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

Numerische Methoden für elliptische und parabolische Differentialgleichungen

Stefan Sauter

Herbstsemester 2023

Version: 19. September 2023

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.

[4] K. Yosida. *Functional Analysis*. Springer-Verlag, 1964.

Inhaltsverzeichnis

1 Partielle Differentialgleichungen

Die allgemeine Differentialgleichung zweiter Ordnung lautet

$$-\operatorname{div}(\mathbf{A}\operatorname{grad} u) + \langle \mathbf{b}, \operatorname{grad} u\rangle + cu = g. \tag{1.1}$$

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

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

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

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

gegeben. Die Koeffizienten in (1.1) sind Matrix- (bzw. Vektor-) wertige (bzw. skalare) Funktionen $\mathbf{A} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ (bzw. $\mathbf{b} : \mathbb{R}^d \to \mathbb{R}^d$ bzw. $c : \mathbb{R}^d \to \mathbb{R}$). Die Matrix $\mathbf{A}$ ist hierbei differenzierbar und symmetrisch.

Eine Differentialgleichung heisst linear, falls die Gleichung linear von der unbekannten Funktion abhängt. Sei Lu eine Abkürzung für die linke Seite der Differentialgleichung (1.1).

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 $Lu = \alpha f_1 + \beta f_2$.

Übungsaufgabe 1.1 *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 (1.1) transformiert werden kann.

Bemerkung 1.2 *Aus der Symmetrie der Matrix $\mathbf{A}(\mathbf{x})$ folgt, dass alle Eigenwerte reell sind.*

Definition 1.3 *(a) Die Gleichung (1.1) heisst **elliptisch** in $\mathbf{x} \in \mathbb{R}^d$, falls $\mathbf{A}(\mathbf{x})$ positiv definit oder negativ definit ist.*

*(b) Die Gleichung (1.1) 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 (1.1) 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 $\operatorname{Rang}(\mathbf{A}(\mathbf{x}), \mathbf{b}(\mathbf{x})) = d$ gilt.*

(d) Die Gleichung (1.1) 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 (1.1) auftreten.

Beispiel 1.4 *Die einfachste elliptische Differentialgleichung ist die* Poisson-Gleichung. *Hier ist $\mathbf{A}(\mathbf{x}) = \mathbf{I}$ die Einheitsmatrix und die Koeffizienten $\mathbf{b}$ und c sind Null:*

$$-\Delta u = g \quad \text{in } \Omega. \tag{1.2}$$

In einer Dimension $d = 1$ reduziert sich die Poisson-Gleichung zu einer gewöhnlichen Differentialgleichung

$$-u'' = g \quad \text{in einem Intervall } \Omega = (a, b). \tag{1.3}$$

Durch zweimalige Integration lässt sich die allgemeine Lösung von (1.3) angeben

$$u(x) = -\int_a^x \left(\int_a^s g(t)\, dt \right) ds + C_1 x + C_0$$

mit beliebigen Konstanten $C_0, C_1 \in \mathbb{R}$.

Übungsaufgabe 1.5 *Sei $\Omega \subset \mathbb{R}^2$ ein zweidimensionales Gebiet. Bestimmen (erraten) Sie drei linear unabhängige Lösungen von*

$$-\Delta u = 4 \quad \text{in } \Omega.$$

Beispiel 1.6 *Die einfachste parabolische Differentialgleichung ist die* Wärmeleitungsgleichung. *Hier ist* $\mathbf{A}(\mathbf{x}) = \mathbf{I}$ *wiederum die Einheitsmatrix,* $\mathbf{b} = (1, 0, \ldots, 0)^T$ *und* $c = 0$*, so dass wir*

$$\partial_t u(t, \mathbf{x}) - \Delta_{\mathbf{x}} u(t, \mathbf{x}) = g(t, \mathbf{x}) \quad \text{in } [0, T] \times \Omega \tag{1.4}$$

erhalten. Üblicherweise wird hier die Dimension in Definition 1.3 durch $d + 1$ *ersetzt und die erste Variable mit* t *(für die Zeit) bezeichnet. Der Index* $\mathbf{x}$ *am Laplace-Operator* $\Delta_{\mathbf{x}}$ *bedeutet, dass nur bezüglich der* d*-dimensionalen* $\mathbf{x}$*-Variablen differenziert wird. Für* $d = 1$ *ergibt sich dann*

$$\partial_t u(t, x) - \partial_x^2 u(t, x) = g(t, x) \quad \text{in } [0, T] \times (a, b).$$

Für die *Entwicklung* numerischer Verfahren zur Diskretisierung von Differentialgleichungen genügt es häufig, die einfachsten Modellprobleme zu betrachten: (1.2) für elliptische Probleme und (1.4) für parabolische Probleme. 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 $x_1, \ldots, x_d$ 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* weitesgehend 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 $p' \in [1, \infty]$, $\frac{1}{p} + \frac{1}{p'} = 1$ mit der Konvention $1/\infty = 0$. Mehrdimensionale Ableitungen lassen sich mittels Multiindizes kompakt schreiben. Für eine Multiindex $\boldsymbol{\alpha} = (\alpha_j)_{j=1}^d \in \mathbb{N}_0^d$ setzen wir $|\boldsymbol{\alpha}| = \alpha_1 + \ldots + \alpha_d$ und

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

Da Ω beschränkt ist, gilt $L^p(\Omega) \subset L^1(\Omega)$ und die triviale Injektion $\iota : L^p(\Omega) \to L^1(\Omega)$ ist stetig. Wir nehmen darüber hinaus immer an, dass Ω ein *Normalgebiet* ist, für das der

Gausssche Integralsatz angewendet werden kann. Daraus folgt für alle $\varphi, \psi \in C_0^\infty(\Omega)$ mit[1]

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

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

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

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

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

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

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

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

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

Dann gilt

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

Aus der Theorie der Lebesgue-Integrale folgt, dass $\psi_1 = \psi_2$ im L^1-Sinne gilt.

zu (2): Folgt aus dem Gausssschen Integralsatz. ■

Beispiel 2.3 *Sei $\Omega =]-1, 1[$ und $\varphi(x) = |x|$. Dann ist φ 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 ρ über 0 impliziert $\rho(0+) = \rho(0-)$, und wir erhalten

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

[1] Der Träger einer Funktion v ist durch

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

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

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

Übungsaufgabe 2.4 *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:

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

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

Definition 2.5 *(1) Für $k \in \mathbb{N}$ und $p \in [1,\infty[$ ist der Sobolev-Raum $W^{k,p}(\Omega)$ und seine Norm durch*

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

gegeben.

(2) Durch

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

ist eine Seminnorm 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 2.6 *(1) Seien φ und Ω wie in Bsp. 2.3. Dann gilt $\varphi \in W^{1,p}(]-1,1[)$ für alle $p \in [1,\infty[$.*

(2) Seien $\Omega = B\left(0,\frac{1}{2}\right)$ die offene $d-$dimensionale Kugel 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

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

und

$$\left\|D^{\boldsymbol{\alpha}}\varphi\right\|^p_{L^p(\Omega)} \sim \omega_{d-1}\int_0^{1/2} r^{(s-|\boldsymbol{\alpha}|)p} r^{d-1} dr < \infty,$$

wobei ω_{d-1} das Volumen der Einheitssphäre in $\mathbb{R}^d$ ist. Der Integrand im letzten Integral ist integrabel, genau dann wenn

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

gilt.

Übungsaufgabe 2.7 *Sei $d \geq 2$ und $\Omega = B\left(0, \frac{1}{2}\right)$ die offene $d-$dimensionale Kugel um den Ursprung mit Radius $\frac{1}{2}$ und $u : \Omega \to \mathbb{R}$ durch*

$$u(\mathbf{x}) = \begin{cases} \log\left(\|\log\|\mathbf{x}\|\|\right) & \mathbf{x} \neq \mathbf{0}, \\ 0 & \mathbf{x} = \mathbf{0} \end{cases}$$

definiert. Zeigen, Sie, das u in $H^1(\Omega)$ gilt.

Satz 2.8 *(1) $W^{k,p}(\Omega)$ versehen mit der Norm $\|\cdot\|_{k,p}$ ist ein Banach-Raum.*
(2) Für $1 \leq p < \infty$ ist $C^\infty(\Omega) \cap W^{k,p}(\Omega)$ dicht in $W^{k,p}(\Omega)$.
(3) $H^k(\Omega)$ ist eine Hilbert-Raum mit Skalarprodukt

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

Der Beweis beruht darauf, dass sich die entsprechenden Eigenschaften der Lebesgue-Räume auf die Sobolev-Räume (als Teilräume) vererbt. Die Details werden hier weggelassen.

Die Betragsfunktion war ein Beispiel einer Funktion, die stückweise glatt ist und die schwache Ableitung mit der stückweisen Ableitung übereinstimmt. Der folgende Satz überträgt dieses Differenzierbarkeitsresultat auf allgemeinere Funktionen.

Satz 2.9 *Sei Ω ein Gebiet und Ω_1, Ω_2 zwei nichtleere, offene, beschränkte und disjunkte Teilmengen von Ω 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 γ bezeichnet: $\Omega = \Omega_1 \cup \Omega_2 \cup \gamma$. Sei n die äussere Normale an Ω_1 und für $x \in \gamma$ wird der Sprung von φ über γ mit

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

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

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

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

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

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

Ist umgekehrt $\varphi \in C(\Omega)$, so verschwindet $[\varphi]$ auf γ, 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 2.10 *$W_0^{k,p}(\Omega)$ 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 2.11 *Das Gebiet Ω besitzt einen **Lipschitz-Rand** bzw. Ω ist ein **Lipschitz-Gebiet**, wenn es ein $N \in \mathbb{N}$ und offene Mengen $U_1, \ldots, U_N \subset \mathbb{R}^d$ 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 2.12 *(1) Ω sei ein Lipschitz-Gebiet. Dann existiert fast überall auf $\partial\Omega$ das äussere Einheitsnormalenfeld n zu Ω.*
(2) $\Omega \subset \mathbb{R}^2$ habe einen stückweise glatten Rand. Zudem gelte für jedes $x_0 \in \partial\Omega$:

i. es existiert ein $\varepsilon > 0$, so dass $B(x_0, \varepsilon) \setminus \partial\Omega$ aus genau zwei Zusammenhangskomponenten besteht, (anschaulich: es existieren keine Punkte, in denen der Rand sich kreuzt oder berührt),

*ii. es existieren zwei nichttriviale Kegel K_0^{in}, K_0^{out} mit Basis x_0, so dass $\Omega \subset \mathbb{R}^d \setminus K_0^{\text{out}}$ und $\mathbb{R}^d \setminus \overline{\Omega} \subset \Omega \setminus K_0^{\text{in}}$ (**Kegelbedingung**). Dann ist Ω 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 $L \subset \Omega$ lässt sich sinnvoll definieren. Das ist auf den ersten Blick überraschend, da niederdimensionale Teilmengen $L \subset \Omega$ Nullmengen in Ω sind und Funktionen, beispielsweise in L^p, beliebig auf Nullmengen abgeändert werden können, ohne die Restklasse zu ändern.

Wie wir jedoch in Satz 2.8 gesehen haben, können wir Sobolev-Funktionen als Grenzwert von Cauchy-Folgen *glatter Funktionen* betrachten. Für glatte Funktionen ist die Einschränkung auf L wohldefiniert und die Frage, die sich stellt, ob die Folge der Einschränkung in einem gewissen Sinne konvergiert und ob der Grenzwert von der Wahl der Folge abhängt. Sei also $u \in H^1(\Omega)$ und $u_n \in C^\infty(\Omega) \cap H^1(\Omega)$ eine beliebige Cauchy-Folge mit

$$\|u - u_n\|_{H^1(\Omega)} \overset{n\to\infty}{\to} 0.$$

Sei $\varphi_n := u_n|_{\partial\Omega}$ die Einschränkung von u_n auf den Rand des Gebiets. Dann bilder φ_n eine Cauchy-Folge in $L^2(\partial\Omega)$, falls wir zeigen

$$\|\varphi_n\|_{L^2(\partial\Omega)} \leq C \|u_n\|_{H^1(\Omega)} \tag{2.1}$$

mit einer Konstanten $C > 0$, die nicht von $(u_n)_n$ abhängt. Dann konvergiert φ_n gegen eine Funktion $\varphi \in L^2(\partial\Omega)$ und wir setzen

$$u|_{\partial\Omega} := \varphi.$$

Wir sehen also, dass das zentrale Konzept in dieser Konstruktion aus den folgenden beiden Aussagen besteht:

1. $C^\infty(\Omega) \cap H^1(\Omega)$ ist dicht in $H^1(\Omega)$,
2. Abschätzung (2.1) gilt für alle glatten Funktionen $C^\infty(\Omega) \cap H^1(\Omega)$.

Der Beweis der Normabschätzung (2.1) beruht auf dem Hauptsatz der Differential- und Integralrechnung. Wir führen ihn lediglich für ein einfaches Gebiet beispielhaft aus – im Allgemeinen muss man die radiale Integrationsrichtung durch eine gekrümmte ersetzen.

Beispiel 2.13 *Sei Ω die offene Einheitskreisscheibe in $\mathbb{R}^2$. Für $u \in C^1\left(\overline{\Omega}\right)$ betrachten wir die Restriktion auf den Rand $\partial\Omega$ in Polarkoordinaten $\hat{u}(r,\varphi) := u \circ \psi(r,\varphi)$ mit $\psi(r,\varphi) := (r\cos\varphi, r\sin\varphi)^T$ und erhalten mit $x = (x_1, x_2)^T$*

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

Integration bezüglich der Winkelvariable liefert

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

Die Cauchy-Schwarzsche-Ungleichung liefert

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

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

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

und schliesslich

$$\|u\|_{L^2(\partial\Omega)} \leq 8^{1/4}\|u\|_{H^1(\Omega)}. \tag{2.2}$$

Das vorige Beispiel ist ein Spursatz für $C^1(\Omega)$-Funktionen. Aussage (2.2) drückt die Stetigkeit des Spuroperators $\gamma_0 : H^1(\Omega) \to L^2(\Omega)$, $\gamma_0(u) := \varphi$ aus.

Wir kommen nun zum allgemeinen Spursatz.

Satz 2.14 (Spursatz) *Seien Ω ein Lipschitz-Gebiet und $p \in [1,\infty[$. Dann gibt es eine stetige lineare Abbildung $\gamma_0 : W^{1,p}(\Omega) \to L^p(\partial\Omega)$ mit der Eigenschaft*

$$\gamma_0(\varphi) = \varphi|_{\partial\Omega} \qquad \forall\varphi \in C^k\left(\overline{\Omega}\right).$$

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 Ω. Wie bereits in der Einleitung gesagt, benötigen wir für die Existenz und Eindeutigkeit Randbedingungen -beispielsweise Dirichlet-Randbedingungen, bei denen die Werte der

unbekannten 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 2.14 ist hierfür nicht geeignet, da man beweisen kann: $\gamma_0\left(W^{1,p}(\Omega)\right) \subsetneq L^p(\partial\Omega)$. Man kann zeigen, dass das Bild von $W^{1,p}(\Omega)$ unter dem Spuroperator einen abgeschlossenen Unterraum von $L^p(\partial\Omega)$ definiert, und wir führen die Bezeichnung

$$\gamma_0\left(H^1(\Omega)\right) =: H^{1/2}(\partial\Omega)$$

ein.

Satz 2.15 $W_0^{1,p}(\Omega) = \{\varphi \in W^{1,p}(\Omega) \mid \gamma_0(\varphi) = 0\}$.

Eine Norm auf $H^{1/2}(\partial\Omega)$ *ist durch*

$$\|\varphi\|_{H^{1/2}(\partial\Omega)} := \left(\|\varphi\|^2_{L^2(\partial\Omega)} + \int_{\partial\Omega}\int_{\partial\Omega} \frac{1}{\|\mathbf{x}-\mathbf{y}\|^{d-1}} \left(\frac{|\varphi(\mathbf{x}) - \varphi(\mathbf{y})|}{\|\mathbf{x}-\mathbf{y}\|^{1/2}} \right)^2 d\mathbf{y} d\mathbf{x} \right)^{1/2}$$

gegeben.

Wir kommen nun zur Definition der Normalen-Ableitung von Funktionen in $H^1(\Omega)$. In diesem Fall lässt sich nicht wie im Vorspann zu Beispiel 2.13 argumentieren. Für glatte Funktionen sind zwar auch die Normalen-Ableitungen wohldefiniert $\psi_n := \partial u/\partial\mathbf{n}$, diese konvergieren aber im Allgemeinen nicht in $L^2(\partial\Omega)$.

Wiederum verwenden wir das Hilfsmittel der partiellen Integration, um die Normalenableitung zu definieren. Betrachten wir dazu das Poisson-Modell-Problem

$$-\Delta u = g \quad \text{in } \Omega. \tag{2.3}$$

Da der Sobolev $H^m(\Omega)$ ein Teilraum von $L^2(\Omega)$ ist, macht es Sinn, die Gleichung (2.3) in $L^2(\Omega)$ zu betrachten, d.h., $g \in L^2(\Omega)$ vorauszusetzen. Die motiviert den Raum

$$H^1(\Omega, \Delta) := \left\{ u \in H^1(\Omega) \mid \Delta u \in L^2(\Omega) \right\}.$$

Übungsaufgabe 2.16 *Sei* Ω *die* d*-dimensionale Einheitskugel. Geben Sie eine Funktion* $u \in H^1(\Omega, \Delta)$ *an, die nicht in* $H^2(\Omega)$ *liegt.*

Für glatte Funktionen $u \in C^2(\Omega) \cap C^1\left(\overline{\Omega}\right)$ folgt durch partielle Integration

$$\int_{\partial\Omega} \frac{\partial u}{\partial\mathbf{n}} v = \int_\Omega \left(\langle \nabla u, \nabla v \rangle + (\Delta u)\, v \right) \quad \forall v \in H^1(\Omega).$$

Die rechte Seite ist auch für Funktionen $u \in H^1(\Omega, \Delta)$ wohldefiniert und motiviert die folgende Definition.

Definition 2.17 *Sei* Ω *ein beschränktes Lipschitz-Gebiet. Für* $u \in H^1(\Omega, \Delta)$ *ist die* (schwache) Normalenableitung $\psi \in \left(H^{1/2}(\partial\Omega)\right)'$ *durch*

$$(\psi, \gamma_0 v)_{L^2(\partial\Omega)} = \int_\Omega \left(\langle \nabla u, \nabla v \rangle + (\Delta u)\, v \right) \quad \forall v \in H^1(\Omega) \tag{2.4}$$

definiert.

Bemerkung 2.18 *Da der Spuroperator $\gamma_0 : H^1(\Omega) \to H^{1/2}(\partial\Omega)$ gemäss Definition surjektiv ist, wird durch (2.4) eindeutig ein Funktional $\psi \in \left(H^{1/2}(\partial\Omega)\right)'$ definiert. Wir setzen*

$$H^{-1/2}(\partial\Omega) := \left(H^{1/2}(\partial\Omega)\right)'$$

und verwenden die Notation für den Normalenspur-Operator $\gamma_1 : H^1(\Omega, \Delta) \to H^{-1/2}(\partial\Omega)$, d.h., $\gamma_1(u) := \psi$ in (2.4).

Für glatte Funktionen Funktionen $u \in C^2(\Omega) \cap C^1\left(\overline{\Omega}\right)$ gilt $\gamma_1(u) = \partial u/\partial \mathbf{n}$.

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

Satz 2.19 (Friedrichssche Ungleichung) *$\|\cdot\|_{k,p}$ und $|\cdot|_{k,p}$ definieren äquivalente Normen auf $W_0^{k,p}(\Omega)$.*

Beweis. Offensichtlich gilt $|v|_{k,p} \le \|v\|_{k,p}$ 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 bezeichnet $|\cdot|_\infty$ die Maximumsnorm auf $\mathbb{R}^d$. Sei $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ mit $|\boldsymbol{\alpha}| = k - 1$ und $\varphi \in C_0^\infty(\Omega)$, $\psi := D^{\boldsymbol{\alpha}}\varphi$. 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

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

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

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

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

$$|\varphi|^p_{W^{k-1,p}(\Omega)} \le (2R)^p |\varphi|^p_{W^{k,p}(\Omega)}. \tag{2.5}$$

Summation über alle k ergibt:

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

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

Bemerkung 2.20 *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 2.21 *Seien $(X, \|\cdot\|_X)$ und $(Y, \|\cdot\|_Y)$ zwei normierte Vektorräume.*

1. *Eine lineare Abbildung $A : X \to Y$ heisst **kompakt**, wenn das Bild $A\left(\overline{B_X(0,1)}\right)$ des abgeschlossenen Einheitsballs in X in Y präkompakt[2] ist.*
2. *X ist **stetig eingebettet** in Y, kurz $X \hookrightarrow Y$, wenn $X \subset Y$ und die kanonische Injektion $\iota : X \to Y$ stetig ist.*
3. *X ist **kompakt eingebettet** in Y, kurz $X \overset{c}{\hookrightarrow} Y$, wenn $X \subset Y$ und die kanonische Injektion $\iota : X \to Y$ kompakt ist.*

Bemerkung 2.22

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

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

zu (2): Sei $A : X \to Y$ ein kompakter, linearer Operator. Dann ist nach Voraussetzung $A\left(\overline{B_X(0,1)}\right)$ präkompakt 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 präkompakten Menge $\iota\left(\overline{B_{\|\cdot\|_X}(0,R)}\right)$ mit $R := \max_{n\in\mathbb{N}} \|\varphi_n\|_X$ enthalten. ■

Satz 2.23 (Sobolevscher Einbettungssatz)
Let Ω be a bounded domain.

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

[2] Eine Teilmenge $K \subset M$ eines metrischen Raumes M heisst *kompakt*, falls jede offene Überdeckung von K eine endliche Teilüberdeckung enthält.
Eine Teilmenge $K \subset M$ eines metrischen Raumes M heisst *präkompakt*, falls ihr Abschluss kompakt ist.

Bemerkung 2.24 *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 2.25 (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 2.19 müssen wir nur zeigen, dass es eine Konstante $C > 0$ gibt mit

$$\|\varphi\|_1 \leq C\,|\varphi|_1 \qquad \forall \varphi \in V. \tag{2.6}$$

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

$$\|\varphi_n\|_1 = 1 \qquad \forall n \in \mathbb{N} \tag{2.7}$$

und

$$\lim_{n\to\infty} |\varphi_n|_1 = 0. \tag{2.8}$$

Wegen Satz 2.23 und Bemerkung 2.22(3) gibt es eine Teilfolge $(\varphi_{n_k})_{k\in\mathbb{N}}$ von $(\varphi_n)_{n\in\mathbb{N}}$ und eine Funktion $\varphi \in L^2(\Omega)$ mit

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

Wegen (2.8) konvergiert $(\varphi_{n_k})_{k\in\mathbb{N}}$ sogar in $H^1(\Omega)$. Mithin ist $\varphi \in H^1(\Omega)$ und $|\varphi|_1 = 0$. Daher ist φ konstant. Aus $\int_\Omega \varphi = 0$ folgt daher $\varphi = 0$ im Widerspruch zu (2.7). ■

Bemerkung 2.26 *Satz 2.25 kann für $H^1(\Omega)$ nicht gelten, da die rechte Seite von (2.6) 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.*

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

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

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

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 $u^\star \in X$. Dieses ist die eindeutige Lösung von

$$a(u^\star, v) = \ell(v) \qquad \forall v \in X. \tag{3.1}$$

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

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

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

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

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.

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

Dann folgt für $n, m \in \mathbb{N}$

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

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

$$a(v, u_\varphi) = (\varphi, v)_H \qquad \forall v \in X. \tag{3.2}$$

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

$$a(u_S, v) = \ell(v) \qquad \forall v \in S. \tag{3.3}$$

Aus (3.1) und (3.3) folgt

$$a(u - u_S, v) = 0 \qquad \forall v \in S. \tag{3.4}$$

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

$$v \to (\varphi, v)_H \qquad \forall v \in X$$

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 $X \overset{c}{\hookrightarrow} Y$ und a_0,$a_1 \in \mathcal{L}^2(X, \mathbb{R})$ zwei stetige Bilinearformen. Die Bilinearform a_0 sei symmetrisch und koerziv. Für die Bilinearform a_1 gebe es eine Konstante $\bar{A} \in \mathbb{R}_{>0}$ mit*

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

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

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

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

$$a(u, u) > 0. \tag{3.6}$$

Dann besitzen die Probleme

$$a(u, v) = \ell(v) \qquad \forall v \in X \tag{3.7}$$

und

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

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

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

Dann ist (3.7) äquivalent zu

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

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

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

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

$$(I + K)u = u_\ell. \tag{3.8}$$

Wegen $X \overset{c}{\hookrightarrow} Y$ und (3.5) ist K kompakt (vgl. Lemma 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. ■

Lemma 3.5 *Der Operator $K : X \to X$ in (3.8) ist kompakt.*

Wir benötigen dafür zwei Hilfsaussagen – die erste ist einfach zu beweisen, für die zweite verweisen wir auf eine Resultat in der Analysis.[3]

Lemma 3.6 *a) Seien $T_1 \in \mathcal{L}(X, Y)$ und $T_2 \in \mathcal{L}(Y, Z)$ stetige lineare Abbildungen und davon (mindestens) eine kompakt. Dann ist $T_2 \circ T_1$ kompakt.*

b) Seien X, Y zwei Banach-Räume kompakt ineinander eingebettet, d.h., $X \underset{c}{\hookrightarrow} Y$. Dann gilt für die Dualräume $Y' \underset{c}{\hookrightarrow} X'$.

Beweis. @a: Sei $K_1 := \overline{B_{\|\cdot\|_X}(0, 1)}$. Ist T_1 kompakt, d.h., $K_2 := T_1(K_1)$ ist präkompakt in Y, so ist auch $T_2(K_2)$ präkompakt in Z.

Ist dagegen T_2 kompakt, beweist man die Behauptung wie folgt. Da eine Skalierung die Kompaktheit nicht ändert, kann ohne Beschränkung der Allgemeinheit $\|T_1\|_{\mathcal{L}(X,Y)} \leq 1$ angenommen werden. Damit ist $K_2 := T_1(K_1)$ eine Teilmenge der Einheitskugel in Y und somit $T_2(K_2)$ präkompakt in Z.

@ b: Siehe [4, §X, Sec. 4]. ■

Beweis von Lemma 3.5.

Wir betrachten

$$a_0(Kw, v) = a_1(w, v) \quad \forall v \in X, \tag{3.9}$$

wobei a_0 die Voraussetzungen von Lax-Milgram erfüllt und

$$|a_1(u, v)| \leq C \|u\|_X \|v\|_Y \qquad \forall u \in X \text{ und } \forall v \in Y$$

[3] Im Folgenden sind X, Y immer Banach-Räume. $\mathcal{L}(X, Y)$ bezeichnet die Menge aller stetigen, linearen Abbildungen von X nach Y. Die Dualräume (Menge der stetigen, linearen Funktionale) von X bzw. Y werden mit $X' := \mathcal{L}(X, \mathbb{R})$ bzw. $Y' := \mathcal{L}(Y, \mathbb{R})$ bezeichnet.

gelten soll. Für gegebenes $w \in X$ definiert $a_1(w, \cdot)$ eine stetige Abbildung $Tw \in Y'$, d.h., das Funktional $Tw \in Y'$ ist definiert durch

$$(Tw)(v) = a_1(w, v) \qquad \forall v \in Y.$$

Weil Y' kompakt in X' eingebettet ist, kann T als kompakte Abbildung von $T : X \to X'$ aufgefasst werden (Hintereinanderausführung der stetigen Abbildung $T \in \mathcal{L}(X, Y')$ mit der kompakten Injektion $\iota : Y' \to X'$. Damit lässt sich (3.9) schreiben als

$$a_0(Kw, v) = (Tw)(v) \quad \forall v \in X.$$

Für $F \in X'$ ist nach Lax-Milgram der Lösungsoperator $S : X' \to X$ wohldefiniert durch

$$a_0(SF, v) = F(v) \qquad \forall v \in X$$

und stetig, d.h. $S \in \mathcal{L}(X', X)$. Das bedeutet $K = S \circ T$. Da S stetig und T kompakt ist, ist K kompakt. ■

Satz 3.7 *Die Bezeichnungen und Voraussetzungen seien wie in Satz 3.4. Zusätzlich sei $S \subset X$ ein endlichdimensionaler Unterraum. Dann besitzt das Problem*

$$a(u_S, v) = \ell(v) \qquad \forall v \in S \tag{3.10}$$

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 (3.10) die Fehlerabschätzung

$$\|u - u_S\|_X \leq \frac{A}{\beta} \inf_{v \in S} \|u - v\|_X . \tag{3.11}$$

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

$$\|u - u_S\|_H \leq A \|u - u_S\|_X \sup_{\varphi \in H \setminus \{0\}} \inf_{v \in S \setminus \{0\}} \frac{\|u_\varphi - v\|_X}{\|\varphi\|_H}. \tag{3.12}$$

Beweis. Die eindeutige Lösbarkeit von (3.10) folgt aus Satz 3.4. Die Fehlerabschätzung (3.11) folgt wie im Beweis von Satz 3.2. Man beachte, dass wir dort nur die zu (3.7) und (3.10) analogen Eigenschaften (3.1) und (3.3) ausgenützt haben. Wegen $X \hookrightarrow H$ definiert jedes $\varphi \in H$ durch $v \to (\varphi, v)_H$ ein stetiges lineares Funktional auf X. Daher folgt die Existenz und Eindeutigkeit von u_φ aus Satz 3.4. Die Fehlerabschätzung (3.12) 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. (1.1))

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

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

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

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

- (homogene) **Dirichlet-Randbedingungen:** $u = 0$ auf Γ,
- (inhomogene) **Neumann-Randbedingungen:** $\langle \mathbf{An}, \operatorname{grad} u\rangle = g$ auf Γ,
- **gemischte Dirichlet-Neumann-Randbedingungen:** $u = 0$ auf Γ_{D} und $\langle \mathbf{An}, \operatorname{grad} u\rangle = g$ auf Γ_{N}.

