1) folgt insgesamt

$$\|u - u_S\|_{H^1(\Omega)} \leq c\eta \quad (9.6)$$

mit

$$\begin{aligned}
\eta &:= \left\{ \sum_{\tau \in \mathcal{G}} \eta_\tau^2 \right\}^{1/2}, \\
\eta_\tau &:= \left\{ h_\tau^2 \|f + \operatorname{div}(\mathbf{A} \nabla u_S) - cu_S\|_{L^2(\tau)}^2 + \frac{1}{2} \sum_{E \subset \partial\tau \setminus \Gamma} h_E \|[\langle n_E, \mathbf{A} \nabla u_S \rangle]_E\|_{L^2(E)}^2 \right\}^{1/2}. \quad (9.7)
\end{aligned}$$

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

Ungleichung (9.6) ist eine **a-posteriori Fehlerabschätzung**. Die Grösse  $\eta$  kann aus den gegebenen Daten  $f$ ,  $\mathbf{A}$ ,  $c$  und der berechneten numerische Lösung  $u_S$  a posteriori berechnet werden. Sie heisst daher auch **a-posteriori Fehlerschätzer**. Ungleichung (9.6) zeigt, dass der Fehlerschätzer **zuverlässig** ist, d.h. ist  $\eta < \varepsilon$ , so ist der Fehler ebenfalls (bis auf einen Faktor) nicht grösser als  $\varepsilon$ . Die Kontrolle von  $\eta$  erlaubt also, eine vorgegebene Toleranz zu erreichen. Um dies mit einem minimalen Aufwand zu erreichen, reicht die obere Schranke (9.6) nicht aus. Wir müssen zusätzlich garantieren, dass  $\eta$  den Fehler nicht überschätzt und die räumliche Verteilung des Fehlers richtig widerspiegelt. Dies nennt man **Effizienz**. Sie ist gegeben, wenn es gelingt, den Fehler auch nach unten durch  $\eta$  abzuschätzen.

Um dies zu erreichen, benötigen wir einige zusätzliche Notationen. Bezeichne mit  $f_S$  irgendeine, im folgenden feste Finite-Elemente-Approximation an  $f$ , z.B. die  $L^2$ -Projektion auf die stückweise konstanten Funktionen  $S^{0,-1}$ . Für ein gegebenes Dreieck  $\tau$  bezeichne mit  $z_1, z_2, z_3$  seine Eckpunkte und setze

$$\psi_\tau := 27b_{z_1}b_{z_2}b_{z_3} \quad \text{auf } \tau.$$

Wie man leicht nachprüft, hat  $\psi_\tau$  folgende Eigenschaften

$$\psi_\tau \in \mathbb{P}_3, \quad \psi_\tau \geq 0 \text{ auf } \tau, \quad \psi_\tau = 0 \text{ auf } \partial\tau, \quad \max_{x \in \tau} \psi_\tau(x) = 1.$$

Insbesondere kann also  $\psi_\tau$  durch Null zu einer Funktion aus  $H_0^1(\Omega)$  fortgesetzt werden. Für eine Kante  $E \in \mathcal{E}_\Omega$  numerieren wir die Eckpunkte der angrenzenden Dreiecke  $\tau_{E_1}$  und  $\tau_{E_2}$  so, dass die Endpunkte von  $E$  zuerst numeriert werden. Bezeichnen  $v_{z_1,i}$ ,  $v_{z_2,i}$ ,  $v_{z_3,i}$  die Basisfunktionen zu  $\tau_{E_i}$ ,  $i = 1, 2$ , so definieren wir

$$\psi_E := 4v_{z_1,i}v_{z_2,i} \quad \text{auf } \tau_{E_i}, \quad i = 1, 2.$$

Offensichtlich hat  $\psi_E$  die folgenden Eigenschaften

$$\psi_E \in C(\omega_E), \quad \psi_E|_{\tau_{E_i}} \in \mathbb{P}_2, \quad i = 1, 2, \quad \psi_E \geq 0 \text{ auf } \omega_E, \quad \psi_E = 0 \text{ auf } \partial\omega_E, \quad \max_{x \in E} \psi_E(x) = 1.$$

Für eine Dreieckskante  $E \in \mathcal{E}_\Omega$  bezeichnen wir mit  $\mathbb{P}_k(E)$  die Polynome von Grad  $\leq k$  in einer Variablen auf  $E$ . Jedes  $\varphi \in \mathbb{P}_k(E)$  kann in kanonischer Weise konstant zu einem Polynom vom Grad  $\leq k$  in zwei Variablen auf  $\mathbb{R}^2$  fortgesetzt werden. Diese Fortsetzung bezeichnen wir mit wieder mit  $\varphi$ .

**Satz 9.2** *Für jedes Dreieck  $\tau \in \mathcal{G}$ , jede Kante  $E \in \mathcal{E}_\Omega$ , jedes  $v \in \mathbb{P}_k$  und jedes  $\sigma \in \mathbb{P}_k(E)$  gilt*

$$\begin{aligned} c_1 \|v\|_{L^2(\tau)} &\leq \left\{ \int_\tau \psi_\tau v^2 \right\}^{1/2} \leq \|v\|_{L^2(\tau)}, \\ c_2 h_\tau^{-1} \|\psi_\tau v\|_{L^2(\tau)} &\leq |\psi_\tau v|_{H^1(\tau)} \leq c_3 h_\tau^{-1} \|\psi_\tau v\|_{L^2(\tau)}, \\ c_4 \|\sigma\|_{L^2(E)} &\leq \left\{ \int_E \psi_E \sigma^2 \right\}^{1/2} \leq \|\sigma\|_{L^2(E)}, \\ c_5 h_E^{-1} \|\psi_E \sigma\|_{L^2(\omega_E)} &\leq |\psi_E \sigma|_{H^1(\omega_E)} \leq c_6 h_E^{-1} \|\psi_E \sigma\|_{L^2(\omega_E)}, \\ \|\psi_E \sigma\|_{L^2(\omega_E)} &\leq c_7 h_E^{1/2} \|\sigma\|_{L^2(E)}. \end{aligned}$$

Die Konstanten  $c_1, c_2, \dots, c_7$  hängen nur von  $k$  und der Grösse  $c_\mathcal{G}$  in der Regularitätsannahme an  $\mathcal{G}$  ab.

**Beweis.** Die obere Schranke in der ersten Abschätzung folgt aus der Cauchy-Schwarzschen Ungleichung. Wie man leicht nachrechnet, definiert  $\int_{\hat{\tau}} \psi_\tau \circ \chi_\tau w^2$  eine (genauer das Quadrat einer) Norm auf  $\mathbb{P}_k$ . Da auf endlichdimensionalen Räumen alle Normen äquivalent sind, gibt es eine Konstante  $\hat{c}$  mit

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

Anwenden des Transformationssatzes liefert

$$\begin{aligned} \hat{c} \|v\|_{L^2(\tau)} &= \hat{c} |\det D\chi_\tau|^{1/2} \|v \circ \chi_\tau\|_{L^2(\hat{\tau})} \leq |\det D\chi_\tau|^{1/2} \left\{ \int_{\hat{\tau}} (\psi_\tau \circ \chi_\tau) (v \circ \chi_\tau) \right\}^{1/2} \\ &= \left\{ \int_\tau \psi_\tau v^2 \right\}^{1/2} \end{aligned}$$

und beweist die untere Schranke der ersten Abschätzung.

Da  $\psi_\tau \circ \chi_\tau$  in den Eckpunkten von  $\hat{\tau}$  verschwindet, definiert  $|(\psi_\tau \circ \chi_\tau) w|_{H^1(\hat{\tau})}$  eine Norm auf  $\mathbb{P}_k$ . Daher existieren zwei Konstanten  $\hat{c}_1$  und  $\hat{c}_2$  mit

$$\hat{c}_1 \|\psi_\tau \circ \chi_\tau w\|_{L^2(\hat{\tau})} \leq |(\psi_\tau \circ \chi_\tau) w|_{H^1(\hat{\tau})} \leq \hat{c}_2 |(\psi_\tau \circ \chi_\tau) w|_{L^2(\hat{\tau})} \quad \forall w \in \mathbb{P}_k.$$

Hieraus uns aus dem Transformationssatz folgt wieder die Behauptung.

Die Abschätzungen für  $\sigma$  folgen mit den gleichen Argumenten. ■

Sei nun  $\tau \in \mathcal{G}$  beliebig. Setze  $w_\tau := \psi_\tau (f_S + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S)$ . Wegen Satz 9.2 ist

$$c_1^2 \|f_S + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S\|_{L^2(\tau)}^2 \leq \int_\tau w_\tau (f_S + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S).$$

Setzen wir  $w_\tau$  als Testfunktion  $v$  in (9.3) ein und berücksichtigen, dass  $w_\tau$  auf  $\partial\tau$  und ausserhalb von  $\tau$  verschwindet, so erhalten wir

$$\begin{aligned} \int_\tau w_\tau (f_S + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S) &= \int_\tau w_\tau (f + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S) + \int_\tau w_\tau (f_S - f) \\ &= a(u - u_S, w_\tau) + \int_\tau w_\tau (f_S - f) \\ &\leq \|a\|_{\mathcal{L}^2} \|u - u_S\|_{H^1(\tau)} \|w_\tau\|_{H^1(\tau)} + \|f_S - f\|_{L^2(\tau)} \|w_\tau\|_{L^2(\tau)}. \end{aligned}$$

Offensichtlich ist

$$\|w_\tau\|_{L^2(\tau)} \leq \|f_S + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S\|_{L^2(\tau)}.$$

Aus Satz 9.2 folgt weiter

$$\begin{aligned} \|w_\tau\|_{H^1(\tau)} &= \left\{ \|w_\tau\|_{L^2(\tau)}^2 + |w_\tau|_{H^1(\tau)}^2 \right\}^{1/2} \leq \{1 + c_3^2 h_\tau^{-2}\}^{1/2} \|w_\tau\|_{L^2(\tau)} \\ &\leq ch_\tau^{-1} \|f_S + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S\|_{L^2(\tau)}. \end{aligned}$$

Aus diesen Abschätzungen und der Dreiecksungleichung ergibt sich

$$h_\tau \|f + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S\|_{L^2(\tau)} \leq c \left\{ \|u - u_S\|_{H^1(\tau)} + h_\tau \|f - f_S\|_{L^2(\tau)} \right\}. \quad (9.8)$$

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

$$w_E := \psi_E [\langle n_E, \mathbf{A}\nabla u_S \rangle]_E.$$

Wegen Satz 9.2 ist

$$c_4^2 \|[\langle n_E, \mathbf{A}\nabla u_S \rangle]_E\|_{L^2(E)}^2 \leq \int_E w_E [\langle n_E, \mathbf{A}\nabla u_S \rangle]_E.$$

Setzen wir  $w_E$  als Testfunktion  $v$  in (9.3) ein und berücksichtigen, dass  $w_E$  auf  $\partial\omega_E$  und ausserhalb  $\omega_E$  verschwindet, so folgt

$$\begin{aligned} \int_E w_E [\langle n_E, \mathbf{A}\nabla u_S \rangle]_E &= \int_{\omega_E} (f + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S) w_E - a(u - u_S, w_E) \\ &\leq \|f + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S\|_{L^2(\omega_E)} \|w_E\|_{L^2(\omega_E)} + \|a\|_{\mathcal{L}^2} \|u - u_S\|_{H^1(\omega_E)} \|w_E\|_{H^1(\omega_E)}. \end{aligned}$$

Wegen Satz 9.2 ist

$$\|w_E\|_{L^2(\omega_E)} \leq c_7 h_E^{1/2} \|[\langle n_E, \mathbf{A}\nabla u_S \rangle]_E\|_{L^2(E)}$$

und

$$\|w_E\|_{H^1(\omega_E)} \leq c_6 h_E^{-1} \|w_E\|_{L^2(\omega_E)} \leq c_6 c_7 h_E^{-1/2} \|[\langle n_E, \mathbf{A}\nabla u_S \rangle]_E\|_{L^2(E)}.$$

Aus diesen Abschätzungen und (9.8) ergibt sich insgesamt

$$h_E^{1/2} \|[\langle n_E, \mathbf{A} \nabla u_S \rangle]_E\|_{L^2(E)} \leq c \left\{ \|u - u_S\|_{H^1(\omega_E)} + h_E \|f - f_S\|_{L^2(\omega_E)} \right\}. \quad (9.9)$$

Aus den Abschätzungen (9.8), (9.9) und der Definition (9.7) erhalten wir folgende **lokale untere Fehlerschranke**

$$\eta_\tau \leq c \left\{ \|u - u_S\|_{H^1(\omega_\tau)} + h_\tau \|f - f_S\|_{L^2(\omega_\tau)} \right\}. \quad (9.10)$$

Summation über alle Dreiecke liefert zudem die **globale untere Fehlerschranke**

$$\eta \leq c \left\{ \|u - u_S\|_{H^1(\Omega)} + \left\{ \sum_{\tau \in \mathcal{G}} h_\tau^2 \|f - f_S\|_{L^2(\tau)}^2 \right\}^{1/2} \right\}. \quad (9.11)$$

In beiden Abschätzungen sind die  $f - f_S$ -Terme Störterme höherer Ordnung, die zudem a priori allein aus der Kenntnis der Daten kontrolliert werden können, ohne eine Differentialgleichung bzw. Diskretisierung zu lösen.

### Bemerkung 9.3

1. Die Abschätzungen (9.6), (9.10) und (9.11) zeigen, dass der Fehlerschätzer zuverlässig und effizient ist. Es ist nicht verwunderlich, dass wir nur in einer Richtung lokale Schranken erhalten. Denn die Abschätzung (9.6) benötigt den inversen Differentialoperator, der ein globaler Operator ist. Die Abschätzung (9.10) dagegen benutzt nur den Differentialoperator, der ein lokaler Operator ist.
2. Die Grösse  $f + \operatorname{div}(\mathbf{A} \nabla u_S) - c u_S$  ist das Residuum der Finite-Elemente-Approximation  $u_S$  bzgl. der starken Form der Differentialgleichung. Die Grösse  $[\langle n_E, \nabla \mathbf{A} u_S \rangle]_E$  ist der Sprung des Spurooperators, der auf kanonische Weise die starke und schwache Form der Differentialgleichung verknüpft.
3. Ähnliche Ergebnisse gelten für die Konvektions-Diffusions-Gleichung. A priori hängen die Konstanten in den Fehlerabschätzungen von der Peclet-Zahl ab und wachsen für abnehmende Diffusion an. Dieser Effekt kann durch eine verbesserte Analyse weitgehend vermieden werden.
4. Es gibt auch andere Fehlerschätzer, die z.B. auf der Lösung lokaler, diskreter Dirichlet- oder Neumann-Probleme beruhen.

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

### Algorithmus 9.4 (Adaptive Gitterverfeinerung)

1. Bestimme eine grobe Triangulierung  $\mathcal{G}_0$  von  $\Omega$ . Setze  $k := 0$ .
2. Löse das diskrete Problem zur Triangulierung  $\mathcal{G}_k$ .
3. Berechne  $\eta_\tau$  für alle  $\tau \in \mathcal{G}_k$  und  $\eta_k := \max_{\tau \in \mathcal{G}_k} \eta_\tau$ .
4. Falls  $\eta_k \leq \varepsilon$  gilt, STOP. Sonst gehe zu 5.
5. Verfeinere alle Dreiecke  $\tau \in \mathcal{G}_k$  mit  $\eta_\tau \geq \gamma \eta_k$ . Verfeinere evtl. weitere Dreiecke, um eine zulässige Triangulierung  $\mathcal{G}_{k+1}$  zu erhalten. Erhöhe  $k$  um 1 und gehe nach 2 zurück.

In Algorithmus 9.4 ist  $\gamma$  ein zu wählender Parameter mit  $0 < \gamma < 1$ . Ist  $\gamma \sim 0$ , werden viele Elemente verfeinert; ist  $\gamma \sim 1$ , werden nur sehr wenige Elemente unterteilt. In der Praxis wählt man häufig  $\gamma = 0.5$ . Algorithmus 9.4 kann auch ggf. durch eine Vergrößerungsstrategie ergänzt werden.

Für die praktische Realisierung von Algorithmus 9.4 müssen wir noch beschreiben, wie Dreiecke verfeinert werden und wie die Zulässigkeit der verfeinerten Triangulierung gesichert werden kann. Dabei müssen wir beachten, dass alle Triangulierungen die Regularitätsbedingung erfüllen sollen, d.h. die Dreieckswinkel sollen nicht zu klein oder zu gross werden. Dazu führen wir folgende Bezeichnungen ein:

- Ein Dreieck wird **rot** unterteilt, wenn sein Kantenmittelpunkte 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) **hängende** Knoten, wenn  $\tau$  nicht unterteilt wurde, aber eines (oder mehrere) der angrenzenden Dreiecke unterteilt wurde. (Dabei bedeutet „angrenzend“, dass die betreffenden Dreiecke eine Kante gemeinsam haben.

Eine rote Unterteilung erzeugt offensichtlich ähnliche Dreiecke und verändert somit die Winkel nicht. Die Vorgabe, primär die längste Kante zu unterteilen, sichert bei der grünen und blauen Unterteilung, dass der kleinste Winkel nicht verkleinert wird. Offensichtlich ist eine Triangulierung genau dann zulässig, wenn kein Dreieck hängende Knoten hat.

Schritt 5 von Algorithmus 9.4 wird nun gemäss der folgenden Regeln durchgeführt:

1. Ist  $\eta_\tau \geq \gamma \eta_k$ , unterteile  $\tau$  rot.
2. Hat  $\tau$  drei hängende Knoten, unterteile  $\tau$  rot.
3. Hat  $\tau$  zwei hängende Knoten, von denen keiner auf der längsten Kante liegt, unterteile  $\tau$  rot.
4. Hat  $\tau$  zwei hängende Knoten, von denen einer auf der längsten Kante liegt, unterteile  $\tau$  blau.



5. Hat  $\tau$  einen hängenden Knoten, unterteile  $\tau$  blau, wenn der hängende Knoten nicht auf der längsten Kante liegt, sonst unterteile  $\tau$  grün.

Man kann zeigen, dass dieser Algorithmus in endlich vielen Schritten eine verfeinerte Triangulierung erzeugt, die den oben diskutierten Kriterien genügt.