Dabei ist $g \in L^2(\Gamma)$, $\Gamma_{\mathrm{D}} \cap \Gamma_{\mathrm{N}} = \emptyset$ und $\Gamma = \Gamma_{\mathrm{D}} \cup \Gamma_{\mathrm{N}}$. Wir werden bei gemischten Randbedingungen stets fordern, dass Γ_{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 Ω und Anwenden des Gaussschen Integralsatzes liefert

$$\begin{aligned}\int_\Omega fv &= -\int_\Omega v \operatorname{div}(\mathbf{A}\operatorname{grad} u) + \int_\Omega \langle \mathbf{b}, \operatorname{grad} u\rangle v + \int_\Omega cuv \\ &= \int_\Omega \left(\langle \operatorname{grad} v, \mathbf{A}\operatorname{grad} u\rangle + \langle \mathbf{b}, \operatorname{grad} u\rangle v + cuv\right).\end{aligned} \tag{4.3}$$

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

$$\int_\Omega \left(\langle \operatorname{grad} v, \mathbf{A}\operatorname{grad} u\rangle + \langle \mathbf{b}, \operatorname{grad} u\rangle v + cuv\right) = \int_\Omega fv \qquad \forall v \in H_0^1(\Omega) \tag{4.4}$$

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 $v \in C^\infty\left(\overline{\Omega}\right)$, so treten in (4.3) zusätzliche Randterme $\int_\Gamma \langle \mathbf{An}, \operatorname{grad} u\rangle v$ auf. Erfüllt u Neumann-Randbedingungen, so lassen sich diese substituieren:

$$\int_\Gamma \langle \mathbf{An}, \operatorname{grad} u\rangle v = \int_\Gamma gv.$$

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*

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

2. *Sei*

$$H^1_{\mathrm{D}}(\Omega) := \left\{\varphi \in H^1(\Omega) : \varphi|_{\Gamma_{\mathrm{D}}} = 0\right\}. \tag{4.5}$$

*Die Funktion $u \in H^1_{\mathrm{D}}(\Omega)$ heisst **schwache Lösung** der Differentialgleichungen (4.1) mit gemischten Randbedingungen, wenn gilt*

$$\int_\Omega \langle \mathbf{A}\operatorname{grad} u, \operatorname{grad} v\rangle + \langle \mathbf{b}, \operatorname{grad} u\rangle v + cuv = \int_\Omega fv + \int_{\Gamma_{\mathrm{N}}} gv \qquad \forall v \in H^1_{\mathrm{D}}(\Omega).$$

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

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

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 $c \in L^\infty(\Omega)$, $\mathbf{b} \in L^\infty\left(\Omega, \mathbb{R}^d\right)$, $\mathbf{A} \in L^\infty\left(\Omega, \mathbb{R}^{d\times d}\right)$.*

3. *Bei inhomogenen Dirichlet-Randbedingungen $u = u_{\mathrm{D}}$ auf Γ bzw. Γ_{D} muss in Definition 4.1 die Bedingung $u \in H^1_0(\Omega)$ bzw. $u \in H^1_{\mathrm{D}}(\Omega)$ durch $u \in u_{\mathrm{D}} + H^1_0(\Omega)$ bzw. $u \in u_{\mathrm{D}} + H^1_{\mathrm{D}}(\Omega)$ ersetzt werden.*

Satz 4.3 (Existenz- und Eindeutigkeitssatz 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 Γ_{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 Γ, 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 Γ 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 $X := H^1_0(\Omega)$, $Y := L^2(\Omega)$ und

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

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 ℓ 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 $H^1_0(\Omega)$:

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

mit der Konstante c_{F} aus der Friedrichsschen Ungleichung.

Aus dem Gaussschen Integralsatz und (4.2) folgt schliesslich

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

Die Bilinearform a ist wegen $-\frac{1}{2} \operatorname{div} \mathbf{b} + c \geq 0$ koerziv auf $H^1_0(\Omega)$.

Die Behauptung folgt daher aus Satz 3.4.

zu 2: In diesem Fall setzen wir $X = H^1_{\mathrm{D}}(\Omega)$ und $\ell(v) = \int_\Omega fv + \int_{\Gamma_{\mathrm{N}}} gv$ und definieren die anderen Grössen wie im ersten Fall. Aus der Cauchy-Schwarzschen-Ungleichung und dem Spursatz 2.14 folgt die Stetigkeit von ℓ

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

Die Koerzivität von a_0 bleibt wegen Bemerkung 2.20 erhalten. Bei der Anwendung des Gaussschen Integralsatzes in der Abschätzung von $a(u,u)$ tritt der zusätzliche Randterm $\int_{\Gamma_{\mathrm{N}}} \langle \mathbf{n}, \mathbf{b} \rangle u^2$ auf. Wegen $\langle \mathbf{n}, \mathbf{b} \rangle \geq 0$ auf Γ_{N} ist er nicht negativ, und die Koerzivität von a bleibt erhalten.

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

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

und somit die Koerzivität von a_0. Die anderen Abschätzungen ändern sich nicht, insbesondere gilt

$$a(u,u) \geq \lambda |u|^2_{H^1(\Omega)} + \int_\Omega \left(-\frac{1}{2} \operatorname{div} \mathbf{b} + c \right) u^2 + \int_\Gamma \langle \mathbf{b}, \mathbf{n} \rangle u^2 \geq \lambda |u|^2_{H^1(\Omega)}$$

für alle $u \in X$ und $a(u,u) = 0$ impliziert $u = \text{const}$, d.h., $u = \rho$ für $\rho \in \mathbb{R}$. Explizit gilt jedoch:

$$a(\rho,\rho) = \int_\Omega c\rho^2 \geq c_0 \|\rho\|^2_{L^2(\Omega)} = c_0 |\Omega| \rho^2$$

mit dem Volumen $|\Omega|$ von Ω. Daraus folgt die Implikation $a(u,u) = 0 \implies u = 0$ und die Behauptung ergibt sich aus der unsymmetrischen Variante des Satzes von Lax-Milgram (cf. Satz 3.4).

zu 4: Alle Grössen sind wie in Fall 3. Wegen $c = 0$ gilt

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

Hieraus und aus der Poincaréschen Ungleichung (Satz 2.25) folgt die Koerzivität von a_0 auf $V := \left\{u \in H^1(\Omega) : \int_\Omega \varphi = 0\right\}$. Die Abschätzung $a(u,u) > 0$ für alle $u \in X \setminus \{0\}$ bleibt gültig, ebenso die Stetigkeit von a_0 und a_1. Lediglich bei der Stetigkeit von ℓ ist Sorgfalt geboten. Da $V \cong H^1(\Omega)/\mathbb{R}$ ist, muss für die Stetigkeit von ℓ die Inklusion $\mathbb{R} \subset \text{Kern}(\ell)$ gelten. Dies ist aber wegen $\int_\Omega f + \int_\Gamma g = 0$ der Fall. ■

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

Beispiel 4.4 *Sei $0 < \alpha < 2\pi$ und Ω_α 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 $v \in \Omega_\alpha \to \mathbb{R}$ durch

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

Dann gilt für jedes $\mathbf{x} \in \Omega_\alpha$

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

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

Definiere

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

Dann gilt

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

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

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

Wie man leicht nachrechnet gilt

$$u \in C^k\left(\overline{\Omega}_\alpha\right) \iff 0 < \alpha \leq \frac{\pi}{k}, \qquad k \geq 1$$

und

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

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

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

und ***keine*** *Abschätzung der Form*

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

erwarten, wie sie für gewöhnliche Differentialgleichungen gelten würde.

Satz 4.5 (Regularitätssatz) *Sei Γ ein C^2-Mannigfaltigkeit oder Ω 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 Ω 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

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

Wir nehmen $A \in C^1(\overline{\Omega})$, $c \in C^0(\overline{\Omega})$ und

$$\begin{aligned} 0 < \lambda_0 \leq A(x) \leq \lambda_1 < \infty, \quad \alpha_1 := \|A'\|_{L^\infty(\Omega)} < \infty, \\ 0 < c_0 \leq c(x) \leq c_1 < \infty \end{aligned}$$

an.

Kontinuierliches Problem:

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

$$a(u,v) = \ell(v) \qquad \forall v \in H_0^1(\Omega). \tag{5.1}$$

Das Ziel dieses Kapitels ist es, dieses Problem mit finite Elementen zu diskretisieren. Die Diskretisierung wird im Rahmen von Satz 3.2 und Satz 3.7 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 $0 = x_0 < x_1 < \ldots < x_{n+1} = 1$ 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 $S^{1,0}_{\mathcal{G}}$ ein:

$$S := S^{1,0}_{\mathcal{G}} := \left\{\varphi \in C^0([0,1]) \mid \forall \tau \in \mathcal{G} : \varphi|_\tau \in \mathbb{P}_1\right\} \cap H^1_0(\Omega).$$

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

$$b_i(x) := \begin{cases} \frac{x - x_{i-1}}{h_{i-1}} & \text{in } \tau_{i-1}, \\ \frac{x_{i+1} - x}{h_i} & \text{in } \tau_i, \\ 0 & \text{sonst.} \end{cases} \tag{5.2}$$

Das diskrete Problem

$$a(u_S, v) = \ell(v) \qquad \forall v \in S \tag{5.3}$$

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[4]

$$\|u - u_S\|_{H^1(\Omega)} \leq C_{\text{qo}} \inf_{v \in S} \|u - v\|_{H^1(\Omega)} \overset{(2.5)}{\leq} C_{\text{qo}} \sqrt{2} \inf_{v \in S} |u - v|_{H^1(\Omega)} \tag{5.4}$$

mit

$$C_{\text{qo}} := \frac{C_{\text{cont}}}{c_{\text{coer}}}, \quad c_{\text{coer}} := \min\{\lambda_0, c_0\}, \quad C_{\text{cont}} := \lambda_1 + c_1$$

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 bewesen.

Sei $\tilde{u}$ die lineare Interpolation von u auf τ. 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

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

[4] Aus (2.5) folgt für ein Gebiet mit Durchmesser ρ die Abschätzung

$$\|\varphi\|_{H^1(\Omega)} \leq \sqrt{1 + \rho^2}\,|\varphi|_{H^1(\Omega)}.$$

Für das Einheitsinterval gilt $\rho = 1$.

Quadrieren und die Hölder-Ungleichung liefert:

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

Integration über τ ergibt

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

Durch Summation über alle Intervalle erhalten wir

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

und haben insgesamt die Approximationseigenschaft:

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

gezeigt. Daraus folgt mit (5.4)

$$\|u - u_S\|_{H^1(\Omega)} \leq C_{\text{qo}} \inf_{v \in S} \|u - v\|_{H^1(\Omega)} \overset{(2.5)}{\leq} C_{\text{qo}} \sqrt{2} h |u|_{H^2(\Omega)}. \tag{5.5}$$

Natürlich muss hierfür $u \in H^2(\Omega)$ vorausgesetzt werden. Diese Bedingung wird im Anschluss diskutiert. Mit Satz 3.2 folgt daraus auch die Fehlerabschätzung

$$\|u - u_S\|_{L^2(\Omega)} \leq \sqrt{2} \frac{C^2_{\text{cont}}}{c_{\text{coer}}} h_{\mathcal{G}} |u|_{H^2(\Omega)} \sup_{\varphi \in L^2(\Omega) \setminus \{0\}} \inf_{v \in S} \frac{\|u_\varphi - v\|_{H^1(\Omega)}}{\|\varphi\|_{L^2(\Omega)}}. \tag{5.6}$$

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

$$\begin{aligned} -\left(A u'_\varphi\right)' + c u_\varphi &= \varphi \quad \text{in } \Omega, \\ u_\varphi(0) = u_\varphi(1) &= 0 \end{aligned} \tag{5.7}$$

Nochmaliges Anwenden des vorigen Approximationsresultats liefert:

$$\sup_{\varphi \in L^2(\Omega) \setminus \{0\}} \inf_{v \in S} \frac{\|u_\varphi - v\|_{H^1(\Omega)}}{\|\varphi\|_{L^2(\Omega)}} \leq \sqrt{2} h_{\mathcal{G}} \sup_{\varphi \in L^2(\Omega) \setminus \{0\}} \frac{|u_\varphi|_{H^2(\Omega)}}{\|\varphi\|_{L^2(\Omega)}}. \tag{5.8}$$

Wie in (5.5) wird auch hier die H^2-Regularität benötigt.

Lemma 5.1 *Das Problem (5.1) ist H^2-regulär.*

Beweis. Wir wählen $v = u$ in (5.1) und erhalten mit Cauchy-Schwarz-Ungleichungen

$$c_{\text{coer}} \|u\|^2_{H^1(\Omega)} \leq a(u,u) \leq \|f\| \|u\| \leq \|f\| \|u\|_{H^1(\Omega)}$$

und damit

$$\|u\|_{H^1(\Omega)} \leq \frac{1}{c_{\text{coer}}} \|f\|. \tag{5.9}$$

Die starke Formulierung von (5.1) liefert die Differentialgleichung

$$-(Au')' + cu = f \quad \text{in } \Omega \quad \text{mit} \quad u(0) = u(1) = 0.$$

Mit Hilfe der Leibniz-Produktregel erhalten wir

$$-u'' = \frac{f + A'u' - cu}{A}.$$

Wir wenden die $L^2(\Omega)$-Norm auf beide Seiten an und erhalten

$$\|u''\| \leq \frac{1}{\lambda_0} \left(\|f\| + \alpha_1 \|u'\| + c_1 \|u\| \right) \overset{(5.9)}{\leq} C_{\text{reg}} \|f\| \quad \text{mit} \quad C_{\text{reg}} := \frac{1}{\lambda_0} \left(1 + \frac{\sqrt{\alpha_1^2 + c_1^2}}{c_{\text{coer}}} \right).$$

und das ist die H^2-Regularität. ■

Die Kombination von diesem Lemma und (5.5), (5.6), (5.8) ergibt

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

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

1. Konstruktion einer Unterteilung $\mathcal{G}$ von Ω 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 Simpliziale finite Elemente in d Dimensionen

Sei $\Omega \subset \mathbb{R}^d$ ein beschränktes Lipschitz-Polytop (Intervall für $d = 1$, polygonales Gebiet für $d = 2$, polyhedrales Gebiet für $d = 3, \ldots$). Wir betrachten die allgemeine, *symmetrische*, elliptische Differentialgleichung mit gemischten-Randbedingungen (siehe (4.5)):

$$\text{Finde } u \in H^1_{\text{D}}(\Omega) \qquad a(u, v) = \ell(v) \qquad \forall v \in H^1_{\text{D}}(\Omega) \tag{6.1}$$

mit

$$a(u, v) = \int_\Omega \left(\mathbf{A} \nabla u, \nabla v + cuv \right) \quad \text{und} \quad \ell(v) = \int_\Omega fv + \int_{\Gamma_{\text{N}}} gv. \tag{6.2}$$

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

$$\mathbf{A} \in L^\infty\left(\Omega, \mathbb{R}^{d \times d}_{\text{symm}}\right), \qquad c \in L^\infty(\Omega, \mathbb{R}_{\geq 0}), \qquad f \in L^2(\Omega), \quad g \in L^2(\Gamma_{\text{N}}) \tag{6.3}$$

und

$$0 < \lambda := \inf_{\mathbf{x} \in \Omega} \inf_{\mathbf{z} \in \mathbb{R}^d \setminus \{0\}} \frac{\mathbf{z}^T \mathbf{A}(\mathbf{x}) \mathbf{z}}{\mathbf{z}^T \mathbf{z}} \leq \sup_{\mathbf{x} \in \Omega} \sup_{\mathbf{z} \in \mathbb{R}^d \setminus \{0\}} \frac{\mathbf{z}^T \mathbf{A}(\mathbf{x}) \mathbf{z}}{\mathbf{z}^T \mathbf{z}} =: \Lambda < \infty. \tag{6.4}$$

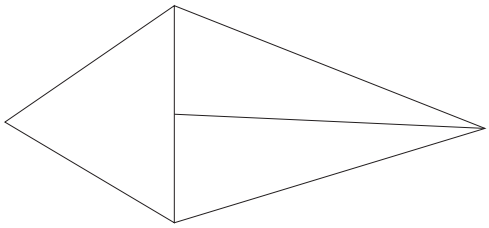


Abbildung 1: Triangulierung mit „hängendem Knoten“.

Definition 6.1 *Eine Unterteilung von Ω in abgeschlossene Simplizes $\mathcal{G} = \{\tau_i : 1 \leq i \leq n_{\mathcal{G}}\}$ ist ein konformes*[5] *simpliziales Finite-Elemente-Gitter, falls $\overline{\Omega} = \bigcup_{\tau\in\mathcal{G}} \tau$ gilt und der Schnitt zweier Simplizes τ_i, τ_j $(i \neq j)$ entweder leer ist oder ein gemeinsamer, niederdimensionaler Oberflächensimplex ist.*

Wir bezeichnen dazu für $0 \leq k \leq d-1$ mit $\mathcal{C}_k(\tau)$ die Menge aller k-dimensionalen (relativ abgeschlossenen) Oberflächensimplizes[6] von τ und setzen $\mathcal{C}_d(\tau) := \{\tau\}$.

Beispiel 6.2 *Die Schnittbedingungen in Definition 6.1 besagen, dass der Schnitt zweier Elemente τ_i, $\tau_j \in \mathcal{G}$, $i \neq j$,*

1. *für $d = 1$ (τ_i sind Intervalle) entweder leer ist oder ein gemeinsamer Endpunkt ist,*
2. *für $d = 2$, (τ_i sind Dreiecke) entweder leer, ein gemeinsamer Eckpunkt oder eine gemeinsame Kante ist,*
3. *für $d = 3$, (τ_i sind Simplizes) entweder leer, ein gemeinsamer Eckpunkt eine gemeinsame Kante oder eine gemeinsame Seitenfläche ist.*

Ein Mass für die Entartung der Simplizes ist die Formregularitätskonstante (shape regularity constant):

$$c_{\mathcal{G}} := \max_{\tau\in\mathcal{G}} \left\{ \frac{h_\tau^d}{|\tau|} \right\} \tag{6.5}$$

mit dem d-dimensionalen Volumen $|\tau|$ von $\tau \in \mathcal{G}$ und

$$h_{\mathcal{G}} := \max_{\tau\in\mathcal{G}} h_\tau \quad \text{mit} \quad h_\tau := \operatorname{diam} \tau.$$

Die Menge aller Eckpunkte der Simplizes werden wieder mit $\Theta_{\mathcal{G}}$ bezeichnet.

Definition 6.3 *Der* Referenzsimplex $\hat{\tau}_d \subset \mathbb{R}^d$ *ist gegeben durch*

$$\hat{\tau}_d := \left\{ \mathbf{x} \in \mathbb{R}^d_{\geq 0} \mid \|\mathbf{x}\|_{\ell_1} \leq 1 \right\}.$$

[5] Die Bezeichnung „konform“ bezieht sich auf die Schnittbedingungen, die sicher stellen werden, dass der endlichdimensionale Finite-Elemente Raum S in $H_0^1(\Omega)$ enthalten (konform) ist. Ein nicht-konformes Finite-Elemente-Gitter kann „hängende Knoten“ enthalten (siehe. Abb. 1).

[6] Für $d = 1$ besteht $\mathcal{C}_0(\tau)$ aus den beiden Intervallenden; für $d = 2$ besteht $\mathcal{C}_0(\tau)$ aus den Dreiecksecken und $\mathcal{C}_1(\tau)$ aus den Dreieckskanten; für $d = 3$ besteht $\mathcal{C}_0(\tau)$ aus den Tetraederecken, $\mathcal{C}_1(\tau)$ aus den Tetraederkanten und $\mathcal{C}_2(\tau)$ aus den Tetraederseitenflächen.

$\hat{\tau}_d$ ist affine äquivalent *zu einem beliebigen Simplex $\tau \subset \mathbb{R}^d$ mit Eckpunkten $\mathbf{x}_i^\tau$, $0 \leq i \leq d$ durch die (nicht eindeutige) affine Transformation*

$$\chi_\tau(\hat{\mathbf{x}}) := \mathbf{x}_0^\tau + \mathbf{m}_\tau \hat{\mathbf{x}}$$

mit der $d \times d$ Jacobi-Matrix $\mathbf{m}_\tau$ deren i-te Spalte durch $\mathbf{x}_i^\tau - \mathbf{x}_0^\tau$ gegeben ist. Falls die Dimension d aus dem Kontext klar ist, schreiben wir kurz $\hat{\tau}$ statt $\hat{\tau}_d$.

Um die Menge $\mathbb{P}_p$ aller d-variaten Polynome zu definieren, benötigen wir zunächst einige Notationen. Für einen Vektor $\mathbf{x} = (x_i)_{i=1}^d \in \mathbb{R}^d$ oder einen Multiindex $\boldsymbol{\mu} = (\mu_i)_{i=1}^d \in \mathbb{N}_0^d$ setzen wir

$$|\mathbf{x}| = \sum_{i=1}^d x_i, \quad |\boldsymbol{\mu}| = \sum_{i=1}^d \mu_i \quad \text{und} \quad \mathbf{x}^{\boldsymbol{\mu}} := \prod_{i=1}^d x_i^{\mu_i}.$$

Die Indexmenge ι_p^d ist gegeben durch

$$\iota_p^d := \left\{ \boldsymbol{\mu} \in \mathbb{N}_0^d \mid |\boldsymbol{\mu}| \leq p \right\}$$

und man beweist durch vollständige Induktion

$$\sharp \iota_p^d = \binom{p+d}{d}.$$

Falls die Dimension aus dem Kontext klar ist, schreiben wir kurz ι_p statt ι_p^d. Der Polynomraum $\mathbb{P}_p$ ist gegeben durch

$$\mathbb{P}_p^d := \left\{ \sum_{\alpha \in \iota_p^d} a_\alpha \mathbf{x}^\alpha \mid \forall \alpha \in \iota_p^d : a_\alpha \in \mathbb{R} \right\}.$$

Wiederum schreiben wir kurz $\mathbb{P}_p$, wenn die Dimension d klar ist und $\mathbb{P}_p^d(\omega)$ für ein Gebiet $\omega \subset \mathbb{R}^d$, falls die Funktionen in $\mathbb{P}_p^d$ nur für $\mathbf{x} \in \omega$ betrachtet werden.

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

$$S := S_{\mathcal{G},\mathrm{D}}^{p,0} := \left\{ \varphi \in C^0\left(\overline{\Omega}\right) \mid \forall \tau \in \mathcal{G} : \varphi|_\tau \in \mathbb{P}_p \right\} \cap H_{\mathrm{D}}^1(\Omega), \tag{6.6}$$

wobei $H_{\mathrm{D}}^1(\Omega)$ wieder wie in Definition 4.1 vom Dirichlet-Teil Γ_{D} des Randes abhängt.

Zunächst werden wir eine Lagrange-Basis für $\mathbb{P}_p$ auf dem Referenzdreieck definieren und dazu eine geeignete Menge von Knotenpunkten. Wir setzen

$$\hat{\Sigma}_p^d := \begin{cases} \frac{1}{d+1}(1)_{i=1}^d & p = 0, \\ \left\{ \frac{1}{p}\boldsymbol{\mu} \mid \boldsymbol{\mu} \in \iota_p^d \right\} & p \geq 1. \end{cases}$$

Für $\mathbf{z} = \frac{1}{p}\boldsymbol{\mu} \in \hat{\Sigma}_p^d$ definieren wir die Funktionen $\hat{B}_{\mathbf{z}}$ durch[7]

$$\hat{B}_{\mathbf{z}}(\mathbf{x}) = \prod_{i=1}^d \prod_{s=0}^{\mu_i - 1} \left(\frac{s - p x_i}{s - p z_i} \right) \prod_{s=|\boldsymbol{\mu}|+1}^p \left(\frac{s - p|\mathbf{x}|}{s - p|\mathbf{z}|} \right). \tag{6.7}$$

[7] Diese Darstellung der Lagrange-Basis für Polynome auf Simplizes stammt von Prof. Charles F. Dunkl, Univ. Virginia.

Satz 6.4 *$\hat{B}_{\mathbf{z}}$, $\mathbf{z} \in \hat{\Sigma}_p^d$, bildet eine Lagrange-Basis von $\mathbb{P}_p$.*

Beweis. 1. Der Polynomgrad von $\hat{B}_{\mathbf{z}}$ ist gegeben durch

$$\left(\sum_{i=1}^{d}\sum_{s=0}^{\mu_i-1} 1\right) + \left(\sum_{s=|\boldsymbol{\mu}|+1}^{p} 1\right) = |\boldsymbol{\mu}| + (p - |\boldsymbol{\mu}|) = p$$

so dass $\hat{B}_{\mathbf{z}} \in \mathbb{P}_p$ gilt.

2. Wir zeigen: $\hat{B}_{\mathbf{z}}$ besitzt die Lagrange-Eigenschaft: $\hat{B}_{\mathbf{z}}(\mathbf{z}) = 1$ und $\hat{B}_{\mathbf{z}}(\mathbf{y}) = 0$ für alle $\mathbf{y} \in \hat{\Sigma}_p^d \backslash \{\mathbf{z}\}$. Für $\mathbf{x} = \mathbf{z}$ stimmen alle Zähler in (6.7) mit den jeweiligen Nennern überein, so dass die Produkte 1 ergeben. Für $\mathbf{y} = \frac{\boldsymbol{\alpha}}{p} \in \hat{\Sigma}_p^d \backslash \{\mathbf{z}\}$ gilt

$$\hat{B}_{\mathbf{z}}(\mathbf{y}) = \left(\prod_{i=1}^{d}\prod_{s=0}^{\mu_i-1}\left(\frac{s-\alpha_i}{s-\mu_i}\right)\right)\prod_{s=|\boldsymbol{\mu}|+1}^{p}\left(\frac{s-|\boldsymbol{\alpha}|}{s-|\boldsymbol{\mu}|}\right).$$

Falls ein $1 \leq i \leq d$ existiert mit $\alpha_i \leq \mu_i - 1$ ist einer der Faktoren $s - \alpha_i$ gleich Null und damit auch $\hat{B}_{\mathbf{z}}(\mathbf{y})$. Sei nun $\alpha_i \geq \mu_i$ für alle i. Da $\boldsymbol{\alpha} \neq \boldsymbol{\mu}$ gilt, muss ein i existieren mit $\alpha_i > \mu_i$ und daher $|\boldsymbol{\alpha}| \geq |\boldsymbol{\mu}| + 1$. Das bedeutet, dass einer der Faktoren $s - |\boldsymbol{\alpha}|$ in diesem Fall gleich Null ist, und somit auch $\hat{B}_{\mathbf{z}}(\mathbf{y}) = 0$ gilt.

3. Die Lagrange-Eigenschaft impliziert, dass alle $\hat{B}_{\mathbf{z}}$ linear unabhängig sind und wegen $\sharp\hat{\Sigma}_p^d = \binom{p+d}{d} = \dim \mathbb{P}_p^d$ folgt, dass $\hat{B}_{\mathbf{z}}$ eine Lagrange-Basis von $\mathbb{P}_p$ ist. ■

Definition 6.5 *Die Menge der Knotenpunkte vom Grad p auf einem beliebigen Simplex $\tau \subset \mathbb{R}^d$ sind durch*

$$\Sigma_p^d(\tau) := \left\{\chi_\tau(\mathbf{z}) : \mathbf{z} \in \hat{\Sigma}_p^d\right\}$$

gegeben.

Bemerkung 6.6

a. *Die „hochgehobenen" Basisfunktionen*

$$B_{\mathbf{z}}^{\tau} := \hat{B}_{\hat{\mathbf{z}}} \circ \chi_\tau^{-1}, \quad \mathbf{z} \in \Sigma_k^d(\tau) \quad \hat{\mathbf{z}} := \chi_\tau^{-1}(\mathbf{z})$$

bilden eine Lagrange-Basis von $\mathbb{P}_p$ für die Knotenmenge $\Sigma_k^d(\tau)$.

b. *Der Polynomraum $\mathbb{P}_p$ ist invariant unter affinen Transformationen.*

c. *Für $1 \leq k \leq d-1$, sei f ein k-dimensionaler Oberflächensimplex von τ oder $f = \tau$. Dann ist für jedes $w \in \mathbb{P}_p^d$, die Einschränkung $w|_f$ durch die Werte in den Knotenpunkten $\mathbf{z} \in f \cap \Sigma_p^d(\tau)$ eindeutig festgelegt.*

Beweis. Aussagen (a) und (b) sind offensichtlich und wir beweisen lediglich Aussage (c). Sei $w \in \mathbb{P}_p$ und $f \in \mathcal{C}_k(\tau)$ ein k-dimensionaler Oberflächensimplex. Wir transportieren beides auf das Referenzelement, so dass $\hat{w} = w \circ \chi_\tau \in \mathbb{P}_p$ und $\hat{f} = \chi_\tau^{-1}(f)$ gilt. Wegen (b) genügt es die Aussage für den Referenzsimplex zu beweisen. O.b.d.A. können wir die affine Transformation χ_τ so wählen, dass $\hat{f} = \hat{\tau}_k$ gilt und wir setzen $\hat{w}_{\hat{f}} := \hat{w}|_{\hat{f}} \in \mathbb{P}_p^k$. Die Konstruktion der Knotenpunkte impliziert, dass $\hat{f} \cap \hat{\Sigma}_p^d = \hat{\Sigma}_p^k$ gilt. Aus Satz 6.4 folgt daher, dass $\hat{w}_{\hat{f}}$ eindeutig durch die Werte in den Knotenpunkten auf $\hat{f}$ festgelegt ist. ■

Korollar 6.7 *Seien $\tau_1, \tau_2 \in \mathcal{G}$ zwei Simplizes eines Finite-Elemente-Gitters $\mathcal{G}$, die disjunktes Inneres und nichtleeren Schnitt $f = \tau_1 \cap \tau_2$ haben. Seien $w_1, w_2 \in \mathbb{P}_p$ gegeben und $w : \tau_1 \cup \tau_2 \to \mathbb{R}$ durch*

$$w := \begin{cases} w_1 & \text{auf } \overset{\circ}{\tau}_1 \\ w_2 & \text{auf } \overset{\circ}{\tau}_2 \end{cases}$$

definiert. Dann stimmen die Spuren $w_1|_f$ und $w_2|_f$ überein, genau dann wenn sie in allen gemeinsamen Knotenpunkten

$$\hat{\Sigma}_p^d(\tau_1) \cap f = \hat{\Sigma}_p^d(\tau_2) \cap f \tag{6.8}$$

übereinstimmen.

Beweis. Die Gleichheit in (6.8) folgt, da affine Abbildungen (hier χ_{τ_1} und χ_{τ_2}) verhältnistreu sind. Daher folgt die Behauptung aus Bemerkung 6.6. ■

Wir definieren nun die Menge $\Sigma_p(\mathcal{G})$ der globalen Knotenpunkte im Finite-Elemente-Gitter $\mathcal{G}$. Diese hängt von den vorgegebenen Randbedingungen ab.

Konvention 6.8 *Im allgemeinen betrachten wir gemischte Randbedingungen, d.h., $\Gamma = \Gamma_{\mathrm{D}} \cup \Gamma_{\mathrm{N}}$ mit disjunkten (messbaren) Teilmengen Γ_{D} und Γ_{N}, wobei Γ_{D} relativ abgeschlossen sein soll. Wie nehmen an, dass auf Γ_{D} homogene Dirichlet-Randbedingungen gegeben sind und auf Γ_{N} (im Allgemeinen) nicht-homogene Neumann-Randbedingungen.*

Definition 6.9 *Die Menge $\Sigma_{p,\mathrm{D}}(\mathcal{G})$ der* globalen Knotenpunkte *im Finite-Elemente-Gitter $\mathcal{G}$ für gemischte Randwertproblem $\Gamma = \Gamma_{\mathrm{D}} \cup \Gamma_{\mathrm{N}}$ (wie in Konvention 6.8) ist gegeben durch*

$$\Sigma_{p,\mathrm{D}}(\mathcal{G}) := \left(\bigcup\nolimits_{\tau \in \mathcal{G}} \Sigma_p^d(\tau) \right) \setminus \Gamma_{\mathrm{D}},$$

wobei angenommen ist, dass Γ_{D} vom Gitter aufgelöst *wird, d.h.*

$$\Gamma_{\mathrm{D}} = \bigcup_{\substack{\tau \in \mathcal{G} \\ |\tau \cap \Gamma_{\mathrm{D}}| > 0}} (\tau \cap \Gamma).$$

Satz 6.10 *Eine Basis für den Raum $S^{p,0}_{\mathcal{G},\mathrm{D}}$ ist durch $b_{\mathbf{z}}$, $\mathbf{z} \in \Sigma_{p,\mathrm{D}}(\mathcal{G})$, gegeben, wobei die Lagrange-Basisfunktionen $b_{\mathbf{z}}$ implizit durch*

$$b_{\mathbf{z}} \in S^{p,0}_{\mathcal{G},\mathrm{D}} \quad \text{und} \quad b_{\mathbf{z}}(\mathbf{z}) = 1 \quad \text{und} \quad b_{\mathbf{z}}(\mathbf{y}) = 0 \quad \forall \mathbf{y} \in \Sigma_{p,\mathrm{D}}(\mathcal{G}) \setminus \{\mathbf{y}\} \tag{6.9}$$

gegeben sind. Sei $\omega_{\mathbf{z}} := \bigcup \{\tau \in \mathcal{G} \mid \mathbf{z} \in \tau\}$. Dann gilt explizit für $\tau \subset \omega_{\mathbf{z}}$

$$b_{\mathbf{z}}|_\tau \circ \chi_\tau = \hat{B}_{\hat{\mathbf{z}}} \quad \text{für } \hat{\mathbf{z}} = \chi_\tau^{-1}(\mathbf{z}) \tag{6.10}$$

und $b_{\mathbf{z}}(\mathbf{x}) = 0$ für alle $\mathbf{x} \in \Omega \setminus \omega_{\mathbf{z}}$. Für den Träger von $b_{\mathbf{z}}$ gilt

$$\operatorname{supp} b_{\mathbf{z}} = \omega_{\mathbf{z}}. \tag{6.11}$$

Beweis. Wir setzen $S = S^{p,0}_{\mathcal{G},\mathrm{D}}$ und $\tilde{S} = \operatorname{span}\{b_{\mathbf{z}} : \mathbf{z} \in \Sigma_{p,\mathrm{D}}(\mathcal{G})\}$. Per Definition gilt $b_{\mathbf{z}} \in S$, so dass $\tilde{S} \subset S$ folgt. Um die umgekehrte Relation $S \subset \tilde{S}$ zu zeigen, wählen wir $w \in S$ und

zeigen $w \in \tilde{S}$. Zunächst definieren wir für stetige Funktionen $q \in C^0\left(\overline{\Omega}\right)$ den Knoteninterpolationsoperator durch

$$I^p_{\mathrm{int,D}} q := \sum_{\mathbf{z} \in \Sigma_{p,\mathrm{D}}(\mathcal{G})} q\left(\mathbf{z}\right) b_{\mathbf{z}} \in \tilde{S}. \tag{6.12}$$

Es genügt daher zu zeigen, dass $\tilde{w} := I^p_{\mathrm{int,D}} w = w$ gilt. Sei $\tau \in \mathcal{G}$ und $\Sigma^d_p\left(\tau\right)$ die Menge aller Knotenpunkte auf τ. Dann gilt $\left(w - \tilde{w}\right)|_\tau \in \mathbb{P}^d_p$. Die Lagrange-Eigenschaft von $b_{\mathbf{z}}$ in (6.9) impliziert

$$\left(w - \tilde{w}\right)\left(\mathbf{z}\right) = 0 \quad \text{für alle } \mathbf{z} \in \Sigma^d_p\left(\tau\right) \setminus \Gamma_{\mathrm{D}}.$$

Wegen $w - \tilde{w} \in S^{p,0}_{\mathcal{G},\mathrm{D}}$ gilt

$$w\left(\mathbf{z}\right) = 0 = \tilde{w}\left(\mathbf{z}\right) \quad \text{für alle } \mathbf{z} \in \Sigma^d_p\left(\tau\right) \cap \Gamma_{\mathrm{D}}$$

so dass

$$\left(w - \tilde{w}\right)\left(\mathbf{z}\right) = 0 \quad \text{für alle } \mathbf{z} \in \Sigma^d_p\left(\tau\right).$$

Da $B_{\mathbf{z}}$, $\mathbf{z} \in \Sigma^d_p\left(\hat{\tau}\right)$, eine Basis für $\mathbb{P}_p$ bildet, folgt

$$\left(w - \tilde{w}\right)|_\tau = \sum_{\mathbf{z} \in \Sigma^d_p(\tau)} \left(w - \tilde{w}\right)\left(\mathbf{z}\right) B^\tau_{\mathbf{z}} = 0 \quad \text{und} \quad w|_\tau = \tilde{w}|_\tau.$$

Da $\tau \in \mathcal{G}$ beliebig war, folgt $w = \tilde{w}$ und $S = \tilde{S}$.

Die Eigenschaften (6.10) und (6.11) folgen analog. ■

Jedes $u \in S^{p,0}_{\mathcal{G},\mathrm{D}}$ lässt sich eindeutig schreiben in der Form

$$u = \sum_{\mathbf{z} \in \Sigma_{p,\mathrm{D}}(\mathcal{G})} u_{\mathbf{z}} b_{\mathbf{z}} \quad \text{mit} \quad u_{\mathbf{z}} = u\left(\mathbf{z}\right). \tag{6.13}$$

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

$$a\left(u_S, v\right) = \ell\left(v\right) \qquad \forall v \in S \tag{6.14}$$

äquivalent zu einem linearen Gleichungssystem mit

$$N := N_{\mathcal{G},p,\mathrm{D}} := \sharp \Sigma_{p,\mathrm{D}}\left(\mathcal{G}\right) = \dim S$$

Gleichungen und Unbekannten. Wir definieren dazu die Systemmatrix $\mathbf{A} = \left(a_{\mathbf{y},\mathbf{z}}\right)_{\mathbf{z},\mathbf{y} \in \Sigma_{p,\mathrm{D}}(\mathcal{G})} \in \mathbb{R}^{N \times N}$ und den rechte-Seite-Vektor $\mathbf{r} = \left(r_{\mathbf{z}}\right)_{\mathbf{z} \in \Sigma_{p,\mathrm{D}}(\mathcal{G})} \in \mathbb{R}^N$ durch

$$a_{\mathbf{y},\mathbf{z}} = a\left(b_{\mathbf{z}}, b_{\mathbf{y}}\right) \quad \text{und} \quad r_{\mathbf{z}} = \ell\left(b_{\mathbf{z}}\right) \qquad \forall \mathbf{z}, \mathbf{y} \in \Sigma_{p,\mathrm{D}}\left(\mathcal{G}\right).$$

Proposition 6.11

a. Die Systemmatrix $\mathbf{A}$ *für Problem (6.2) ist symmetrisch und positiv definit.*

b. Das lineare Gleichungssystem

$$\mathbf{A}\mathbf{u} = \mathbf{r} \tag{6.15}$$

mit $\mathbf{u} = \left(u_{\mathbf{z}}\right)_{\mathbf{z} \in \Sigma_{p,\mathrm{D}}(\mathcal{G})}$ *besitzt eine eindeutige Lösung. Die zugehörige Finite-Elemente-Funktion* $u_S = \sum_{\mathbf{z} \in \Sigma_{p,\mathrm{D}}(\mathcal{G})} u_{\mathbf{z}} b_{\mathbf{z}}$ *ist die eindeutige Lösung von (6.14). Ist umgekehrt* u_S *die eindeutige Lösung von (6.14), so erfüllt der durch (6.13) definierte Koeffizientenvektor* $\mathbf{u}$ *das Gleichungssystem (6.15).*

c. *Die Matrix ist schwach besetzt, d.h. die Anzahl der Nicht-Nullelemente pro Zeile „$\mathbf{z}$“ ist beschränkt durch*

$$\text{„}1+\binom{p+d}{d}\sharp\left\{\tau\in\mathcal{G}\ \textit{mit}\ \mathbf{z}\in\tau\ \right\}\text{“}.$$

Diese Anzahl ist unabhängig von $\sharp\mathcal{G}$, hängt aber von der Formregularität der Simplizes ab und vom Polynomgrad wie $\mathcal{O}\left(p^d\right)$.

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 $\mathbf{u}=(u_{\mathbf{z}})_{\mathbf{z}\in\Sigma_{p,\mathrm{D}}(\mathcal{G})}$ assoziieren wir die Finite-Elemente-Funktion u wie in (6.13). Da $(b_z)_{\mathbf{z}\in\Sigma_{p,\mathrm{D}}(\mathcal{G})}$ eine Basis von $S^{p,0}_{\mathcal{G},\mathrm{D}}$ ist, gilt die Äquivalenz $\mathbf{u}=\mathbf{0}\iff u=0$.

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

$$\begin{aligned}\langle\mathbf{u},\mathbf{Au}\rangle&=\sum_{\mathbf{z},\mathbf{y}\in\Sigma_{p,\mathrm{D}}(\mathcal{G})}u_{\mathbf{z}}u_{\mathbf{y}}a_{\mathbf{z},\mathbf{y}}=\sum_{\mathbf{z},\mathbf{y}\in\Sigma_{p,\mathrm{D}}(\mathcal{G})}u_{\mathbf{z}}u_{\mathbf{y}}a\left(b_{\mathbf{z}},b_{\mathbf{y}}\right)\\&=a\left(\sum_{\mathbf{z}\in\Sigma_{p,\mathrm{D}}(\mathcal{G})}u_{\mathbf{z}}b_{\mathbf{z}},\sum_{\mathbf{y}\in\Sigma_{p,\mathrm{D}}(\mathcal{G})}u_{\mathbf{y}}b_{\mathbf{y}}\right)=a(u,u)\geq C\left\|u\right\|^2_{H^1(\Omega)}\end{aligned}$$

und wegen der oben erklärten Äquivalenz die positive Definitheit

$$\langle\mathbf{u},\mathbf{Au}\rangle>0\qquad\forall\mathbf{u}=(u_{\mathbf{z}})_{\mathbf{z}\in\Sigma_{p,\mathrm{D}}(\mathcal{G})}\quad\text{mit}\quad\mathbf{u}\neq\mathbf{0}.$$

@ b: Das lineare Gleichungssystem (6.15) ist die Basisdarstellung der Variationsaufgabe (6.14). Daher folgt die Aussage aus elementaren Sätzen der linearen Algebra.

@ c: Sei $\mathbf{z}\in\Sigma_{p,\mathrm{D}}(\mathcal{G})$ ein Knotenpunkt. Der Träger von $b_{\mathbf{z}}$ (und damit auch von $\nabla b_{\mathbf{z}}$) ist durch $\omega_{\mathbf{z}}$ (cf. Satz 6.10) gegeben. Das Matrixelement $a_{\mathbf{y},\mathbf{z}}$ ist also von Null verschieden, falls die $\omega_{\mathbf{z}}\cap\omega_{\mathbf{y}}$ positives $d-$dimensionales Mass besitzt. Dies ist aber genau dann der Fall, falls $\mathbf{z}$ und $\mathbf{y}$ in einem gemeinsamen Simplex $\tau\in\mathcal{G}$ liegen. ■

Für die Fehlerabschätzung – genauer die Abschätzung des Bestapproximationsfehlers

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

verwenden wir den Knoteninterpolationsoperator aus (6.12). Man beachte, dass gilt

$$I^p_{\mathcal{G},\mathrm{D}}:\left(H^1_{\mathrm{D}}(\Omega)\cap C^0\left(\overline{\Omega}\right)\right)\to S^{p,0}_{\mathcal{G},\mathrm{D}}.$$

Die Abschätzung des Interpolationsfehlers geschieht zunächst auf dem Referenzelement. Dazu definieren wir $\hat{\Pi}_k:C_0\left(\widehat{\tau}\right)\to\mathbb{P}_p$ durch

$$\left(\hat{\Pi}_p u\right)(\mathbf{z})=u(\mathbf{z})\qquad\forall\mathbf{z}\in\Sigma_p.\tag{6.16}$$

Man beachte, dass für jedes $u\in C^0\left(\overline{\Omega}\right)$ und jedes $\tau\in\mathcal{G}$ die Identität gilt:

$$\left(I^p_{\mathcal{G},\mathrm{D}}u\right)\big|_\tau=\left(\hat{\Pi}_p\hat{u}\right)\circ\chi_\tau^{-1}\quad\text{mit}\quad\hat{u}=u|_\tau\circ\chi_\tau.$$

Satz 6.12 *Sei $p \in \mathbb{N}$ mit $p+1 > d/2$. Dann wird durch*

$$[u]_{p+1} := |u|_{p+1} + \sum_{\mathbf{z} \in \hat{\Sigma}_p^d} |u(\mathbf{z})|$$

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

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

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

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

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

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

$$\|u_n\|_{H^{p+1}(\hat{\tau})} = 1 \qquad \forall n \in \mathbb{N} \tag{6.17}$$

und

$$\lim_{n\to\infty} [u_n]_{p+1} = 0. \tag{6.18}$$

Wegen Bemerkung 2.22 und Satz 2.23 gibt es eine Teilfolge $(u_{n_m})_{m\in\mathbb{N}}$ und ein $u \in H^p(\hat{\tau})$ mit

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

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

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

gilt, so dass $u \in H^{p+1}(\hat{\tau})$ folgt. Wegen (6.18) ist insbesondere

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

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

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

Wegen $|u|_{H^{p+1}(\hat{\tau})} = 0$ ist $u \in \mathbb{P}_p^d$. Wegen Satz 2.23 gilt

$$u(\mathbf{z}) = \lim_{m\to\infty} u_{n_m}(\mathbf{z}) \qquad \forall \mathbf{z} \in \hat{\Sigma}_p^d.$$

Hieraus und aus (6.18) folgt aber

$$u(\mathbf{z}) = 0 \qquad \forall \mathbf{z} \in \hat{\Sigma}_p^d.$$

Wegen Bemerkung 6.6c ist also $u = 0$ im Widerspruch zu (6.17). ■

Satz 6.13 *Sei $k \in \mathbb{N}$. Dann gibt es eine Konstante c, die von k abhängt, mit*

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

Beweis. Aus Satz 6.12 folgt für beliebiges $u \in H^{k+1}(\hat{\tau})$

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

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 (vgl. Def. 6.3)

$$\chi_\tau(\hat{\mathbf{x}}) = \mathbf{x}_0^\tau + \mathbf{m}_\tau \hat{\mathbf{x}} \tag{6.19}$$

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

Lemma 6.14 *Für jedes $\tau \in \mathcal{G}$ gilt für die Matrix $\mathbf{m}_\tau$ aus (6.19)*

$$\|\mathbf{m}_\tau\| \leq \frac{d + \sqrt{d}}{2} h_\tau, \qquad \left\|\mathbf{m}_\tau^{-1}\right\| \leq \frac{\sqrt{2}\tilde{c}_\mathcal{G}}{h_\tau},$$

mit einer Konstanten $\tilde{c}_\mathcal{G}$, welche lediglich (linear) von der Formregularitätskonstanten $c_\mathcal{G}$ (siehe (6.5)) abhängt.

Beweis. Bezeichne mit $h_{\hat{\tau}}$ den Durchmesser von $\hat{\tau}_d$ und $\rho_{\hat{\tau},d}$ den Durchmesser des grössten, in $\hat{\tau}_d$ eingeschriebenen Balles. Man rechnet leicht nach, dass $h_{\hat{\tau}} = \sqrt{2}$ gilt. Für $\rho_{\hat{\tau},d}$ verwenden wir die Formel

$$\rho_{\hat{\tau},d} = 2\frac{d\left|\hat{\tau}_d\right|}{\sum_{f \in \mathcal{C}_{d-1}(\hat{\tau}_d)} |f|} = 2\frac{d\left|\hat{\tau}_d\right|}{d\left|\hat{\tau}_{d-1}\right| + |f_d|}$$

mit dem Volumen $|f_d|$ des grössten $(d-1)$-dimensionalen Oberflächensimplex an $\hat{\tau}_d$. Wir verwenden $|\hat{\tau}_d| = 1/d!$ und $|f_d| := \frac{\sqrt{d}}{(d-1)!}$ und erhalten

$$\rho_{\hat{\tau},d} = \frac{2}{d + \sqrt{d}}.$$

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 $\chi_{\hat{\tau}} : \hat{\tau} \to \tau$ bijektiv ist, folgt

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

Also ist

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

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

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

Wir verwenden nochmals die Formel für den Inkreisdurchmesser und erhalten mit der Formregularität

$$\rho_\tau = 2\frac{d\,|\tau|}{\sum\limits_{f\in\mathcal{C}_{d-1}(\tau)}|f|} \geq \frac{2d}{c_\mathcal{G}}\frac{h_\tau^d}{(d+1)\max\limits_{f\in\mathcal{C}_{d-1}(\tau_d)}|f|} \geq \frac{2d}{c_\mathcal{G}}\frac{h_\tau^d}{(d+1)\,\frac{\sqrt{d}}{(d-1)!}\left(\frac{h_\tau}{\sqrt{2}}\right)^{d-1}} \geq \tilde{c}_\mathcal{G}^{-1}h_\tau$$

mit

$$\tilde{c}_\mathcal{G} := \frac{\sqrt{d}\,(d+1)}{d!2^{\frac{d+1}{2}}}c_\mathcal{G}.$$

■

Satz 6.15 *Sei $p\in\mathbb{N}$. Dann gelten für alle $u\in H^{p+1}(\Omega)$ die Interpolationsfehlerabschätzungen*

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

Die Konstanten c_1 und c_2 hängen von Ω, p und der Konstanten $c_\mathcal{G}$ in der Regularitätsbedingung an $\mathcal{G}$ ab.

Beweis. Sei $\tau\in\mathcal{G}$ und $\chi_\tau(\hat{\mathbf{x}}) = \mathbf{x}_0^\tau + \mathbf{m}_\tau\hat{\mathbf{x}}$ eine affine Transformation des Referenzelements $\hat{\tau}$ auf τ. Die Transformationsformel für Integrale liefert[8]

$$\begin{aligned}\left\|u-I_\mathcal{G}^p u\right\|_{L^2(\tau)} &= |\det\mathbf{m}_\tau|^{1/2}\left\|\left(u-I_\mathcal{G}^p u\right)\circ\chi_\tau\right\|_{L^2(\hat{\tau})},\\ \left|u-I_\mathcal{G}^p u\right|_{H^1(\tau)} &\leq |\det\mathbf{m}_\tau|^{1/2}\left\|\mathbf{m}_\tau^{-1}\right\|\left|\left(u-I_\mathcal{G}^p u\right)\circ\chi_\tau\right|_{H^1(\hat{\tau})},\\ |u\circ\chi_\tau|_{H^{p+1}(\hat{\tau})} &\leq C_{p,d}\,|\det\mathbf{m}_\tau|^{-1/2}\,\|\mathbf{m}_\tau\|^{p+1}\,|u|_{H^{p+1}(\tau)}.\end{aligned} \tag{6.21}$$

[8]Für (6.21), setzen wir $\hat{v} = v\circ\chi_\tau$ und verwenden die Kettenregel für $\boldsymbol{\alpha}\in\mathbb{N}_0^d$ mit $|\boldsymbol{\alpha}| = k+1$

$$\left(\left(\partial^{\boldsymbol{\alpha}}\hat{v}\right)\circ\chi_\tau^{-1}\right) = \left(\mathbf{m}_\tau^\intercal\nabla\right)^{\boldsymbol{\alpha}}v.$$

Man kann sich einfach klar machen, dass sich die rechte Seite schreiben lässt als

$$\sum_{\boldsymbol{\alpha}\in\mathbb{N}_0^d} q_{\boldsymbol{\alpha}}\partial^{\boldsymbol{\alpha}}v$$

mit reellen Koeffizienten $q_{\boldsymbol{\alpha}}$. Für einen Matrixmultiindex $\boldsymbol{\mu}\in\mathbb{N}_0^{d\times d}$, setzen wir $|\boldsymbol{\mu}| = \sum_{i,j=1}^d \mu_{i,j}$ und $\mathbf{m}^{\boldsymbol{\mu}} = \prod\limits_{i,j=1}^d m_{i,j}^{\mu_{i,j}}$. Dann besitzen die Koeffizienten $q_{\boldsymbol{\alpha}}$ die Form

$$q_{\boldsymbol{\alpha}} = \sum_{\substack{\boldsymbol{\mu}\in\mathbb{N}_0^{d\times d}\\ |\boldsymbol{\mu}|=k+1}} c_{\boldsymbol{\mu}}\mathbf{m}^{\boldsymbol{\mu}}$$

mit Koeffizienten $c_{\boldsymbol{\mu}}$ die nicht von $\mathbf{m}$ abhängen. Um den Betrag punktweise abzuschätzen verwenden wir

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

und erhalten mit $|\boldsymbol{\alpha}| = k+1$

$$\left|\left(\left(\partial^\alpha\hat{v}\right)\circ F_\tau^{-1}\right)(\mathbf{x})\right|^2 = \left|\left(\mathbf{m}_\tau^\intercal\nabla\right)^\alpha v(\mathbf{x})\right|^2 \leq C\,\|\mathbf{m}_\tau\|^{2(k+1)}\sum_{\substack{\boldsymbol{\mu}\in\mathbb{N}_0^d\\ |\boldsymbol{\mu}|=k+1}}\left|\partial^{\boldsymbol{\mu}}v(\mathbf{x})\right|^2. \tag{6.20}$$

Summation über alle $|\alpha| = k+1$ und Integration liefert (6.21).

Aus Lemma 6.14 folgt

$$\|\mathbf{m}_\tau\| \le h_\tau/\rho_{\hat\tau}, \qquad \left\|\mathbf{m}_\tau^{-1}\right\| \le h_{\hat\tau}/\rho_\tau.$$

Aus diesen Abschätzungen und Satz 6.13 ergibt sich wegen $I_{\mathcal{G}}^p u \circ \chi_\tau = \hat\Pi_\tau (u \circ \chi_\tau)$

$$\left\|u - I_{\mathcal{G}}^p u\right\|_{L^2(\tau)} \le c\,(h_\tau/\rho_{\hat\tau})^{p+1}\,|u|_{H^{p+1}(\tau)}\,, \tag{6.22a}$$

$$\left|u - I_{\mathcal{G}}^p u\right|_{H^1(\tau)} \le c h_{\hat\tau}/\rho_\tau\,(h_\tau/\rho_{\hat\tau})^{p+1}\,|u|_{H^{p+1}(\tau)} = c h_{\hat\tau}\rho_{\hat\tau}^{-p-1}\left(\frac{h_\tau}{\rho_\tau}\right) h_\tau^p\,|u|_{H^{p+1}(\tau)}\,. \tag{6.22b}$$

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

Theorem 6.16 *Wir betrachten das Problem (6.1) mit der Bilinearform a und Linearform ℓ aus (6.2). Die Daten erfüllen die Voraussetzungen (6.3) und (6.4). Dann besitzt das Finite-Elemente-Galerkin-Verfahren:*

$$\textit{Finde } u_S \in S^{p,0}_{\mathcal{G},\mathrm{D}} \qquad a(u_S, v) = \ell(v) \qquad \forall v \in S^{p,0}_{\mathcal{G},\mathrm{D}}$$

eine eindeutige Lösung, die die Fehlerabschätzungen

$$\|u - u_S\|_{H^1(\Omega)} \le \frac{A}{\alpha} \inf_{v \in S^{p,0}_{\mathcal{G},\mathrm{D}}} \|u - v\|_{H^1(\Omega)}$$

$$\|u - u_S\|_{L^2(\Omega)} \le \frac{A^2}{\alpha} \inf_{v \in S^{p,0}_{\mathcal{G},\mathrm{D}}} \|u - v\|_{H^1(\Omega)} \sup_{\varphi \in L^2\setminus\{0\}} \inf_{v \in S^{p,0}_{\mathcal{G},\mathrm{D}}} \frac{\|u_\varphi - v\|_{H^1(\Omega)}}{\|\varphi\|_{L^2(\Omega)}}$$

erfüllt. Hier bezeichnet $u_\varphi \in H^1_{\mathrm{D}}(\Omega)$ die eindeutige Lösung des Problems

$$a(v, u_\varphi) = (\varphi, v)_{L^2(\Omega)} \qquad \forall v \in H^1_{\mathrm{D}}(\Omega)\,. \tag{6.23}$$

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

$$\|u - u_S\|_{H^1(\Omega)} \le C h^p\,|u|_{H^{p+1}(\Omega)}\,.$$

Ist das Problem darüber hinaus $H^2(\Omega)$-regulär[9]*, erhalten wir*

$$\|u - u_S\|_{L^2(\Omega)} \le C h^{p+1}\,|u|_{H^{p+1}(\Omega)}\,.$$

Für die Bewertung von Verfahren spielen verschiedene Kriterien ein 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 $\operatorname{div}(\mathbf{A}\operatorname{grad} u)$. 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.