## 10 Numerische Lösung der diskreten Probleme

Um die wesentlichen Punkte besser herausarbeiten zu können und um technische Schwierigkeiten zu vermeiden, betrachten wir im folgenden die Reaktions-Diffusions-Gleichung, d.h., zum Differentialoperator  $-\operatorname{div}(\mathbf{A} \operatorname{grad} u) + cu$ , mit homogenen Dirichlet-Randbedingungen und Dreieckselemente der Ordnung  $k \geq 1$ , d.h.  $\mathcal{G}$  besteht aus Dreiecken und  $S := S_0^{k,0}$ .

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

$$(\varphi, \psi)_S := \sum_{\tau \in \mathcal{G}} \sum_{\mathbf{z} \in \Sigma_k(\tau)} h_\tau^2 \varphi(\mathbf{z}) \psi(\mathbf{z}),$$

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

$\|\cdot\|_S$  ist eine skalierte euklidische Norm auf  $\mathbb{R}^N$ ,  $N = \#\overset{\circ}{K}_{\mathcal{G}}^k$ . Die Matrix, die zu  $(\cdot, \cdot)_S$  und der Knoten-Basis gehört, ist diagonal. Insbesondere können Gleichungssystem der Form

$$(u, v)_S = \ell(v) \quad \forall v \in S$$

leicht gelöst werden.

### Satz 10.1

1.  $\|\cdot\|_S$  und  $\|\cdot\|_{L^2(\Omega)}$  sind äquivalente Normen auf  $S$ . Die entsprechenden Konstanten hängen nur von  $k$  und der Konstanten  $c_{\mathcal{G}}$  in der Regularitätsbedingung an  $\mathcal{G}$  ab.
2. Für alle  $v \in S$  gilt die **inverse Abschätzung**

$$|v|_{H^1(\Omega)} \leq ch_{\min}^{-1} \|v\|_{L^2(\Omega)}.$$

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

**Beweis. Zu (1):** Wegen Satz 7.4 ist  $\left\{ \sum_{z \in \hat{\Sigma}_k} |\varphi(z)|^2 \right\}^{1/2}$  eine Norm auf  $\mathbb{P}_k$ . Da  $\mathbb{P}_k$  endlich dimensional ist, gibt es zwei Konstanten  $\hat{c}_1, \hat{c}_2$ , die nur von  $k$  abhängen, mit

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

Sei nun  $\tau \in \mathcal{G}$  beliebig und  $\chi_\tau : \hat{\tau} \rightarrow \tau$  eine affine Transformation. Wegen

$$c_1 h_\tau^2 \leq |\det D\chi_\tau| \leq c_2 h_\tau^2$$

folgt aus obiger Abschätzung mit dem Transformationssatz

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

Hieraus folgt die Behauptung **(1)** durch Quadrieren und Summieren über alle Dreiecke. Man beachte, dass die Konstanten  $c_1, c_2$  von  $c_{\mathcal{G}}$  abhängen.

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

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

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

$$\begin{aligned} |\varphi|_{H^1(\tau)} &\leq |\det D\chi_\tau|^{1/2} \|D\chi_\tau^{-1}\| |\varphi \circ \chi_\tau|_{H^1(\hat{\tau})} \\ &\leq \hat{c} |\det D\chi_\tau|^{1/2} \|D\chi_\tau^{-1}\| \|\varphi \circ \chi_\tau\|_{L^2(\hat{\tau})} \\ &= \hat{c} \|D\chi_\tau^{-1}\| \|\varphi\|_{L^2(\tau)} \leq \hat{c} h_{\hat{\tau}}/\rho_\tau \|\varphi\|_{L^2(\tau)} \leq c' h_\tau^{-1} \|\varphi\|_{L^2(\tau)}. \end{aligned}$$

Dabei hängt  $c'$  von  $c_{\mathcal{G}}$  ab. Quadrieren und Summieren dieser Ungleichung beweist Teil **(2)**. ■

Aus Satz 10.1(1), dem Beweis des Existenzsatzes 4.3 und der Friedrichsschen Ungleichung folgt für alle  $v \in S$

$$a(v, v) \geq \beta |v|_{H^1(\Omega)}^2 \geq \beta' \|v\|_{L^2(\Omega)}^2 \geq \beta'' \|v\|_S^2.$$

Ebenso folgt aus dem Beweis von Satz 4.3 und Satz 10.1 für alle  $v, w \in S$

$$\begin{aligned} a(v, w) &\leq B |v|_{H^1(\Omega)} |w|_{H^1(\Omega)} \leq c^2 B \left( \min_{\tau \in \mathcal{G}} h_\tau \right)^{-2} \|v\|_{L^2(\Omega)} \|w\|_{L^2(\Omega)} \\ &\leq c' \left( \min_{\tau \in \mathcal{G}} h_\tau \right)^{-2} \|v\|_S \|w\|_S. \end{aligned}$$

Also hat die Steifigkeitsmatrix die Kondition  $O(\min_{\tau \in \mathcal{G}} h_\tau)^{-2}$ . Daher scheiden das Jacobi- und das Gauss-Seidel-Verfahren als Löser aus. Da die Bilinearform  $a$  symmetrisch ist, ist die Steifigkeitsmatrix symmetrisch, und ein CG-Verfahren oder besser ein vorkonditioniertes CG-Verfahren können als Löser benutzt werden.

Die effizientesten Methoden zur Lösung elliptischer Differentialgleichung sind Mehrgitterverfahren. Wir werden im folgenden ein Mehrgitterverfahren für Finite-Elemente-Diskretisierungen von elliptischen partiellen Differentialgleichungen analysieren. Wie der Name *Mehrgitterverfahren* besagt, wird nicht nur ein Gitter benötigt, sondern eine *Gitterhierarchie*  $(\mathcal{G}_m)_{m=0}^M$ , bestehend aus zulässigen Finite-Elemente-Gittern  $\mathcal{G}_m$  für das Gebiet  $\Omega$ . Wir nehmen an, dass

die Gitter geschachtelt sind, d.h., für alle  $0 \leq m \leq M - 1$  und alle  $\tau \in \mathcal{G}_m$  existiert eine Teilmenge (Söhne)  $\text{sons}(\tau) \subset \mathcal{G}_{m+1}$  mit der Eigenschaft, dass

$$\bar{\tau} = \bigcup_{t \in \text{sons}(\tau)} \bar{t}$$

gilt. Wir machen die folgenden Annahmen (die mit aufwendigeren Mittel der Funktionalanalysis abgeschwächt werden können).

1.  $\Omega$  ist konvex.
2. Jede Triangulierung  $\mathcal{G}_m$  ist **quasi-uniform**, d.h.

$$h_m := \max_{\tau \in \mathcal{G}_m} h_\tau \leq c \min_{\tau \in \mathcal{G}_m} h_\tau$$

mit einer von  $m$  unabhängigen Konstanten  $c$ .

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

### Algorithmus 10.2 (MG-Verfahren mit $V$ -Zyklus und Jacobi-Glättung)

0. Gegeben sei eine Näherung  $u_m^0 \in S_m$  für die Lösung des diskreten Problems.

1. (**Vorglättung**) Für  $i = 1, \dots, \nu_1$  berechne  $u_m^i$  aus  $u_m^{i-1}$  als Lösung von<sup>8</sup>

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

2. (**Grobitterkorrektur**) Berechne

$$\ell_{m-1}(v) = \ell_m(v) - a(u_m^{\nu_1}, v) \quad \forall v \in S_{m-1}.$$

**Falls**  $m > 1$ , wende das MG-Verfahren mit Startwerten  $u_{m-1}^0 = 0$  auf das Problem

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

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

**Setze**

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

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

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

### Bemerkung 10.3

1. Es ist  $\ell_M(v) = \ell(v) = \int_\Omega f v$ . Die rechten Seiten  $\ell_m$  zu den gröberen Triangulierungen werden rekursiv berechnet.
2. Die Dämpfungsparameter  $\omega_m$  werden später bestimmt.

---

<sup>8</sup>Abkürzung:  $(\cdot, \cdot)_m := (\cdot, \cdot)_{S_m}$

3. Die Gleichungssysteme in den Glättungsschritten haben eine diagonale Koeffizientenmatrix.
4. Die Grobitterkorrektur nutzt aus, dass die Finite-Elemente-Räume geschachtelt sind, d.h.,  $S_{m-1} \subset S_m$ . In der Praxis stellt man  $u_m$  und  $u_{m-1}$  als Vektoren dar, deren Komponenten die Werte in den entsprechenden Gitterpunkten sind. Insbesondere muss  $u_{m-1}$  vom Gitter  $\mathcal{G}_{m-1}$  auf das Gitter  $\mathcal{G}_m$  interpoliert werden.

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

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

O.B.d.A. können wir die Eigenwerte der Größe nach ordnen

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

Da die Triangulierungen quasi-uniform sind, folgt aus Satz 10.1  $\Lambda_m \sim h_m^{-2}$ . Wir nehmen im folgenden an, dass

$$\omega_m = \Lambda_m$$

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

Jedes  $v \in S_m$  lässt sich eindeutig darstellen als

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

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

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

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

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

Dabei bedeutet „ $\simeq$ “ die Äquivalenz von Normen mit von  $m$  unabhängigen Konstanten. Sind  $v, w \in S_m$  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

$$a(v, w) = \sum_{\mu=1}^{N_m} \lambda_{m,\mu} c_\mu d_\mu \leq \left\{ \sum_{\mu=1}^{N_m} c_\mu^2 \right\}^{1/2} \left\{ \sum_{\mu=1}^{N_m} \lambda_{m,\mu}^2 d_\mu^2 \right\}^{1/2} = \| \| v \| \|_{0,m} \| \| w \| \|_{2,m}. \quad (10.1)$$

**Satz 10.4** Bezeichne mit  $Q_m : S_m \rightarrow S_{m-1}$  die **Ritz-Projektion**, d.h.  $Q_m v \in S_{m-1}$  und

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

Dann gilt für alle  $v \in S_m$

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

Die Konstante  $c$  hängt nicht von  $m$  ab.

**Beweis.** Aus der Definition der Ritz-Projektion und (10.1) folgt

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

Da  $\Omega$  konvex ist, folgt aus Satz 3.2

$$\|v - Q_m v\|_{L^2(\Omega)} \leq c h_{m-1} \|v - Q_m v\|_{H^1(\Omega)} \leq c' h_{m-1} \|v - Q_m v\|_{1,m}.$$

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

Als nächstes definieren wir einen Operator  $J : S_m \rightarrow S_m$  durch

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

$J$  beschreibt die Fehlerfortpflanzung in den Glättungsschritten von Algorithmus 10.2. Ist  $v = \sum c_\mu \psi_{m,\mu}$ , so folgt

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

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

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

Dann gilt offensichtlich für  $v \in S_m$

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

**Satz 10.5** Sei  $v \in S_m$  und  $\rho = \rho(J^\nu v)$ . Dann gilt

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

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

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

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

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

Hieraus folgt

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

■

**Satz 10.6** Sei  $v \in S_m$  und  $\rho = \rho(v)$ . Dann gilt

$$\|\| (I - Q_m) v \|\|_{1,m} \leq \min \left\{ 1, c\sqrt{1-\rho} \right\} \|\| v \|\|_{1,m}.$$

**Beweis.** Da  $I - Q_m$  eine Orthogonalprojektion ist, ist

$$\|\| (I - Q_m) v \|\|_{1,m} \leq \|\| v \|\|_{1,m}.$$

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

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

Hieraus und aus Satz 10.4 folgt

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

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

Nach diesen Vorbereitungen können wir nun die Konvergenz von Algorithmus 10.2 beweisen.

**Satz 10.7** Bezeichne mit  $\delta_m$  die Konvergenzrate von Algorithmus 10.2 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 10.6

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

**Beweis.** Bezeichne mit  $u_m^*$  die Lösung der Finite-Elemente-Probleme auf dem  $m$ -ten Gitter und mit  $e_m^i = u_m^* - u_m^i$  „Naherung nach dem  $i$ -ten Mehrgitterschritt“ den Fehler im  $i$ -ten Schritt von Algorithmus 10.2. Dann gilt

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

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

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

Wegen der Galerkin-Orthogonalitat

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

folgt

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

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

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

Da  $I - Q_m$  ein Projektor bzgl. des Skalarproduktes  $a(\cdot, \cdot)$  ist, folgt fur den ersten Summanden

$$a((I - Q_m) e_m^\nu, J^\nu w) = a((I - Q_m) e_m^\nu, (I - Q_m) J^\nu w) \leq \|(I - Q_m) e_m^\nu\|_{1,m} \|(I - Q_m) J^\nu w\|_{1,m}.$$

Fur den zweiten Summanden gilt

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

Aus diesen beiden Abschätzungen folgt mit der Cauchy-Schwarzschen Ungleichung

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

Wegen der Sätze 10.5 und 10.6 ist

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

und

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

Da

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

ist, folgt hieraus

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

Also ist

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

Da auf dem gröbsten Gitter exakt gelöst wird, gilt

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

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

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

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

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



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

Der Beweis von Satz 10.7 beruht im wesentlichen auf den Annahmen, dass  $\Omega$  konvex ist und die Gitter quasiuniform sind. Die Konvexität von  $\Omega$  wird in Satz 10.4 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 Systemmatrix 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 10.2 wesentlich bessere Konvergenzresultate erzielen kann.

## 11 Parabolische partielle Differentialgleichung

Sei  $\Omega$  wieder ein Gebiet im  $\mathbb{R}^d$ . Wir betrachten das Problem der Wärmeleitung und die folgende physikalische Fragestellung: Die anfängliche Temperaturverteilung im Gebiet wird beeinflusst durch angebrachte Wärmequellen und die Aussentemperatur (z.B. Sonneneinstrahlung). Physikalisch nehmen wir an, dass die Wärmeänderung durch *Diffusion* erfolgt, d.h., der Wärmestrom entsteht durch Konzentrationsausgleich der Partikel, die die Wärme tragen und *nicht* durch *Konvektion*, d.h., der Wärmestrom wird aktiv in einem Behälter eingesogen (z.B. Warmwasser wird in ein Kaltwasserbecken gepumpt.)

Die Temperatur  $u$  ist eine Funktion, die von jedem Ortspunkt abhängt und von der Zeit:  $u : \Omega \times [0, T] \rightarrow \mathbb{R}$ . Die Gleichung, zur Charakterisierung von  $u$  ist die parabolische partielle Differentialgleichung

$$\begin{aligned} \dot{u} - \Delta u &= f && \text{in } \Omega && \text{für alle } t \in [0, T], \\ u &= 0 && \text{auf } \partial\Omega && \text{für alle } t \in [0, T] \text{ (zeitabhängige Randbedingung),} \\ u(\cdot, 0) &= v && \text{in } \Omega && \text{Anfangsbedingung.} \end{aligned} \quad (11.1)$$

Hier bezeichnet  $\dot{u}$  die Ableitung von  $u$  nach der Zeit und  $\Delta$  den  $d$ -dimensionalen Laplace-Operator:

$$\Delta u(x, t) = \sum_{i=1}^d \frac{\partial^2 u(x, t)}{\partial x_i^2}.$$

Indem wir für die Gleichung (11.1) die Ortsabhängigkeit mit finiten Elementen diskretisieren, entsteht ein System gewöhnlicher Differentialgleichung bezüglich der Zeit mit vorgegebenen Anfangsbedingungen.

### 11.1 Ortsdiskretisierung des parabolischen Problems

Sei  $\mathcal{G}$  wieder ein Finite-Elemente-Gitter für das Gebiet  $\Omega$ . Die Menge der Gitterpunkte wird wieder mit  $\Theta$  bezeichnet und die Menge der inneren Gitterpunkte mit  $\Theta^0$ . Die Basis für den Raum aller stetigen, stückweise affinen Finite-Elemente-Funktionen zu den inneren Knotenpunkten wird wieder mit  $(b_i)_{i=1}^N$ ,  $N = \#\Theta^0$  bezeichnet. Mit  $S$  bezeichnen wir den Raum aller stetigen, stückweise affinen Funktionen mit Null-Randbedingungen.

$$S = \text{span} \{b_i : 1 \leq i \leq N\}.$$

Für die semidiskrete Lösung von (11.1) verwenden wir den Ansatz

$$u_S(x, t) := \sum_{i=1}^N u_{S,i}(t) b_i(x),$$

wobei wir zunächst voraussetzen, dass die Koeffizienten  $u_{S,i}(t)$  hinreichend glatt sind, so dass die folgenden Umformungen erlaubt sind. Das zu (11.1) gehörende *semidiskrete* Problem ist gegeben durch die Aufgabe: Finde Funktionen  $\mathbf{u}_S \in C^1([0, T], \mathbb{R}^N)$ , so dass für alle  $t \in [0, T]$  gilt

$$\sum_{j=1}^N \int_{\Omega} \dot{u}_{S,j} b_j b_i + u_{S,j} \langle \nabla b_j, \nabla b_i \rangle = \int_{\Omega} f b_i \quad \forall 1 \leq i \leq N \quad (11.2)$$

$$u(\cdot, 0) = v_S$$

ist. Da  $u_S(\cdot, 0)$  eine Finite-Elemente-Funktion ist, muss daher auch  $v_S \in S$  gelten. Falls  $v$  aus (11.1) nicht aus  $S$  ist, muss daher noch eine geeignete Projektion oder allgemeinere Abbildung zwischengeschaltet werden. Mit den (zeitunabhängigen)  $N \times N$ -Matrizen  $\mathbf{M}$  und  $\mathbf{A}$  und dem rechte-Seite-Vektor  $\mathbf{r} \in \mathbb{R}^N$

$$\mathbf{M} = \left( \int_{\Omega} b_j b_i \right)_{i,j=1}^N, \quad \mathbf{A} = \left( \int_{\Omega} \langle \nabla b_j, \nabla b_i \rangle \right)_{i,j=1}^N, \quad \mathbf{r}(t) = \left( \int_{\Omega} f(\cdot, t) b_i \right)_{i=1}^N$$

lässt sich (11.2) als gewöhnliche Differentialgleichung für den zeitabhängigen Koeffizientenvektor  $\mathbf{u}_S$  auffassen

$$\mathbf{M} \dot{\mathbf{u}}_S(t) + \mathbf{A} \mathbf{u}_S(t) = \mathbf{r}(t) \quad \forall t \in [0, T]$$

$$\mathbf{u}_S(0) = \mathbf{v}_S.$$

## 11.2 Einschrittverfahren für die homogene Wärmeleitungsgleichung

Wir betrachten das homogene parabolische Problem

$$\begin{aligned} u_t &= \Delta u & \text{in } \Omega & \quad \text{für alle } t > 0, \\ u &= 0 & \text{auf } \partial\Omega & \quad \text{für alle } t > 0 \text{ (zeitabhängige Randbedingung),} \\ u(\cdot, 0) &= v & \text{in } \Omega & \quad \text{Anfangsbedingung.} \end{aligned} \quad (11.3)$$

Der Einfachheit halber nehmen wir für das Folgende an, dass  $v \in S$  gilt.

### 11.2.1 Exkurs in die Funktionalanalysis

Um die Diskretisierung bezüglich der Zeit mehr in den Vordergrund zu stellen, betrachten wir eine allgemeine Evolutionsgleichung in einem Hilbert-Raum.

**Zur Erinnerung:** Ein **Banach**-Raum  $B$  ist ein normierter Vektorraum, der vollständig ist, d.h., jede Cauchy-Folge in  $B$  konvergiert gegen ein Element in  $B$ . Ein **Hilbert**-Raum  $H$  ist ein Banach-Raum, wenn ein Skalarprodukt  $(\cdot, \cdot)_H$  auf  $H$  erklärt ist, und die Norm in  $H$  durch  $\|\cdot\|_H := (\cdot, \cdot)^{1/2}$  definiert ist. Ein Hilbert-Raum heisst **separabel**, falls eine dichte Teilmenge  $A \subset H$  existiert mit abzählbarer Basis  $A = \text{span} \{a_n : n \in \mathbb{N}\}$ .

**Konvention 11.1** Wir nehmen generell an, dass  $H$  ein separabler Hilbert-Raum ist. Für unsere Anwendungen wird immer  $H = L^2(\Omega)$  oder  $H = S$  gelten.

Im folgenden betrachten wir immer die abstrakte Situation, eines abstrakten, linearen Operators  $A$ , der auf einem Teilraum  $W \subset H$  definiert ist. Die folgenden Annahmen sichern, dass viele Eigenschaften, die wie aus der linearen Algebra für endlichdimensionale Vektorräume kennengelernt haben, sich auch auf unendlichdimensionale Vektorräume übertragen lassen.

Der Operator  $A$  ist **positiv definit**, falls

$$(Av, v)_H > 0 \quad \forall v \in W \setminus \{0\}$$

gilt. Der Operator  $A$  ist **selbstadjungiert**<sup>9</sup>, falls

$$(Av, w)_H = (v, Aw)_H \quad \forall v, w \in W.$$

**Konvention 11.2** Generell nehmen wir an, dass  $A$  ein linearer, selbstadjungierter, positiv definit, nicht-notwendigerweise beschränkter Operator auf einem Teilraum  $W \subset H$  ist, dessen Inverse  $A^{-1} : H \rightarrow H$  kompakt ist. Das Spektrum von  $A$ , d.h., die Menge aller Zahlen  $\lambda \in \mathbb{C}$ , für die  $A - \lambda I$  entweder nicht invertierbar ist oder die Inverse  $(A - \lambda I)^{-1}$  unbeschränkt ist, wird mit  $\sigma(A)$  bezeichnet.

**Satz 11.3** Der Operator  $A$  besitzt Eigenwerte  $(\lambda_j)_{j=1}^N$  und eine zugehörige Basis orthonormaler Eigenfunktionen  $(\varphi_j)_{j=1}^N$  für ein  $N \in \mathbb{N} \cup \{\infty\}$ .

**Definition 11.4** Für eine beliebige Funktion  $g$ , die auf dem Spektrum  $\sigma(A)$  von  $A$  definiert ist, setzen wir

$$g(A)v = \sum_{j=1}^N g(\lambda_j) (v, \varphi_j)_H \varphi_j \quad \forall v \in H. \quad (11.4)$$

**Bemerkung 11.5** Satz 11.3 und Definition 11.4 sind aus der linearen Algebra endlichdimensionaler Vektorräume bekannt. Die Voraussetzung an den Raum  $H$  und den Operator  $A$  sichern, dass sich diese Eigenschaften und Begriffe auch auf unendlichdimensionale Vektorräume übertragen lassen.

## 11.2.2 Zeitdiskretisierung abstrakter Evolutionsprobleme

Wir formulieren das Ausgangsproblem (11.1) zunächst als abstraktes Problem.

Sei  $H$  ein separabler Hilbert-Raum, und wir nehmen an, die Konventionen 11.1 und 11.2 sind erfüllt. Wir betrachten das Anfangswertproblem

$$u' + Au = 0 \quad \text{für } t > 0 \quad \text{mit } u(0) = v. \quad (11.5)$$

**Satz 11.6** Die Lösung des Problems (11.5) besitzt die Darstellung

$$u(t) = E(t)v = \sum_{j=1}^N e^{-\lambda_j t} (v, \varphi_j)_H \varphi_j. \quad (11.6)$$

---

<sup>9</sup>Im endlichdimensionalen Vektorräumen über  $\mathbb{R}$  entspricht der Begriff „selbstadjungiert“ dem Begriff „symmetrisch“ und für die darstellende Matrix  $\mathbf{A}$  bezüglich einer Basis gilt dann  $\mathbf{A} = \mathbf{A}^\top$ .

Im Sinn von Definition 11.4 ist der Lösungsoperator  $E(t)$  durch die Wahl  $g(\lambda) := e^{-t\lambda}$  charakterisiert. Diese Überlegung verwenden wir zur Konstruktion von Einschrittverfahren. Sei dazu  $k$  eine Zeitschrittweite und für  $n = 0, 1, 2, \dots$  die Zeitpunkte  $t_n := nk$  definiert.

Die Lösung (11.5) besitzt die folgende Halbgruppeneigenschaft:

$$u(t_{n+1}) = E(k)u(t_n).$$

Es ist daher naheliegend, ein Einschrittverfahren zu konstruieren, indem die Funktion  $e^{-\lambda}$  durch eine rationale Funktion  $r(\lambda)$  approximiert wird.

**Bemerkung 11.7** *Da im allgemeinen das Eigensystem der Operators  $A$  nicht explizit bekannt ist, ist die Darstellung (11.6) ungeeignet zur Berechnung der Lösung. Alternativ könnte man die Darstellung*

$$e^{At}v = \sum_{i=0}^{\infty} \frac{t^i}{i!} A^i v$$

verwenden. Zwar benötigt man nun zur Auswertung nicht mehr das Eigensystem von  $A$ , da aber die Summe unendlich ist, lässt sich diese natürlich genausowenig in endlicher Zeit auswerten.

Ist die Funktion  $g$  allerdings eine rationale Funktion

$$g(\lambda) = \frac{p(\lambda)}{q(\lambda)}$$

mit Polynomen  $p(x) = \sum_{i=0}^n \alpha_i x^i$  und  $q(x) = \sum_{i=0}^m \beta_i x^i$ , so lässt sich die Anwendung  $g(A)v$  durch Berechnung von

$$B := \sum_{i=0}^n \alpha_i A^i, \quad C := \sum_{i=0}^m \beta_i A^i \quad \text{und} \quad C^{-1}Bv$$

in endlicher Zeit auswerten.

Sei also  $r(\lambda)$  eine rationale Approximation von  $e^{-\lambda}$ . Dann ist die Approximation  $U^{n+1}$  des Anfangswertproblems zum Zeitpunkt  $t_{n+1}$  rekursiv durch

$$U^{n+1} := E_k U^n \quad \text{für } n = 0, 1, 2, \dots \quad (11.7)$$

definiert, wobei  $E_k := r(kA)$  gesetzt wird, und die Startnäherung gemäß  $U^0 := v$  definiert ist.

Um die Genauigkeit dieser Methode zu analysieren, betrachten wir zunächst das skalare Problem

$$\forall t > 0: \quad u' + au = 0 \quad \text{mit} \quad u(0) = 1. \quad (11.8)$$

Die zugehörige diskrete Lösung ist dann durch

$$U^{n+1} = r(ka)U^n \quad (11.9)$$

gegeben. Die Methode besitzt die Konsistenzordnung  $q$ , falls die exakte Lösung von (11.8), eingesetzt in (11.9), eine Genauigkeit von  $O(k^{q+1})$  besitzt. Da die exakte Lösung von (11.8) durch  $e^{-at}$  gegeben ist, kann diese Aussage auch gemäß

$$r(ka) = e^{-ka} + O(k^{q+1})$$

beschrieben werden, bzw. durch

$$r(\lambda) = e^{-\lambda} + O(\lambda^{q+1}) \quad \text{für } \lambda \rightarrow 0. \quad (11.10)$$

Neben der Konsistenz des Verfahrens spielt dessen Stabilität eine wesentliche Rolle. Mit der Spektraldarstellung (11.4) folgert man

$$U^n = E_k^n v = \sum_{j=1}^N r(k\lambda_j)^n (v, \varphi_j)_H \varphi_j. \quad (11.11)$$

Der Lösungsoperator  $E_k$  wird stabil im Hilbert-Raum  $H$  genannt, falls  $\|E_k^n\|_H \leq C$  für  $n \geq 1$  gilt. Diese Bedingung lässt sich auch alternativ beschreiben. Da  $(\varphi_j)$  ein Orthonormalsystem ist, gilt die Parsevalsche Gleichung

$$\|v\|_H^2 = (v, v)_H = \sum_{i,j=1}^N (v, \varphi_j)_H (v, \varphi_i)_H (\varphi_i, \varphi_j)_H = \sum_{i=1}^N (v, \varphi_i)_H^2.$$

Daraus folgt

$$\begin{aligned} \|E_k^n\|_H &= \sup_{v \in H \setminus \{0\}} \frac{\|E_k^n v\|_H}{\|v\|_H} = \sqrt{\sup_{v \in H \setminus \{0\}} \frac{\left\| \sum_{j=1}^N r(k\lambda_j)^n (v, \varphi_j)_H \varphi_j \right\|_H^2}{\sum_{i=1}^N (v, \varphi_i)_H^2}} \\ &= \sqrt{\sup_{v \in H \setminus \{0\}} \frac{\sum_{j=1}^N r(k\lambda_j)^{2n} (v, \varphi_j)_H^2}{\sum_{i=1}^N (v, \varphi_i)_H^2}} = \left( \max_{1 \leq j \leq N} |r(k\lambda_j)| \right)^n. \end{aligned} \quad (11.12)$$

Die Bedingung  $\|E_k^n\|_H \leq C$  für alle  $n$  ist daher äquivalent zur Bedingung

$$\max_{\lambda \in \sigma(kA)} |r(\lambda)| \leq 1. \quad (11.13)$$

Daraus folgt mit der Darstellung (11.11)

$$\|U^n\|_H^2 \leq \sum_{i=1}^N (v, \varphi_i)_H^2 = \|v\|_H^2.$$

Die Stabilitätsbedingung (11.13) wird für alle folgenden Methoden erfüllt sein.

Wir illustrieren dieses abstrakte Vorgehen für zwei weitverbreitete Zeitdiskretisierungsverfahren für die Wärmeleitungsgleichung. Für das ortsdiskretisierte Problem gilt  $H = \mathring{S}$  und  $A = -\Delta_S$  und lautet: Finde  $u \in \mathring{S} \times \mathbb{R}_{>0}$ , so dass für alle  $t > 0$  gilt:

$$\int_{\Omega} \dot{u}(x, t) \chi(x) dx = - \int_{\Omega} \langle \nabla u(x, t), \nabla \chi(x) \rangle dx \quad \forall \chi \in \mathring{S}$$

mit den Anfangswerten  $u(\cdot, 0) = v$ .

Das implizite Euler-Verfahren ist gegeben durch

$$(U^{n+1}, \chi)_{L^2(\Omega)} + k (\nabla U^{n+1}, \nabla \chi)_{L^2(\Omega)} = (U^n, \chi)_{L^2(\Omega)} \quad \forall \chi \in \mathring{S}, \quad n \geq 0 \quad (11.14)$$

und das Crank-Nicolson-Verfahren durch

$$(U^{n+1}, \chi)_{L^2(\Omega)} + \frac{1}{2}k (\nabla U^{n+1}, \nabla \chi)_{L^2(\Omega)} = (U^n, \chi)_{L^2(\Omega)} - \frac{1}{2}k (\nabla U^n, \nabla \chi)_{L^2(\Omega)}$$

für alle  $\chi \in \overset{\circ}{S}$ ,  $n \geq 0$ . In kompakter Operatorschreibweise gilt für (11.14)

$$(I - k\Delta_S)U^{n+1} = U^n \quad \text{bzw.} \quad U^{n+1} = (I - k\Delta_S)^{-1}U^n$$

und für das Crank-Nicolson-Verfahren

$$U^{n+1} = \left(I - \frac{k}{2}\Delta_S\right)^{-1} \left(I + \frac{k}{2}\Delta_S\right)U^n.$$

Mit  $A = -\Delta_S$  sind beide Schemata von der Form (11.11) mit

$$r(\lambda) = \frac{1}{1 + \lambda} \quad \text{und} \quad r(\lambda) = \frac{1 - \frac{1}{2}\lambda}{1 + \frac{1}{2}\lambda}.$$

In beiden Fällen gilt  $|r(\lambda)| \leq 1$  für alle  $\lambda \geq 0$  und da  $-\Delta_S$  positiv definit ist (ohne Beweis) gilt  $\sigma(kA) = \sigma(-k\Delta_S) \subset [0, \infty[$ . Daher ist die Stabilitätsbedingung (11.13) erfüllt.

Wir beginnen die Fehleranalyse der Zeitdiskretisierung mit einer Abschätzung in der Hilbert-Raumnorm für den Fall glatter Anfangsdaten. Um die Glattheit zu beschreiben, verwenden wir den Operator  $-\Delta$  und setzen für  $s \geq 1$

$$H^s := \left\{v \in L^2(\Omega) \wedge v|_{\partial\Omega} = 0 \wedge (-\Delta)^{s/2}v \in L^2(\Omega)\right\}.$$

Je grösser  $s$  ist, desto grösser ist die Differenzierbarkeitsanforderung an die Funktionen  $v \in H^s$ . Die Norm auf  $H^s$  ist durch

$$\|v\|_s := (A^s v, v)_{L^2(\Omega)}^{1/2} = \|A^{s/2}v\|_{L^2(\Omega)} = \left\{ \sum_{j=1}^N \lambda_j^s (v, \varphi_j)_{L^2(\Omega)}^2 \right\}^{1/2}$$

definiert.

**Hilfssatz 11.8** *Das Verfahren sei von Konsistenzordnung  $q$ , so dass (11.10) gilt. Dann gilt*

1.  $\exists \lambda_0, C > 0$  :

$$|r(\lambda) - e^{-\lambda}| \leq C\lambda^{q+1} \quad \forall 0 \leq \lambda \leq \lambda_0.$$

2.  $\exists \lambda_1 > 0, \exists c \in ]0, 1[$  :

$$|r(\lambda)| \leq e^{-c\lambda} \quad \forall 0 \leq \lambda \leq \lambda_1. \quad (11.15)$$

3. Falls  $|r(\lambda)| < 1$  für alle  $0 \leq \lambda \leq \bar{\lambda}$  gilt, kann in (11.15)  $\lambda_1 \geq \bar{\lambda}$  gewählt werden.

**Beweis.** Übungsaufgabe. ■

**Satz 11.9** *Das Verfahren sei von Konsistenzordnung  $q$  und stabil, so dass (11.10) und (11.13) gelten. Dann erfüllen die Lösungen von (11.5) und (11.7) die Abschätzung*

$$\|U^n - u(t_n)\|_{L^2(\Omega)} \leq Ck^q \|v\|_{2q}, \quad \forall t_n \geq 0.$$

**Beweis.** Wir führen die Funktion  $F_n(\lambda) = r(\lambda)^n - e^{-n\lambda}$  und erinnern an die Definition (11.4). Daher gilt

$$U^n - u(t_n) = r(kA)^n v - e^{-nkA} v = F_n(kA) v. \quad (11.16)$$

Die Behauptung lässt sich in der Form schreiben

$$\|F_n(kA) v\|_{L^2(\Omega)} \leq Ck^q \|v\|_{2q} \quad \forall t_n \geq 0.$$

Falls  $F_n(kA)$  als Operator von  $H^{2q}$  nach  $L^2(\Omega)$  aufgefasst wird, ist diese Abschätzung äquivalent zur Operatornormabschätzung

$$\begin{aligned} \|F_n(kA)\|_{L^2(\Omega) \leftarrow H^{2q}} &= \sup_{v \in H^{2q} \setminus \{0\}} \frac{\|F_n(kA) v\|_{L^2(\Omega)}}{\|v\|_{2q}} = \sup_{v \in H^{2q} \setminus \{0\}} \frac{\|F_n(kA) v\|_{L^2(\Omega)}}{\|A^q v\|_{L^2(\Omega)}} \\ &= \sup_{v \in L^2 \setminus \{0\}} \frac{\|F_n(kA) A^{-q} v\|_{L^2(\Omega)}}{\|v\|_{L^2(\Omega)}} = \|F_n(kA) A^{-q}\|_{L^2(\Omega) \leftarrow L^2(\Omega)} \\ &= \|A^{-q} F_n(kA)\|_{L^2(\Omega) \leftarrow L^2(\Omega)} \stackrel{!}{\leq} Ck^q \quad \forall t_n \geq 0. \end{aligned}$$

Analog wie (11.12) ergibt sich

$$\|A^{-q} F_n(kA)\|_{L^2(\Omega) \leftarrow L^2(\Omega)} = \sup_{\lambda \in \sigma(A)} |\lambda^{-q} F_n(k\lambda)| = k^q \sup_{\lambda \in \sigma(kA)} |\lambda^{-q} F_n(\lambda)| \stackrel{!}{\leq} Ck^q.$$

Daher zeigen wir im folgenden

$$|F_n(\lambda)| \leq C\lambda^q \quad \forall \lambda \in \sigma(kA).$$

Die Abschätzung (11.10) ergibt für hinreichend kleines  $\lambda_0 > 0$

$$|r(\lambda) - e^{-\lambda}| \leq C\lambda^{q+1} \quad \forall 0 \leq \lambda \leq \lambda_0.$$