[9] $H^2(\Omega)$-regulär bedeutet, dass die Lösung von (6.23) für *jede* rechte Seite $\varphi \in L^2(\Omega)$ in $H^2(\Omega)$ ist und $\|u\|_{H^2(\Omega)} \le C\,\|f\|_{L^2(\Omega)}$ gilt (vgl. Satz 4.5).

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.

Im Rahmen von Satz 3.1 ist jetzt

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

Der diskrete Raum S ist wieder durch

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

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

$$a_S(u,v) := a(u,v) + \sum_{\tau \in \mathcal{G}} \delta_\tau \int_\tau \langle \mathbf{b}, \nabla u \rangle \langle \mathbf{b}, \nabla v \rangle \tag{6.24a}$$

$$\ell_S(v) := \ell(v) + \sum_{\tau \in \mathcal{G}} \delta_\tau \int_\tau f \langle \mathbf{b}, \nabla v \rangle \tag{6.24b}$$

$$|u|_{1,S} := \left\{ \varepsilon |u|^2_{H^1(\Omega)} + \sum_{\tau \in \mathcal{G}} \delta_\tau \|\langle \mathbf{b}, \nabla u \rangle\|^2_{L^2(\tau)} \right\}^{1/2}. \tag{6.24c}$$

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

$$a_S(u_S, v) = \ell_S(v) \qquad \forall v \in S. \tag{6.25}$$

Dabei sind die δ_τ nicht-negative Parameter, die wir später fixieren werden. Die δ_τ 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

$$\begin{array}{cc} -\varepsilon\Delta u + \partial_1 u = \lambda u & \text{in } \Omega = (0,1)^2, \\ u = 0 & \text{auf } \Gamma = \partial\Omega. \end{array} \tag{6.26}$$

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

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

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 (6.24) 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

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

in a_S aufzunehmen.

Satz 6.17 *Die Bilinearform a_S ist koerziv bzgl. der $|\cdot|_{1,S}$-Norm, d.h.*

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

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

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

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

Bemerkung 6.18

1. *Aus Satz 6.17 folgt, dass das diskrete Problem (6.25) eindeutig lösbar ist.*
2. *Satz 6.17 benötigt ausser der Annahme $\delta_\tau \geq 0$ keine weiteren Voraussetzungen an die δ_τ.*
3. *Im allgemeinen Fall bleibt Satz 6.17 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 (6.25) benötigen wir eine Abschätzung der L^2-Norm des Interpolationsfehlers, wie wir sie bereits in (6.22) bewiesen haben:

1. Für alle $v \in H^2(\hat{\tau})$ gilt
$$\left\| v - \hat{\Pi}_1 v \right\|_{L^2(\hat{\tau})} \leq C |v|_{H^2(\hat{\tau})}.$$
2. Für alle $\tau \in \mathcal{G}$ und alle $v \in H^2(\tau)$ gilt
$$\|v - I_{\mathcal{G}} v\|_{L^2(\tau)} \leq C h_\tau^2 |v|_{H^2(\tau)}.$$

Satz 6.19 *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 (6.25). Es sei $u \in H^2(\Omega)$. Dann gilt die Fehlerabschätzung*

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

Dabei ist $c_{\mathcal{G}}$ wie in (6.5) 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 (6.22) 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 6.17 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, ℓ_S und (6.25)

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

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

wir weiterhin

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

Aus den letzten beiden Abschätzungen folgt

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

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

Bemerkung 6.20 *Es gilt $\delta_\tau \sim h_\tau$, wenn $\varepsilon \ll h_\tau$ ist. In diesem Fall liefert Satz 6.19 eine optimale Fehlerabschätzung der Form*

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

in Stromrichtung. Im Fall $h_\tau \ll \varepsilon$, in dem auch die Diskretisierung (6.14) gut funktioniert, ist $\delta_\tau \sim h_\tau^2\varepsilon^{-1}$, und Satz 6.19 liefert eine Fehlerabschätzung der Form

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

die vergleichbar ist zu derjenigen von Satz 6.16.

Die folgenden beiden Beispiele illustrieren das verbesserte Konvergenzverhalten von SUPG für konvektionsdominierte Probleme verglichen zur Standard-Finite-Elemente-Methode.[10]

[10]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.

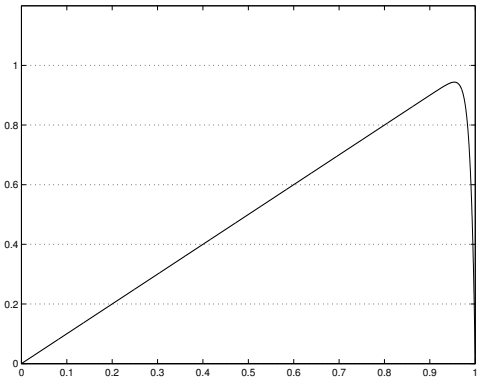


Abbildung 2: Die exakte Lösung für die eindimensionale Konvektions-Diffusionsgleichung. Der kritische Bereich ist die Rand-Grenzschicht der Breite $O(\varepsilon)$ bei $x = 1$.

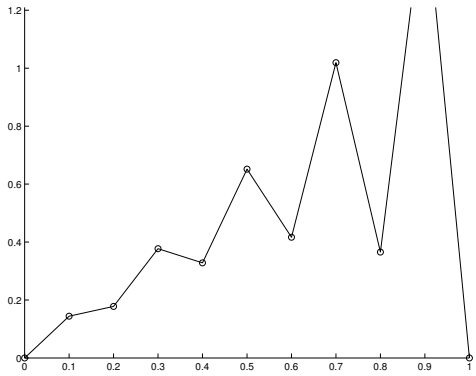


Abbildung 3: Numerische Lösung mit linearen finiten Elementen auf einem äquidistanten Gitter.

Beispiel 6.21 *Wir betrachten zunächst ein eindimensionales Modellproblem. Sei* $\Omega = (0,1)$ *und* $0 < \varepsilon \ll 1$ *ein Parameter. Aufgabe:*

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

Die exakte Lösung

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

ist in Abb. 2 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 3). Abbildung 4 zeigt den Fehlerverlauf bei kleiner werdender Schrittweite. Man sieht, dass die asymptotische Konvergenzrate für kleiner werdende Diffusion immer später einstellt.*

Beispiel 6.22 *Wir betrachten nun ein zweidimensionales Modellproblem. Sei* $\Omega = (0,1)^2$ *und* $0 < \varepsilon \ll 1$ *wieder ein Parameter. Aufgabe:*

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

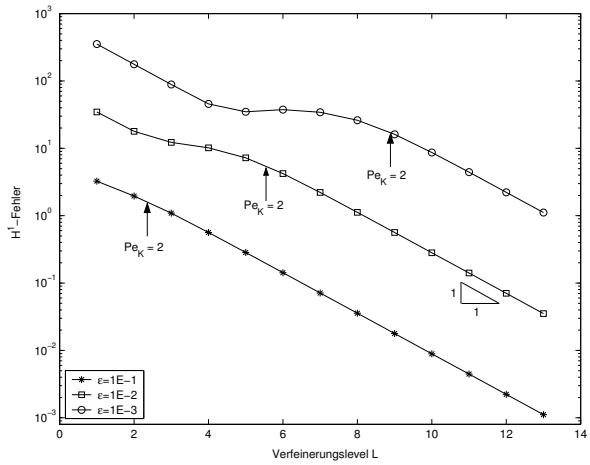


Abbildung 4: H^1-Fehler für verschiedene, moderate Parameter.

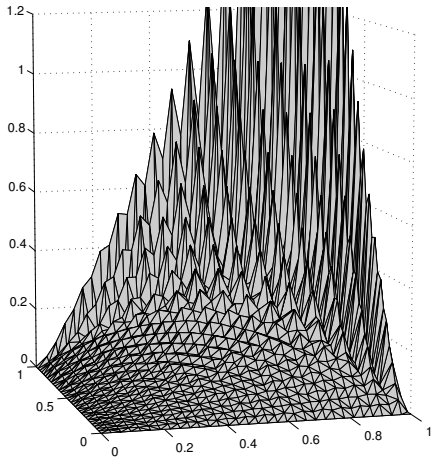


Abbildung 5: Numerische Lösung mit linearen finiten Elementen auf einem regelmässigen Dreiecksgitter.

Wir setzen $\mathbf{b} = (1,1)^T$ *und verwenden die Funktion* η *aus Beispiel 6.21. Die exakte Lösung für die rechte Seite* $f(x,y) = \eta(x) + \eta(y)$ *ist gegeben durch*

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

Diskretisiert man dieses Problem mit linearen finiten Elementen auf einem regelmässigen Dreiecksgitter zeigen sich wiederum unphysikalische Oszillationen (siehe Abbildung 5) Die Oszillationen werden noch deutlicher, wenn man die Funktion entlang der Diagonalen $\overline{\binom{0}{0}, \binom{1}{0}}$ *zeichnet (siehe Abbildung 6). Qualitativ zeigt der Fehler das gleiche Verhalten wie im eindimensionalen Fall. Für kleiner werdende Diffusion stellt sich die asymptotische Konvergenzrate immer später ein.*

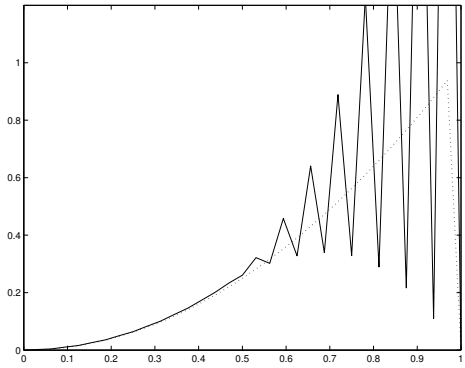
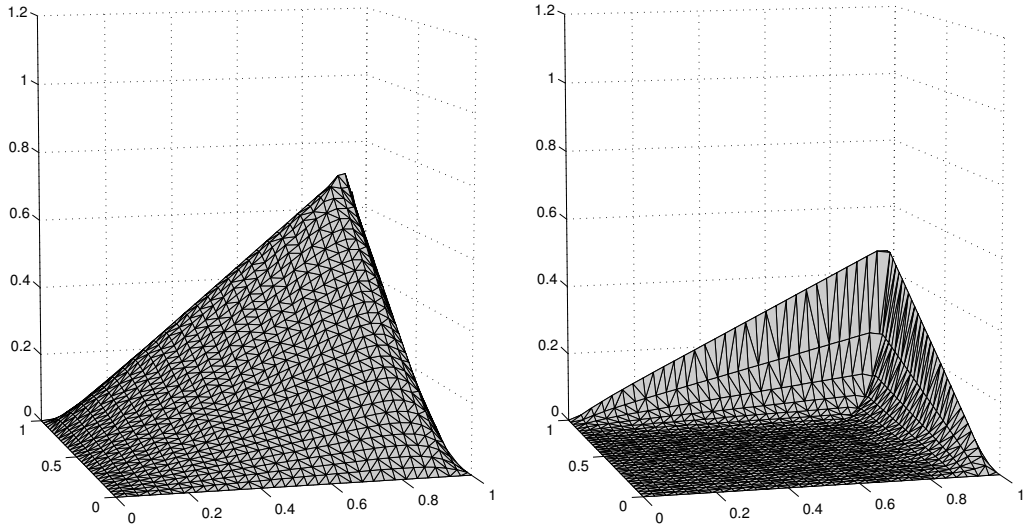


Abbildung 6: Vergleich der exakten und der numerischen Lösung mit linearen finiten Elementen auf einem regelmässigen Dreiecksgitter.

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 $\overline{\binom{0}{0}, \binom{1}{0}}$ *deutlich zeigt (siehe Abb. 7)*

7 Implementierung

In diesem Kapitel gehen wir auf die Implementierung der Methoden der letzten Kapitel und die dafür benötigten Datenstrukturen ein. Wir beschränken uns auf den Fall, dass $\mathcal{G}$ eine *Triangulierung* des polygonalen Gebietes $\Omega \subset \mathbb{R}^2$ ist. 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

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

und die Koordinaten der Ecken werden im Feld

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

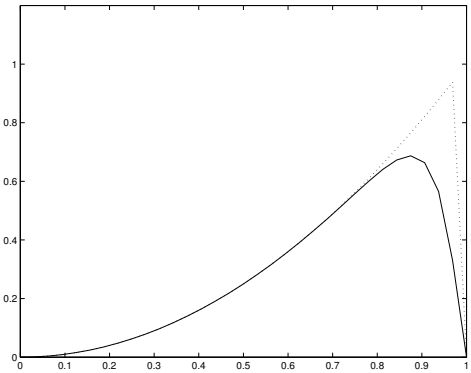


Abbildung 7: SUPG-Lösung und exakte Lösung über der Diagonalen. Es treten keine unphysikalischen Oszillationen auf und das qualitative Verhalten wird korrekt wiedergegeben.

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

$$\begin{aligned}\Delta = \text{record}\\ &\text{loc2glob} : \text{array}\,[1..3]\ \text{of}\ \texttt{integer}\\ &\text{neighbor} : \text{array}\,[1..3]\ \text{of}\ \texttt{integer}\\ &\text{mark} : \text{array}\,[1..2]\ \text{of}\ \text{int3D}\\ &p : \texttt{integer;}\\ &\text{loc2globdof} : \text{array}\left[1..\tfrac{(p+1)(p+2)}{2}\right]\ \text{of}\ \texttt{integer}\\ \text{end},\end{aligned} \tag{7.1}$$

so dass die Dreiecke im Feld

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

abgespeichert werden. Die einzelnen Grössen sind wie folgt definiert:

1. $\tau(i)\,.\,\text{loc2glob}\,(j) = k$: Die j-te Ecke des Dreiecks i hat die globale Nummer k und die Koordinaten $\mathbf{x}\,(k)$

2. $\tau(i)\,.\,\text{neighbor}\,(j) = k$: Das Nachbardreieck von $\tau(i)$, welches die j-te Kante[11] mit $\tau(i)$ gemeinsam hat, ist $\tau(k)$. Falls die j-te Kante von $\tau(i)$ eine Randkante ist wird $\tau(i)\,.\,\text{neighbor}\,(j) = 0$ gesetzt[12].

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

[11] Die j-te Kante eines Dreiecks ist diejenige, welche der j-ten Ecke gegenüber liegt.

[12] Für Kanten auf dem Rand von Γ kann neighbor (j) noch verfeinert genutzt werden, um den Typ des Randes, (Dirichlet/Neumann-Kante oder/und die gegebenen Randdaten) zu beschreiben.

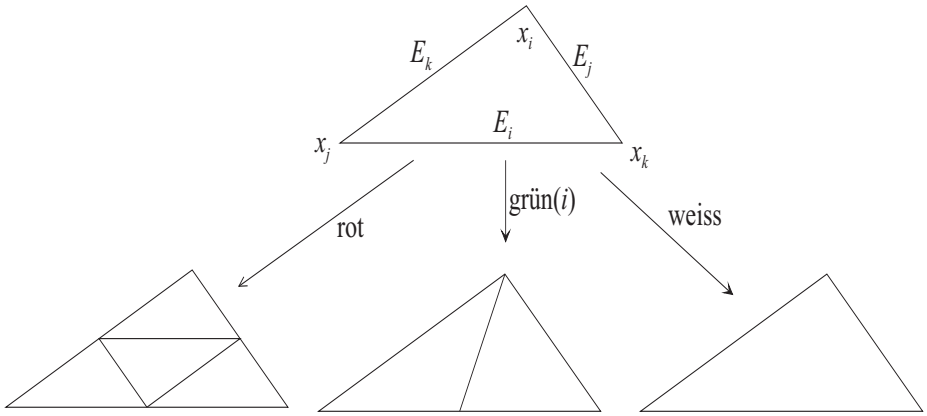


Abbildung 8: Verschiedene Verfeinerungsmuster für Dreiecke: $(4,4,4)$: Zerlegung in vier kongruente Dreiecke, (i,j,k): Verbindung des i-ten Eckpunkts mit der gegenüberliegenden Kantenmitte $\mathbf{m}_i$ und (für $j \neq 0$) Verbindung $\overline{\mathbf{m}_i, \mathbf{m}_j}$ und (für $k \neq 0$) Verbindung $\overline{\mathbf{m}_i, \mathbf{m}_k}$. Für $(0,0,0)$: keine Verfeinerung.

mehrere Gitter $(\mathcal{G}_\ell)_{\ell=0}^L$ und geometrische Abhängigkeiten zwischen diesen Gittern verwaltet werden. Für Dreiecke existieren verschiedene Verfeinerungsmuster, die in Abbildung 8 dargestellt sind.

Wir definieren die Grösse mark in (7.1) wie folgt

1. $\tau(k)\,.\,\mathrm{mark}\,(1) = (j,\ell)$. Das Dreieck $\tau(k)$ ist durch Unterteilung des Dreiecks τ_j der vorigen Stufe entstanden und das Dreieck $\tau(\ell)$ ist ein „Kind“ von $\tau(k)$ auf der feineren Stufe.

2. $\tau(k)\,.\,\mathrm{mark}\,(2) = (\ell,m,n)$. Wir unterscheiden die Fälle[13]

   (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 $\ell_{\mathrm{glob}} := \tau(k)\,.\,\mathrm{loc2glob}\,(\ell)$. Das Dreieck $\tau(k)$ wird durch die Strecke $\overline{\mathbf{x}\,(\ell_{\mathrm{glob}})\,\mathbf{m}_{k,\ell}}$ unterteilt.

   (c) $(\ell,m,0)$, $1 \leq \ell,m \leq 3$, $\ell \neq m$. Wir setzen $\ell_{\mathrm{glob}} := \tau(k)\,.\,\mathrm{loc2glob}\,(\ell)$. Das Dreieck $\tau(k)$ wird unterteilt, indem die Strecken $\overline{\mathbf{x}\,(\ell_{\mathrm{glob}})\,\mathbf{m}_{k,\ell}}$ und $\overline{\mathbf{m}_{k,\ell}\mathbf{m}_{k,m}}$ eingefügt werden.

   (d) (ℓ,m,n), $1 \leq \ell,m,n \leq 3$, mit paarweise verschiedenen ℓ,m,n. Wir setzen $\ell_{\mathrm{glob}} := \tau(k)\,.\,\mathrm{loc2glob}\,(\ell)$ und zerteilen $\tau(k)$ durch Einfügen der Strecken $\overline{\mathbf{x}\,(\ell_{\mathrm{glob}})\,\mathbf{m}_{k,\ell}}$ 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 gröbste Triangulierung kann auf zwei Weisen erzeugt werden. Bei der ersten Möglichkeit gibt man die Daten für die Dreiecke und Eckpunkte des gröbsten Gitters von Hand ein.

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

Nach Eingabe der gröbsten Triangulierung werden weitere Triangulierungen durch gleichmässige oder adaptive Verfeinerung erzeugt.

Bei der zweiten Variante erzeugt man die gröbste 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:

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 Ω 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:

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

Man beachte, dass zur Berechnung von c_τ 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 τ(k) ∈ 𝒢 do
      for j := 1 to 3 do (*Schleife über alle Nachbarn von τ(k)*)
        if "τ(k).mark(2) ist derart, dass m_{k,j} beim Verfeinern nicht
        generiert werden würde" then begin
          ℓ := τ(k).neighbor(j); (*τ(ℓ) ist der Nachbar von τ(k) über E_{k,j}*)
          berechne m ∈ {1,2,3} mit k = τ(ℓ).neighbor(m);
          if "τ(ℓ).mark(2) ist derart, dass m_{ℓ,m} beim Verfeinern
          generiert werden würde" then
            ändere τ(k).mark(2) in eine "nächsthöhere" Strategie,
            die m_{k,j} als Verfeinerungspunkt enthält;
        end
    if "Markierung von 𝒢 enthält noch hängende Knoten" then goto 1
end
```

Für das Aufstellen des linearen Gleichungssytems 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 ℓ_τ und a_τ die Einschränkung des Funktionals ℓ bzw. der Bilinearform a auf das Element τ :

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

Hier sind wir von Dirichlet-Randbedingungen ausgegangen; im Fall inhomogener Neumann-Randbedingungen muss das Funktional ℓ 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 τ_k berechnet. Eine Möglichkeit ist beispielsweise die Schwerpunktsregel

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

oder die Kantenmittelpunktsregel

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

Übungsaufgabe 7.1 *Berechnen Sie den Exaktheitsgrad der Schwerpunktsregel und der Kantenmittelpunktsregel.*

Die numerische Quadratur lässt sich als Störung der Gesamtdiskretisierung interpretieren, genauso wie der Fehler durch die Approximation eines gekrümmten Gebiets durch einen Polyeder/ein Polygongebiet oder auch durch iteratives, näherungsweises Lösen des entstehenden linearen Gleichungssystems. Die Genauigkeit dieses Störungen sollte so gewählt werden, dass die Konvergenzrate des ungestörten Verfahrens nicht verschlechtert wird (sondern höchstens die Konstante C in der Fehlerabschätzung). Das bedeutet, dass für Finite-Elemente-Verfahren hoher Ordnung (Polynomgrad p) die Genauigkeitsanforderungen an die Störungen steigen. Wir geben im Folgenden eine *Familie* von Quadraturverfahren für Dreiecke an, die von der eindimensionalen Gauss-Quadratur abgeleitet sind und daher stabil ist (positive Quadraturgewichte) und exponentiell in der Ordnung konvergieren. Sei dazu Q^n die Gauss-Quadratur (mit Gewichtsfunktion 1) für das Intervall $(0,1)$

$$\int_0^1 g \approx Q^n(g) = \sum_{i=1}^{n} \omega_{n,i} g(\xi_{n,i}).$$

Diese sollten für $n = 1, 2, \ldots, 10$ abgespeichert vorliegen (Gewichte und Stützstellen). Sei nun τ ein Dreieck mit Eckpunkten x_i^τ, $i = 0, 1, 2$, und Elementabbildung

$$\chi_\tau : \hat{\tau} \to \tau, \quad \chi_\tau(\hat{x}) = x_0^\tau + \mathbf{m}_\tau \hat{x}$$

mit der Matrix $\mathbf{m}_\tau$, deren Spaltenvektoren durch $x_1^\tau - x_0^\tau$ und $x_2^\tau - x_0^\tau$ gegeben sind. Wegen

$$\int_\tau g = 2|\tau| \int_{\hat{\tau}} \hat{g} \quad \text{mit} \quad \hat{g} = g \circ \chi_\tau$$

genügt es, eine Quadraturformel für das Referenzdreieck zu definieren. Dazu verwendet man, dass die Abbildung

$$q : (0,1)^2 \to \hat{\tau}, \quad q(\xi,\eta) = \begin{pmatrix} \xi(1-\eta) \\ \xi\eta \end{pmatrix}$$

surjektiv ist und die Determinante der Jacobimatrix gleich ξ ist. Also gilt

$$\int_{\hat{\tau}} \hat{g} = \int_0^1 \int_0^1 \xi \hat{g} \circ q(\xi,\eta)\, d\xi d\eta.$$

Approximiert man die Integrale $\int_0^1 \ldots$ jeweils durch die Gauss-Quadratur Q^n erhält man

$$\int_{\hat{\tau}} \hat{g} \approx Q_{\hat{\tau}}^n(\hat{g}) := \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{n,i}\omega_{n,k}\xi_{n,i} \hat{g}(\zeta_{n,i,j}, \lambda_{n,i,j})$$

mit

$$\zeta_{n,i,j} := q_1(\xi_{n,i}, \xi_{n,j}) = \xi_{n,i}(1 - \xi_{n,j}) \quad \text{und} \quad \lambda_{n,i,j} := \xi_{n,i}\xi_{n,j}.$$

Die Gewichte $\omega_{n,i}\omega_{n,j}$ sind positiv, so dass die Quadraturformel für $Q_{\hat{\tau}}^n$ stabil ist und für $n \to \infty$ für hinreichend glatte Funktionen exponentiell konvergiert.

Bemerkung 7.2 *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 $C^2(\overline{\tau})$.*

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**
    $n := \tau(k).\text{loc2globdof}(i)$;
    $c := \text{dof}(n).\text{sys}.\text{count}$;
    for $j = 1$ **to** $\binom{p_{\text{loc}}+2}{2}$ **do begin**
      $m := \tau(k).\text{loc2globdof}(j)$;
      if "es gibt $1 \le \ell \le c$ mit $m = \text{dof}(n).\text{sys}.\text{el}(\ell).\text{pos}$" **then**
        $\text{dof}(n).\text{sys}.\text{el}(\ell).\text{value} := \text{dof}(n).\text{sys}.\text{el}(\ell).\text{value} + a_{\tau_k}(b_{k,j}, b_{k,i})$
      else begin
        $c := c + 1$;
        $\text{dof}(n).\text{sys}.\text{el}(c).\text{value} := a_{\tau_k}(b_{k,j}, b_{k,i})$;
        $\text{dof}(n).\text{sys}.\text{el}(c).\text{pos} := m$;
        $\text{dof}(n).\text{sys}.\text{count} := c$;
      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 $\mathbf{y} = \mathbf{Au}$, 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**
    $\text{dof}(i).y = 0$;
    for $j := 1$ **to** $\text{dof}(i).\text{sys}.\text{count}$ **do**
      $\text{dof}(i).y := \text{dof}(i).y + \text{dof}(i).\text{sys}.\text{el}(j).\text{value} * \text{dof}(\text{dof}(i).\text{sys}.\text{el}(j).\text{pos}).u$
  end;
end;

8 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 **aymptotisch**, d.h., sie sagen etwas über die Konvergenzgeschwindigkeit des

Fehlers aus, wenn die Elementgrössen gegen Null streben. Für ein gegebenes Problem und eine gegebene Unterteilung sind sie aber unbrauchbar, da sie u.a. von Normen der unbekannten Lösung der Differentialgleichung abhängen. Für die Praxis stellt sich aber natürlich die Frage nach dem tatsächlichen Fehler der berechneten Finite-Elemente-Lösung. Zudem will man i.a. eine bestimmt 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 $S = S^{k,0}_{\mathcal{G},0}$. 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 Ω 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 φ mit $[\varphi]_E$ den Sprung von φ über E in Richtung n_E, d.h.

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

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_τ/h_E mit $\overline{\tau} \cap \overline{\tau'} \neq \emptyset$ und $\overline{E} \cap \overline{\tau} \neq \emptyset$ 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 $\tau \in \mathcal{G}$:

$$\begin{aligned} \omega_z := \operatorname{Tr} b_z = \overline{\bigcup_{\substack{\tau'\in\mathcal{G}\\ z\in\overline{\tau'}}} \tau'}, \quad \omega_\tau = \overline{\bigcup_{\overline{\tau}\cap\overline{\tau'}\in\mathcal{E}_\Omega} \tau'}, \quad \tilde{\omega}_\tau := \overline{\bigcup_{\overline{\tau}\cap\overline{\tau'}\neq\emptyset} \tau'}, \\ \omega_E := \bigcup_{E\subset\partial\tau'} \overline{\tau'}, \qquad \tilde{\omega}_E := \overline{\bigcup_{\overline{E}\cap\overline{\tau'}\neq\emptyset} \tau'}. \end{aligned} \tag{8.1}$$

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

Im folgenden bezeichnen $u \in H^1_0(\Omega)$ 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

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

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

$$\|u - u_S\|_{H^1(\Omega)} \leq \frac{1}{\alpha} \sup_{\substack{v\in H^1_0(\Omega)\\ \|v\|_{H^1(\Omega)}=1}} a(u - u_S, v). \tag{8.2}$$

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

$$a(u - u_S, v) = 0 \qquad \forall v \in S. \tag{8.3}$$

Die Abbildung $H_0^1(\Omega) \ni v \to a(u - u_S, v) = \ell(v) - a(u_S, v) \in \mathbb{R}$ 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 Gaussschen Integralsatzes auf jedem Element $\tau \in \mathcal{G}$ liefert

$$\begin{aligned}
a(u - u_S, v) = \ell(v) - a(u_S, v) &= \int_\Omega fv - \int_\Omega \{\langle \nabla v_S, \mathbf{A}\nabla u_S\rangle + cu_S v\} \\
&= \sum_{\tau\in\mathcal{G}} \left\{ \int_\tau fv - \int_\tau \langle \nabla u_S, \mathbf{A}\nabla v\rangle - \int_\tau cu_S v \right\} \\
&= \sum_{\tau\in\mathcal{G}} \left\{ \int_\tau fv + \int_\tau \operatorname{div}(\mathbf{A}\nabla u_S)\, v - \int_{\partial\tau} \langle n_\tau, \mathbf{A}\nabla u_S\rangle\, v - \int_\tau cu_S v \right\} \\
&= \sum_{\tau\in\mathcal{G}} \int_\tau \{f + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S\}\, v + \sum_{E\in\mathcal{E}_\Omega} \int_E [\langle n_E, \mathbf{A}\nabla u_S\rangle]_E\, v.
\end{aligned} \tag{8.4}$$

Dabei bezeichnet n_τ die äussere Normale zu τ. Wegen der Galerkin-Orthogonalität (8.3) können wir auf der rechten Seite von (8.4) ein beliebiges Element $v_S \in S$ von v subtrahieren. Aus der Cauchy-Schwarzschen Ungleichung für Integrale folgt dann

$$\begin{aligned}
a(u - u_S, v) \le &\sum_{\tau\in\mathcal{G}} \|f + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S\|_{L^2(\tau)} \|v - v_S\|_{L^2(\tau)} \\
&+ \sum_{E\in\mathcal{E}_\Omega} \|[\langle n_E, \mathbf{A}\nabla u_S\rangle]_E\|_{L^2(E)} \|v - v_S\|_{L^2(E)}.
\end{aligned} \tag{8.5}$$

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** $R_S : H_0^1(\Omega) \to S^{1,0}_{\mathcal{G},0}$ wie folgt. Für jedes $z \in \Theta_\Omega$ bezeichnen wir mit $\pi_z : L^2(\omega_z) \to \mathbb{R}$ die orthogonale Projektion, d.h.

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

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

$$R_S \varphi := \sum_{z\in\Theta_\Omega} (\pi_z \varphi)\, b_z. \tag{8.6}$$

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

Satz 8.1 *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*

$$\begin{aligned}
\|v - R_S v\|_{L^2(\tau)} &\le c_1 h_\tau |v|_{H^1(\tilde{\omega}_\tau)}, \\
\|v - R_S v\|_{L^2(E)} &\le c_2 h_E^{1/2} |v|_{H^1(\tilde{\omega}_E)}.
\end{aligned}$$

Beweis. 1. Schritt: Wir beweisen: für jedes $z \in \Theta_\Omega$ und jedes $\varphi \in H^1(\omega_z)$ gilt

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

Sei dazu

$$n_T := \max_{z \in \Theta_\tau} n_T(z) \quad \text{mit} \quad n_T(z) := \sharp\{\tau \in \mathcal{G} \mid \tau \subset \omega_z\}.$$

Diese Zahl hängt lediglich von der Formregularität des Gitters $\mathcal{G}$ ab. Wir definieren dann für $3 \leq \ell \leq n_T$ *Referenzsterne* $\hat{\omega}_\ell$ die aus ℓ Dreiecken bestehen mit paarweise disjunktem Innern, mit 0 als einem Eckpunkt, die beiden Schenkel an 0 die Länge 1 haben und dort den Öffnungswinkel $2\pi/\ell$ besitzen. Sei nun ω_z wie in (8.1) und ℓ die Anzahl der Dreiecke in ω_z. Wir definieren eine Lipschitz-stetige, stückweise affine Abbildung $\chi_z : \hat{\omega}_\ell \to \omega_z$ implizit für $\tau \subset \omega_z$ und $\hat{\tau} := \chi_z^{-1}(\tau)$ durch "$\chi_\tau := \chi_z|_{\hat{\tau}} : \hat{\tau} \to \tau$ ist affin". Wir setzen $\hat{\varphi} = \varphi \circ \chi_z$ und $\hat{\varphi}_\tau := \varphi|_\tau \circ \chi_\tau$. Dann gilt mit $\hat{\pi}_{\hat{\tau}} \varphi_{\hat{\tau}} := \frac{1}{|\hat{\tau}|} \int_{\hat{\tau}} \hat{\varphi}_{\hat{\tau}}$

$$\|\varphi - \pi_z \varphi\|^2_{L^2(\omega_z)} = \sum_{t \subset \omega_z} \|\varphi - \pi_z \varphi\|^2_{L^2(t)} = \sum_{t \subset \omega_z} \frac{|t|}{|\hat{t}|} \|\hat{\varphi}_t - \pi_z \varphi\|^2_{L^2(\hat{t})} = \sum_{t \subset \omega_z} \frac{|t|}{|\hat{t}|} \left\| \hat{\varphi}_t - \sum_{\tau \subset \omega_z} \frac{|\tau|}{|\omega_z|} \hat{\pi}_{\hat{\tau}} \hat{\varphi}_{\hat{\tau}} \right\|^2_{L^2(\hat{t})} \tag{8.7}$$

$$= \sum_{t \subset \omega_z} \frac{|t|}{|\hat{t}|} \left\| \sum_{\tau \subset \omega_z} \frac{|\tau|}{|\omega_z|} (\hat{\varphi}_t - \hat{\pi}_{\hat{\tau}} \hat{\varphi}_{\hat{\tau}}) \right\|^2_{L^2(\hat{t})} \leq \sum_{t \subset \omega_z} \frac{|t|}{|\hat{t}|} \sum_{\tau \subset \omega_z} \|\hat{\varphi}_t - \hat{\pi}_{\hat{\tau}} \hat{\varphi}_{\hat{\tau}}\|^2_{L^2(\hat{t})}. \tag{8.8}$$

Wie definieren $\hat{\pi}_0 \hat{\varphi} := \int_{\hat{\omega}_\ell} \hat{\varphi} / |\hat{\omega}_\ell|$. Um die letzte Norm in (8.8) abzuschätzen, verwenden wir

$$\|\hat{\varphi}_t - \hat{\pi}_{\hat{\tau}} \hat{\varphi}_{\hat{\tau}}\|_{L^2(\hat{t})} \leq \|\hat{\varphi} - \pi_{\hat{\tau}} \hat{\varphi}_{\hat{\tau}}\|_{L^2(\hat{\omega}_\ell)} = \|(I - \hat{\pi}_{\hat{\tau}})(I - \hat{\pi}_0) \hat{\varphi}\|_{L^2(\hat{\omega}_\ell)}$$

da $(I - \hat{\pi}_{\hat{\tau}}) \hat{\pi}_0 \hat{\varphi} = 0$ gilt. Daraus folgt

$$\|\hat{\varphi}_t - \pi_{\hat{\tau}} \hat{\varphi}_{\hat{\tau}}\|_{L^2(\hat{t})} \leq \left(1 + \sup_{\psi \in L^2(\hat{\omega}_\ell) \setminus \{0\}} \frac{\|\hat{\pi}_{\hat{\tau}} \psi\|_{L^2(\hat{\omega}_\ell)}}{\|\psi\|_{L^2(\hat{\omega}_\ell)}} \right) \|(I - \hat{\pi}_0) \hat{\varphi}\|_{L^2(\hat{\omega}_\ell)}. \tag{8.9}$$

Die Abschätzung $\|(I - \hat{\pi}_0) \hat{\varphi}\|_{L^2(\hat{\omega}_\ell)} \leq C \|\nabla \hat{\varphi}\|_{L^2(\hat{\omega}_\ell)}$ folgt aus der Poincaré-Ungleichung. Das Supremum (8.9) lässt sich abschätzen durch

$$\|\hat{\pi}_{\hat{\tau}} \psi\|_{L^2(\hat{\omega}_\ell)} = \sqrt{|\hat{\omega}_\ell|} \, |\hat{\pi}_{\hat{\tau}} \psi| = \sqrt{\frac{|\hat{\omega}_\ell|}{|\hat{\tau}|}} \|\hat{\pi}_{\hat{\tau}} \psi\|_{L^2(\hat{\tau})} \leq \sqrt{\ell} \, \|\psi\|_{L^2(\hat{\tau})},$$

da $\hat{\pi}_{\hat{\tau}}$ die $L^2(\hat{\tau})$-Orthogonalprojektion ist. Die Kombination dieser Ungleichungen liefert

$$\|\hat{\varphi}_t - \pi_{\hat{\tau}} \hat{\varphi}_{\hat{\tau}}\|_{L^2(\hat{t})} \leq \left(1 + \sqrt{\ell}\right) \|\nabla \hat{\varphi}\|_{L^2(\hat{\omega}_\ell)} \leq C h_t^{-1} \sqrt{\frac{|\hat{\omega}_\ell|}{|\omega_\ell|}} \|\nabla \varphi\|_{L^2(\omega_\ell)}.$$

Zusammen mit (8.7) ergibt sich die erste Zwischenbehauptung.

Wegen der Regularitätsannahme an $\mathcal{G}$ gibt es eine Konstante c' mit $\operatorname{diam} \omega_z \leq c' h_\tau$ für jedes Dreieck τ, 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 τ und $\varphi \in H^1(\tau)$. Wähle die affine Transformation $\chi_\tau : \hat{\tau} \to \tau$ 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

$$h_E^{1/2} |\det D\chi_\tau|^{-1/2} \leq c h_\tau^{-1/2}, \qquad h_E^{1/2} |\det D\chi_\tau|^{-1/2} \|D\chi_\tau\| \leq c h_\tau^{1/2}.$$

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 Γ hat. Bezeichne die Menge der Eckpunkte von τ mit Θ_τ. Dann ist

$$\sum_{z \in \Theta_\tau} b_z = 1 \qquad \text{auf } \tau.$$

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 τ, das mindestens einen Eckpunkt auf dem Rand Γ hat. Dann gilt auf τ

$$v - R_S v = \sum_{z \in \Theta_\tau} b_z v - \sum_{z \in \Theta_\tau \cap \Theta_\Omega} b_z \pi_z v = \sum_{z \in \Theta_\tau} b_z (v - \pi_z v) + \sum_{z \in \Theta_\tau \setminus \Theta_\Omega} b_z \pi_z v$$

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 τ, der auf Γ liegt. Dann ist

$$\|b_z \pi_z v\|_{L^2(\tau)} \leq c h_\tau |\pi_z v|.$$

Da $z \in \Gamma$ ist, gibt es ein Dreieck $\tau' \in \mathcal{G}$ und eine Kante E' von τ', 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 τ.

5 Schritt: Sei $E \in \mathcal{E}_\Omega$ eine Kante die keinen Eckpunkt auf dem Rand Γ hat. Bezeichne mit Θ_E die Menge der Eckpunkte von E. Dann ist

$$\sum_{z \in \Theta_E} b_z = 1 \qquad \text{auf } E.$$

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 τ_E ein Dreieck, das E als Kante hat.

6. Schritt: Betrachte nun eine Kante E, die einen Eckpunkt auf dem Rand Γ 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 (8.5) wieder auf. Wir setzen $v_S = R_S v$, benutzen Satz 8.1 und wenden die Cauchy-Schwarzsche Ungleichung für endliche Summen an:

$$\begin{aligned}
a(u - u_S, v) &\leq c_1 \sum_{\tau \in \mathcal{G}} h_\tau \|f + \operatorname{div}(\mathbf{A} \nabla u_S) - c u_S\|_{L^2(\tau)} |v|_{H^1(\tilde{\omega}_\tau)} \\
&\quad + c_2 \sum_{E \in \mathcal{E}_\Omega} h_E^{1/2} \|[\langle n_E, \mathbf{A} \nabla u_S \rangle]_E\|_{L^2(E)} |v|_{H^1(\tilde{\omega}_E)} \\
&\leq c_1 \left\{ \sum_{\tau \in \mathcal{G}} h_\tau^2 \|f + \operatorname{div}(\mathbf{A} \nabla u_S) - c u_S\|_{L^2(\tau)}^2 \right\}^{1/2} \left\{ \sum_{\tau \in \mathcal{G}} |v|_{H^1(\tilde{\omega}_\tau)}^2 \right\}^{1/2} \\
&\quad + c_2 \left\{ \sum_{E \in \mathcal{E}_\Omega} h_E \|[\langle n_E, \mathbf{A} \nabla u_S \rangle]_E\|_{L^2(E)}^2 \right\}^{1/2} \left\{ \sum_{E \in \mathcal{E}_\Omega} |v|_{H^1(\tilde{\omega}_E)}^2 \right\}^{1/2} \\
&\leq c' |v|_{H^1(\Omega)} \left\{ \sum_{\tau \in \mathcal{G}} h_\tau^2 \|f + \operatorname{div}(\mathbf{A} \nabla u_S) - c u_S\|_{L^2(\tau)}^2 + \sum_{E \in \mathcal{E}_\Omega} h_E \|[\langle n_E, \mathbf{A} \nabla u_S \rangle]_E\|_{L^2(E)}^2 \right\}^{1/2}.
\end{aligned}$$

Dabei haben wir im letzten Schritt die Formregularität von $\mathcal{G}$ ausgenutzt. Hieraus und aus (8.2) folgt insgesamt

$$\|u - u_S\|_{H^1(\Omega)} \leq c\eta \tag{8.10}$$

mit

$$\eta := \left\{ \sum_{\tau \in \mathcal{G}} \eta_\tau^2 \right\}^{1/2},$$

$$\eta_\tau := \left\{ h_\tau^2 \left\| f + \operatorname{div}(\mathbf{A}\nabla u_S) - c u_S \right\|_{L^2(\tau)}^2 + \frac{1}{2} \sum_{E \subset \partial\tau \setminus \Gamma} h_E \left\| [\langle n_E, \mathbf{A}\nabla u_S \rangle]_E \right\|_{L^2(E)}^2 \right\}^{1/2}. \tag{8.11}$$

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

Ungleichung (8.10) ist eine **a-posteriori Fehlerabschätzung**. Die Grösse η 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 (8.10) zeigt, dass der Fehlerschätzer **zuverlässig** ist, d.h. ist $\eta < \varepsilon$, so ist der Fehler ebenfalls (bis auf einen Faktor) nicht grösser als ε. Die Kontrolle von η erlaubt also, eine vorgegebene Toleranz zu erreichen. Um dies mit einem minimalen Aufwand zu erreichen, reicht die obere Schranke (8.10) nicht aus. Wir müssen zusätzlich garantieren, dass η 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 η 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 τ bezeichne mit z_1, z_2, z_3 seine Eckpunkte und setze

$$\psi_\tau := 27 b_{z_1} b_{z_2} b_{z_3} \qquad \text{auf } \tau.$$

Wie man leicht nachprüft, hat ψ_τ 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 ψ_τ durch Null zu einer Funktion aus $H_0^1(\Omega)$ forstgesetzt werden. Für eine Kante $E \in \mathcal{E}_\Omega$ numerieren wir die Eckpunkte der angrenzenden Dreiecke τ_{E_1} und τ_{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 τ_{E_i}, $i = 1, 2$, so definieren wir

$$\psi_E := 4 v_{z_1,i} v_{z_2,i} \qquad \text{auf } \tau_{E_i},\ i = 1, 2.$$

Offensichtlich hat ψ_E die folgenden Eigenschaften

$$\psi_E \in C(\omega_E), \quad \psi_E|_{\tau_{E_i}} \in \mathbb{P}_2,\ i = 1, 2, \quad \psi_E \geq 0 \text{ auf } \omega_E, \quad \psi_E = 0 \text{ auf } \partial\omega_E, \quad \max_{x \in E} \psi_E(x) = 1.$$

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 φ.

Satz 8.2 *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*

$$\begin{aligned}
c_1 \|v\|_{L^2(\tau)} &\le \left\{ \int_\tau \psi_\tau v^2 \right\}^{1/2} \le \|v\|_{L^2(\tau)}, \\
c_2 h_\tau^{-1} \|\psi_\tau v\|_{L^2(\tau)} &\le |\psi_\tau v|_{H^1(\tau)} \le c_3 h_\tau^{-1} \|\psi_\tau v\|_{L^2(\tau)}, \\
c_4 \|\sigma\|_{L^2(E)} &\le \left\{ \int_E \psi_E \sigma^2 \right\}^{1/2} \le \|\sigma\|_{L^2(E)}, \\
c_5 h_E^{-1} \|\psi_E \sigma\|_{L^2(\omega_E)} &\le |\psi_E \sigma|_{H^1(\omega_E)} \le c_6 h_E^{-1} \|\psi_E \sigma\|_{L^2(\omega_E)}, \\
\|\psi_E \sigma\|_{L^2(\omega_E)} &\le c_7 h_E^{1/2} \|\sigma\|_{L^2(E)}.
\end{aligned}$$

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

$$\hat c \|w\|_{L^2(\hat\tau)} \le \left\{ \int_{\hat\tau} (\psi_\tau \circ \chi_\tau) w^2 \right\}^{1/2} \qquad \forall w \in \mathbb{P}_k.$$

Anwenden des Transformationssatzes liefert

$$\begin{aligned}
\hat c \|v\|_{L^2(\tau)} = \hat c |\det D\chi_\tau|^{1/2} \|v \circ \chi_\tau\|_{L^2(\hat\tau)} &\le |\det D\chi_\tau|^{1/2} \left\{ \int_{\hat\tau} (\psi_\tau \circ \chi_\tau)(v \circ \chi_\tau) \right\}^{1/2} \\
&= \left\{ \int_\tau \psi_\tau v^2 \right\}^{1/2}
\end{aligned}$$

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

$$\hat c_1 \|\psi_\tau \circ \chi_\tau w\|_{L^2(\hat\tau)} \le |(\psi_\tau \circ \chi_\tau) w|_{H^1(\hat\tau)} \le \hat c_2 |(\psi_\tau \circ \chi_\tau) w|_{L^2(\hat\tau)} \qquad \forall w \in \mathbb{P}_k.$$

Hieraus uns aus dem Transformationssatz folgt wieder die Behauptung.

Die Abschätzungen für σ folgen mit den gleichen Argumenten. ■

Sei nun $\tau \in \mathcal{G}$ beliebig. Setze $w_\tau := \psi_\tau (f_S + \operatorname{div}(\mathbf{A}\nabla u_S) - c u_S)$. Wegen Satz 8.2 ist

$$c_1^2 \|f_S + \operatorname{div}(\mathbf{A}\nabla u_S) - c u_S\|_{L^2(\tau)}^2 \le \int_\tau w_\tau (f_S + \operatorname{div}(\mathbf{A}\nabla u_S) - c u_S).$$

Setzen wir w_τ als Testfunktion v in (8.4) ein und berücksichtigen, dass w_τ auf $\partial\tau$ und ausserhalb von τ verschwindet, so erhalten wir

$$\begin{aligned}
\int_\tau w_\tau (f_S + \operatorname{div}(\mathbf{A}\nabla u_S) - c u_S) &= \int_\tau w_\tau (f + \operatorname{div}(\mathbf{A}\nabla u_S) - c u_S) + \int_\tau w_\tau (f_S - f) \\
&= a(u - u_S, w_\tau) + \int_\tau w_\tau (f_S - f) \\
&\le \|a\|_{\mathcal{L}^2} \|u - u_S\|_{H^1(\tau)} \|w_\tau\|_{H^1(\tau)} + \|f_S - f\|_{L^2(\tau)} \|w_\tau\|_{L^2(\tau)}.
\end{aligned}$$

Offensichtlich ist

$$\|w_\tau\|_{L^2(\tau)} \leq \|f_S + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S\|_{L^2(\tau)}.$$

Aus Satz 8.2 folgt weiter

$$\begin{aligned}\|w_\tau\|_{H^1(\tau)} &= \left\{\|w_\tau\|^2_{L^2(\tau)} + |w_\tau|^2_{H^1(\tau)}\right\}^{1/2} \leq \left\{1 + c_3^2 h_\tau^{-2}\right\}^{1/2} \|w_\tau\|_{L^2(\tau)} \\ &\leq c h_\tau^{-1} \|f_S + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S\|_{L^2(\tau)}.\end{aligned}$$

Aus diesen Abschätzungen und der Dreiecksungleichung ergibt sich

$$h_\tau \|f + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S\|_{L^2(\tau)} \leq c\left\{\|u - u_S\|_{H^1(\tau)} + h_\tau \|f - f_S\|_{L^2(\tau)}\right\}. \tag{8.12}$$

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

$$w_E := \psi_E \left[\langle n_E, \mathbf{A}\nabla u_S\rangle\right]_E.$$

Wegen Satz 8.2 ist

$$c_4^2 \|[\langle n_E, \mathbf{A}\nabla u_S\rangle]_E\|^2_{L^2(E)} \leq \int_E w_E \left[\langle n_E, \mathbf{A}\nabla u_S\rangle\right]_E.$$

Setzen wir w_E als Testfunktion v in (8.4) ein und berücksichtigen, dass w_E auf $\partial\omega_E$ und ausserhalb w_E verschwindet, so folgt

$$\begin{aligned}\int_E w_E \left[n_E, \mathbf{A}\nabla u_S\right]_E &= \int_{\omega_E} (f + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S) w_E - a(u - u_S, w_E) \\ &\leq \|f + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S\|_{L^2(\omega_E)} \|w_E\|_{L^2(\omega_E)} + \|a\|_{\mathcal{L}^2} \|u - u_S\|_{H^1(\omega_E)} \|w_E\|_{H^1(\omega_E)}.\end{aligned}$$

Wegen Satz 8.2 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 (8.12) ergibt sich insgesamt

$$h_E^{1/2} \|[\langle n_E, \mathbf{A}\nabla u_S\rangle]_E\|_{L^2(E)} \leq c\left\{\|u - u_S\|_{H^1(\omega_E)} + h_E \|f - f_S\|_{L^2(\omega_E)}\right\}. \tag{8.13}$$

Aus den Abschätzungen (8.12), (8.13) und der Definition (8.11) erhalten wir folgende **lokale untere Fehlerschranke**

$$\eta_\tau \leq c\left\{\|u - u_S\|_{H^1(\omega_\tau)} + h_\tau \|f - f_S\|_{L^2(\omega_\tau)}\right\}. \tag{8.14}$$

Summation über alle Dreiecke liefert zudem die **globale untere Fehlerschranke**

$$\eta \leq c\left\{\|u - u_S\|_{H^1(\Omega)} + \left\{\sum_{\tau\in\mathcal{G}} h_\tau^2 \|f - f_S\|^2_{L^2(\tau)}\right\}^{1/2}\right\}. \tag{8.15}$$

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 8.3

1. *Die Abschätzungen (8.10), (8.14) und (8.15) zeigen, dass der Fehlerschätzer zuverlässig und effizient ist. Es ist nicht verwunderlich, dass wir nur in einer Richtung lokale Schranken erhalten. Denn die Abschätzung (8.10) benötigt den inversen Differentialoperator, der ein globaler Operator ist. Die Abschätzung (8.14) dagegen benutzt nur den Differentialoperator, der ein lokaler Operator ist.*

2. *Die Grösse $f + \operatorname{div}(\mathbf{A}\nabla u_S) - cu_S$ 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 Spuroperators, der auf kanonische Weise die starke und schwache Form der Differentialgleichung verknüpft.*

3. *Ähnliche Ergebnisse gelten für die Konvektions-Diffusions-Gleichung. A priori hängen die Konstanten in den Fehlerabschätzungen von der Peclet-Zahl ab und wachsen für abnehmende Diffusion an. Dieser Effekt kann durch ein 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 η 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 η liefern. Dies legt es nahe, Elemente, die einen zu grossen Beitrag η_τ liefern, zu unterteilen, und führt auf folgenden Algorithmus.

Algorithmus 8.4 (Adaptive Gitterverfeinerung)

1. *Bestimme eine grobe Triangulierung $\mathcal{G}_0$ von Ω. Setze $k := 0$.*

2. *Löse das diskrete Problem zur Triangulierung $\mathcal{G}_k$.*

3. *Berechne η_τ 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 8.4 ist γ ein zu wählender Parameter mit $0 < \gamma < 1$. Ist $\gamma \sim 0$, werden viele Elemente verfeinert; ist $\gamma \sim 1$, werden nur sehr wenige Elemente unterteilt. In der Praxis wählt man häufig $\gamma = 0.5$. Algorithmus 8.4 kann auch ggf. durch eine Vergröberungsstrategie ergänzt werden.

Für die praktische Realisierung von Algorithmus 8.4 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 wir **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 τ 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 8.4 wird nun gemäss der folgenden Regeln durchgeführt:

1. Ist $\eta_\tau \geq \gamma \eta_k$, unterteile τ rot.
2. Hat τ drei hängende Knoten, unterteile τ rot.
3. Hat τ zwei hängende Knoten, von denen keiner auf der längsten Kante liegt, unterteile τ rot.
4. Hat τ zwei hängende Knoten, von denen einer auf der längsten Kante liegt, unterteile τ blau.
5. Hat τ einen hängenden Knoten, unterteile τ blau, wenn der hängende Knoten nicht auf der längsten Kante liegt, sonst unterteile τ grün.

Man kann zeigen, dass dieser Algorithmus in endlich vielen Schritten eine verfeinerte Triangulierung erzeugt, die den oben diskutierten Kriterien genügt.

9 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 $\mathbb{R}^N$, $N = \sharp \overset{\circ}{K}{}^k_{\mathcal{G}}$. Die Matrix, die zu $(\cdot,\cdot)_S$ und der Knoten-Basis gehört, ist diagonal. Insbesondere können Gleichungssystem der Form

$$(u,v)_S = \ell(v) \qquad \forall v \in S$$

leicht gelöst werden.

Satz 9.1

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***

$$|v|_{H^1(\Omega)} \le c h_{\min}^{-1} \|v\|_{L^2(\Omega)}.$$

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

Beweis. Zu (1): Wegen Satz **??** 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

$$\hat c_1 \|\varphi\|_{L^2(\hat\tau)} \le \left\{\sum_{z\in\hat\Sigma_k} |\varphi(z)|^2\right\}^{1/2} \le \hat c_2 \|\varphi\|_{L^2(\hat\tau)} \qquad \forall \varphi\in\mathbb{P}_k.$$

Sei nun $\tau\in\mathcal{G}$ beliebig und $\chi_\tau : \hat\tau \to \tau$ eine affine Transformation. Wegen

$$c_1 h_\tau^2 \le |\det D\chi_\tau| \le c_2 h_\tau^2$$

folgt aus obiger Abschätzung mit dem Transformationssatz

$$\begin{aligned}\hat c_1 \|\varphi\|_{L^2(\tau)} &= \hat c_1 |\det D\chi_\tau|^{1/2} \|\varphi\circ\chi_\tau\|_{L^2(\hat\tau)} \le c_2^{1/2} h_\tau \left\{\sum_{z\in\hat\Sigma_k} |\varphi\circ\chi_\tau(z)|^2\right\}^{1/2} \\ &\le c_2^{1/2} h_\tau \hat c_2 \|\varphi\circ\chi_\tau\|_{L^2(\hat\tau)} = c_2^{1/2}\hat c_2 h_\tau |\det D\chi_\tau|^{-1/2} \|\varphi\|_{L^2(\tau)} \le c_2^{1/2}\hat c_2 c_1^{-1/2} \|\varphi\|_{L^2(\tau)}.\end{aligned}$$

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

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

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

$$\begin{aligned}|\varphi|_{H^1(\tau)} &\le |\det D\chi_\tau|^{1/2} \left\|D\chi_\tau^{-1}\right\| |\varphi \circ \chi_\tau|_{H^1(\hat\tau)} \\ &\le \hat c\, |\det D\chi_\tau|^{1/2} \left\|D\chi_\tau^{-1}\right\| \|\varphi \circ \chi_\tau\|_{L^2(\hat\tau)} \\ &= \hat c \left\|D\chi_\tau^{-1}\right\| \|\varphi\|_{L^2(\tau)} \le \hat c h_{\hat\tau}/\rho_\tau \|\varphi\|_{L^2(\tau)} \le c' h_\tau^{-1} \|\varphi\|_{L^2(\tau)}\,.\end{aligned}$$

Dabei hängt c' von $c_{\mathcal{G}}$ ab. Quadrieren und Summieren dieser Ungleichung beweist Teil **(2)**. ■

Aus Satz 9.1(1), dem Beweis des Existenzsatzes 4.3 und der Friedrichsschen Ungleichung folgt für alle $v \in S$

$$a(v,v) \ge \beta |v|^2_{H^1(\Omega)} \ge \beta' \|v\|^2_{L^2(\Omega)} \ge \beta'' \|v\|^2_S\,.$$

Ebenso folgt aus dem Beweis von Satz 4.3 und Satz 9.1 für alle $v, w \in S$

$$\begin{aligned}a(v,w) &\le B |v|_{H^1(\Omega)} |w|_{H^1(\Omega)} \le c^2 B \left(\min_{\tau\in\mathcal{G}} h_\tau\right)^{-2} \|v\|_{L^2(\Omega)} \|w\|_{L^2(\Omega)} \\ &\le c' \left(\min_{\tau\in\mathcal{G}} h_\tau\right)^{-2} \|v\|_S \|w\|_S\,.\end{aligned}$$

Also hat die Steifigkeitsmatrix die Kondition $O\left(\min_{\tau\in\mathcal{G}} h_\tau\right)^{-2}$. 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. Wir 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 Ω. Wir nehmen an, dass die Gitter geschachtelt sind, d.h., für alle $0 \le m \le M-1$ und alle $\tau \in \mathcal{G}_m$ existiert eine Teilmenge (*Söhne*) $sons(\tau) \subset \mathcal{G}_{m+1}$ mit der Eigenschaft, dass

$$\overline{\tau} = \bigcup_{t\in \mathrm{sons}(\tau)} \overline{t}$$

gilt. Wir machen die folgenden Annahmen (die mit aufwendigeren Mittel der Funktionalanalysis abgeschwächt werden können).

1. Ω ist konvex.
2. Jede Triangulierung $\mathcal{G}_m$ ist **quasi-uniform**, d.h.

   $$h_m := \max_{\tau\in\mathcal{G}_m} h_\tau \le c \min_{\tau\in\mathcal{G}_m} h_\tau$$

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

Algorithmus 9.2 (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 $i = 1, \ldots, \nu_1$ berechne u_m^i aus u_m^{i-1} als Lösung von*[14]

$$\left(u_m^i - u_m^{i-1}, v\right)_m = \omega_m^{-1} \left\{\ell_m(v) - a\left(u_m^{i-1}, v\right)\right\} \qquad \forall v \in S_m.$$

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

$$\ell_{m-1}(v) = \ell_m(v) - a\left(u_m^{\nu_1}, v\right) \qquad \forall v \in S_{m-1}.$$

Falls *$m > 1$, wende das MG-Verfahren mit Startwerten $u_{m-1}^0 = 0$ auf das Problem*

$$a\left(u_{m-1}^\star, v\right) = \ell_{m-1}(v) \qquad \forall v \in S_{m-1}$$

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

Setze

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

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

$$\left(u_m^i - u_m^{i-1}, v\right)_m = \omega^{-1} \left\{\ell_m(v) - a\left(u_m^{i-1}, v\right)\right\} \qquad \forall v \in S_m.$$

Bemerkung 9.3

1. *Es ist $\ell_M(v) = \ell(v) = \int_\Omega fv$. Die rechten Seiten ℓ_m zu den gröberen Triangulierungen werden rekursiv berechnet.*

2. *Die Dämpfungsparameter ω_m werden später bestimmt.*

3. *Die Gleichungssysteme in den Glättungsschritten haben eine diagonale Koeffizientenmatrix.*

4. *Die Grobgitterkorrektur nutzt aus, dass die Finite-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 $\lambda_{m,1}, \ldots, \lambda_{m,N_m}$, $N_m := \dim S_m$, von Eigenwerten und zugehörigen, bzgl $(\cdot,\cdot)_m$ orthonormierten Eigenfunktionen $\psi_{m,1}, \ldots, \psi_{m,N_m}$:

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

O.B.d.A. können wir die Eigenwerte der Grösse nach ordnen

$$0 < \lambda_{m,1} \leq \ldots \leq \lambda_{m,N_m} = \Lambda_m.$$

[14] Abkürzung: $(\cdot,\cdot)_m := (\cdot,\cdot)_{S_m}$

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

$$\omega_m = \Lambda_m$$

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

Jedes $v \in S_m$ 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

$$a(v,w) = \sum_{\mu=1}^{N_m} \lambda_{m,\mu} c_\mu d_\mu \leq \left\{ \sum_{\mu=1}^{N_m} c_\mu^2 \right\}^{1/2} \left\{ \sum_{\mu=1}^{N_m} \lambda_{m,\mu}^2 d_\mu^2 \right\}^{1/2} = |||v|||_{0,m} \, |||w|||_{2,m}. \tag{9.1}$$

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

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

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 (9.1) 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 Ω 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 $J : S_m \to S_m$ durch

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

J beschreibt die Fehlerfortpflanzung in den Glättungsschritten von Algorithmus 9.2. 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 $v \in S_\ell$

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

Dann gilt offensichtlich für $v \in S_m$

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

Satz 9.5 *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

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

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

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

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 9.6 *Sei $v \in S_m$ und $\rho = \rho(v)$. Dann gilt*

$$\left|\!\left|\!\left|(I - Q_m)\, v\right|\!\right|\!\right|_{1,m} \leq \min\left\{1, c\sqrt{1-\rho}\right\} \left|\!\left|\!\left|v\right|\!\right|\!\right|_{1,m}.$$

Beweis. Da $I - Q_m$ eine Orthogonalprojektion ist, ist

$$\left|\!\left|\!\left|(I - Q_m)\, v\right|\!\right|\!\right|_{1,m} \leq \left|\!\left|\!\left|v\right|\!\right|\!\right|_{1,m}.$$

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

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

Hieraus und aus Satz 9.4 folgt

$$\left|\!\left|\!\left|(I - Q_m)\, v\right|\!\right|\!\right|_{1,m}^2 \leq c^2 h_m^2 \left|\!\left|\!\left|v\right|\!\right|\!\right|_{2,m}^2 = c^2 h_m^2 \Lambda_m (1-\rho) \left|\!\left|\!\left|v\right|\!\right|\!\right|_{1,m}^2.$$

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

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

Satz 9.7 *Bezeichne mit δ_m die Konvergenzrate von Algorithmus 9.2 mit $\nu_1 = \nu_2 = \nu$ auf dem m-ten Gitter gemessen in der $\left|\!\left|\!\left|\cdot\right|\!\right|\!\right|_{1,m}$-Norm. Dann gilt mit der Konstanten c aus Satz 9.6*

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

Beweis. Bezeichne mit $u_m^\star$ die Lösung der Finite-Elemente-Probleme auf dem m-ten Gitter und mit $e_m^i = u_m^\star - u_m^i$ „Näherung nach dem $i$ten Mehrgitterschritt“ den Fehler im i-ten Schritt von Algorithmus 9.2. Dann gilt

$$\begin{aligned}e_m^{\nu+1} &= e_m^\nu - u_{m-1}^\star + u_{m-1}^\star - \tilde{u}_{m-1} = e_m^\nu - u_{m-1}^\star + \delta_{m-1} \frac{1}{\delta_{m-1}} \left(u_{m-1}^\star - \tilde{u}_{m-1}\right) \\ &= (I - Q_m)\, e_m^\nu + \delta_{m-1} w_{m-1}\end{aligned}$$

mit $w_{m-1} := \delta_{m-1}^{-1}\left(u_{m-1}^\star - \tilde{u}_{m-1}\right) \in S_{m-1}$ und (per Induktion)

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

Wegen der Galerkin-Orthogonalität

$$a\left((I - Q_m)\, v, w\right) = 0 \qquad \forall v \in S_m, w \in S_{m-1}$$

folgt

$$\begin{aligned}\left|\!\left|\!\left|(I - Q_m)\, e_m^\nu + w_{m-1}\right|\!\right|\!\right|_{1,m}^2 &= \left|\!\left|\!\left|(I - Q_m)\, e_m^\nu\right|\!\right|\!\right|_{1,m}^2 + \left|\!\left|\!\left|w_{m-1}\right|\!\right|\!\right|_{1,m}^2 \\ &= \left|\!\left|\!\left|(I - Q_m)\, e_m^\nu\right|\!\right|\!\right|_{1,m}^2 + \left|\!\left|\!\left|w_{m-1}\right|\!\right|\!\right|_{1,m-1}^2 \\ &\leq \left|\!\left|\!\left|(I - Q_m)\, e_m^\nu\right|\!\right|\!\right|_{1,m}^2 + \left|\!\left|\!\left|u_{m-1}^\star\right|\!\right|\!\right|_{1,m-1}^2 \\ &= \left|\!\left|\!\left|(I - Q_m)\, e_m^\nu\right|\!\right|\!\right|_{1,m}^2 + \left|\!\left|\!\left|u_{m-1}^\star\right|\!\right|\!\right|_{1,m}^2 \\ &= \left|\!\left|\!\left|(I - Q_m)\, e_m^\nu + \underbrace{u_{m-1}^\star}_{Q_m e_m^\nu}\right|\!\right|\!\right|_{1,m}^2 = \left|\!\left|\!\left|e_m^\nu\right|\!\right|\!\right|_{1,m}^2.\end{aligned}$$

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

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

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

$$a\left(\left(I - Q_m\right) e_m^\nu, J^\nu w\right) = a\left(\left(I - Q_m\right) e_m^\nu, \left(I - Q_m\right) J^\nu w\right) \leq \|\|\|\left(I - Q_m\right) e_m^\nu\|\|\|_{1,m} \|\|\|\left(I - Q_m\right) J^\nu w\|\|\|_{1,m}.$$

Für den zweiten Summanden gilt

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

Aus diesen beiden Abschätzungen folgt mit der Cauchy-Schwarzschen Ungleichung

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

Wegen der Sätze 9.5 und 9.6 ist

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

und

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

Da

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

ist, folgt hieraus

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

Also ist

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

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

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

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

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

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

Der Beweis von Satz 9.7 beruht im wesentlichen auf den Annahmen, dass Ω konvex ist und die Gitter quasiuniform sind. Die Konvexität von Ω wird in Satz 9.4 benötigt, da dort das Dualitätsargument von Aubin-Nitsche benutzt wird, das die H^2.Regularität der Differentialgleichung voraussetzt. Die Uniformität der Gitter wird für die inverse Abschätzung zur Kontrolle des grössten Eigenwertes der 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 9.2 wesentlich bessere Konvergenzresultate erzielen kann.

10 Parabolische partielle Differentialgleichung

Sei Ω 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: $u : \Omega \times [0,T] \to \mathbb{R}$. Die Gleichung, zur Charakterisierung von u ist die parabolische partielle Differentialgleichung

$$\begin{array}{rll}
\dot{u} - \Delta u = f & \text{in } \Omega & \text{für alle } t \in [0,T]\,, \\
u = 0 & \text{auf } \partial\Omega & \text{für alle } t \in [0,T] \text{ (zeitabhängige Randbedingung)}, \\
u\,(\cdot,0) = v & \text{in } \Omega & \text{Anfangsbedingung.}
\end{array} \tag{10.1}$$

Hier bezeichnet $\dot{u}$ die Ableitung von u nach der Zeit und Δ 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 (10.1) die Ortsabhängigkeit mit finiten Elementen diskretisieren, entsteht ein System gewöhnlicher Differentialgleichung bezüglich der Zeit mit vorgegebenen Anfangsbedingungen.

10.1 Ortsdiskretisierung des parabolischen Problems

Sei $\mathcal{G}$ wieder ein Finite-Elemente-Gitter für das Gebiet Ω. Die Menge der Gitterpunkte wird wieder mit Θ bezeichnet und die Menge der inneren Gitterpunkte mit Θ^0. Die Basis für den Raum aller stetigen, stückweise affinen Finite-Elemente-Funktionen zu den inneren Knotenpunkten wird wieder mit $(b_i)_{i=1}^N$, $N = \sharp\Theta^0$ bezeichnet. Mit S bezeichnen wir den Raum aller stetigen, stückweise affinen Funktionen mit Null-Randbedingungen.

$$S = \operatorname{span}\left\{b_i : 1 \leq i \leq N\right\}.$$

Für die semidiskrete Lösung von (10.1) 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 (10.1) gehörende *semidiskrete* Problem ist gegeben durch die Aufgabe: Finde Funktionen $\mathbf{u}_S \in C^1\left([0,T], \mathbb{R}^N\right)$, so dass für alle $t \in [0,T]$ gilt

$$\begin{gathered}\sum_{j=1}^{N} \int_\Omega \dot{u}_{S,j} b_j b_i + u_{S,j} \langle \nabla b_j, \nabla b_i \rangle = \int_\Omega f b_i \quad \forall 1 \leq i \leq N \\ u(\cdot,0) = v_S\end{gathered} \tag{10.2}$$

ist. Da $u_S(\cdot,0)$ eine Finite-Elemente-Funktion ist, muss daher auch $v_S \in S$ gelten. Falls v aus (10.1) 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 (10.2) als gewöhnliche Differentialgleichung für den zeitabhängigen Koeffizientenvektor $\mathbf{u}_S$ auffassen

$$\begin{gathered}\mathbf{M}\dot{\mathbf{u}}_S(t) + \mathbf{A}\mathbf{u}_S(t) = \mathbf{r}(t) \quad \forall t \in [0,T] \\ \mathbf{u}_S(0) = \mathbf{v}_S.\end{gathered}$$

10.2 Einschrittverfahren für die homogene Wärmeleitungsgleichung

Wir betrachten das homogene parabolische Problem

$$\begin{array}{rlll} u_t = \Delta u & \text{in } \Omega & \text{für alle } t > 0, & \\ u = 0 & \text{auf } \partial\Omega & \text{für alle } t > 0 \text{ (zeitabhängige Randbedingung)}, & \\ u(\cdot,0) = v & \text{in } \Omega & \text{Anfangsbedingung.} & \end{array} \tag{10.3}$$

Der Einfachheit halber nehmen wir für das Folgende an, dass $v \in S$ gilt.

10.2.1 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 $A = \operatorname{span}\{a_n : n \in \mathbb{N}\}$.

Konvention 10.1 *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 *endlich*dimensionale Vektorräume kennengelernt haben, sich auch auf *unendlich*dimensionale Vektorräume übertragen lassen.

Der Operator A ist **positiv definit**, falls

$$(Av, v)_H > 0 \qquad \forall v \in W \setminus \{0\}$$

gilt. Der Operator A ist **selbstadjungiert**[15], falls

$$(Av, w)_H = (v, Aw)_H \qquad \forall v, w \in W.$$

Konvention 10.2 *Generell nehmen wir an, dass A ein linearer, selbstadjungierter, positiv definiter, nicht-notwendigerweise beschränkter Operator auf einem Teilraum $W \subset H$ ist, dessen Inverse $A^{-1} : H \to H$ 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 10.3 *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 10.4 *Für eine beliebige Funktion g, die auf dem Spektrum $\sigma(A)$ von A definiert ist, setzen wir*

$$g(A)\, v = \sum_{j=1}^{N} g(\lambda_j)\, (v, \varphi_j)_H\, \varphi_j \quad \forall v \in H. \tag{10.4}$$

Bemerkung 10.5 *Satz 10.3 und Definition 10.4 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.*

[15] 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 $\mathbf{A} = \mathbf{A}^\intercal$.

10.2.2 Zeitdiskretisierung abstrakter Evolutionsprobleme

Wir formulieren das Ausgangsproblem (10.1) zunächst als abstraktes Problem.

Sei H ein separabler Hilbert-Raum, und wir nehmen an, die Konventionen 10.1 und 10.2 sind erfüllt. Wir betrachten das Anfangswertproblem

$$u' + Au = 0 \qquad \text{für } t > 0 \quad \text{mit} \quad u(0) = v. \tag{10.5}$$

Satz 10.6 *Die Lösung des Problems (10.5) besitzt die Darstellung*

$$u(t) = E(t)\, v = \sum_{j=1}^{N} e^{-\lambda_j t} (v, \varphi_j)_H \varphi_j. \tag{10.6}$$

Im Sinn von Definition 10.4 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 $n = 0, 1, 2, \ldots$ die Zeitpunkte $t_n := nk$ definiert.

Die Lösung (10.5) 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 10.7 *Da im allgemeinen das Eigensystem der Operators A nicht explizit bekannt ist, ist die Darstellung (10.6) 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 Anwendunge $g(A)\, v$ durch Berechnung von

$$B := \sum_{i=0}^{n} \alpha_i A^i, \quad C := \sum_{i=0}^{m} \beta_i A^i \quad \textit{und} \quad C^{-1} B v$$

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

$$U^{n+1} := E_k U^n \quad \text{für } n = 0, 1, 2, \ldots \tag{10.7}$$

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

$$\forall t > 0: \quad u' + au = 0 \quad \text{mit} \quad u(0) = 1. \tag{10.8}$$

Die zugehörige diskrete Lösung ist dann durch

$$U^{n+1} = r(ka)\, U^n \tag{10.9}$$

gegeben. Die Methode besitzt die Konsistenzordnung q, falls die exakte Lösung von (10.8), eingesetzt in (10.9), eine Genauigkeit von $O(k^{q+1})$ besitzt. Da die exakte Lösung von (10.8) durch e^{-at} gegeben ist, kann diese Aussage auch gemäss

$$r(ka) = e^{-ka} + O\left(k^{q+1}\right)$$

beschrieben werden, bzw. durch

$$r(\lambda) = e^{-\lambda} + O\left(\lambda^{q+1}\right) \quad \text{für } \lambda \to 0. \tag{10.10}$$

Neben der Konsistenz des Verfahrens spielt dessen Stabilität eine wesentliche Rolle. Mit der Spektraldarstellung (10.4) folgert man

$$U^n = E_k^n v = \sum_{j=1}^{N} r(k\lambda_j)^n (v, \varphi_j)_H \varphi_j. \tag{10.11}$$

Der Lösungsoperator E_k wird stabil im Hilbert-Raum H genannt, falls $\|E_k^n\|_H \le C$ für $n \ge 1$ gilt. Diese Bedingung lässt sich auch alternativ beschreiben. Da (φ_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

$$\begin{aligned} \|E_k^n\|_H &= \sup_{v\in H\setminus\{0\}} \frac{\|E_k^n v\|_H}{\|v\|_H} = \sqrt{\sup_{v\in H\setminus\{0\}} \frac{\left\|\sum_{j=1}^N r(k\lambda_j)^n (v,\varphi_j)_H \varphi_j\right\|_H^2}{\sum_{i=1}^N (v,\varphi_i)_H^2}} \\ &= \sqrt{\sup_{v\in H\setminus\{0\}} \frac{\sum_{j=1}^N r(k\lambda_j)^{2n} (v,\varphi_j)_H^2}{\sum_{i=1}^N (v,\varphi_i)_H^2}} = \left(\max_{1\le j\le N} |r(k\lambda_j)|\right)^n . \end{aligned} \tag{10.12}$$

Die Bedienung $\|E_k^n\|_H \le C$ für alle n ist daher äquivalent zur Bedingung

$$\max_{\lambda\in\sigma(kA)} |r(\lambda)| \le 1. \tag{10.13}$$

Daraus folgt mit der Darstellung (10.11)

$$\|U^n\|_H^2 \le \sum_{i=1}^{N} (v,\varphi_i)_H^2 = \|v\|_H^2 .$$

Die Stabilitätsbedingung (10.13) wird für alle folgenden Methoden erfüllt sein.

Wir illustrieren dieses abstrakte Vorgehen für zwei weitverbreitete Zeitdiskretisierungsmethoden für die Wärmeleitungsgleichung. Für das ortsdiskretisierte Problem gilt $H = \overset{\circ}{S}$ und $A = -\Delta_S$ und lautet: Finde $u \in \overset{\circ}{S} \times \mathbb{R}_{>0}$, so dass für alle $t > 0$ gilt:

$$\int_\Omega \dot{u}(x,t)\,\chi(x)\,dx = -\int_\Omega \langle \nabla u(x,t), \nabla\chi(x)\rangle\,dx \qquad \forall \chi \in \overset{\circ}{S}$$