Die Dreiecksungleichung liefert

$$|r(\lambda)| \leq |e^{-\lambda} - r(\lambda)| + |e^{-\lambda}| \leq C\lambda^{q+1} + e^{-\lambda}.$$

Eine Taylorentwicklung um  $\lambda = 0$  und festes  $c \in ]0, 1[$  liefert

$$\begin{aligned} C\lambda^{q+1} + e^{-\lambda} - e^{-c\lambda} &= C\lambda^{q+1} + (1 - \lambda) - (1 - c\lambda) + O(\lambda^2) \\ &= (c - 1)\lambda + O(\lambda^2). \end{aligned}$$

Aus  $c - 1 < 0$  folgt die Existenz eines  $\lambda_1 > 0$ , so dass

$$C\lambda^{q+1} + e^{-\lambda} - e^{-c\lambda} \leq 0 \quad \forall 0 \leq \lambda \leq \lambda_1.$$

Das impliziert

$$|r(\lambda)| \leq e^{-c\lambda} \quad \forall 0 \leq \lambda \leq \lambda_1. \quad (11.17)$$

Setzen wir  $\bar{\lambda} := \min\{\lambda_0, \lambda_1\}$ , erhalten wir

$$|F_n(\lambda)| = \left| (r(\lambda) - e^{-\lambda}) \sum_{j=0}^{n-1} \underbrace{r(\lambda)^{n-1-j}}_{\leq e^{-c\lambda(n-1-j)}} e^{-j\lambda} \right| \leq Cn\lambda^{q+1} e^{-c(n-1)\lambda} \leq \tilde{C}\lambda^q \quad (11.18)$$

für alle  $0 \leq \lambda \leq \bar{\lambda}$ . Aus der Stabilität von  $r$  folgt für  $\lambda \geq \bar{\lambda}$  und  $\lambda \in \sigma(kA)$  :

$$|F_n(\lambda)| \leq |r(\lambda)^n| + e^{-n\lambda} \leq 2 \leq C\lambda^q$$

mit  $C = 2\bar{\lambda}^{-q}$ . ■

Satz 11.9 erfordert glatte Anfangsdaten, d.h.,  $v \in H^{2q}$  und macht keine Aussage für weniger glatte Funktionen  $v$ . Um für diese Fälle Fehlerabschätzungen herzuleiten, klassifizieren wir die erzeugenden Funktionen  $r$  in vier Klassen:

- I.  $|r(\lambda)| < 1 \quad \forall 0 < \lambda < \alpha \quad \text{für ein } \alpha > 0,$
- II.  $|r(\lambda)| < 1 \quad \forall 0 < \lambda < \infty,$
- III.  $|r(\lambda)| < 1 \quad \forall 0 < \lambda \text{ und } |r(\infty)| < 1.$
- IV.  $|r(\lambda)| < 1 \quad \forall 0 < \lambda \text{ und } |r(\infty)| = 0.$

In unserem Fall ist  $A = -\Delta_S$  ein beschränkter linearer Operator, der vom Ortsdiskretisierungsparameter  $h$  abhängt. Sei  $\lambda_{\max}$  der grösste Eigenwert von  $A$  und  $k$  die Zeitschrittweite. Dann ist das grösste in (11.13) auftretende  $\lambda$  durch  $k\lambda_{\max}$  gegeben. Das führt zur Definition der Unterklassen:

I'  $r(\lambda)$  ist von Typ I und  $k\lambda_{\max} \leq \alpha_0$  für ein  $\alpha_0 \in ]0, \alpha[$ ,

II'  $r(\lambda)$  ist von Typ II und  $k\lambda_{\max} \leq \alpha_1$  für ein  $\alpha_1 \in ]0, \infty[$ .

Diese Bedingungen sind als Einschränkungen der Zeitschrittweite in Abhängigkeit der Ortsschrittweite zu interpretieren.

**Bemerkung 11.10** Für Verfahren vom Typ I' bzw. II', welche die Konsistenzbedingung (11.10) für ein  $q \geq 1$  erfüllen, gilt mit  $\lambda_1 = \alpha_0$  bzw.  $\lambda_1 = \alpha_1$

$$|r(\lambda)| < 1 \quad \forall 0 < \lambda \leq \lambda_1$$

und daher auch (11.17).

Insbesondere gilt wegen  $k\lambda \leq \lambda_1$  die Abschätzung  $|r(\lambda)| \leq e^{-c\lambda}$  für  $\lambda \in \sigma(kA)$  mit einem  $c \in ]0, 1[$ .

Bevor wir zur Fehleranalyse dieser Verfahren kommen, geben wir Hinweise zur Konstruktion von Verfahren zu den eingeführten Klassen.

In der Vorlesung Numerik I haben wir die Approximation von Funktionen durch (stückweise) Polynome behandelt. Effizienter ist die rationale Approximation, deren Konstruktion jedoch komplizierter ist. (Das liegt daran, dass die Menge

$$\left\{ \frac{p}{q} : p \in \mathbb{P}_n, q \in \mathbb{P}_m \setminus \{0\} \right\}$$

kein Vektorraum ist.) Eine optimale Approximation der Funktion  $e^{-\lambda}$  ist durch die Padé-Approximation gegeben:

$$r_{m,n}(\lambda) = \frac{p_{m,n}(\lambda)}{d_{m,n}(\lambda)} \quad \text{mit} \quad p_{m,n} \in \mathbb{P}_n \text{ und } d_{m,n} \in \mathbb{P}_m.$$



Genauer gilt explizit

$$p_{m,n}(\lambda) := \sum_{j=0}^n \binom{n}{j} / \binom{n+m}{j} \frac{(-\lambda)^j}{j!},$$

$$d_{m,n}(\lambda) = \sum_{j=0}^m \binom{m}{j} / \binom{n+m}{j} \frac{\lambda^j}{j!}.$$

Die Wahl der Koeffizienten ist dadurch bestimmt, dass in der Taylorentwicklung von  $e^{-\lambda}$  um 0 möglichst viele Terme übereinstimmen. Es gilt

$$r_{m,n}(\lambda) = e^{-\lambda} + O(\lambda^{m+n+1}) \quad \text{für } \lambda \rightarrow 0,$$

d.h.  $r_{m,n}$  approximiert  $e^{-\lambda}$  mit Ordnung  $q = m + n$ .

Offensichtlich ist  $r_{m,n}$  vom Typ II für  $m = n \geq 1$ , vom Typ IV für  $m > n$  und vom Typ I für  $m < n$ . Insbesondere ergibt sich für  $r_{0,1}(\lambda) = 1 - \lambda$  das explizite Euler-Verfahren:

$$U^{n+1} = (I - kA)U^n.$$

Diese rationale Funktion ist vom Typ I mit  $\alpha = 2$ . Für  $A = -\Delta_S$  kann man  $\|-\Delta_S\|_{L^2(\Omega) \leftarrow L^2(\Omega)} \leq Ch^{-2}$  zeigen. Daher ist das explizite Euler-Verfahren vom Typ I' unter der Voraussetzung  $k/h^2 \leq \alpha_0/C$  mit  $\alpha_0 < 2$ .

Die subdiagonale bzw. diagonale Padé-Approximation mit linearem Nenner sind

$$r_{1,0} = \frac{1}{1 + \lambda} \quad \text{bzw.} \quad r_{1,1}(\lambda) = \frac{1 - \lambda/2}{1 + \lambda/2},$$

d.h. das implizite Euler-Verfahren bzw. das Crank-Nicolson-Schema. Diese Verfahren sind vom Typ IV bzw. vom Typ II.

Wir kommen nun zur Fehlerabschätzungen für nichtglatte Daten.

**Satz 11.11** *Das Zeitschrittverfahren besitze die Konsistenzordnung  $q$  und sei vom Typ I', II' oder III. Dann erfüllt die Lösungen von (11.5) und (11.7) die Fehlerabschätzung*

$$\|U^n - u(t_n)\|_{L^2(\Omega)} \leq Ck^q t_n^{-q} \|v\|_{L^2(\Omega)} \quad \text{für } t_n > 0.$$

Für Verfahren vom Typ III ist die Konstante  $C$  ist unabhängig von  $A$  und in den Fällen I' und II' hängt die Konstante lediglich von den Parametern  $\alpha_0$  und  $\alpha_1$  ab.

**Beweis.** Wir verwenden die Notationen des Beweises von Satz 11.9 und müssen daher zeigen, dass

$$\|F_n(kA)\|_{L^2(\Omega) \leftarrow L^2(\Omega)} \leq Ck^q t_n^{-q} \quad \text{für } t_n > 0.$$

Mit Hilfe der Spektraldarstellung folgert man, dass dies äquivalent ist zur folgenden Abschätzung

$$|F_n(\lambda)| \leq Ck^q t_n^{-q} = Cn^{-q} \quad \forall \lambda \in \sigma(kA), \quad n \geq 1. \quad (11.19)$$

Wir erinnern daran (vgl. Bemerkung 11.10), dass (11.17) gilt mit  $\lambda_1 = \alpha_0$  bzw.  $\lambda_1 = \alpha_1$  für Verfahren vom Typ I' bzw. II' und für Verfahren vom Typ III für jedes  $\lambda_1 > 0$ . Daraus folgt mit (11.18)

$$|F_n(\lambda)| \leq Cn^{-q} (n\lambda)^{q+1} e^{-cn\lambda} \leq Cn^{-q} \quad \forall 0 \leq \lambda \leq \lambda_1.$$

Für Verfahren vom Typ I' und II' ist damit der Beweis von (11.19) abgeschlossen, da in diesem Fall  $k\lambda_{\max} \leq \lambda_1$  gilt. Für Verfahren vom Typ III müssen wir auch grosse Werte von  $\lambda$  betrachten. Für  $\lambda \geq \lambda_1 = 1$  (man beachte, dass (11.17) gilt für beliebig grosses, aber festes  $\lambda_1$ ) gilt  $e^{-n\lambda} \leq e^{-cn} \leq Cn^{-q}$  für jedes  $0 < c < 1$ . Des weiteren gilt wegen  $|r(\infty)| < 1$  die Gleichheit  $\sup_{\lambda \geq 1} |r(\lambda)| = e^{-c}$  mit einem  $0 < c < 1$ , so dass  $\sup_{\lambda \geq 1} |r(\lambda)^n| \leq e^{-cn} \leq Cn^{-q}$  und daher

$$\sup_{\lambda \geq 1} |F_n(\lambda)| \leq Cn^{-q}.$$

Damit ist auch für Verfahren vom Typ III die Behauptung bewiesen. ■

Die Fehlerschranke in Satz 11.11 wird gross für kleine Zeiten  $t$ . Daher ist es naheliegend, zu Beginn der Berechnungen kleinere Zeitschritte zu verwenden, um dadurch uniformere Abschätzungen zu erhalten. Wir analysieren dieses Vorgehen an Hand des impliziten Euler-Verfahrens.

Wir betrachten eine Zerlegung der positiven Zeitachse gemäss

$$0 = t_0 < t_1 < \dots < t_n < \dots$$

Die Intervalle bezeichnen wir mit  $J_n = ]t_{n-1}, t_n[$  und die Intervalllänge mit  $k_n = t_n - t_{n-1}$ . Die Approximation  $U^n$  der exakten Lösung (11.5) ist definiert durch

$$\bar{\partial}_n U^n + AU^n = 0 \quad \forall n \geq 1 \quad \text{mit} \quad U^0 = v. \quad (11.20)$$

Hier bezeichnet  $\bar{\partial}_n U^n$  die Rückwärtsdifferenz  $(U^n - U^{n-1})/k_n$ .

**Satz 11.12** *Für die Lösungen (11.20) und (11.5) gilt*

$$\|U^n - u(t_n)\|_{L^2(\Omega)} \leq \sum_{j=1}^n k_j \int_{J_j} \|\ddot{u}\|_{L^2(\Omega)} dt, \quad \forall t_n \geq 0.$$

**Beweis.** Die Lösung lässt sich in der Form schreiben

$$U^n = E_{k_n} U^{n-1}, \quad \text{für } n \geq 1 \text{ mit } E_k = (I + kA)^{-1}, \quad U^0 = v$$

bzw. in der prägnanteren Form

$$U^n = E_{n,1} v \quad \text{mit } E_{n,j} = E_{k_n} E_{k_{n-1}} \cdots E_{k_j} \quad \text{für } j \leq n.$$

Der Fehler  $\eta^n = U^n - u^n$  erfüllt

$$\bar{\partial}_n \eta^n + A\eta^n = \omega^n := -\bar{\partial}_n u^n - Au^n = \dot{u}^n - \bar{\partial}_n u^n.$$

Daraus folgt

$$\eta^n = E_{k_n} \eta^{n-1} + k_n E_{k_n} \omega^n.$$

Iteration dieser Rekursion ergibt mit  $\eta^0 = 0$

$$\eta^n = \sum_{j=1}^n k_j E_{n,j} \omega^j, \quad \text{für } n \geq 1.$$

Wie zuvor gilt  $\|E_k\|_{L^2(\Omega) \leftarrow L^2(\Omega)} \leq 1$ , so dass auch  $\|E_{n,j}\|_{L^2(\Omega) \leftarrow L^2(\Omega)} \leq 1$  gilt. Daraus folgt

$$\|\eta^n\|_{L^2(\Omega)} \leq \sum_{j=1}^n k_j \|\omega^j\|_{L^2(\Omega)}.$$

Die Behauptung folgt schliesslich aus

$$\|\omega^j\|_{L^2(\Omega)} = \|\dot{u}(t_j) - \bar{\partial}_j u(t_j)\|_{L^2(\Omega)} \leq \int_{J_j} \|\ddot{u}\|_{L^2(\Omega)} dt.$$

■

Bislang bezogen sich die Fehlerabschätzungen lediglich auf die reine Zeitdiskretisierung. Im folgenden wollen wir Fehlerabschätzungen herleiten, die sich auf das *volldiskrete* Galerkin-Zeitschrittverfahren beziehen und explizit ist in der Zeitschrittweite  $k$  und Ortsschrittweite  $h$ . Wir beginnen wieder mit dem Fall von glatten Daten. Die Glattheit wird mit Hilfe des Raumes  $\dot{H}^s(\Omega)$  ausgedrückt:

$$\dot{H}^s(\Omega) := \{v \in H^s(\Omega) \mid \forall j < s/2: \Delta^j v = 0 \text{ im Sinne der Spurbildung}\}.$$

Für unsere Anwendungen wird der Fall  $s = 1, 2$  eine wesentliche Rolle spielen, für den gilt

$$\dot{H}^s(\Omega) := H_0^1(\Omega) \cap H^s(\Omega).$$

Für das Folgende müssen wir erst eine Reihe von Bezeichnungen einführen. Der Finite-Elemente-Raum zum Poisson-Modellproblem wird wieder mit  $S \subset H_0^1(\Omega)$  bezeichnet und  $T_S : L^2(\Omega) \rightarrow S$  bezeichnet den Lösungsoperator, der jeder rechten Seite  $f \in L^2(\Omega)$  die eindeutige Galerkin-Lösung mittels

$$a(T_S f, v) = (f, v)_{L^2(\Omega)} \quad \forall v \in S$$

zuordnet. Analog ist der kontinuierliche Lösungsoperator  $T : L^2(\Omega) \rightarrow H_0^1(\Omega)$  durch

$$a(Tf, v) = (f, v)_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega)$$

definiert. Man beachte, dass die Annahme  $f \in L^2(\Omega)$  und die Voraussetzung, dass  $\Omega$  glatt berandet ist, implizieren, dass  $Tf \in \dot{H}^2(\Omega)$  gilt. Wir setzen  $T_S^0 := I$  und bezeichnen die  $j$ -fache Hintereinanderausführung mit  $T_S^j$ . Man beachte auch, dass  $T(-\Delta) = I$  und  $T_S(-\Delta_S) = I$  gilt. Die folgende Annahme ist im folgenden Kapitel immer vorausgesetzt.

### Annahme 11.13

1. Der Operators  $T_S$  ist selbstadjungert und positiv semidefinit auf  $L^2(\Omega)$  und positiv definit auf  $S$ .
2. Es existiert eine positive natürliche Zahl  $r \geq 2$ , so dass

$$\|(T_S - T)f\|_{L^2(\Omega)} \leq Ch^s \|f\|_{H^{s-2}(\Omega)} \quad \forall 2 \leq s \leq r \quad \forall f \in H^{s-2}(\Omega).$$

Wir verwenden die Notationen:

$U^n$ : volldiskrete Lösung,

$u_S(t)$ : semidiskrete Lösung von (11.2),

$u(t)$ : exakte Lösung der parabolischen Differentialgleichung.

**Satz 11.14** *Es gelte Annahme 11.13. Sei die Zeitdiskretisierung ein Einschrittverfahren vom Typ I' oder II und wir nehmen an, dass*

1. die Anfangswerte  $v \in \dot{H}^{\max\{r, 2q\}}$  erfüllen,
2.  $\|v_S - v\|_{L^2(\Omega)} \leq Ch^r |v|_r$  gilt. Dann erfüllt die volldiskrete Galerkin-Lösung die Fehlerabschätzung

$$\|U^n - u(t_n)\|_{L^2(\Omega)} \leq C \left( h^r |v|_{H^r(\Omega)} + k^q |v|_{2q} \right) \quad \forall t_n \geq 0.$$

Für den Beweis benötigen wir die folgenden beiden Lemmata.

**Lemma 11.15** *Es gilt*

$$v = \sum_{j=0}^{q-1} T_S^j (T - T_S) (-\Delta)^{j+1} v + T_S^q (-\Delta)^q v \quad \forall v \in \dot{H}^{2q}(\Omega). \quad (11.21)$$

**Beweis.** Wegen  $T(-\Delta) = I$  folgt die Gleichheit durch einfaches Ausmultiplizieren. ■

**Lemma 11.16** *Sei die Zeitdiskretisierung ein Einschrittverfahren vom Typ I' oder II. Dann gilt*

$$\|F_n(-k\Delta_S) P_S T_S^j\|_{L^2(\Omega) \leftarrow L^2(\Omega)} \leq \|F_n(-k\Delta_S) P_S T_S^j\|_{S \leftarrow S} \leq Ck^j \quad \forall 0 \leq j \leq q, \quad \forall n \geq 0 \quad (11.22)$$

mit  $F_n$  wie in (11.16) und der (bezüglich  $(\cdot, \cdot)_{L^2(\Omega)}$ ) orthogonalen Projektion  $P_S : L^2(\Omega) \rightarrow S$ ,

**Beweis.** Da  $P_S$  eine Projektion ist, gilt  $T_S = T_S P_S$  und auch  $P_S T_S^j = T_S^j = (-\Delta_S)^{-j}$  für all  $j > 0$ . Daraus folgt mit Hilfe einer Entwicklung in ein orthonormales Eigensystem  $(\lambda_i, \varphi_i)_{i=1}^N$  von  $-\Delta_S$  für beliebiges  $v \in S$

$$\begin{aligned} \|F_n(-k\Delta_S) P_S T_S^j v\|_{L^2(\Omega)} &= \left\| \sum_{i=1}^N (v, \varphi_i)_{L^2(\Omega)} F_n(-k\Delta_S) \lambda_i^{-j} \varphi_i \right\|_{L^2(\Omega)} \\ &= \left\| k^j \sum_{i=1}^N (v, \varphi_i)_{L^2(\Omega)} F_n(k\lambda_i) (k\lambda_i)^{-j} \varphi_i \right\|_{L^2(\Omega)} \leq k^j \sup_{\lambda \in \sigma(-k\Delta_S)} |\lambda^{-j} F_n(\lambda)| \|v\|_{L^2(\Omega)}. \end{aligned}$$

Sei  $\lambda_0$  eine positive Zahl, so dass  $|r(\lambda)| < 1$  gilt für alle  $0 < \lambda \leq \lambda_0$ . Dann implizieren die Annahmen des Satzes, dass für derartige  $\lambda$  gilt

$$|r(\lambda) - e^{-\lambda}| \leq C\lambda^{j+1} \quad 0 \leq j \leq q \quad \text{und} \quad |r(\lambda)| \leq e^{-c\lambda} \quad \text{für ein geeignetes } c \in ]0, 1[.$$

Daraus folgt für alle  $0 < \lambda \leq \lambda_0$

$$|\lambda^{-j} F_n(\lambda)| = |\lambda^{-j} (r(\lambda) - e^{-\lambda})| \left| \sum_{\ell=0}^{n-1} r(\lambda)^{n-1-\ell} e^{-\ell\lambda} \right| \leq Cn\lambda e^{-cn\lambda} \leq C.$$

Für Einschrittverfahren vom Typ I ist damit der Beweis gegeben. Für Verfahren vom Typ II, ist die gewünschte Ungleichung trivial für  $\lambda > \lambda_0$ .

Damit ist in beiden Fällen die rechte Ungleichung in (11.22) gezeigt. Die linke Ungleichung folgt wegen

$$\begin{aligned} \sup_{v \in L^2(\Omega) \setminus \{0\}} \frac{\|F_n(-k\Delta_S) P_S T_S^j v\|_{L^2(\Omega) \leftarrow L^2(\Omega)}}{\|v\|_{L^2(\Omega)}} &= \sup_{v \in L^2(\Omega) \setminus \{0\}} \frac{\|F_n(-k\Delta_S) P_S T_S^j P_S v\|_{L^2(\Omega) \leftarrow L^2(\Omega)}}{\|v\|_{L^2(\Omega)}} \\ &= \left( \sup_{w \in S \setminus \{0\}} \frac{\|F_n(-k\Delta_S) P_S T_S^j w\|_{L^2(\Omega) \leftarrow L^2(\Omega)}}{\|w\|_{L^2(\Omega)}} \right) \left( \sup_{v \in L^2(\Omega) \setminus \{0\}} \frac{\|P_S v\|_{L^2(\Omega)}}{\|v\|_{L^2(\Omega)}} \right). \end{aligned}$$

Da  $P_S$  die  $L^2(\Omega)$ -orthogonal-Projektion ist, ist der letzte Faktor nach oben durch 1 beschränkt.

■

#### Beweis von Satz 11.14:

Die Dreiecksungleichung liefert:

$$\|U^n - u(t_n)\|_{L^2(\Omega)} \leq \|U^n - u_S(t_n)\|_{L^2(\Omega)} + \|u_S(t_n) - u(t_n)\|_{L^2(\Omega)}$$

und wir schätzen beide Terme separat ab.

#### Abschätzung von $\|U^n - u_S(t_n)\|_{L^2(\Omega)}$

Unsere Stabilitätsannahmen an das Einschrittverfahren implizieren mit  $E_{kS}v := e^{-k\Delta_S}v$

$$\|E_{kS}^n(v_S - P_S v)\|_{L^2(\Omega)} \leq \|v_S - v\|_{L^2(\Omega)} + \|v - P_S v\|_{L^2(\Omega)} \leq Ch^r |v|_{H^r(\Omega)}$$

und daher genügt es im folgenden den Fall  $v_S = P_S v$  zu untersuchen. Sei  $U^n$  die volldiskrete Lösung und  $u_S(t_n)$  die Lösung des semidiskreten Problems (11.2) zum Zeitpunkt  $t_n$ . Dann gilt

$$U^n - u_S(t_n) = F_n(-k\Delta_S) P_S v.$$

Sei nur  $(\lambda_\ell, \varphi_\ell)_{\ell=1}^N$  ein orthonormales Eigensystem für  $-\Delta$  mit Nullrandwerten. Wir setzen

$$v_k = \sum_{\ell: k\lambda_\ell < 1} (v, \varphi_\ell)_{L^2(\Omega)} \varphi_\ell.$$

Beachte, dass wegen  $-\Delta\varphi_\ell = \lambda_\ell\varphi_\ell$  die Nullrandwerte bei Anwendung von  $-\Delta$  erhalten bleiben und somit  $v_k \in \dot{H}^s(\Omega)$  für alle  $s \geq 0$  gilt. Darüber hinaus folgt

1.

$$\|v - v_k\|_{L^2(\Omega)} = \left\| \sum_{\ell: k\lambda_\ell \geq 1} (v, \varphi_\ell)_{L^2(\Omega)} \varphi_\ell \right\|_{L^2(\Omega)} = \left\| \sum_{\ell: k\lambda_\ell \geq 1} (v, \varphi_\ell)_{L^2(\Omega)} \lambda_\ell^{-q} (-\Delta)^q \varphi_\ell \right\|_{L^2(\Omega)} \quad (11.23a)$$

$$= \left\| \sum_{\ell: k\lambda_\ell \geq 1} (v, \varphi_\ell)_{L^2(\Omega)} \lambda_\ell^{-q} \varphi_\ell \right\|_{2q} = \left( \sup_{k\lambda_\ell \geq 1} \lambda_\ell^{-q} \right) \|v\|_{2q} \leq k^q \|v\|_{2q}. \quad (11.23b)$$

2.

$$|v_k|_{2q} \leq |v|_{2q} \quad \text{da } v_k \text{ und } v - v_k \text{ orthogonal sind,} \quad (11.23c)$$

3.

$$|v_k|_{r+2j} = \left\| (-\Delta)^{r/2+j} v_k \right\|_{L^2(\Omega)} = \left\| \sum_{\ell: k\lambda_\ell < 1} (v, \varphi_\ell)_{L^2(\Omega)} \lambda_\ell^{r/2+j} \varphi_\ell \right\|_{L^2(\Omega)} \quad (11.23d)$$

$$\leq \sup_{\ell: k\lambda_\ell < 1} \lambda_\ell^j \left\| \sum_{\ell: k\lambda_\ell < 1} (v, \varphi_\ell)_{L^2(\Omega)} \lambda_\ell^{r/2} \varphi_\ell \right\|_{L^2(\Omega)} \leq \left( \sup_{\ell: k\lambda_\ell < 1} \lambda_\ell^j \right) \|v\|_r \leq k^{-j} \|v\|_r, \quad (11.23e)$$

für alle  $j = 0, 1, \dots, q-1$ .

Wir wenden nun die Identität (11.21) auf  $v_k$  an und verwenden die Abkürzung  $F_n = F_n(-k\Delta_S)P_S$ . Damit ergibt sich

$$F_n v_k = \sum_{j=0}^{q-1} F_n T_S^j (T - T_S) (-\Delta)^{j+1} v_k + F_n T_S^q (-\Delta)^q v_k.$$

Aus Lemma 11.16 und (11.23c) folgt

$$\|F_n T_S^q (-\Delta)^q v_k\|_{L^2(\Omega)} \leq C k^q \|\Delta^q v_k\|_{L^2(\Omega)} = C k^q |v_k|_{2q} \stackrel{(11.23c)}{\leq} C k^q |v|_{2q}.$$

Indem wir die Eigenschaft aus Annahme 11.13(2) ausnützen und mit (11.23e) verbinden, ergibt sich

$$\begin{aligned} \left\| F_n T_S^j (T - T_S) (-\Delta)^{j+1} v_k \right\|_{L^2(\Omega)} &\stackrel{\text{Lem. 11.16}}{\leq} C k^j \left\| (T - T_S) (-\Delta)^{j+1} v_k \right\|_{L^2(\Omega)} \\ &\stackrel{\text{Ann. 11.13(2)}}{\leq} C k^j h^r \left\| (-\Delta)^{j+1} v_k \right\|_{H^{r-2}(\Omega)} \\ &\stackrel{(11.23e)}{\leq} C k^j h^r |v_k|_{r+2j} \leq C h^r |v_k|_r \end{aligned}$$

für alle  $0 \leq j \leq q-1$ . Zusammen ergeben diese Abschätzungen

$$\|F_n v_k\|_{L^2(\Omega)} \leq C \left( h^r |v|_r + k^q |v|_{2q} \right).$$

Die Stabilität des Einschrittverfahrens und (11.23b) ergeben

$$\|F_n (v - v_k)\|_{L^2(\Omega)} \leq 2 \|v - v_k\|_{L^2(\Omega)} \stackrel{(11.23b)}{\leq} C k^q |v|_{2q}.$$

**Abschätzung von**  $\|u_S(t_n) - u(t_n)\|_{L^2(\Omega)}$

Diese Abschätzung ergibt sich als Folge von

**Hilfsaussage 1:** Es gelte

$$T_S \dot{e} + e = \rho \quad \text{für } t \geq 0 \text{ mit } T_S e(0) = 0. \quad (11.24)$$

Dann gilt

$$\|e(t)\|_{L^2(\Omega)}^2 \leq C \|\rho(t)\|_{L^2(\Omega)}^2 + \frac{1}{t} \int_0^t \left( \|\rho\|_{L^2(\Omega)}^2 + s^2 \|\dot{\rho}\|_{L^2(\Omega)}^2 \right) ds.$$

**Beweis der Hilfsaussage 1:**

Wir multiplizieren (11.24) skalar mit  $2\dot{e}$  und erhalten

$$2(T_S \dot{e}, \dot{e})_{L^2(\Omega)} + \frac{d}{dt} \|e\|_{L^2(\Omega)}^2 = 2(\rho, \dot{e})_{L^2(\Omega)}.$$

Da  $T_S$  nach Annahme positiv semidefinit ist, folgt

$$\frac{d}{dt} \|e\|_{L^2(\Omega)}^2 \leq 2(\rho, \dot{e})_{L^2(\Omega)} = 2\frac{d}{dt}(\rho, e)_{L^2(\Omega)} - 2(\dot{\rho}, e).$$

Wir multiplizieren mit  $t$  und erhalten

$$\frac{d}{dt} (t \|e\|_{L^2(\Omega)}^2) \leq 2\frac{d}{dt} (t(\rho, e)_{L^2(\Omega)}) - 2t(\dot{\rho}, e) + \|e\|_{L^2(\Omega)}^2 - 2(\rho, e)_{L^2(\Omega)}.$$

Integration nach  $t$  liefert

$$t \|e\|_{L^2(\Omega)}^2 \leq 2t \|\rho\|_{L^2(\Omega)} \|e\|_{L^2(\Omega)} + \int_0^t \left( 2s \|\dot{\rho}\|_{L^2(\Omega)} \|e\|_{L^2(\Omega)} + \|e\|_{L^2(\Omega)}^2 + 2\|\rho\|_{L^2(\Omega)} \|e\|_{L^2(\Omega)} \right) ds,$$

d.h.

$$\|e\|_{L^2(\Omega)}^2 \leq C \left( \|\rho(t)\|_{L^2(\Omega)}^2 + \frac{1}{t} \int_0^t s^2 \|\dot{\rho}\|_{L^2(\Omega)}^2 + \|e\|_{L^2(\Omega)}^2 + \|\rho\|_{L^2(\Omega)}^2 \right). \quad (11.25)$$

Multipliziert man (11.24) mit  $e$ , ergibt sich

$$\frac{d}{dt} (T_S e, e)_{L^2(\Omega)} + 2\|e\|_{L^2(\Omega)}^2 = \|\rho\|_{L^2(\Omega)}^2 + \|e\|_{L^2(\Omega)}^2.$$

Integration liefert

$$(T_S e, e)_{L^2(\Omega)} + \int_0^t \|e\|_{L^2(\Omega)}^2 ds \leq (T_S e(0), e(0))_{L^2(\Omega)} + \int_0^t \|\rho\|_{L^2(\Omega)}^2 ds.$$

Da  $T_S$  positiv semidefinit ist und  $T_S e(0)$  wegen (11.24) verschwindet folgt

$$\int_0^t \|e\|_{L^2(\Omega)}^2 ds \leq \int_0^t \|\rho\|_{L^2(\Omega)}^2 ds.$$

Zusammen mit (11.25) ergibt sich die Hilfsbehauptung 1

$$\|e\|_{L^2(\Omega)}^2 \leq C \left( \|\rho(t)\|_{L^2(\Omega)}^2 + \frac{1}{t} \int_0^t s^2 \|\dot{\rho}\|_{L^2(\Omega)}^2 + \|\rho\|_{L^2(\Omega)}^2 \right).$$

Aus Hilfsbehauptung 1 folgt direkt

$$\|e(t)\|_{L^2(\Omega)} \leq C \sup_{s \leq t} \left( s \|\dot{\rho}(s)\|_{L^2(\Omega)} + \|\rho(s)\|_{L^2(\Omega)} \right) \quad \forall t > 0. \quad (11.26)$$

Der Fehler  $e(t) = u_S(t) - u(t)$  erfüllt die Gleichung (11.24) mit

$$\rho(t) = -(T_S - T) \Delta u = -(T_S - T) \dot{u}.$$

Weiter gilt für die Anfangsbedingung in (11.24) mit Annahme 11.13(1) (genauer:  $T_S : S \rightarrow S$  ist bijektiv)

$$(T_S e(0), w)_{L^2(\Omega)} = (P_S v - v, T_S w)_{L^2(\Omega)} = 0 \quad \forall w \in S,$$

da  $P_S$  die  $L^2(\Omega)$ -Orthogonalprojektion auf  $S$  ist und damit  $T_S e(0) = 0$  gilt. Daher lässt sich (11.26) anwenden, und es bleibt die beiden  $\rho$ -Terme auf der rechten Seite anzuschätzen. Für die  $L^2(\Omega)$ -Norm von  $\rho$  erhalten wir

$$\begin{aligned} \|\rho(s)\|_{L^2(\Omega)} &\leq \|(T_S - T) \dot{u}\|_{L^2(\Omega)} \stackrel{\text{Ann. 11.13(2)}}{\leq} Ch^r \|\dot{u}\|_{H^{r-2}(\Omega)} \\ &\leq Ch^r \|u(s)\|_{H^r(\Omega)} \end{aligned}$$

und analog

$$s \|\dot{\rho}(s)\|_{L^2(\Omega)} \leq Ch^r s \|\dot{u}(s)\|_{H^r(\Omega)}.$$

Schliesslich müssen die Normen von  $u(s)$  und  $s\dot{u}(s)$  noch auf die Anfangswerte zurücktransportiert werden. Dies folgt aus der Spektraldarstellung der Anfangswerte: Für

$$v = \sum_{i=1}^N (v, \varphi_i)_{L^2(\Omega)} \varphi_i$$

ist die Lösung  $u$  durch

$$u(s) = E(s) v = \sum_{i=1}^N (v, \varphi_i)_{L^2(\Omega)} e^{-\lambda_i s} \varphi_i \quad (11.27)$$

gegeben. Die Orthonormalität der Eigenvektorbasis und  $|e^{-\lambda_i s}| \leq 1$  implizieren direkt

$$|u(s)|_r \leq |v|_r.$$

Man beachte, dass  $\dot{u} = \Delta u$  gilt und daraus folgt

$$\begin{aligned} s \|u(s)\|_{r+2} &= \left\| \sum_{i=1}^N (v, \varphi_i)_{L^2(\Omega)} \lambda_i^{r/2+1} s e^{-\lambda_i s} \varphi_i \right\|_{L^2(\Omega)} \\ &\leq \sqrt{\sum_{i=1}^N (v, \varphi_i)_{L^2(\Omega)} \lambda_i^r (s \lambda_i)^2 e^{-2\lambda_i s}}. \end{aligned}$$

Man beachte, dass  $x^2 e^{-2x}$  für  $x \geq 0$  uniform durch eine Konstante  $C$  beschränkt ist so, dass die Stabilität

$$s \|\dot{u}(s)\|_r \leq C \|v\|_r$$

schliesslich folgt. ■

## 12 Das unstetige Galerkin-Zeitschrittverfahren

Im vorigen Abschnitt haben wir die Galerkin-Finite-Elemente-Methode zur Ortsdiskretisierung verwendet und Einschrittverfahren für das ortsdiskretisierte Problem betrachtet. In diesem Abschnitt werden wir auch für die Zeitdiskretisierung ein *Galerkin*-Verfahren verwenden.



Die Ansatzfunktionen werden in der Regel unstetig sein, und man spricht daher vom unstetigen Galerkin-Verfahren. Die Abkürzung „DG“ oder „DGM“ ist vom Englischen: „Discontinuous Galerkin Method“ abgeleitet.

Die Idee hierbei ist, für die Lösung einen stückweisen (möglicherweise unstetigen) Polynomansatz (in Raum und Zeit) zu verwenden und die darin vorkommenden Koeffizienten mittels der parabolischen Differentialgleichung zu bestimmen.