mit den Anfangswerten $u(\cdot,0) = v$.

Das implizite Euler-Verfahren ist gegeben durch

$$\left(U^{n+1},\chi\right)_{L^2(\Omega)} + k\left(\nabla U^{n+1}, \nabla\chi\right)_{L^2(\Omega)} = \left(U^n,\chi\right)_{L^2(\Omega)} \qquad \forall \chi \in \overset{\circ}{S},\quad n \geq 0 \tag{10.14}$$

und das Crank-Nicolson-Verfahren durch

$$\left(U^{n+1},\chi\right)_{L^2(\Omega)} + \frac{1}{2}k\left(\nabla U^{n+1}, \nabla\chi\right)_{L^2(\Omega)} = \left(U^n,\chi\right)_{L^2(\Omega)} - \frac{1}{2}k\left(\nabla U^n, \nabla\chi\right)_{L^2(\Omega)}$$

für alle $\chi \in \overset{\circ}{S}$, $n \geq 0$. In kompakter Operatorschreibweise gilt für (10.14)

$$\left(I - k\Delta_S\right)U^{n+1} = U^n \quad \text{bzw.} \quad U^{n+1} = \left(I - k\Delta_S\right)^{-1}U^n$$

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 (10.11) 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 (10.13) 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

$$\|v\|_s := (A^s v, v)^{1/2}_{L^2(\Omega)} = \left\|A^{s/2}v\right\|_{L^2(\Omega)} = \left\{ \sum_{j=1}^N \lambda_j^s \left(v,\varphi_j\right)^2_{L^2(\Omega)} \right\}^{1/2}$$

definiert.

Hilfssatz 10.8 *Das Verfahren sei von Konsistenzordnung q, so dass (10.10) gilt. Dann gilt*

1. $\exists \lambda_0, C > 0$:
$$\left| r(\lambda) - e^{-\lambda} \right| \leq C\lambda^{q+1} \qquad \forall 0 \leq \lambda \leq \lambda_0.$$

2. $\exists \lambda_1 > 0$, $\exists c \in]0,1[$:
$$|r(\lambda)| \leq e^{-c\lambda} \qquad \forall 0 \leq \lambda \leq \lambda_1. \tag{10.15}$$

3. *Falls* $|r(\lambda)| < 1$ *für alle* $0 \leq \lambda \leq \overline{\lambda}$ *gilt, kann in (10.15)* $\lambda_1 \geq \overline{\lambda}$ *gewählt werden.*

Beweis. Übungsaufgabe. ■

Satz 10.9 *Das Verfahren sei von Konsistenzordnung q und stabil, so dass (10.10) und (10.13) gelten. Dann erfüllen die Lösungen von (10.5) und (10.7) die Abschätzung*

$$\left\| U^n - u(t_n) \right\|_{L^2(\Omega)} \leq Ck^q \left\| v \right\|_{2q}, \qquad \forall t_n \geq 0.$$

Beweis. Wir führen die Funktion $F_n(\lambda) = r(\lambda)^n - e^{-n\lambda}$ und erinnern an die Definition (10.4). Daher gilt
$$U^n - u(t_n) = r(kA)^n v - e^{-nkA} v = F_n(kA) v. \tag{10.16}$$
Die Behauptung lässt sich in der Form schreiben

$$\left\| F_n(kA) v \right\|_{L^2(\Omega)} \leq Ck^q \left\| v \right\|_{2q} \qquad \forall t_n \geq 0.$$

Falls $F_n(kA)$ als Operator von H^{2q} nach $L^2(\Omega)$ aufgefasst wird, ist diese Abschätzung äquivalent zur Operatornormabschätzung

$$\begin{aligned}
\left\| F_n(kA) \right\|_{L^2(\Omega) \leftarrow H^{2q}} &= \sup_{v \in H^{2q} \setminus \{0\}} \frac{\left\| F_n(kA) v \right\|_{L^2(\Omega)}}{\left\| v \right\|_{2q}} = \sup_{v \in H^{2q} \setminus \{0\}} \frac{\left\| F_n(kA) v \right\|_{L^2(\Omega)}}{\left\| A^q v \right\|_{L^2(\Omega)}} \\
&= \sup_{v \in L^2 \setminus \{0\}} \frac{\left\| F_n(kA) A^{-q} v \right\|_{L^2(\Omega)}}{\left\| v \right\|_{L^2(\Omega)}} = \left\| F_n(kA) A^{-q} \right\|_{L^2(\Omega) \leftarrow L^2(\Omega)} \\
&= \left\| A^{-q} F_n(kA) \right\|_{L^2(\Omega) \leftarrow L^2(\Omega)} \overset{!}{\leq} Ck^q \qquad \forall t_n \geq 0.
\end{aligned}$$

Analog wie (10.12) ergibt sich

$$\left\| A^{-q} F_n(kA) \right\|_{L^2(\Omega) \leftarrow L^2(\Omega)} = \sup_{\lambda \in \sigma(A)} \left| \lambda^{-q} F_n(k\lambda) \right| = k^q \sup_{\lambda \in \sigma(kA)} \left| \lambda^{-q} F_n(\lambda) \right| \overset{!}{\leq} Ck^q.$$

Daher zeigen wir im folgenden

$$|F_n(\lambda)| \leq C\lambda^q \qquad \forall \lambda \in \sigma(kA).$$

Die Abschätzung (10.10) ergibt für hinreichend kleines $\lambda_0 > 0$

$$\left| r(\lambda) - e^{-\lambda} \right| \leq C\lambda^{q+1} \qquad \forall 0 \leq \lambda \leq \lambda_0.$$

Die Dreiecksungleichung liefert

$$|r(\lambda)| \leq \left| e^{-\lambda} - r(\lambda) \right| + \left| e^{-\lambda} \right| \leq C\lambda^{q+1} + e^{-\lambda}.$$

Eine Taylorentwicklung um $\lambda = 0$ und festes $c \in]0,1[$ liefert

$$\begin{aligned} C\lambda^{q+1} + e^{-\lambda} - e^{-c\lambda} &= C\lambda^{q+1} + (1-\lambda) - (1-c\lambda) + O\left(\lambda^2\right) \\ &= (c-1)\,\lambda + O\left(\lambda^2\right). \end{aligned}$$

Aus $c - 1 < 0$ folgt die Existenz eines $\lambda_1 > 0$, so dass

$$C\lambda^{q+1} + e^{-\lambda} - e^{-c\lambda} \leq 0 \qquad \forall 0 \leq \lambda \leq \lambda_1.$$

Das impliziert

$$|r(\lambda)| \leq e^{-c\lambda} \qquad \forall 0 \leq \lambda \leq \lambda_1. \tag{10.17}$$

Setzen wir $\overline{\lambda} := \min\{\lambda_0, \lambda_1\}$, erhalten wir

$$|F_n(\lambda)| = \left| \left(r(\lambda) - e^{-\lambda}\right) \sum_{j=0}^{n-1} \underbrace{r(\lambda)^{n-1-j}}_{\leq e^{-c\lambda(n-1-j)}} e^{-j\lambda} \right| \leq Cn\lambda^{q+1} e^{-c(n-1)\lambda} \leq \tilde{C}\lambda^q \tag{10.18}$$

für alle $0 \leq \lambda \leq \overline{\lambda}$. Aus der Stabilität von r folgt für $\lambda \geq \overline{\lambda}$ und $\lambda \in \sigma(kA)$:

$$|F_n(\lambda)| \leq |r(\lambda)^n| + e^{-n\lambda} \leq 2 \leq C\lambda^q$$

mit $C = 2\overline{\lambda}^{-q}$. ■

Satz 10.9 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. $|r(\lambda)| < 1 \qquad \forall 0 < \lambda < \alpha \qquad$ für ein $\alpha > 0$,

II. $|r(\lambda)| < 1 \qquad \forall 0 < \lambda < \infty$,

III. $|r(\lambda)| < 1 \qquad \forall 0 < \lambda$ und $|r(\infty)| < 1$.

IV. $|r(\lambda)| < 1 \qquad \forall 0 < \lambda$ 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 (10.13) auftretende λ 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 10.10 *Für Verfahren vom Typ* I' *bzw.* II', *welche die Konsistenzbedingung (10.10) für ein $q \geq 1$ erfüllen, gilt mit $\lambda_1 = \alpha_0$ bzw. $\lambda_1 = \alpha_1$*

$$|r(\lambda)| < 1 \qquad \forall 0 < \lambda \leq \lambda_1$$

und daher auch (10.17).

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

$$\left\{ \frac{p}{q} : p \in \mathbb{P}_n,\, q \in \mathbb{P}_m \backslash \{0\} \right\}$$

kein Vektorraum ist.) Eine optimale Approximation der Funktion $e^{-\lambda}$ ist durch die Padé-Approximation gegeben:

$$r_{m,n}(\lambda) = \frac{p_{m,n}(\lambda)}{d_{m,n}(\lambda)} \qquad \text{mit} \qquad p_{m,n} \in \mathbb{P}_n \text{ und } d_{m,n} \in \mathbb{P}_m.$$

Genauer gilt explizit

$$p_{m,n}(\lambda) := \sum_{j=0}^{n} \tbinom{n}{j} / \tbinom{n+m}{j} \frac{(-\lambda)^j}{j!},$$
$$d_{m,n}(\lambda) = \sum_{j=0}^{m} \tbinom{m}{j} / \tbinom{n+m}{j} \frac{\lambda^j}{j!}.$$

Die Wahl der Koeffizienten ist dadurch bestimmt, dass in der Taylorentwicklung von $e^{-\lambda}$ um 0 möglichst viele Terme übereinstimmen. Es gilt

$$r_{m,n}(\lambda) = e^{-\lambda} + O\left(\lambda^{m+n+1}\right) \qquad \text{für } \lambda \to 0,$$

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 10.11 *Das Zeitschrittverfahren besitze die Konsistenzordnung q und sei vom Typ I', II' oder III. Dann erfüllt die Lösungen von (10.5) und (10.7) die Fehlerabschätzung*

$$\|U^n - u(t_n)\|_{L^2(\Omega)} \leq Ck^q t_n^{-q} \|v\|_{L^2(\Omega)} \quad \textit{für } t_n > 0.$$

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 α_0 und α_1 ab.

Beweis. Wir verwenden die Notationen des Beweises von Satz 10.9 und müssen daher zeigen, dass

$$\|F_n(kA)\|_{L^2(\Omega)\leftarrow L^2(\Omega)} \leq Ck^q t_n^{-q} \qquad \text{für } t_n > 0.$$

Mit Hilfe der Spektraldarstellung folgert man, dass dies äquivalent ist zur folgenden Abschätzung

$$|F_n(\lambda)| \leq Ck^q t_n^{-q} = Cn^{-q} \qquad \forall \lambda \in \sigma(kA), \quad n \geq 1. \tag{10.19}$$

Wir erinnern daran (vgl. Bemerkung 10.10), dass (10.17) 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 (10.18)

$$|F_n(\lambda)| \leq Cn^{-q}(n\lambda)^{q+1} e^{-cn\lambda} \leq Cn^{-q} \quad \forall 0 \leq \lambda \leq \lambda_1.$$

Für Verfahren vom Typ I' und II' ist damit der Beweis von (10.19) abgeschlossen, da in diesem Fall $k\lambda_{\max} \leq \lambda_1$ gilt. Für Verfahren vom Typ III müssen wir auch grosse Werte von λ betrachten. Für $\lambda \geq \lambda_1 = 1$ (man beachte, dass (10.17) gilt für beliebig grosses, aber festes λ_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 10.11 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

$$0 = t_0 < t_1 < \ldots < t_n < \ldots$$

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 (10.5) ist definiert durch

$$\overline{\partial}_n U^n + AU^n = 0 \quad \forall n \geq 1 \quad \text{mit} \quad U^0 = v. \tag{10.20}$$

Hier bezeichnet $\overline{\partial}_n U^n$ die Rückwärtsdifferenz $(U^n - U^{n-1})/k_n$.

Satz 10.12 *Für die Lösungen (10.20) und (10.5) 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

$$\overline{\partial}_n \eta^n + A\eta^n = \omega^n := -\overline{\partial}_n u^n - Au^n = \dot{u}^n - \overline{\partial}_n u^n.$$

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

$$\|\eta^n\|_{L^2(\Omega)} \leq \sum_{j=1}^{n} k_j \left\|\omega^j\right\|_{L^2(\Omega)}.$$

Die Behauptung folgt schliesslich aus

$$\left\|\omega^j\right\|_{L^2(\Omega)} = \left\|\dot{u}\left(t_j\right) - \overline{\partial}_j u\left(t_j\right)\right\|_{L^2(\Omega)} \leq \int_{J_j} \|\ddot{u}\|_{L^2(\Omega)}\, dt.$$

■

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:

$$\dot{H}^s(\Omega) := \left\{v \in H^s(\Omega) \mid \forall j < s/2 : \quad \Delta^j v = 0 \text{ im Sinne der Spurbildung}\right\}.$$

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 $T_S : L^2(\Omega) \to S$ bezeichnet den Lösungsoperator, der jeder rechten Seite $f \in L^2(\Omega)$ die eindeutige Galerkin-Lösung mittels

$$a\left(T_S f, v\right) = \left(f, v\right)_{L^2(\Omega)} \quad \forall v \in S$$

zuordnet. Analog ist der kontinuierliche Lösungsoperator $T : L^2(\Omega) \to H_0^1(\Omega)$ durch

$$a\left(Tf, v\right) = \left(f, v\right)_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega)$$

definiert. Man beachte, dass die Annahme $f \in L^2(\Omega)$ und die Voraussetzung, dass Ω 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 10.13

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*

$$\|(T_S - T) f\|_{L^2(\Omega)} \leq C h^s \|f\|_{H^{s-2}(\Omega)} \qquad \forall 2 \leq s \leq r \quad \forall f \in H^{s-2}(\Omega).$$

Wir verwenden die Notationen:
U^n: volldiskrete Lösung,
$u_S(t)$: semidiskrete Lösung von (10.2),
$u(t)$: exakte Lösung der parabolischen Differentialgleichung.

Satz 10.14 *Es gelte Annahme 10.13. 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. $\|v_S - v\|_{L^2(\Omega)} \leq C h^r |v|_r$ *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 10.15 *Es gilt*

$$v = \sum_{j=0}^{q-1} T_S^j (T - T_S) (-\Delta)^{j+1} v + T_S^q (-\Delta)^q v \quad \forall v \in \dot{H}^{2q}(\Omega). \tag{10.21}$$

Beweis. Wegen $T(-\Delta) = I$ folgt die Gleichheit durch einfaches Ausmultiplizieren. ■

Lemma 10.16 *Sei die Zeitdiskretisierung ein Einschrittverfahren vom Typ* I′ *oder* II. *Dann gilt*

$$\left\|F_n(-k\Delta_S) P_S T_S^j\right\|_{L^2(\Omega) \leftarrow L^2(\Omega)} \leq \left\|F_n(-k\Delta_S) P_S T_S^j\right\|_{S \leftarrow S} \leq C k^j \qquad \forall 0 \leq j \leq q, \quad \forall n \geq 0 \tag{10.22}$$

mit F_n *wie in (10.16) und der (bezüglich* $(\cdot,\cdot)_{L^2(\Omega)}$*) orthogonalen Projektion* $P_S : L^2(\Omega) \to S$,

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 $v \in S$

$$\begin{aligned}\left\|F_n(-k\Delta_S) P_S T_S^j v\right\|_{L^2(\Omega)} &= \left\| \sum_{i=1}^N (v, \varphi_i)_{L^2(\Omega)} F_n(-k\Delta_S) \lambda_i^{-j} \varphi_i \right\|_{L^2(\Omega)} \\ &= \left\| k^j \sum_{i=1}^N (v, \varphi_i)_{L^2(\Omega)} F_n(k\lambda_i) (k\lambda_i)^{-j} \varphi_i \right\|_{L^2(\Omega)} \leq k^j \sup_{\lambda \in \sigma(-k\Delta_S)} \left|\lambda^{-j} F_n(\lambda)\right| \|v\|_{L^2(\Omega)}.\end{aligned}$$

Sei λ_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 λ gilt

$$\left|r(\lambda) - e^{-\lambda}\right| \leq C\lambda^{j+1} \qquad 0 \leq j \leq q \quad \text{und} \quad |r(\lambda)| \leq e^{-c\lambda} \quad \text{für ein geeignetes } c \in]0,1[.$$

Daraus folgt für alle $0 < \lambda \leq \lambda_0$

$$\left|\lambda^{-j} F_n(\lambda)\right| = \left|\lambda^{-j}\left(r(\lambda) - e^{-\lambda}\right)\right| \left|\sum_{\ell=0}^{n-1} r(\lambda)^{n-1-\ell} e^{-\ell\lambda}\right| \leq Cn\lambda e^{-cn\lambda} \leq C.$$

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 (10.22) gezeigt. Die linke Ungleichung folgt wegen

$$\begin{aligned}\sup_{v\in L^2(\Omega)\setminus\{0\}} \frac{\left\|F_n(-k\Delta_S) P_S T_S^j v\right\|_{L^2(\Omega)\leftarrow L^2(\Omega)}}{\|v\|_{L^2(\Omega)}} &= \sup_{v\in L^2(\Omega)\setminus\{0\}} \frac{\left\|F_n(-k\Delta_S) P_S T_S^j P_S v\right\|_{L^2(\Omega)\leftarrow L^2(\Omega)}}{\|v\|_{L^2(\Omega)}} \\ &= \left(\sup_{w\in S\setminus\{0\}} \frac{\left\|F_n(-k\Delta_S) P_S T_S^j w\right\|_{L^2(\Omega)\leftarrow L^2(\Omega)}}{\|w\|_{L^2(\Omega)}}\right)\left(\sup_{v\in L^2(\Omega)\setminus\{0\}} \frac{\|P_S v\|_{L^2(\Omega)}}{\|v\|_{L^2(\Omega)}}\right).\end{aligned}$$

Da P_S die $L^2(\Omega)$-orthogonal-Projektion ist, ist der letzte Faktor nach oben durch 1 beschränkt.

■

Beweis von Satz 10.14:

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 (10.2) 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.

$$\|v - v_k\|_{L^2(\Omega)} = \left\|\sum_{\ell: k\lambda_\ell \geq 1} (v, \varphi_\ell)_{L^2(\Omega)} \varphi_\ell\right\|_{L^2(\Omega)} = \left\|\sum_{\ell: k\lambda_\ell \geq 1} (v, \varphi_\ell)_{L^2(\Omega)} \lambda_\ell^{-q} (-\Delta)^q \varphi_\ell\right\|_{L^2(\Omega)} \tag{10.23a}$$

$$= \left\|\sum_{\ell: k\lambda_\ell \geq 1} (v, \varphi_\ell)_{L^2(\Omega)} \lambda_\ell^{-q} \varphi_\ell\right\|_{2q} = \left(\sup_{k\lambda_\ell \geq 1} \lambda_\ell^{-q}\right) \|v\|_{2q} \leq k^q \|v\|_{2q}. \tag{10.23b}$$

2.

$$|v_k|_{2q} \leq |v|_{2q} \qquad \text{da } v_k \text{ und } v - v_k \text{ orthogonal sind,} \tag{10.23c}$$

3.

$$|v_k|_{r+2j} = \left\| (-\Delta)^{r/2+j} v_k \right\|_{L^2(\Omega)} = \left\| \sum_{\ell : k\lambda_\ell < 1} (v, \varphi_\ell)_{L^2(\Omega)} \lambda_\ell^{r/2+j} \varphi_\ell \right\|_{L^2(\Omega)} \tag{10.23d}$$

$$\leq \sup_{\ell : k\lambda_\ell < 1} \lambda_\ell^j \left\| \sum_{\ell : k\lambda_\ell < 1} (v, \varphi_\ell)_{L^2(\Omega)} \lambda_\ell^{r/2} \varphi_\ell \right\|_{L^2(\Omega)} \leq \left(\sup_{\ell : k\lambda_\ell < 1} \lambda_\ell^j \right) \|v\|_r \leq k^{-j} \|v\|_r , \tag{10.23e}$$

für alle $j = 0, 1, \ldots, q-1$.

Wir wenden nun die Identität (10.21) 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 10.16 und (10.23c) folgt

$$\|F_n T_S^q (-\Delta)^q v_k\|_{L^2(\Omega)} \leq Ck^q \|\Delta^q v_k\|_{L^2(\Omega)} = Ck^q |v_k|_{2q} \overset{(10.23c)}{\leq} Ck^q |v|_{2q} .$$

Indem wir die Eigenschaft aus Annahme 10.13(2) ausnützen und mit (10.23e) verbinden, ergibt sich

$$\begin{aligned} \left\| F_n T_S^j (T - T_S) (-\Delta)^{j+1} v_k \right\|_{L^2(\Omega)} &\overset{\text{Lem. 10.16}}{\leq} Ck^j \left\| (T - T_S) (-\Delta)^{j+1} v_k \right\|_{L^2(\Omega)} \\ &\overset{\text{Ann. 10.13(2)}}{\leq} Ck^j h^r \left\| (-\Delta)^{j+1} v_k \right\|_{H^{r-2}(\Omega)} \\ &\leq Ck^j h^r |v_k|_{r+2j} \overset{(10.23e)}{\leq} Ch^r |v_k|_r \end{aligned}$$

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 (10.23b) ergeben

$$\|F_n (v - v_k)\|_{L^2(\Omega)} \leq 2 \|v - v_k\|_{L^2(\Omega)} \overset{(10.23b)}{\leq} Ck^q |v|_{2q} .$$

Abschätzung von $\|u_S(t_n) - u(t_n)\|_{L^2(\Omega)}$
Diese Abschätzung ergibt sich als Folge von
Hilfsaussage 1: Es gelte

$$T_S \dot{e} + e = \rho \quad \text{für } t \geq 0 \text{ mit } T_S e(0) = 0. \tag{10.24}$$

Dann gilt

$$\|e(t)\|^2_{L^2(\Omega)} \leq C \|\rho(t)\|^2_{L^2(\Omega)} + \frac{1}{t}\int_0^t \left(\|\rho\|^2_{L^2(\Omega)} + s^2 \|\dot{\rho}\|^2_{L^2(\Omega)}\right) ds.$$

Beweis der Hilfsaussage 1:
Wir multiplizieren (10.24) skalar mit $2\dot{e}$ und erhalten

$$2(T_S\dot{e}, \dot{e})_{L^2(\Omega)} + \frac{d}{dt}\|e\|^2_{L^2(\Omega)} = 2(\rho, \dot{e})_{L^2(\Omega)}.$$

Da T_S nach Annahme positiv semidefinit ist, folgt

$$\frac{d}{dt}\|e\|^2_{L^2(\Omega)} \leq 2(\rho, \dot{e})_{L^2(\Omega)} = 2\frac{d}{dt}(\rho, e)_{L^2(\Omega)} - 2(\dot{\rho}, e).$$

Wir multiplizieren mit t und erhalten

$$\frac{d}{dt}\left(t\|e\|^2_{L^2(\Omega)}\right) \leq 2\frac{d}{dt}\left(t(\rho, e)_{L^2(\Omega)}\right) - 2t(\dot{\rho}, e) + \|e\|^2_{L^2(\Omega)} - 2(\rho, e)_{L^2(\Omega)}.$$

Integration nach t liefert

$$t\|e\|^2_{L^2(\Omega)} \leq 2t\|\rho\|_{L^2(\Omega)}\|e\|_{L^2(\Omega)} + \int_0^t \left(2s\|\dot{\rho}\|_{L^2(\Omega)}\|e\|_{L^2(\Omega)} + \|e\|^2_{L^2(\Omega)} + 2\|\rho\|_{L^2(\Omega)}\|e\|_{L^2(\Omega)}\right) ds,$$

d.h.

$$\|e\|^2_{L^2(\Omega)} \leq C\left(\|\rho(t)\|^2_{L^2(\Omega)} + \frac{1}{t}\int_0^t s^2\|\dot{\rho}\|^2_{L^2(\Omega)} + \|e\|^2_{L^2(\Omega)} + \|\rho\|^2_{L^2(\Omega)}\right). \tag{10.25}$$

Multipliziert man (10.24) mit e, ergibt sich

$$\frac{d}{dt}(T_S e, e)_{L^2(\Omega)} + 2\|e\|^2_{L^2(\Omega)} = \|\rho\|^2_{L^2(\Omega)} + \|e\|^2_{L^2(\Omega)}.$$

Integration liefert

$$(T_S e, e)_{L^2(\Omega)} + \int_0^t \|e\|^2_{L^2(\Omega)} ds \leq (T_S e(0), e(0))_{L^2(\Omega)} + \int_0^t \|\rho\|^2_{L^2(\Omega)} ds.$$

Da T_S positiv semidefinit ist und $T_S e(0)$ wegen (10.24) verschwindet folgt

$$\int_0^t \|e\|^2_{L^2(\Omega)} ds \leq \int_0^t \|\rho\|^2_{L^2(\Omega)} ds.$$

Zusammen mit (10.25) ergibt sich die Hilfsbehauptung 1

$$\|e\|^2_{L^2(\Omega)} \leq C\left(\|\rho(t)\|^2_{L^2(\Omega)} + \frac{1}{t}\int_0^t s^2\|\dot{\rho}\|^2_{L^2(\Omega)} + \|\rho\|^2_{L^2(\Omega)}\right).$$

Aus Hilfsbehauptung 1 folgt direkt

$$\|e(t)\|_{L^2(\Omega)} \leq C \sup_{s\leq t}\left(s\|\dot{\rho}(s)\|_{L^2(\Omega)} + \|\rho(s)\|_{L^2(\Omega)}\right) \quad \forall t > 0. \tag{10.26}$$

Der Fehler $e(t) = u_S(t) - u(t)$ erfüllt die Gleichung (10.24) mit

$$\rho(t) = -(T_S - T)\Delta u = -(T_S - T)\dot{u}.$$

Weiter gilt für die Anfangsbedingung in (10.24) mit Annahme 10.13(1) (genauer: $T_S : S \to S$ 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 (10.26) anwenden, und es bleibt die beiden ρ-Terme auf der rechten Seite anzuschätzen. Für die $L^2(\Omega)$-Norm von ρ erhalten wir

$$\begin{aligned}\|\rho(s)\|_{L^2(\Omega)} &\leq \|(T_S - T)\dot{u}\|_{L^2(\Omega)} \overset{\text{Ann. } 10.13(2)}{\leq} Ch^r \|\dot{u}\|_{H^{r-2}(\Omega)} \\ &\leq Ch^r \|u(s)\|_{H^r(\Omega)}\end{aligned}$$

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

$$u(s) = E(s)v = \sum_{i=1}^{N} (v, \varphi_i)_{L^2(\Omega)} e^{-\lambda_i s} \varphi_i \tag{10.27}$$

gegeben. Die Orthonormalität der Eigenvektorbasis und $\left|e^{-\lambda_i s}\right| \leq 1$ 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. ■

11 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 10.2.2 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 11.1 *In unserer Anwendung gilt $H = L^2(\Omega)$ oder $H = S \subset H_0^1(\Omega)$. Für hinreichend glatte*[16] *Funktionen $u, v : [0,T] \times L^2(\Omega) \to \mathbb{R}$ erhalten wir mit $t_N := T$*

$$\begin{aligned}
\frac{d}{dt}\int_\Omega u(t,x)\, v(t,x)\, dx &= \int_\Omega \dot{u}(t,x)\, v(t,x) + u(t,x)\, \partial_t v(t,x)\, dx,\\
\int_0^{t_N}\int_\Omega v(t,x)\, \dot{u}(t,x)\, dxdt &= \int_\Omega \int_0^{t_N} v(t,x)\, \dot{u}(t,x)\, dtdx\\
&= -\int_\Omega \int_0^{t_N} u(t,x)\, \dot{v}(t,x)\, dtdx + \int_\Omega \left(v(t,x)\, u(t,x)|_{t=0}^{t_N}\right) dx.
\end{aligned}$$

bzw. kompakter geschrieben:

$$\frac{d}{dt}(u,v)_H = (\dot{u},v)_H + (u,\dot{v})_H \tag{11.1a}$$

$$\int_0^{t_N} (\dot{u},v)\, dt = -\int_0^{t_N} (u,\dot{v})_H + \left((u,v)_H|_{t=0}^{t_N}\right). \tag{11.1b}$$

Im folgenden nehmen wir generell an, dass die Relationen (11.1) erfüllt sind.

Wir betrachten das Anfangswertproblem:

$$\dot{u} + Au = f \quad \text{für } t > 0 \text{ mit } u(0) = v. \tag{11.2}$$

Um in der Zeit zu diskretisieren, zerlegen wir die Zeitachse gemäss

$$0 = t_0 < t_1 < \ldots < t_n < \ldots$$

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 gegebener maximaler Polynomgrad für den Lösungsansatz auf den einzelnen Zeitintervallen. Die

[16] Hinreichend glatt in dem Sinne, dass die Vertauschung von Differentiation und Integration zulässig ist und auch partielle Integration angewendet werden kann.

Koeffizienten dieses Ansatzes sind in H, d.h., der Ansatz liegt im Raum

$$S := \left\{ u : [0,\infty[\to H : \forall J_n : \exists \left(\psi_j\right)_{j=0}^{q} \in H^{q+1} : u\left(t\right) = \sum_{j=0}^{q} \psi_j t^j, \quad \forall t \in J_n \right\}.$$

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 $u\left(0\right)$ 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 (11.2) für glatte Testfunktionen $w : [0, t_N] \to H$ die folgende Gleichung erfüllt:

$$\int_0^{t_N} \left(\dot{u}, w\right)_H + \left(Au, w\right)_H dt = \int_0^{t_N} \left(f, w\right)_H dt.$$

Nach partieller Integration bezüglich der Zeit erhalten wir für alle Funktionen w mit $w\left(t_N\right) = 0$

$$\int_0^{t_N} -\left(u, \dot{w}\right)_H + \left(Au, w\right)_H dt + \left(u\left(t\right), w\left(t\right)\right)_H \Big|_{t=0}^{t_N} = \int_0^{t_N} \left(f, w\right)_H dt$$

d.h.

$$\int_0^{t_N} -\left(u, \dot{w}\right)_H + \left(Au, w\right)_H dt = \left(u\left(0\right), w\left(0\right)\right)_H + \int_0^{t_N} \left(f, w\right)_H dt. \tag{11.3}$$

Wir ersetzen nun u in der schwachen Formulierung (11.3) durch eine Funktion $U \in S$ und integrieren partiell auf jedem Zeitintervall J_n :

$$\sum_{n=1}^{N} \int_{J_n} -\left(U, \dot{w}\right)_H dt = \int_0^{t_N} \left(\partial_{\mathrm{stw}} U, w\right)_H dt + \sum_{n=1}^{N-1} \left(\left[U\right]_n, w_+^n\right)_H + \left(U_+^0, w_+^0\right)_H, \tag{11.4}$$

wobei der Sprung der Funktion U im Gitterpunkt t_n durch $U^n = U_+^n - U^n$ definiert ist. In (11.4) 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 ∂_{stw} ist für $t \in \overset{\circ}{J}_n$ durch

$$\partial_{\mathrm{stw}} U\left(t\right) = \dot{U}\left(t\right)$$

und auf (der Nullmenge) der Gitterpunkte beliebig definiert. Beispielsweise gilt für $q = 0$ und jedes $V \in S$ die Gleichung $\partial_{\mathrm{stw}} V = 0$.

Die unstetige Galerkin-Diskretisierung des Problems (11.2) lautet: Finde $U \in S$, so dass für alle $W \in S$ gilt

$$\int_0^{t_N} \left(\partial_{\mathrm{stw}} U, W\right)_H + \left(AU, W\right)_H dt + \sum_{n=1}^{N-1} \left(\left[U\right]_n, W_+^n\right)_H + \left(U_+^0, W_+^0\right)_H = \left(v, W_+^0\right)_H + \int_0^{t_N} \left(f, W\right)_H dt, \tag{11.5}$$

$$U^0 = v.$$

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[17]

$$\int_{J_n} (\partial_t U, W)_H + (AU, W)_H \, dt + \left(U_+^{n-1}, W_+^{n-1}\right)_H = \left(U^{n-1}, W_+^{n-1}\right)_H + \int_{J_n} (f, W)_H \, dt, \quad (11.6)$$