Wie im vorigen Unterkapitel 11.2.2 konzentrieren wir uns hier auf die Zeitdiskretisierung und verwenden wieder den folgenden abstrakten Rahmen:

Sei  $H$  ein Hilbert-Raum und  $A$  ein selbstadjungierter, positiv definiter, nicht notwendigerweise beschränkter Operator mit kompakter Inversen, dessen Definitionsbereich  $D(A) \subset H$  erfüllt.

**Bemerkung 12.1** *In unserer Anwendung gilt  $H = L^2(\Omega)$  oder  $H = S \subset H_0^1(\Omega)$ . Für hinreichend glatte<sup>10</sup> Funktionen  $u, v : [0, T] \times L^2(\Omega) \rightarrow \mathbb{R}$  erhalten wir mit  $t_N := T$*

$$\begin{aligned} \frac{d}{dt} \int_{\Omega} u(t, x) v(t, x) dx &= \int_{\Omega} \dot{u}(t, x) v(t, x) + u(t, x) \partial_t v(t, x) dx, \\ \int_0^{t_N} \int_{\Omega} v(t, x) \dot{u}(t, x) dx dt &= \int_{\Omega} \int_0^{t_N} v(t, x) \dot{u}(t, x) dt dx \\ &= - \int_{\Omega} \int_0^{t_N} u(t, x) \dot{v}(t, x) dt dx + \int_{\Omega} (v(t, x) u(t, x)) \Big|_{t=0}^{t_N} dx. \end{aligned}$$

bzw. kompakter geschrieben:

$$\frac{d}{dt} (u, v)_H = (\dot{u}, v)_H + (u, \dot{v})_H \quad (12.1a)$$

$$\int_0^{t_N} (\dot{u}, v) dt = - \int_0^{t_N} (u, \dot{v})_H + ((u, v)_H) \Big|_{t=0}^{t_N}. \quad (12.1b)$$

Im folgenden nehmen wir generell an, dass die Relationen (12.1) erfüllt sind.

Wir betrachten das Anfangswertproblem:

$$\dot{u} + Au = f \quad \text{für } t > 0 \text{ mit } u(0) = v. \quad (12.2)$$

Um in der Zeit zu diskretisieren, zerlegen wir die Zeitachse gemäss

$$0 = t_0 < t_1 < \dots < t_n < \dots$$

und setzen  $J_n = ]t_{n-1}, t_n]$  für  $n \in \mathbb{N}$ . Die Schrittweite  $k_n = t_n - t_{n-1}$  muss nicht notwendig uniform sein. Die maximale Schrittweite wird mit  $k := \max k_i$  bezeichnet. Sei  $q$  ein geeigneter maximaler Polynomgrad für den Lösungsansatz auf den einzelnen Zeitintervallen. Die Koeffizienten dieses Ansatzes sind in  $H$ , d.h., der Ansatz liegt im Raum

$$S := \left\{ u : [0, \infty[ \rightarrow H : \forall J_n : \exists (\psi_j)_{j=0}^q \in H^{q+1} : u(t) = \sum_{j=0}^q \psi_j t^j, \quad \forall t \in J_n \right\}.$$

---

<sup>10</sup>Hinreichend glatt in dem Sinne, dass die Vertauschung von Differentiation und Integration zulässig ist und auch partielle Integration angewendet werden kann.

Man beachte, dass diese Funktionen (im Gegensatz zu den Finite-Elemente-Funktionen für die Ortsdiskretisierung) unstetig in den Zeitgitterpunkten  $t_i$  sein können. Da die Zeitintervalle links offen sind, muss  $u(0)$  separat definiert werden. Für eine Funktion  $u \in S$  bezeichnen wir mit  $u^n$  bzw.  $u_+^n$  den Wert der Funktion  $u$  im Gitterpunkt  $t_n$  bzw. den rechtsseitigen Grenzwert in  $t_n$ .

Um das unstetige Galerkin-Zeitschrittverfahren zu definieren, betrachten wir ein festes Intervall  $[0, t_N]$  und bemerken, dass die exakte Lösung von (12.2) für glatte Testfunktionen  $w : [0, t_N] \rightarrow H$  die folgende Gleichung erfüllt:

$$\int_0^{t_N} (\dot{u}, w)_H + (Au, w)_H dt = \int_0^{t_N} (f, w)_H dt.$$

Nach partieller Integration bezüglich der Zeit erhalten wir für alle Funktionen  $w$  mit  $w(t_N) = 0$

$$\int_0^{t_N} -(u, \dot{w})_H + (Au, w)_H dt + (u(t), w(t))_H|_{t=0}^{t_N} = \int_0^{t_N} (f, w)_H dt$$

d.h.

$$\int_0^{t_N} -(u, \dot{w})_H + (Au, w)_H dt = (u(0), w(0))_H + \int_0^{t_N} (f, w)_H dt. \quad (12.3)$$

Wir ersetzen nun  $u$  in der schwachen Formulierung (12.3) durch eine Funktion  $U \in S$  und integrieren partiell auf jedem Zeitintervall  $J_n$ :

$$\sum_{n=1}^N \int_{J_n} -(U, \dot{w})_H dt = \int_0^{t_N} (\partial_{\text{stw}} U, w)_H dt + \sum_{n=1}^{N-1} ([U]_n, w_+^n)_H + (U_+^0, w_+^0)_H, \quad (12.4)$$

wobei der Sprung der Funktion  $U$  im Gitterpunkt  $t_n$  durch  $U^n = U_+^n - U^n$  definiert ist. In (12.4) bezeichnet  $w_+^n$  den rechtsseitigen Grenzwert von  $w$  in  $t_n$ . Da  $w$  als hinreichend glatt (mindestens stetig) angenommen wurde, gilt  $w^n = w_+^n$ . Für die *Diskretisierung* mit unstetigen Funktionen muss jedoch angegeben werden, welcher einseitige Grenzwert zu wählen ist. Die stückweise Ableitung  $\partial_{\text{stw}}$  ist für  $t \in \overset{\circ}{J}_n$  durch

$$\partial_{\text{stw}} U(t) = \dot{U}(t)$$

und auf (der Nullmenge) der Gitterpunkte beliebig definiert. Beispielsweise gilt für  $q = 0$  und jedes  $V \in S$  die Gleichung  $\partial_{\text{stw}} V = 0$ .

Die unstetige Galerkin-Diskretisierung des Problems (12.2) lautet: Finde  $U \in S$ , so dass für alle  $W \in S$  gilt

$$\int_0^{t_N} (\partial_{\text{stw}} U, W)_H + (AU, W)_H dt + \sum_{n=1}^{N-1} ([U]_n, W_+^n)_H + (U_+^0, W_+^0)_H = (v, W_+^0)_H + \int_0^{t_N} (f, W)_H dt, \quad (12.5)$$

$$U^0 = v.$$

Da die Funktion  $W$  aus  $S$  in den Gitterpunkten im allgemeinen nicht stetig ist, dürfen wir ihre Werte auf den unterschiedlichen Zeitintervallen unabhängig voneinander wählen. Das bedeutet, wir können für  $S$  eine Basis angeben, so dass der Träger jeder Basisfunktion immer

nur in einem Intervall  $J_n$  enthalten ist. Mit der Bezeichnung  $S^n := S|_{J_n}$  ergibt sich also: Finde  $U \in S$ , so dass für alle  $W \in S^n$  und alle  $1 \leq n \leq N$  gilt<sup>11</sup>

$$\int_{J_n} (\partial_t U, W)_H + (AU, W)_H dt + (U_+^{n-1}, W_+^{n-1})_H = (U^{n-1}, W_+^{n-1})_H + \int_{J_n} (f, W)_H dt, \quad (12.6)$$

$$U^0 = v.$$

Im nächsten Schritt zeigen wir, dass das lokale Problem (12.6) eine eindeutige Lösung in  $S$  auf  $J_n$  besitzt für gegebenes  $U^{n-1}$  und  $f|_{J_n}$ .

### Eindeutigkeit:

Es genügt zu zeigen, dass die homogene Gleichung nur die triviale Lösung  $U = 0$  besitzt:

$$\int_{J_n} (\dot{U}, W)_H + (AU, W)_H dt + (U_+^{n-1}, W_+^{n-1})_H = 0 \quad \forall W \in S^n.$$

Sei  $U$  eine homogene Lösung und wähle  $W = U$  auf  $J_n$ . Dann gilt wegen  $2(\dot{U}, U)_H = \frac{d}{dt} \|U\|_H^2$  (vgl. (12.1))

$$\|U^n\|_H^2 - \|U_+^{n-1}\|_H^2 + 2 \int_{J_n} (AU, U)_H dt + 2 \|U_+^{n-1}\|_H^2 = 0,$$

das heisst (Erinnerung:  $\|u\|_1 = (Au, u)_H^{1/2}$ )

$$\|U^n\|_H^2 + \|U_+^{n-1}\|_H^2 + 2 \int_{J_n} \|U\|_1^2 dt = 0.$$

Daraus folgt  $\|U\|_1 = 0$  auf  $J_n$  und daher  $U = 0$  auf  $J_n$ . Dies ergibt die Eindeutigkeit.

### Existenz:

Da das Problem endlichdimensional ist, folgt die Existenz aus der Eindeutigkeit.

Die einfachste Situation liegt für den Fall  $q = 0$  vor, d.h. stückweise konstanten Ansätzen. Dann gilt  $\partial_{\text{stw}} U \equiv 0$  und  $U(t) = U_+^{n-1} = U^n$  auf  $J_n$ . Damit ergibt sich

$$(U^n, W)_H + k_n (AU^n, W)_H = (U^{n-1}, W)_H + \left( \int_{J_n} f dt, W \right)_H \quad \forall W \in S,$$

$$U^0 = v$$

bzw in Kurzschreibweise

$$(I + k_n A) U^n = U^{n-1} + \int_{J_n} f dt. \quad (12.7)$$

Gleichung (12.7) kann auch in der Form geschrieben werden

$$\bar{\partial}_n U^n + AU^n = \frac{1}{k_n} \int_{J_n} f(t) dt \quad \text{mit} \quad \bar{\partial}_n U^n = \frac{U^n - U^{n-1}}{k_n}.$$

Daran erkennt man, dass es sich bei diesem Verfahren um das modifizierte implizite Euler-Verfahren handelt, wobei die Punktauswertung  $f(t_n)$  durch einen Integralmittelwert ersetzt

<sup>11</sup>Hinweis: Wir betrachten Gleichung (12.5) mit Testfunktionen  $W \in S$ , die  $W \equiv 0$  auf  $[0, t_N] \setminus J_n$  erfüllen.

wurde. Fasst man die Punktauswertung als Ein-Punkt-Quadraturapproximation des Integralmittelwertes auf, so stimmt das implizite Euler-Verfahren mit der stückweise konstanten DGM mit Quadratur überein.

Wir betrachten noch den Fall der stückweise linearen Approximation, d.h.,  $q = 1$ . Es gilt

$$U(t) = \tilde{U}_0^n + \tilde{U}_1^n (t - t_{n-1}) / k_n \text{ auf } J_n.$$

Für die Bestimmung von  $\tilde{U}_0^n = \tilde{U}_0$  und  $\tilde{U}_1^n = \tilde{U}_1$  erhalten wir das gekoppelte System:

$$\begin{aligned} (\tilde{U}_0, \psi)_H + k_n A (\tilde{U}_0, \psi)_H + (\tilde{U}_1, \psi)_H + \frac{1}{2} k_n A (\tilde{U}_1, \psi)_H &= (U^{n-1}, \psi)_H + \left( \int_{J_n} f(t) dt, \psi \right)_H \\ \frac{1}{2} k_n A (\tilde{U}_0, \eta)_H + \frac{1}{2} (\tilde{U}_1, \eta)_H + \frac{1}{3} k_n (A \tilde{U}_1, \eta)_H &= \left( k_n^{-1} \int_{J_n} (t - t_n) f(t) dt, \eta \right)_H \end{aligned}$$

für alle  $\psi, \eta \in H$ . In Operatorschreibweise erhalten wir die kompaktere Form eines gekoppelten Systems

$$\begin{bmatrix} I + k_n A & I + \frac{1}{2} k_n A \\ \frac{1}{2} k_n A & \frac{1}{2} I + \frac{1}{3} k_n A \end{bmatrix} \begin{pmatrix} \tilde{U}_0 \\ \tilde{U}_1 \end{pmatrix} = \begin{pmatrix} U^{n-1} + \int_{J_n} f(t) dt \\ k_n^{-1} \int_{J_n} (t - t_n) f(t) dt \end{pmatrix}.$$

**Satz 12.2** Die Lösungen von (12.6) (mit  $q \geq 0$ ) und (12.2) erfüllen die Fehlerabschätzung

$$\|U^N - u(t_N)\|_H \leq C \left( \sum_{n=1}^N k_n^{2q+2} \int_{J_n} \|u^{(q+1)}\|_1^2 dt \right)^{1/2} \quad \text{für alle } t_N \geq 0. \quad (12.8)$$

**Beweis.** Wir definieren den Interpolanten  $\tilde{u}$  der exakten Lösung  $u$  von (12.2) durch die Bedingungen

$$\begin{aligned} \tilde{u}(t_n) &= u(t_n) \quad \forall n \geq 0, \\ \int_{J_n} (\tilde{u}(t) - u(t)) t^\ell dt &= 0 \quad \forall 0 \leq \ell \leq q-1, \quad n \geq 1. \end{aligned} \quad (12.9)$$

Damit interpoliert  $\tilde{u}$  in den Knotenpunkten und der Interpolationsfehler ist orthogonal auf  $\mathbb{P}_{q-1}(J_n)$ . Um einzusehen, dass diese Gleichungen ein eindeutiges  $\tilde{u} \in \mathbb{P}_q(J_n)$  definieren, genügt es, eine Entwicklung in  $H$  bzgl. eines Orthonormalsystems zu verwenden. Da die Anzahl der Gleichungen und die Anzahl der Freiheitsgrade gleich  $q+1$  ist, folgt die Behauptung, wenn wir die Implikation zeigen:  $u|_{J_n} \equiv 0 \implies \tilde{u}|_{J_n} \equiv 0$ . Indem wir  $J_n$  auf das Einheitsintervall  $(0, 1)$  transformieren, so dass  $t_n$  auf Null abgebildet wird, genügt es einzusehen, dass aus der Bedingung: „ $e = u - \tilde{u} = t \sum_{j=0}^{q-1} a_j t^j$  ist orthogonal zu  $\mathbb{P}_{q-1}(0, 1)$ “ folgt:  $e = 0$ . Diese Folgerung ergibt sich jedoch aus

$$0 = \int_0^1 e(t) \sum_{j=0}^{q-1} a_j t^j dt = \int_0^1 t \left( \sum_{j=0}^{q-1} a_j t^j dt \right)^2 dt \implies \sum_{j=0}^{q-1} a_j t^j dt \equiv 0.$$

Ausserdem folgt hieraus, dass  $\tilde{u}$  mit  $u$  auf  $J_n$  übereinstimmt, falls  $u \in \mathbb{P}_q$ , m.a.W.: die Interpolation besitzt den Exaktheitsgrad  $q$ .

Standardabschätzungen für den Interpolationsfehler ergeben

$$\|\tilde{u}(t) - u(t)\|_j^2 \leq C k_n^{2q+1} \int_{J_n} \|u^{(q+1)}\|_j^2 dt \quad \forall t \in J_n, \quad j = 0, 1. \quad (12.10)$$

Wir zerlegen den Fehler gemäss

$$U - u = (U - \tilde{u}) + (\tilde{u} - u) = \theta + \rho$$

und beachten  $\rho^n := \rho(t_n) = 0$  für alle  $n \geq 0$ . Es genügt daher, die Grösse  $\theta^n$  durch die rechte Seite von (12.8) zu beschränken. Aus (12.6) und (12.2) erhalten wir<sup>12</sup>

$$\begin{aligned} & \int_{J_n} \left( \dot{\theta}, X \right)_H + (A\theta, X) dt + ([\theta]_{n-1}, X_+^{n-1})_H \\ &= - \int_{J_n} (\dot{\rho}, X)_H + (A\rho, X)_H dt - ([\rho]_{n-1}, X_+^{n-1})_H \quad \forall X \in S. \end{aligned} \quad (12.11)$$

Wir verwenden die Eigenschaften der Interpolierenden  $\tilde{u}$ , um die folgende Gleichheit herzuleiten

$$\begin{aligned} & \int_{J_n} (\dot{\rho}, X)_H dt + ([\rho]_{n-1}, X_+^{n-1})_H \\ &= (\rho, X)|_{t_{n-1}+0}^{t_n} - \int_{J_n} (\rho, \dot{X})_H dt + ([\rho]_{n-1}, X_+^{n-1})_H \\ &= - (\rho_+^{n-1}, X_+^{n-1})_H + (\rho_+^{n-1}, X_+^{n-1})_H = 0 \quad \forall X \in S. \end{aligned}$$

Wählen wir  $X = 2\theta$  in (12.11) ergibt sich zunächst

$$\begin{aligned} & 2 \int_{J_n} \left( \dot{\theta}, \theta \right)_H dt + 2 ([\theta]_{n-1}, \theta_+^{n-1})_H \\ &= \|\theta^n\|_H^2 - \|\theta_+^{n-1}\|_H^2 + 2 \|\theta_+^{n-1}\|_H^2 - 2 (\theta^{n-1}, \theta_+^{n-1})_H \\ &\geq \|\theta^n\|_H^2 - \|\theta_+^{n-1}\|_H^2 + \|\theta_+^{n-1}\|_H^2 - \|\theta^{n-1}\|_H^2 \\ &= \|\theta^n\|_H^2 - \|\theta^{n-1}\|_H^2 \end{aligned}$$

und schliesslich daraus

$$\begin{aligned} \|\theta^n\|_H^2 + 2 \int_{J_n} \|\theta\|_1^2 dt &\leq \|\theta^{n-1}\|_H^2 + 2 \int_{J_n} |(A\rho, \theta)_H| dt \\ &\leq \|\theta^{n-1}\|_H^2 + \int_{J_n} \|\theta\|_1^2 dt + \int_{J_n} \|\rho\|_1^2 dt. \end{aligned}$$

Insgesamt haben wir gezeigt:

$$\|\theta^n\|_H^2 + \int_{J_n} \|\theta\|_1^2 dt \leq \|\theta^{n-1}\|_H^2 + \int_{J_n} \|\rho\|_1^2 dt.$$

Summation liefert unter Verwendung  $\theta^0 = U^0 - \tilde{u}^0 = v - v = 0$  die Abschätzung

$$\|\theta^N\|_H^2 + \int_0^{t_N} \|\theta\|_1^2 dt \leq \int_0^{t_N} \|\rho\|_1^2 dt. \quad (12.12)$$

Abschätzung (12.10) kombiniert mit (12.12) ergibt schliesslich

$$\|\theta^N\|_H^2 \leq \sum_{n=1}^N \int_{J_n} \|\rho\|_1^2 dt \leq C \sum_{n=1}^N k^{2q+2} \int_{J_n} \|u^{(q+1)}\|_1^2 dt$$

---

<sup>12</sup>Diese Beziehung wird *Galerkin-Orthogonalität* genannt.

und damit ist der Beweis gegeben. ■

Für konstante Schrittweite  $k_n = k$  vereinfacht sich die Abschätzung aus Satz 12.2 zu

$$\|U^n - u(t_n)\|_H \leq Ck^{q+1} \left( \int_0^{t_N} \|u^{(q+1)}\|_1^2 dt \right)^{1/2}.$$

Wir betonen, dass im Falle des modifizierten, impliziten Euler-Verfahrens (12.7) die Fehlerschranke nur erste Ableitungen bezüglich der Zeit beinhaltet, im Gegensatz zum (standard) impliziten Euler-Verfahren, für das zweite Ableitungen auftreten. Das liegt daran, dass mehr Regularität verlangt werden muss, wenn das Integral noch durch eine Ein-Punkt-Gauss-Formel approximiert wird.

Wir wenden nun die unstetige Galerkin-Methode auf die Lösung der partiellen Differentialgleichung an

$$\begin{aligned} \dot{u} - \Delta u &= f && \text{in } \Omega \quad \forall t > 0, \\ u &= 0 && \text{auf } \partial\Omega, \\ u(\cdot, 0) &= v && \text{in } \Omega. \end{aligned} \tag{12.13}$$

Um technische Schwierigkeiten zu vermeiden, nehmen wir an, dass  $\Omega$  ein konvexes Polygongebiet ist.

Mit  $S$  bezeichnen wir den Finite-Elemente-Raum, der über einer Triangulierung von  $\Omega$  definiert ist und die Nullrandbedingung erfüllt.

Das halbdiskrete Problem (diskret im Ort und kontinuierlich bezüglich der Zeit) lautet: Finde  $u_S(t) \in S$  für  $t > 0$ , so dass

$$\begin{aligned} (\dot{u}_S, \chi)_{L^2(\Omega)} + (\nabla u_S, \nabla \chi)_{L^2(\Omega)} &= (f, \chi)_{L^2(\Omega)} \quad \forall \chi \in S, \quad t > 0 \\ u_S(0) &= v_S, \end{aligned}$$

wobei  $v_S$  eine Approximation von  $v$  ist.

Für dieses Problem wenden wir nun die unstetige Galerkin-Zeitdiskretisierung an. Der abstrakte Hilbert-Raum  $H$  ist der Finite-Elemente-Raum  $S$  und das Skalarprodukt auf  $S$  durch das  $L^2(\Omega)$ -Skalarprodukt gegeben. Der abstrakte Operator  $A$  ist in diesem Fall der diskrete Laplace-Operator  $-\Delta_S$ . Der endlichdimensionale Raum  $\mathbf{S}$  wird für die Ansätze in *Ort und Zeit* verwendet:

$$\mathbf{S} = \left\{ X : [0, \infty[ \rightarrow S \mid \forall n : X|_{J_n} = \sum_{j=0}^q X_j t^j, \quad X_j \in S \right\}.$$

Die Raum-Zeit-Diskretisierung lässt sich nun in der Form schreiben

$$B_N(U, X) = (v_S, X_+^0)_{L^2(\Omega)} + \int_0^{t_N} (f, X)_{L^2(\Omega)} dt \quad \forall X \in \mathbf{S}, \quad N \geq 0, \tag{12.14}$$

wobei diesmal gilt

$$\begin{aligned} B_N(V, W) &= \int_0^{t_N} (\partial_{\text{stw}} V, W)_{L^2(\Omega)} + (\nabla V, \nabla W)_{L^2(\Omega)} dt + \sum_{n=1}^{N-1} ([V]_n, W_+^n)_{L^2(\Omega)} + (V_+^0, W_+^0)_{L^2(\Omega)} \\ &= \int_0^{t_N} \left( - (V, \partial_{\text{stw}} W)_{L^2(\Omega)} + \langle \nabla V, \nabla W \rangle_{L^2(\Omega)} \right) dt - \sum_{n=1}^{N-1} (V^n, [W]_n)_{L^2(\Omega)} + (V^N, W^N)_{L^2(\Omega)}. \end{aligned} \tag{12.15}$$

Die Gleichung, welche der Fehler  $e = U - u$  erfüllt, lautet

$$B_N(e, X) = (v_S - v, X_+^0)_{L^2(\Omega)} \quad \forall X \in \mathbf{S}.$$

Wir bemerken, dass die rechte Seite für  $v_S = P_S v$  verschwindet. (Hier bezeichnet  $P_S$  die  $L^2$ -Orthogonalprojektion der Anfangsdaten auf  $S$ .)

Der folgende Satz gibt eine Fehlerabschätzung für das volldiskrete Problem an, das heisst, ist explizit bezüglich der Orts- und Zeitdiskretisierungsparameter.

Wir erinnern an ein Konvergenzresultat, welches wir im folgenden verwenden werden. Dazu definieren wir die Ritz-Projektion  $R_S : H^1 \rightarrow S$  durch

$$(\nabla R_S v, \nabla \chi)_{L^2(\Omega)} = (\nabla v, \nabla \chi)_{L^2(\Omega)} \quad \forall \chi \in S.$$

**Satz 12.3** *Die Ortsdiskretisierung basiere auf finite Elementen, die stückweise Polynome vom Grad  $r$  mit  $r \geq 1$  sind. Die Lösung  $u$  des Problems (12.13) erfülle  $u(t_i) \in H^{r+1}$ . Dann gelten die Fehlerabschätzungen*

$$\begin{aligned} \|\nabla(u - R_S u)(t_i)\|_{L^2(\Omega)} &\leq Ch^r \|u(t_i)\|_{r+1}, \\ \|(u - R_S u)(t_i)\|_{L^2(\Omega)} &\leq Ch^{r+1} \|u(t_i)\|_{r+1}. \end{aligned}$$

Im der folgenden Fehlerabschätzung tritt die Grösse

$$L_N := 1 + \sqrt{\log \frac{t_N}{k_N}}$$

auf. Die Abhängigkeit von  $L_N$  von der Schrittweite  $k_N$  ist für praktische Anwendungen vernachlässigbar klein. Um den Fehler  $e = U - u$  der Orts-Zeit-Diskretisierung abzuschätzen, werden wir die Aufspaltung

$$e = U - u = (U - R_S \tilde{u}) + (R_S \tilde{u} - v) =: \theta + \rho$$

verwenden, wobei  $\tilde{u}$  wieder die Interpolierende von  $u$  bezeichnet (vgl. (12.9)). Die Fehleranteile zum Zeitpunkt  $t_N$  werden mit  $\theta^N$  und  $\rho^N$  bezeichnet.

Wir benötigen ein Stabilitätsresultat, welches wir hier lediglich zitieren.

**Lemma 12.4** *Sei  $k_{n+1}/k_n \geq c > 0$  und  $\varphi \in L^2(\Omega)$ . Dann gilt für die Lösung  $Z_S \in \mathbf{S}$  von*

$$B_N(X, Z_S) = (X^N, \varphi)_{L^2(\Omega)} \quad \forall X \in \mathbf{S}$$

die Abschätzung

$$\int_0^{t_N} \|\Delta_S Z_S\|_{L^2(\Omega)} dt + \sum_{n=1}^{N-1} \|[Z_S]_n\|_{L^2(\Omega)} \leq CL_N \|\varphi\|_{L^2(\Omega)}. \quad (12.16)$$

**Satz 12.5** *Für aufeinanderfolgende Zeitschrittweiten gelte  $k_{n+1}/k_n \geq c > 0$ ,  $\forall n \geq 0$ . Sei  $q = 0$ . Dann gilt für die Lösungen von (12.14) und (12.13) mit  $v_S = P_S v$  die Fehlerabschätzung<sup>13</sup>*

$$\|U^N - u(t_N)\|_{L^2(\Omega)} \leq CL_N \max_{n \leq N} \left( h^{r+1} \|u\|_{r+1, J_n} + k_n \|\dot{u}\|_{0, J_n} \right).$$

<sup>13</sup>Hier und im folgenden verwenden wir die Bezeichnung  $\|u\|_{r+1, J_n} = \sup_{t \in J_n} \|u(t)\|_{r+1}$ .

**Beweis.** Mit  $\tilde{u}$  bezeichnen wir die (bzgl.  $t$ ) stückweise konstante Funktion, welche durch die Bedingung  $\tilde{u}(t) = u(t_n)$  für alle  $t \in J_n$  charakterisiert ist. Damit erhalten wir die Fehlerdarstellung

$$e = U - u = (U - R_S \tilde{u}) + (R_S \tilde{u} - u) = \theta + \rho.$$

Da  $\tilde{u}(t_N) = u(t_N)$  gilt, ergibt sich

$$\|\rho^N\|_{L^2(\Omega)} = \|(R_S u - u)(t_N)\|_{L^2(\Omega)} \leq Ch^{r+1} \|u(t_N)\|_{r+1}.$$

Im folgenden beschäftigen wir uns daher mit der Abschätzung von  $\theta$ . Sei  $\varphi \in L^2(\Omega)$  beliebig und  $Z_S$  die Lösung von

$$B_N(X, Z_S) = (X^N, P_S \varphi)_{L^2(\Omega)} = (X^N, \varphi)_{L^2(\Omega)} \quad \forall X \in \mathbf{S}.$$

Da  $\dot{Z}_S(t) \equiv 0$  auf  $J_n$  gilt, erhalten wir

$$\begin{aligned} (\theta^N, \varphi)_{L^2(\Omega)} &= B_N(\theta, Z_S) = -B_N(\rho, Z_S) \\ &= -\sum_{n=1}^N \int_{J_n} (\nabla \rho, \nabla Z_S)_{L^2(\Omega)} dt + \sum_{n=1}^{N-1} (\rho^n, [Z_S]_n)_{L^2(\Omega)} - (\rho^N, P_S \varphi)_{L^2(\Omega)}. \end{aligned}$$

Wir erinnern an die Definition  $(-\Delta_S V, W)_{L^2(\Omega)} := (\nabla V, \nabla W)_{L^2(\Omega)}$ ,  $\forall V, W \in S$ , woraus

$$-(R_S \rho, \Delta_S Z_S)_{L^2(\Omega)} = (\nabla R_S \rho, \nabla Z_S)_{L^2(\Omega)} = (\nabla \rho, \nabla Z_S)_{L^2(\Omega)}$$

folgt. Damit ergibt sich weiter

$$\begin{aligned} \left| (\theta^N, \varphi)_{L^2(\Omega)} \right| &\leq \max_{n \leq N} \left( \|\rho\|_{0, J_n} + \|R_S \rho\|_{0, J_n} \right) \\ &\quad \times \left( \int_0^{t_N} \|\Delta_S Z_S\|_{L^2(\Omega)} dt + \sum_{n=1}^{N-1} \|[Z_S]_n\|_{L^2(\Omega)} + \|\varphi\|_{L^2(\Omega)} \right). \end{aligned}$$

Das Stabilitätsresultat (12.16) lässt sich anwenden und wir erhalten

$$\|\theta^N\|_{L^2(\Omega)} \leq CL_N \max_{n \leq N} \left( \|\rho\|_{0, J_n} + \|R_S \rho\|_{0, J_n} \right).$$

Des weiteren erhalten wir die folgende Abschätzung

$$\begin{aligned} \|\rho\|_{0, J_n} &= \|R_S \tilde{u} - u\|_{0, J_n} \leq \|(R_S - I) \tilde{u}\|_{0, J_n} + \|\tilde{u} - u\|_{0, J_n} \\ &\leq Ch^{r+1} \|u\|_{r+1, J_n} + Ck_n \|\dot{u}\|_{0, J_n}. \end{aligned}$$

Da  $R_S \rho = R_S \tilde{u} - R_S u = \rho - (R_S u - u)$  gilt, erfüllt diese Funktion die gleiche Fehlerabschätzung. ■

Diese Fehlerabschätzungen enthalten Grössen, die von der exakten Lösung abhängen und die durch die gegebenen Daten  $(f, \Omega, u(0))$  abschätzbar sind, *vorausgesetzt* die Lösung ist hinreichend glatt. Derartige Fehlerabschätzungen werden *a priori*-Fehlerabschätzungen genannt. Da jedoch die exakte Lösung unbekannt ist, liefern diese Fehlerabschätzungen keine scharfen oberen Schranken für den Fehler.



Im folgenden werden wir daher eine *a posteriori*-Fehlerabschätzung herleiten, bei der der Fehler durch die Daten des Problems und durch die bereits *berechnete* diskrete Lösung abgeschätzt wird. Derartige Fehlerabschätzungen können verwendet werden, um problemangepasste (adaptive) Diskretisierungen zu erzeugen. Genauer soll für eine vorgegebene Fehlertoleranz der jeweilig neue Zeitschritt so bestimmt werden, dass die neue Lösung dieser vorgegebenen Fehlertoleranz genügt.

Wir werden uns hier auf den Fall des unstetigen Galerkin-Zeitschrittverfahrens beschränken mit stückweise konstanter Approximation der Zeitabhängigkeit, d.h.,  $q = 0$ .

Wir betrachten daher zunächst das Anfangswertproblem (12.2) und eine Näherungslösung in

$$S := \{X : [0, \infty[ \rightarrow H \mid \forall n : X|_{J_n} = \psi \in H\},$$

die durch

$$B_N(U, X) = (v, X_+^0)_{L^2(\Omega)} + \int_0^{t_N} (f, X)_{L^2(\Omega)} dt \quad \forall X \in S \quad (12.17)$$

gegeben ist, wobei wir an die Definition von  $B_N(\cdot, \cdot)$  erinnern:

$$B_N(V, W) = \int_0^{t_N} (\partial_{\text{stw}} V, W)_H + (AV, W)_H dt + \sum_{n=1}^{N-1} ([V]_n, W_+^n)_H + (V_+^0, W_+^0)_H. \quad (12.18)$$

Wie wir bereits in (12.7) gesehen haben, kann diese Gleichung in der folgenden Form geschrieben werden:

$$U^n + k_n A U^n = U^{n-1} + \int_{J_n} f dt \quad \text{für } n \geq 1, U^0 = v. \quad (12.19)$$

**Satz 12.6** *Für die Lösungen von (12.2) und (12.19) gilt die Fehlerabschätzung*

$$\|U^N - u(t_N)\|_H \leq CL_N \max_{n \leq N} \left( k_n \|f\|_{0, J_n} + k_n \|\bar{\partial}_n U^n\|_H \right).$$

Der Beweis erfordert etwas Vorbereitung. Er verwendet die Lösung des Rückwärtsproblems

$$\begin{aligned} -\dot{z} + Az &= 0 & \text{für } t < t_N \\ z(t_N) &= \varphi. \end{aligned} \quad (12.20)$$

Mit Hilfe des Rückwärtsproblems erhalten wir die folgende Fehlerdarstellung.

**Lemma 12.7** *Seien  $U$  und  $u$  die Lösungen von (12.19) und (12.2). Sei  $z$  die Lösung des Rückwärtsproblems (12.20). Dann gilt für den Fehler  $e = U - u$  die Darstellung*

$$(e^N, \varphi)_H = \int_0^{t_N} (AU, z - X)_H dt + \sum_{n=0}^{N-1} ([U]_n, z^n - X_+^n)_H - \int_0^{t_N} (f, z - X) dt \quad \forall X \in S.$$

**Beweis.** Wir erinnern, dass der Fehler wegen der Galerkin-Orthogonalität die Gleichung

$$B_N(e, X) = 0 \quad \forall X \in S$$

erfüllt. Die Definition der Rückwärtslösung  $z$  und die Darstellung der Bilinearform  $B_N(\cdot, \cdot)$  liefern:

$$\begin{aligned} (e^N, \varphi)_H &= B_N(e, z) = B_N(e, z - X) = B_N(U, z - X) - B_N(u, z - X) \\ &= \int_0^{t_N} (AU, z - X)_H dt + \sum_{n=1}^{N-1} ([U]_n, (z - X)_+^n)_H + (U_+^0, (z - X)_+^0)_H \\ &\quad - (v, (z - X)_+^0)_H - \int_0^{t_N} (f, z - X)_H dt. \end{aligned}$$

Dies ist die gewünschte Fehlerdarstellung. ■

Zur weiteren Vorbereitung benötigen wir ein Stabilitätsresultat für die exakte Lösung.