$$U^0 = v.$$

Im nächsten Schritt zeigen wir, dass das lokale Problem (11.6) 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:

$$\int_{J_n} \left(\dot{U}, W\right)_H + (AU, W)_H \, dt + \left(U_+^{n-1}, W_+^{n-1}\right)_H = 0 \qquad \forall W \in S^n.$$

Sei U eine homogene Lösung und wähle $W = U$ auf J_n. Dann gilt wegen $2\left(\dot{U}, U\right)_H = \frac{d}{dt}\|U\|_H^2$ (vgl. (11.1))

$$\|U^n\|_H^2 - \left\|U_+^{n-1}\right\|_H^2 + 2\int_{J_n} (AU, U)_H \, dt + 2\left\|U_+^{n-1}\right\|_H^2 = 0,$$

das heisst (Erinnerung: $\|u\|_1 = (Au, u)_H^{1/2}$)

$$\|U^n\|_H^2 + \left\|U_+^{n-1}\right\|_H^2 + 2\int_{J_n} \|U\|_1^2 \, dt = 0.$$

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

$$(U^n, W)_H + k_n (AU^n, W)_H = \left(U^{n-1}, W\right)_H + \left(\int_{J_n} f dt, W\right)_H \qquad \forall W \in S,$$
$$U^0 = v$$

bzw in Kurzschreibweise

$$(I + k_n A) U^n = U^{n-1} + \int_{J_n} f dt. \quad (11.7)$$

Gleichung (11.7) kann auch in der Form geschrieben werden

$$\overline{\partial}_n U^n + AU^n = \frac{1}{k_n} \int_{J_n} f(t) \, dt \quad \text{mit} \quad \overline{\partial}_n U^n = \frac{U^n - U^{n-1}}{k_n}.$$

[17] Hinweis: Wir betrachten Gleichung (11.5) mit Testfunktionen $W \in S$, die $W \equiv 0$ auf $[0, t_N] \setminus J_n$ erfüllen.

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 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:

$$\left(\tilde{U}_0, \psi\right)_H + k_n A\left(\tilde{U}_0, \psi\right)_H + \left(\tilde{U}_1, \psi\right)_H + \frac{1}{2} k_n A\left(\tilde{U}_1, \psi\right)_H = \left(U^{n-1}, \psi\right)_H + \left(\int_{J_n} f(t)\, dt, \psi\right)_H$$

$$\frac{1}{2} k_n A\left(\tilde{U}_0, \eta\right)_H + \frac{1}{2}\left(\tilde{U}_1, \eta\right)_H + \frac{1}{3} k_n \left(A\tilde{U}_1, \eta\right)_H = \left(k_n^{-1} \int_{J_n} (t - t_n) f(t)\, dt, \eta\right)_H$$

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 11.2 *Die Lösungen von (11.6) (mit $q \geq 0$) und (11.2) erfüllen die Fehlerabschätzung*

$$\left\| U^N - u(t_N) \right\|_H \leq C \left(\sum_{n=1}^{N} k_n^{2q+2} \int_{J_n} \left\| u^{(q+1)} \right\|_1^2 dt \right)^{1/2} \qquad \text{für alle } t_N \geq 0. \tag{11.8}$$

Beweis. Wir definieren den Interpolanten $\tilde{u}$ der exakten Lösung u von (11.2) durch die Bedingungen

$$\begin{array}{ll} \tilde{u}(t_n) = u(t_n) & \forall n \geq 0, \\ \int_{J_n} (\tilde{u}(t) - u(t)) t^\ell dt = 0 & \forall 0 \leq \ell \leq q-1, \quad n \geq 1. \end{array} \tag{11.9}$$

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

$$\|\tilde{u}(t) - u(t)\|_j^2 \leq C k_n^{2q+1} \int_{J_n} \left\| u^{(q+1)} \right\|_j^2 dt \quad \forall t \in J_n, \quad j = 0, 1. \tag{11.10}$$

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 θ^n durch die rechte Seite von (11.8) zu beschränken. Aus (11.6) und (11.2) erhalten wir[18]

$$\begin{aligned} &\int_{J_n} \left(\dot{\theta}, X\right)_H + (A\theta, X)\, dt + \left([\theta]_{n-1}, X_+^{n-1}\right)_H \qquad (11.11) \\ &\quad = -\int_{J_n} (\dot{\rho}, X)_H + (A\rho, X)_H\, dt - \left([\rho]_{n-1}, X_+^{n-1}\right)_H \qquad \forall X \in S. \end{aligned}$$

Wir verwenden die Eigenschaften der Interpolierenden $\tilde{u}$, um die folgende Gleichheit herzuleiten

$$\begin{aligned} &\int_{J_n} (\dot{\rho}, X)_H\, dt + \left([\rho]_{n-1}, X_+^{n-1}\right)_H \\ &\quad = (\rho, X)|_{t_{n-1}+0}^{t_n} - \int_{J_n} \left(\rho, \dot{X}\right)_H dt + \left([\rho]_{n-1}, X_+^{n-1}\right)_H \\ &\quad = -\left(\rho_+^{n-1}, X_+^{n-1}\right)_H + \left(\rho_+^{n-1}, X_+^{n-1}\right)_H = 0 \qquad \forall X \in S. \end{aligned}$$

Wählen wir $X = 2\theta$ in (11.11) ergibt sich zunächst

$$\begin{aligned} &2\int_{J_n} \left(\dot{\theta}, \theta\right)_H dt + 2\left([\theta]_{n-1}, \theta_+^{n-1}\right)_H \\ &\quad = \|\theta^n\|_H^2 - \left\|\theta_+^{n-1}\right\|_H^2 + 2\left\|\theta_+^{n-1}\right\|_H^2 - 2\left(\theta^{n-1}, \theta_+^{n-1}\right)_H \\ &\quad \geq \|\theta^n\|_H^2 - \left\|\theta_+^{n-1}\right\|_H^2 + \left\|\theta_+^{n-1}\right\|_H^2 - \left\|\theta^{n-1}\right\|_H^2 \\ &\quad = \|\theta^n\|_H^2 - \left\|\theta^{n-1}\right\|_H^2 \end{aligned}$$

und schliesslich daraus

$$\begin{aligned} \|\theta^n\|_H^2 + 2\int_{J_n} \|\theta\|_1^2\, dt &\leq \left\|\theta^{n-1}\right\|_H^2 + 2\int_{J_n} |(A\rho, \theta)_H|\, dt \\ &\leq \left\|\theta^{n-1}\right\|_H^2 + \int_{J_n} \|\theta\|_1^2\, dt + \int_{J_n} \|\rho\|_1^2\, dt. \end{aligned}$$

Insgesamt haben wir gezeigt:

$$\|\theta^n\|_H^2 + \int_{J_n} \|\theta\|_1^2\, dt \leq \left\|\theta^{n-1}\right\|_H^2 + \int_{J_n} \|\rho\|_1^2\, dt.$$

Summation liefert unter Verwendung $\theta^0 = U^0 - \tilde{u}^0 = v - v = 0$ die Abschätzung

$$\left\|\theta^N\right\|_H^2 + \int_0^{t_N} \|\theta\|_1^2\, dt \leq \int_0^{t_N} \|\rho\|_1^2\, dt. \qquad (11.12)$$

Abschätzung (11.10) kombiniert mit (11.12) ergibt schliesslich

$$\left\|\theta^N\right\|_H^2 \leq \sum_{n=1}^{N} \int_{J_n} \|\rho\|_1^2\, dt \leq C \sum_{n=1}^{N} k^{2q+2} \int_{J_n} \left\|u^{(q+1)}\right\|_1^2 dt$$

[18] Diese Beziehung wird *Galerkin-Orthogonalität* genannt.

und damit ist der Beweis gegeben. ■

Für konstante Schrittweite $k_n = k$ vereinfacht sich die Abschätzung aus Satz 11.2 zu

$$\|U^n - u(t_n)\|_H \leq Ck^{q+1} \left(\int_0^{t_N} \left\| u^{(q+1)} \right\|_1^2 dt \right)^{1/2}.$$

Wir betonen, dass im Falle des modifizierten, impliziten Euler-Verfahrens (11.7) 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

$$\begin{array}{rl} \dot{u} - \Delta u = f & \text{in } \Omega \quad \forall t > 0, \\ u = 0 & \text{auf } \partial\Omega, \\ u(\cdot, 0) = v & \text{in } \Omega. \end{array} \tag{11.13}$$

Um technische Schwierigkeiten zu vermeiden, nehmen wir an, dass Ω ein konvexes Polygongebiet ist.

Mit S bezeichnen wir den Finite-Elemente-Raum, der über einer Triangulierung von Ω 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

$$\begin{gathered} (\dot{u}_S, \chi)_{L^2(\Omega)} + (\nabla u_S, \nabla \chi)_{L^2(\Omega)} = (f, \chi)_{L^2(\Omega)} \quad \forall \chi \in S, \quad t > 0 \\ u_S(0) = v_S, \end{gathered}$$

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:

$$\mathbf{S} = \left\{ X : [0, \infty[\to S \mid \forall n : X|_{J_n} = \sum_{j=0}^{q} X_j t^j, \quad X_j \in S \right\}.$$

Die Raum-Zeit-Diskretisierung lässt sich nun in der Form schreiben

$$B_N(U, X) = \left(v_S, X_+^0\right)_{L^2(\Omega)} + \int_0^{t_N} (f, X)_{L^2(\Omega)} \, dt \quad \forall X \in \mathbf{S}, \qquad N \geq 0, \tag{11.14}$$

wobei diesmal gilt

$$B_N(V, W) = \int_0^{t_N} (\partial_{\text{stw}} V, W)_{L^2(\Omega)} + (\nabla V, \nabla W)_{L^2(\Omega)} \, dt + \sum_{n=1}^{N-1} \left([V]_n, W_+^n \right)_{L^2(\Omega)} + \left(V_+^0, W_+^0\right)_{L^2(\Omega)} \tag{11.15}$$

$$= \int_0^{t_N} \left(-(V, \partial_{\text{stw}} W)_{L^2(\Omega)} + \langle \nabla V, \nabla W \rangle_{L^2(\Omega)} \right) dt - \sum_{n=1}^{N-1} (V^n, [W]_n)_{L^2(\Omega)} + \left(V^N, W^N\right)_{L^2(\Omega)}.$$

Die Gleichung, welche der Fehler $e = U - u$ erfüllt, lautet

$$B_N(e, X) = \left(v_S - v, X_+^0\right)_{L^2(\Omega)} \qquad \forall X \in \mathbf{S}.$$

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 $R_S : H^1 \to S$ durch

$$(\nabla R_S v, \nabla \chi)_{L^2(\Omega)} = (\nabla v, \nabla \chi)_{L^2(\Omega)} \qquad \forall \chi \in S.$$

Satz 11.3 *Die Ortsdiskretisierung basiere auf finite Elementen, die stückweise Polynome vom Grad r mit $r \geq 1$ sind. Die Lösung u des Problems (11.13) erfülle $u(t_i) \in H^{r+1}$. Dann gelten die Fehlerabschätzungen*

$$\begin{aligned} \|\nabla (u - R_S u)(t_i)\|_{L^2(\Omega)} &\leq C h^r \|u(t_i)\|_{r+1}, \\ \|(u - R_S u)(t_i)\|_{L^2(\Omega)} &\leq C h^{r+1} \|u(t_i)\|_{r+1}. \end{aligned}$$

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. (11.9)). Die Fehleranteile zum Zeitpunkt t_N werden mit θ^N und ρ^N bezeichnet.

Wir benötigen ein Stabilitätsresultat, welches wir hier lediglich zitieren.

Lemma 11.4 *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*

$$B_N(X, Z_S) = \left(X^N, \varphi\right)_{L^2(\Omega)} \qquad \forall X \in \mathbf{S}$$

die Abschätzung

$$\int_0^{t_N} \|\Delta_S Z_S\|_{L^2(\Omega)} \, dt + \sum_{n=1}^{N-1} \|[Z_S]_n\|_{L^2(\Omega)} \leq C L_N \|\varphi\|_{L^2(\Omega)}. \tag{11.16}$$

Satz 11.5 *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 (11.14) und (11.13) mit $v_S = P_S v$ die Fehlerabschätzung*[19]

$$\left\|U^N - u(t_N)\right\|_{L^2(\Omega)} \leq C L_N \max_{n \leq N} \left(h^{r+1} \|u\|_{r+1, J_n} + k_n \|\dot{u}\|_{0, J_n}\right).$$

[19] 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

$$\left\|\rho^N\right\|_{L^2(\Omega)} = \left\|(R_S u - u)(t_N)\right\|_{L^2(\Omega)} \leq Ch^{r+1} \left\|u(t_N)\right\|_{r+1}.$$

Im folgenden beschäftigen wir uns daher mit der Abschätzung von θ. Sei $\varphi \in L^2(\Omega)$ beliebig und Z_S die Lösung von

$$B_N(X, Z_S) = \left(X^N, P_S\varphi\right)_{L^2(\Omega)} = \left(X^N, \varphi\right)_{L^2(\Omega)} \qquad \forall X \in \mathbf{S}.$$

Da $\dot{Z}_S(t) \equiv 0$ auf J_n gilt, erhalten wir

$$\begin{aligned}\left(\theta^N, \varphi\right)_{L^2(\Omega)} &= B_N(\theta, Z_S) = -B_N(\rho, Z_S) \\ &= -\sum_{n=1}^{N} \int_{J_n} (\nabla\rho, \nabla Z_S)_{L^2(\Omega)}\, dt + \sum_{n=1}^{N-1} (\rho^n, [Z_S]_n)_{L^2(\Omega)} - \left(\rho^N, P_S\varphi\right)_{L^2(\Omega)}.\end{aligned}$$

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

$$\begin{aligned}\left|\left(\theta^N, \varphi\right)_{L^2(\Omega)}\right| &\leq \max_{n \leq N} \left(\|\rho\|_{0,J_n} + \|R_S\rho\|_{0,J_n}\right) \\ &\quad \times \left(\int_0^{t_N} \|\Delta_S Z_S\|_{L^2(\Omega)}\, dt + \sum_{n=1}^{N-1} \|[Z_S]_n\|_{L^2(\Omega)} + \|\varphi\|_{L^2(\Omega)}\right).\end{aligned}$$

Das Stabilitätsresultat (11.16) lässt sich anwenden und wir erhalten

$$\left\|\theta^N\right\|_{L^2(\Omega)} \leq CL_N \max_{n \leq N} \left(\|\rho\|_{0,J_n} + \|R_S\rho\|_{0,J_n}\right).$$

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 (11.2) und eine Näherungslösung in

$$S := \left\{ X : [0, \infty[\to H \mid \forall n : X|_{J_n} = \psi \in H \right\},$$

die durch

$$B_N(U, X) = \left(v, X_+^0\right)_{L^2(\Omega)} + \int_0^{t_N} (f, X)_{L^2(\Omega)} \, dt \qquad \forall X \in S \tag{11.17}$$

gegeben ist, wobei wir an die Definition von $B_N(\cdot, \cdot)$ erinnern:

$$B_N(V, W) = \int_0^{t_N} (\partial_{\text{stw}} V, W)_H + (AV, W)_H \, dt + \sum_{n=1}^{N-1} \left([V]_n, W_+^n \right)_H + \left(V_+^0, W_+^0 \right)_H. \tag{11.18}$$

Wir wir bereits in (11.7) gesehen haben, kann diese Gleichung in der folgenden Form geschrieben werden:

$$U^n + k_n A U^n = U^{n-1} + \int_{J_n} f dt \qquad \text{für } n \geq 1,\ U^0 = v. \tag{11.19}$$

Satz 11.6 *Für die Lösungen von (11.2) und (11.19) gilt die Fehlerabschätzung*

$$\left\| U^N - u(t_N) \right\|_H \leq C L_N \max_{n \leq N} \left(k_n \|f\|_{0, J_n} + k_n \left\| \overline{\partial}_n U^n \right\|_H \right).$$

Der Beweis erfordert etwas Vorbereitung. Er verwendet die Lösung des Rückwärtsproblems

$$\begin{aligned} -\dot{z} + Az &= 0 \qquad \text{für } t < t_N \\ z(t_N) &= \varphi. \end{aligned} \tag{11.20}$$

Mit Hilfe des Rückwärtsproblems erhalten wir die folgende Fehlerdarstellung.

Lemma 11.7 *Seien U und u die Lösungen von (11.19) und (11.2). Sei z die Lösung des Rückwärtsproblems (11.20). Dann gilt für den Fehler $e = U - u$ die Darstellung*

$$\left(e^N, \varphi\right)_H = \int_0^{t_N} (AU, z - X)_H \, dt + \sum_{n=0}^{N-1} \left([U]_n, z^n - X_+^n \right)_H - \int_0^{t_N} (f, z - X) \, dt \quad \forall X \in S.$$

Beweis. Wir erinnern, dass der Fehler wegen der Galerkin-Orthogonalität die Gleichung

$$B_N(e, X) = 0 \qquad \forall X \in S$$

erfüllt. Die Definition der Rückwärtslösung z und die Darstellung der Bilinearform $B_N(\cdot,\cdot)$ liefern:

$$\begin{aligned}\left(e^N,\varphi\right)_H &= B_N(e,z) = B_N(e,z-X) = B_N(U,z-X) - B_N(u,z-X)\\ &= \int_0^{t_N} (AU,z-X)_H\,dt + \sum_{n=1}^{N-1}\left([U]_n,(z-X)_+^n\right)_H + \left(U_+^0,(z-X)_+^0\right)_H\\ &- \left(v,(z-X)_+^0\right)_H - \int_0^{t_N}(f,z-X)_H\,dt.\end{aligned}$$

Dies ist die gewünschte Fehlerdarstellung. ■

Zur weiteren Vorbereitung benötigen wir ein Stabilitätsresultat für die exakte Lösung.

Lemma 11.8 *Die Lösung des Rückwärtsproblems (11.20) erfüllt*

$$\int_0^{t_{N-1}} \|\dot z\|\,dt + \|z\|_{0,J_N} \le CL_N \|\varphi\|_H.$$

Beweis. Zum Beweis dieser Aussage verwenden wir die Stabilität des Vorwärtsproblems (11.2) mit $f\equiv 0$:

$$\int_{k_1}^{t_N} \|\dot u\|_H\,dt + \|u\|_{0,J_1} \le C\tilde L_N \|v\|_H, \qquad \text{mit} \quad \tilde L_N = 1+\sqrt{\log\frac{t_N}{k_1}}.$$

Um diese Ungleichung einzusehen, beachten wir zunächst $\|u(t)\|_H \le \|v\|_H$ für $t\ge 0$ (Dies folgt aus der Spekraldarstellung der Lösung des homogenen Problems.) Speziell gilt

$$\|u\|_{0,J_1} \le \|v\|_H.$$

Die Spektraldarstellung der Lösung von (11.2) 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

$$\int_0^\infty t\,\|\dot u\|_H^2\,dt = \sum_{n\in N}\lambda_n^2\left(\int_0^\infty te^{-2\lambda_n t}dt\right)(v,\psi_n)_H^2 = \frac14\sum_{n\in N}(v,\psi_n)_H^2 = \frac14\|v\|_H^2.$$

Daraus schliessen wir

$$\left(\int_{k_1}^{t_N}\|\dot u\|_H^2\,dt\right)^2 \le \int_{k_1}^{t_N}\frac{dt}{t}\int_{k_1}^{t_N} t\,\|\dot u\|_H^2\,dt \le \frac14\left(\log\frac{t_N}{k_1}\right)\|v\|_H^2,$$

und daraus folgt die Behauptung. ■

Damit haben wir alle Hilfsmittel zusammengestellt um Satz 11.6 zu beweisen.

Beweis von Satz 11.6:

Wir bezeichnen die drei Terme aus der Fehlerdarstellung von Lemma 11.7 mit I, II, III. Wir wählen

$$X(t) := \bar{z}^n := k_n^{-1} \int_{J_n} z(t)\, dt \qquad \forall t \in J_n, \quad n \geq 1.$$

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 $X_+^n = \bar{z}^{n+1}$

$$|II| \leq \max_{n \leq N-1} \|[U]_n\|_H \sum_{n=1}^{N} \left\|\bar{z}^n - z^{n-1}\right\|_H .$$

Die rechte Seite lässt sich abschätzen gemäss

$$\left\|\bar{z}^n - z^{n-1}\right\|_H = \left\| k_n^{-1} \int_{J_n} \left(z - z^{n-1}\right) dt \right\|_H \leq \int_{J_n} \|\dot{z}\|_H\, dt \quad \forall n < N$$
$$\left\|\bar{z}^n - z^{n-1}\right\|_H \leq 2 \|z\|_{0,J_n} .$$

Mit Lemma 11.8 ergibt sich

$$\sum_{n=1}^{N} \left\|\bar{z}^n - z^{n-1}\right\|_H \leq \int_0^{t_{N-1}} \|\dot{z}\|_H\, dt + 2 \|z\|_{0,J_N} \leq C L_N \|\varphi\|_H .$$

Da $[U]_n = k_n \overline{\partial}_n U^n$ gilt, folgt die gewünschte Abschätzung für den Term II.

In ähnlicher Weise erhalten wir für den Term III die Abschätzung

$$\begin{aligned} |III| &\leq \max_{n \leq N} \left(k_n \|f\|_{0,J_n}\right) \sum_{n=1}^{N} \left(k_n^{-1} \int_{J_n} \|\bar{z}^n - z\|_H\, dt \right) \\ &\leq \max_{n \leq N} \left(k_n \|f\|_{0,J_n}\right) \left(\int_0^{t_{N-1}} \|\dot{z}\|_H\, dt + 2 \|z\|_{0,J_n} \right) \\ &\leq C L_N \max_{n \leq N} \left(k_n \|f\|_{0,J_n}\right) \|\varphi\|_H . \end{aligned}$$

Damit ist der Beweis gegeben. ■

Zum Schluss dieses Abschnitts werden wir eine *a posteriori*-Fehlerabschätzung für das unstetige 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 (11.13) gegeben und die volldiskrete Diskretisierung durch: Finde $U \in \mathbf{S}$ mit

$$B_N(U, X) = \left(v, X_+^0\right)_{L^2(\Omega)} + \int_0^{t_N} (f, X)\, dt \quad \forall X \in \mathbf{S},\ N \geq 1. \tag{11.21}$$

Die Bilinearform $B_N(\cdot, \cdot)$ ist wiederum durch (11.15) gegeben.

Das Gebiet Ω sei zerlegt in eine Finite-Elemente-Triangulierung $\mathcal{G}$ mit maximalem Dreiecksdurchmesser h. Der Finite-Elemente-Raum für die Ortsdiskretisierung ist durch

$$S := \left\{ u \in C^0(\Omega) : u|_{\partial\Omega} \equiv 0 \wedge u|_\tau \in \mathbb{P}_1, \quad \forall \tau \in \mathcal{G} \right\}$$

gegeben. Die Formulierung (11.21) 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

$$\int_{J_n} (\nabla U, \nabla X)_{L^2(\Omega)} + \left([U]_{n-1}, X_+^{n-1} \right)_{L^2(\Omega)} = \int_{J_n} (f, X)_{L^2(\Omega)} \, dt \quad \forall X \in \mathbf{S},\, n \geq 1$$

bzw.

$$\left(U^n, \chi_+^{n-1} \right)_{L^2(\Omega)} + k_n (\nabla U, \nabla \chi)_{L^2(\Omega)} = \left(U^{n-1}, \chi \right)_{L^2(\Omega)} + \left(\int_{J_n} f(t)\, dt, \chi \right)_{L^2(\Omega)} \qquad \forall \chi \in S,\ n \geq 1$$

mit der Anfangsbedingung $U^0 = P_S v$.

Wir erinnern an Satz 11.5: Unter der Voraussetzung, dass aufeinanderfolgende Schrittweiten $k_{n+1}/k_n \geq c > 0$ erfüllen, gilt die Fehlerabschätzung

$$\|U - u\|_{0,J_n} \leq C L_N \max_{n \leq N} \left(k_n \|\dot{u}\|_{0,J_n} + h^2 \|u\|_{2,J_n} \right).$$

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 $\overline{\partial}_n U^n$ 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_γ, der senkrecht zu γ ist und fixieren eine der beiden möglichen Richtungen.

Das erlaubt uns, den Sprung des Gradienten $\nabla\chi$ in Richtung der Kantennormalen n_γ zu definieren:

$$[\partial\chi/\partial n_\gamma]_\gamma := \lim_{\varepsilon \to +0} \langle n_\gamma, \nabla\chi(x + \varepsilon n_\gamma) - \nabla\chi(x - \varepsilon n_\gamma) \rangle .$$

Man beachte, dass der Normalenvektor n_γ auf der Kante γ 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_γ.

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 11.9 *Die Lösungen von (11.21) und (11.13) erfüllen die a posteriori-Fehlerabschätzung*

$$\left\|U^N - u\left(t_N\right)\right\|_{L^2(\Omega)} \leq CL_N \left\{ \max_{n \leq N} \left(h^2 + k_n\right) \|f\|_{0,J_n} + h^2 \|U^n\|_{2,S} + k_n \left\|\overline{\partial}_n U^n\right\|_{L^2(\Omega)} \right\}.$$

Der Beweis benötigt einen vorbereitenden Hilfssatz.

Hilfssatz 11.10 *Für alle $W \in S$ und $v \in H^2$ gilt*

$$\left|\left(\nabla W, \nabla\left(P_S - I\right) v\right)_{L^2(\Omega)}\right| \leq Ch^2 \|W\|_{2,S} \|v\|_2 .$$

Beweis. (Skizze) Ohne Beweis verwenden wir die folgenden beiden Aussagen:

1. **Spurabschätzung:** Sei $\tau \in \mathcal{G}$ ein beliebiges Dreieck mit Durchmesser h_τ. Dann gilt für alle $w \in H^1$
$$\|w\|^2_{L^2(\gamma)} \leq C \left(h_\tau \|\nabla v\|^2_{L^2(\tau)} + h_\tau^{-1} \|v\|^2_{L^2(\tau)} \right). \tag{11.22}$$

2. **Fehlerabschätzung für L^2-Projektion:** Für alle $v \in H^2$ gilt
$$\|P_S v - v\|_{L^2(\Omega)} + h \|\nabla\left(P_S v - v\right)\|_{L^2(\Omega)} \leq Ch^2 \|v\|_2 . \tag{11.23}$$

Zunächst zeigen wir, dass für $W \in S$ und $v \in H^1$ gilt

$$\left|\left(\nabla W, \nabla v\right)_{L^2(\Omega)}\right| \leq C \|W\|_{2,S} \left(\|v\|_{L^2(\Omega)} + h \|\nabla v\|_{L^2(\Omega)} \right). \tag{11.24}$$

Wir betrachten ein beliebiges Dreieck $\tau \in \mathcal{G}$ mit Kanten $\gamma_{\tau,j}$, $j = 1, 2, 3$, und wenden die Greenschen Formel an:

$$\int_\tau \langle \nabla W, \nabla v \rangle \, dx = \sum_{j=1}^{3} \left. \frac{\partial W}{\partial n_{\tau,j}} \right|_{\gamma_{\tau,j}} \int_{\gamma_{\tau,j}} v ds.$$

Beim Summieren über alle Dreiecke τ treten alle Kanten zweimal auf, wobei die Kantennormalenrichtungen entgegengesetzt orientiert sind. Fixiert man eine dieser Normalenrichtungen n_γ, erhält man

$$\left(\nabla W, \nabla v\right)_{L^2(\Omega)} \leq C \|W\|_{2,S} \sqrt{\sum_\gamma \left(\int_\gamma v ds \right)^2}.$$

Die *Höldersche Ungleichung* und die *Spurabschätzung* (11.22) ergeben

$$\sum_\gamma \left(\int_\gamma v ds \right)^2 \leq C \left(\|v\|^2_{L^2(\tau)} + h^2 \|\nabla v\|^2_{L^2(\tau)} \right).$$

Damit haben wir (11.24) gezeigt. Zusammen mit der Approximationseigenschaft (11.23) der L^2-Projektion ergibt sich die Behauptung. ■

Beweis von Satz 11.9:

Die Fehlerdarstellung $e = U - u$ aus Lemma 11.7 gilt auch in diesem Fall und daraus folgt

$$\left(e^N, \varphi\right)_{L^2(\Omega)} = -\int_0^{t_N} \left(\nabla U, \nabla\left(X - z\right)\right)_H dt - \sum_{n=0}^{N-1} \left(\left[U\right]_n, X_+^n - z^n\right)_{L^2(\Omega)} + \int_0^{t_N} \left(f, X - z\right) dt \tag{11.25}$$

$$=: I + II + III$$

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 $L^2\left(\Omega \times J_n\right)$ 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

$$X - z = \left(P_S \bar{z} - P_S z\right) + \left(P_S z - z\right).$$

Nun gilt für den ersten Term

$$\int_{J_n} \left(\nabla U, \nabla\left(P_S \bar{z} - P_S z\right)\right)_{L^2(\Omega)} dt = -\int_{J_n} \left(\Delta_S U, \bar{z} - z\right) dt = 0$$

und für den zweiten (wegen Hilfssatz 11.10)

$$\left|\int_{J_n} \left(\nabla U, \nabla\left(P_S \bar{z} - P_S z\right)\right)_{L^2(\Omega)} dt\right| = \left|\left(\nabla U^n, \nabla\left(P_S - I\right) \int_{J_n} z dt\right)\right|$$

$$\leq C h^2 \left\|U^n\right\|_{2,S} \left\|\int_{J_n} z dt\right\|_2 \leq C h^2 \left\|U^n\right\|_{2,S} \left\|\Delta \int_{J_n} z dt\right\|_{L^2(\Omega)}.$$

Da $\int_{J_N} \Delta z = \int_{J_n} \dot{z} = z\left(t_N\right) - z\left(t_{N-1}\right)$ zeigt Lemma 11.8

$$\begin{aligned} |I| &\leq C h^2 \left(\max_{n \leq N} \left\|U^n\right\|_{2,S}\right) \sum_{n=1}^{N} \left\|\int_{J_n} \Delta z dt\right\|_{L^2(\Omega)} \\ &\leq C h^2 \left(\max_{n \leq N} \left\|U^n\right\|_{2,S}\right) \left(\int_0^{t_{N-1}} \left\|\dot{z}\right\| dt + \left\|z\right\|_{0,J_N}\right) \\ &\leq C L_N h^2 \max_{n \leq N} \left\|U^n\right\|_{2,S} \left\|\varphi\right\|_{L^2(\Omega)}. \end{aligned}$$

Wir betrachten nun den zweiten Term in (11.25) und verwenden $\left[U\right]_{n-1} = U^n - U^{n-1} = k_n \bar{\partial}_n U^n$. Daraus folgt

$$\begin{aligned} II &= -\sum_{n=1}^{N} \left(\left[U\right]_{n-1}, X^n - z^{n-1}\right)_{L^2(\Omega)} = \sum_{n=1}^{N} \left(\left[U\right]_{n-1}, X^n - P_S z^{n-1}\right)_{L^2(\Omega)} \\ &\leq \sum_{n=1}^{N} k_n \left\|\bar{\partial}_n U\right\|_{L^2(\Omega)} \left\|\bar{z}^n - z^{n-1}\right\|_{L^2(\Omega)}. \end{aligned}$$

Verwenden wir nochmals Lemma 11.8, folgt die gewünschte Abschätzung für II :

$$|II| \leq C \max_{n \leq N} \left(k_n \left\|\bar{\partial}_n U^m\right\|_{L^2(\Omega)}\right) \left(\int_0^{t_{N-1}} \left\|\dot{z}\right\| dt + \left\|z\right\|_{0,J_n}\right) \leq C L_N \max_{n \leq N} \left(k_n \left\|\bar{\partial}_n U^n\right\|\right) \left\|\varphi\right\|_{L^2(\Omega)}.$$

Es bleibt, den dritten Term in (11.25) 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

$$\int_{J_n} \|z - P_S \bar{z}\|_{L^2(\Omega)} \, dt \leq C \left(h^2 + k_n\right) \int_{J_n} \|\dot{z}\| \, dt \qquad \forall n < N.$$

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 11.8:

$$\begin{aligned} |III| &\leq C \max_{n \leq N} \left(h^2 + k_n\right) \|f\|_{0,J_n} \left(\int_0^{t_{N-1}} \|\dot{z}\|_{L^2(\Omega)} \, dt + \|z\|_{0,J_N} \right) \\ &\leq C L_N \left(\max_{n \leq N} \left(h^2 + k_n\right) \|f\|_{0,J_n} \right) \|\varphi\|_{L^2(\Omega)}. \end{aligned}$$

Diese drei Abschätzungen zusammen mit (11.25) ergeben die Behauptung. ■