**Lemma 12.8** *Die Lösung des Rückwärtsproblems (12.20) erfüllt*

$$\int_0^{t_{N-1}} \|\dot{z}\| dt + \|z\|_{0, J_N} \leq CL_N \|\varphi\|_H.$$

**Beweis.** Zum Beweis dieser Aussage verwenden wir die Stabilität des Vorwärtsproblems (12.2) mit  $f \equiv 0$ :

$$\int_{k_1}^{t_N} \|\dot{u}\|_H dt + \|u\|_{0, J_1} \leq C\tilde{L}_N \|v\|_H, \quad \text{mit} \quad \tilde{L}_N = 1 + \sqrt{\log \frac{t_N}{k_1}}.$$

Um diese Ungleichung einzusehen, beachten wir zunächst  $\|u(t)\|_H \leq \|v\|_H$  für  $t \geq 0$  (Dies folgt aus der Spektraldarstellung der Lösung des homogenen Problems.) Speziell gilt

$$\|u\|_{0, J_1} \leq \|v\|_H.$$

Die Spektraldarstellung der Lösung von (12.2) liefert mit dem orthonormalen Eigensystem  $(\psi_n)_n$  von  $A$ :

$$\dot{u}(t) = \sum_{n \in N} \lambda_n e^{-\lambda_n t} (v, \psi_n)_H \psi_n.$$

Daraus folgt

$$\|\dot{u}\|_H^2 = \sum_{n \in N} \lambda_n^2 e^{-2\lambda_n t} (v, \psi_n)_H^2$$

und weiter

$$\int_0^\infty t \|\dot{u}\|_H^2 dt = \sum_{n \in N} \lambda_n^2 \left( \int_0^\infty t e^{-2\lambda_n t} dt \right) (v, \psi_n)_H^2 = \frac{1}{4} \sum_{n \in N} (v, \psi_n)_H^2 = \frac{1}{4} \|v\|_H^2.$$

Daraus schliessen wir

$$\left( \int_{k_1}^{t_N} \|\dot{u}\|_H^2 dt \right)^2 \leq \int_{k_1}^{t_N} \frac{dt}{t} \int_{k_1}^{t_N} t \|\dot{u}\|_H^2 dt \leq \frac{1}{4} \left( \log \frac{t_N}{k_1} \right) \|v\|_H^2,$$

und daraus folgt die Behauptung. ■

Damit haben wir alle Hilfsmittel zusammengestellt um Satz 12.6 zu beweisen.

**Beweis von Satz 12.6:**

Wir bezeichnen die drei Terme aus der Fehlerdarstellung von Lemma 12.7 mit  $I$ ,  $II$ ,  $III$ . Wir wählen

$$X(t) := \bar{z}^n := k_n^{-1} \int_{J_n} z(t) dt \quad \forall t \in J_n, \quad n \geq 1.$$

Da  $U(t)$  konstant auf  $J_n$  ist, gilt

$$\int_{J_n} (AU, z - \bar{z}^n) dt = 0 \quad \forall n \geq 1,$$

und daher ist der Term  $I = 0$ .

Für  $II$  gilt wegen  $X_+^n = \bar{z}^{n+1}$

$$|II| \leq \max_{n \leq N-1} \|[U]_n\|_H \sum_{n=1}^N \|\bar{z}^n - z^{n-1}\|_H.$$

Die rechte Seite lässt sich abschätzen gemäss

$$\begin{aligned} \|\bar{z}^n - z^{n-1}\|_H &= \left\| k_n^{-1} \int_{J_n} (z - z^{n-1}) dt \right\|_H \leq \int_{J_n} \|\dot{z}\|_H dt \quad \forall n < N \\ \|\bar{z}^n - z^{n-1}\|_H &\leq 2 \|z\|_{0, J_n}. \end{aligned}$$

Mit Lemma 12.8 ergibt sich

$$\sum_{n=1}^N \|\bar{z}^n - z^{n-1}\|_H \leq \int_0^{t_{N-1}} \|\dot{z}\|_H dt + 2 \|z\|_{0, J_N} \leq CL_N \|\varphi\|_H.$$

Da  $[U]_n = k_n \bar{\partial}_n U^n$  gilt, folgt die gewünschte Abschätzung für den Term  $II$ .

In ähnlicher Weise erhalten wir für den Term  $III$  die Abschätzung

$$\begin{aligned} |III| &\leq \max_{n \leq N} \left( k_n \|f\|_{0, J_n} \right) \sum_{n=1}^N \left( k_n^{-1} \int_{J_n} \|\bar{z}^n - z\|_H dt \right) \\ &\leq \max_{n \leq N} \left( k_n \|f\|_{0, J_n} \right) \left( \int_0^{t_{N-1}} \|\dot{z}\|_H dt + 2 \|z\|_{0, J_n} \right) \\ &\leq CL_N \max_{n \leq N} \left( k_n \|f\|_{0, J_n} \right) \|\varphi\|_H. \end{aligned}$$

Damit ist der Beweis gegeben. ■

Zum Schluss dieses Abschnitts werden wir eine *a posteriori*-Fehlerabschätzung für das un-stetige Galerkin-Verfahren angeben im Fall der Wärmeleitungsgleichung in einem beschränkten, konvexen, polygonalen Gebiet  $\Omega \subset \mathbb{R}^2$  und Dirichlet-Randbedingungen.

Das kontinuierliche Problem ist durch (12.13) gegeben und die volldiskrete Diskretisierung durch: Finde  $U \in \mathbf{S}$  mit

$$B_N(U, X) = (v, X_+^0)_{L^2(\Omega)} + \int_0^{t_N} (f, X) dt \quad \forall X \in \mathbf{S}, \quad N \geq 1. \quad (12.21)$$

Die Bilinearform  $B_N(\cdot, \cdot)$  ist wiederum durch (12.15) gegeben.

Das Gebiet  $\Omega$  sei zerlegt in eine Finite-Elemente-Triangulierung  $\mathcal{G}$  mit maximalem Dreiecksdurchmesser  $h$ . Der Finite-Elemente-Raum für die Ortsdiskretisierung ist durch

$$S := \{u \in C^0(\Omega) : u|_{\partial\Omega} \equiv 0 \wedge u|_{\tau} \in \mathbb{P}_1, \quad \forall \tau \in \mathcal{G}\}$$

gegeben. Die Formulierung (12.21) impliziert bereits, dass wir annehmen, die Anfangsdaten erfüllen  $v \in S$  oder sind durch  $v_S := P_S v$  mit der  $L^2$ -Orthogonalprojektion  $P_S$  definiert.

Wir beschränken uns hier auf den Fall, dass das Ortsgitter **nicht** mit der Zeit problemangepasst (adaptiv) verändert wird. Für die Zeitdiskretisierung verwenden wir einen stückweise konstanten Ansatz auf den Zeitintervallen  $J_n$ .

Die volldiskrete Formulierung lautet damit: Finde  $U \in \mathbf{S}$ , so dass gilt

$$\int_{J_n} (\nabla U, \nabla X)_{L^2(\Omega)} + ([U]_{n-1}, X_+^{n-1})_{L^2(\Omega)} = \int_{J_n} (f, X)_{L^2(\Omega)} dt \quad \forall X \in \mathbf{S}, n \geq 1$$

bzw.

$$(U^n, \chi_+^{n-1})_{L^2(\Omega)} + k_n (\nabla U, \nabla \chi)_{L^2(\Omega)} = (U^{n-1}, \chi)_{L^2(\Omega)} + \left( \int_{J_n} f(t) dt, \chi \right)_{L^2(\Omega)} \quad \forall \chi \in S, n \geq 1$$

mit der Anfangsbedingung  $U^0 = P_S v$ .

Wir erinnern an Satz 12.5: Unter der Voraussetzung, dass aufeinanderfolgende Schrittweiten  $k_{n+1}/k_n \geq c > 0$  erfüllen, gilt die Fehlerabschätzung

$$\|U - u\|_{0, J_n} \leq CL_N \max_{n \leq N} \left( k_n \|\dot{u}\|_{0, J_n} + h^2 \|u\|_{2, J_n} \right).$$

Für eine *a posteriori* Fehlerabschätzung sollte die rechte Seite der Fehlerabschätzung durch berechenbare Grössen ersetzt werden. Zu diesem Zweck erscheint es natürlich, zu versuchen  $\dot{u}$  auf den Zeitintervallen  $J_n$  durch die Differenzenapproximation  $\bar{\partial}_n U^n$  zu ersetzen. Darüber hinaus benötigen wir einen Ersatz für die zweiten Ortsableitungen, die in der Norm  $\|u\|_{2, J_n}$  enthalten sind. Aus diesem Grund führen wir die Menge der inneren Kanten  $\mathcal{E}$  von  $\mathcal{G}$  ein. Für jede Kante  $\gamma \in \mathcal{E}$  wählen wir einen Vektor  $n_\gamma$ , der senkrecht zu  $\gamma$  ist und fixieren eine der beiden möglichen Richtungen.

Das erlaubt uns, den Sprung des Gradienten  $\nabla \chi$  in Richtung der Kantennormalen  $n_\gamma$  zu definieren:

$$[\partial \chi / \partial n_\gamma]_\gamma := \lim_{\varepsilon \rightarrow +0} \langle n_\gamma, \nabla \chi(x + \varepsilon n_\gamma) - \nabla \chi(x - \varepsilon n_\gamma) \rangle.$$

Man beachte, dass der Normalenvektor  $n_\gamma$  auf der Kante  $\gamma$  konstant ist. Da  $\nabla \chi$  auf jedem Dreieck konstant ist, ist der Normalensprung eine konstante Funktion auf jeder Kante. Man überlegt sich leicht, dass die Grösse  $[\partial \chi / \partial n_\gamma]_\gamma$  unabhängig ist von der gewählten Richtung der Normalen  $n_\gamma$ .

Das diskrete Analogon zur  $H^2$ -Norm definieren wir gemäss

$$\|\chi\|_{2, S} := \left( \sum_{\gamma \in \mathcal{E}} \left| \left[ \frac{\partial \chi}{\partial n} \right]_\gamma \right|^2 \right)^{1/2}.$$

Damit haben wir alle Vorbereitungen zur Formulierung des *a posteriori*-Fehlerschätzers getroffen.

**Satz 12.9** Die Lösungen von (12.21) und (12.13) erfüllen die a posteriori-Fehlerabschätzung

$$\|U^N - u(t_N)\|_{L^2(\Omega)} \leq CL_N \left\{ \max_{n \leq N} (h^2 + k_n) \|f\|_{0, J_n} + h^2 \|U^n\|_{2, S} + k_n \|\bar{\partial}_n U^n\|_{L^2(\Omega)} \right\}.$$

Der Beweis benötigt einen vorbereitenden Hilfssatz.

**Hilfssatz 12.10** Für alle  $W \in S$  und  $v \in H^2$  gilt

$$\left| (\nabla W, \nabla (P_S - I)v)_{L^2(\Omega)} \right| \leq Ch^2 \|W\|_{2, S} \|v\|_2.$$

**Beweis.** (Skizze) Ohne Beweis verwenden wir die folgenden beiden Aussagen:

1. **Spurabschätzung:** Sei  $\tau \in \mathcal{G}$  ein beliebiges Dreieck mit Durchmesser  $h_\tau$ . Dann gilt für alle  $w \in H^1$

$$\|w\|_{L^2(\gamma)}^2 \leq C \left( h_\tau \|\nabla w\|_{L^2(\tau)}^2 + h_\tau^{-1} \|w\|_{L^2(\tau)}^2 \right). \quad (12.22)$$

2. **Fehlerabschätzung für  $L^2$ -Projektion:** Für alle  $v \in H^2$  gilt

$$\|P_S v - v\|_{L^2(\Omega)} + h \|\nabla (P_S v - v)\|_{L^2(\Omega)} \leq Ch^2 \|v\|_2. \quad (12.23)$$

Zunächst zeigen wir, dass für  $W \in S$  und  $v \in H^1$  gilt

$$\left| (\nabla W, \nabla v)_{L^2(\Omega)} \right| \leq C \|W\|_{2, S} \left( \|v\|_{L^2(\Omega)} + h \|\nabla v\|_{L^2(\Omega)} \right). \quad (12.24)$$

Wir betrachten ein beliebiges Dreieck  $\tau \in \mathcal{G}$  mit Kanten  $\gamma_{\tau, j}$ ,  $j = 1, 2, 3$ , und wenden die Greenschen Formel an:

$$\int_\tau \langle \nabla W, \nabla v \rangle dx = \sum_{j=1}^3 \frac{\partial W}{\partial n_{\tau, j}} \Big|_{\gamma_{\tau, j}} \int_{\gamma_{\tau, j}} v ds.$$

Beim Summieren über alle Dreiecke  $\tau$  treten alle Kanten zweimal auf, wobei die Kantennormalenrichtungen entgegengesetzt orientiert sind. Fixiert man eine dieser Normalenrichtungen  $n_\gamma$ , erhält man

$$(\nabla W, \nabla v)_{L^2(\Omega)} \leq C \|W\|_{2, S} \sqrt{\sum_\gamma \left( \int_\gamma v ds \right)^2}.$$

Die Höldersche Ungleichung und die Spurabschätzung (12.22) ergeben

$$\sum_\gamma \left( \int_\gamma v ds \right)^2 \leq C \left( \|v\|_{L^2(\tau)}^2 + h^2 \|\nabla v\|_{L^2(\tau)}^2 \right).$$

Damit haben wir (12.24) gezeigt. Zusammen mit der Approximationseigenschaft (12.23) der  $L^2$ -Projektion ergibt sich die Behauptung. ■

**Beweis von Satz 12.9:**

Die Fehlerdarstellung  $e = U - u$  aus Lemma 12.7 gilt auch in diesem Fall und daraus folgt

$$(e^N, \varphi)_{L^2(\Omega)} = - \int_0^{t_N} (\nabla U, \nabla (X - z))_H dt - \sum_{n=0}^{N-1} ([U]_n, X_+^n - z^n)_{L^2(\Omega)} + \int_0^{t_N} (f, X - z) dt \quad (12.25)$$

$$=: I + II + III$$

für alle  $X \in \mathbf{S}$ . (Hier haben wir  $U^0 = P_S v$  ausgenützt. Wir wählen  $X \in \mathbf{S}$  als Orthogonalprojektion von  $z$  auf  $L^2(\Omega \times J_n)$  für alle  $n \geq 1$ , i.e.,  $X = P_S \bar{z}$  mit der  $L^2$ -Orthogonalprojektion  $P_S$  im Ort und  $\bar{z}|_{J_n} = k_n^{-1} \int_{J_n} z dt$ . Dies führt auf die Aufspaltung

$$X - z = (P_S \bar{z} - P_S z) + (P_S z - z).$$

Nun gilt für den ersten Term

$$\int_{J_n} (\nabla U, \nabla (P_S \bar{z} - P_S z))_{L^2(\Omega)} dt = - \int_{J_n} (\Delta_S U, \bar{z} - z) dt = 0$$

und für den zweiten (wegen Hilfssatz 12.10)

$$\begin{aligned} \left| \int_{J_n} (\nabla U, \nabla (P_S \bar{z} - P_S z))_{L^2(\Omega)} dt \right| &= \left| \left( \nabla U^n, \nabla (P_S - I) \int_{J_n} z dt \right) \right| \\ &\leq Ch^2 \|U^n\|_{2,S} \left\| \int_{J_n} z dt \right\|_2 \leq Ch^2 \|U^n\|_{2,S} \left\| \Delta \int_{J_n} z dt \right\|_{L^2(\Omega)}. \end{aligned}$$

Da  $\int_{J_N} \Delta z = \int_{J_n} \dot{z} = z(t_N) - z(t_{N-1})$  zeigt Lemma 12.8

$$\begin{aligned} |I| &\leq Ch^2 \left( \max_{n \leq N} \|U^n\|_{2,S} \right) \sum_{n=1}^N \left\| \int_{J_n} \Delta z dt \right\|_{L^2(\Omega)} \\ &\leq Ch^2 \left( \max_{n \leq N} \|U^n\|_{2,S} \right) \left( \int_0^{t_{N-1}} \|\dot{z}\| dt + \|z\|_{0,J_N} \right) \\ &\leq CL_N h^2 \max_{n \leq N} \|U^n\|_{2,S} \|\varphi\|_{L^2(\Omega)}. \end{aligned}$$

Wir betrachten nun den zweiten Term in (12.25) und verwenden  $[U]_{n-1} = U^n - U^{n-1} = k_n \bar{\partial}_n U^n$ . Daraus folgt

$$\begin{aligned} II &= - \sum_{n=1}^N ([U]_{n-1}, X^n - z^{n-1})_{L^2(\Omega)} = \sum_{n=1}^N ([U]_{n-1}, X^n - P_S z^{n-1})_{L^2(\Omega)} \\ &\leq \sum_{n=1}^N k_n \|\bar{\partial}_n U\|_{L^2(\Omega)} \|\bar{z}^n - z^{n-1}\|_{L^2(\Omega)}. \end{aligned}$$

Verwenden wir nochmals Lemma 12.8, folgt die gewünschte Abschätzung für  $II$  :

$$|II| \leq C \max_{n \leq N} \left( k_n \|\bar{\partial}_n U^m\|_{L^2(\Omega)} \right) \left( \int_0^{t_{N-1}} \|\dot{z}\| dt + \|z\|_{0,J_n} \right) \leq CL_N \max_{n \leq N} (k_n \|\bar{\partial}_n U^n\|) \|\varphi\|_{L^2(\Omega)}.$$

Es bleibt, den dritten Term in (12.25) zu analysieren. Wir beginnen mit

$$\left| \int_{J_n} (f, X - z)_{L^2(\Omega)} dt \right| \leq \|f\|_{0, J_n} \int_{J_n} \|P_S \bar{z} - z\| dt.$$

Indem wir den Term  $P_S z$  einschieben, ergibt sich auf  $J_n$

$$\|P_S \bar{z} - z\|_{L^2(\Omega)} \leq \|P_S z - z\|_{L^2(\Omega)} + \|z - \bar{z}\|_{L^2(\Omega)} \leq Ch^2 \|z\|_2 + \int_{J_n} \|\dot{z}\|_{L^2(\Omega)} dt.$$

Daraus folgt

$$\int_{J_n} \|z - P_S \bar{z}\|_{L^2(\Omega)} dt \leq C (h^2 + k_n) \int_{J_n} \|\dot{z}\| dt \quad \forall n < N.$$

Des weiteren folgt unter Verwendung von  $\|z - P_S \bar{z}\|_{L^2(\Omega)} \leq 2 \|z\|_{0, J_n}$  die Abschätzung

$$\int_{J_n} \|z - P_S \bar{z}\|_{L^2(\Omega)} dt \leq 2k_N \|z\|_{0, J_N}.$$

Zusammen ergibt sich die Abschätzung von Term *III* unter Verwendung von Lemma 12.8:

$$\begin{aligned} |III| &\leq C \max_{n \leq N} (h^2 + k_n) \|f\|_{0, J_n} \left( \int_0^{t_{N-1}} \|\dot{z}\|_{L^2(\Omega)} dt + \|z\|_{0, J_N} \right) \\ &\leq CL_N \left( \max_{n \leq N} (h^2 + k_n) \|f\|_{0, J_n} \right) \|\varphi\|_{L^2(\Omega)}. \end{aligned}$$

Diese drei Abschätzungen zusammen mit (12.25) ergeben die Behauptung. ■