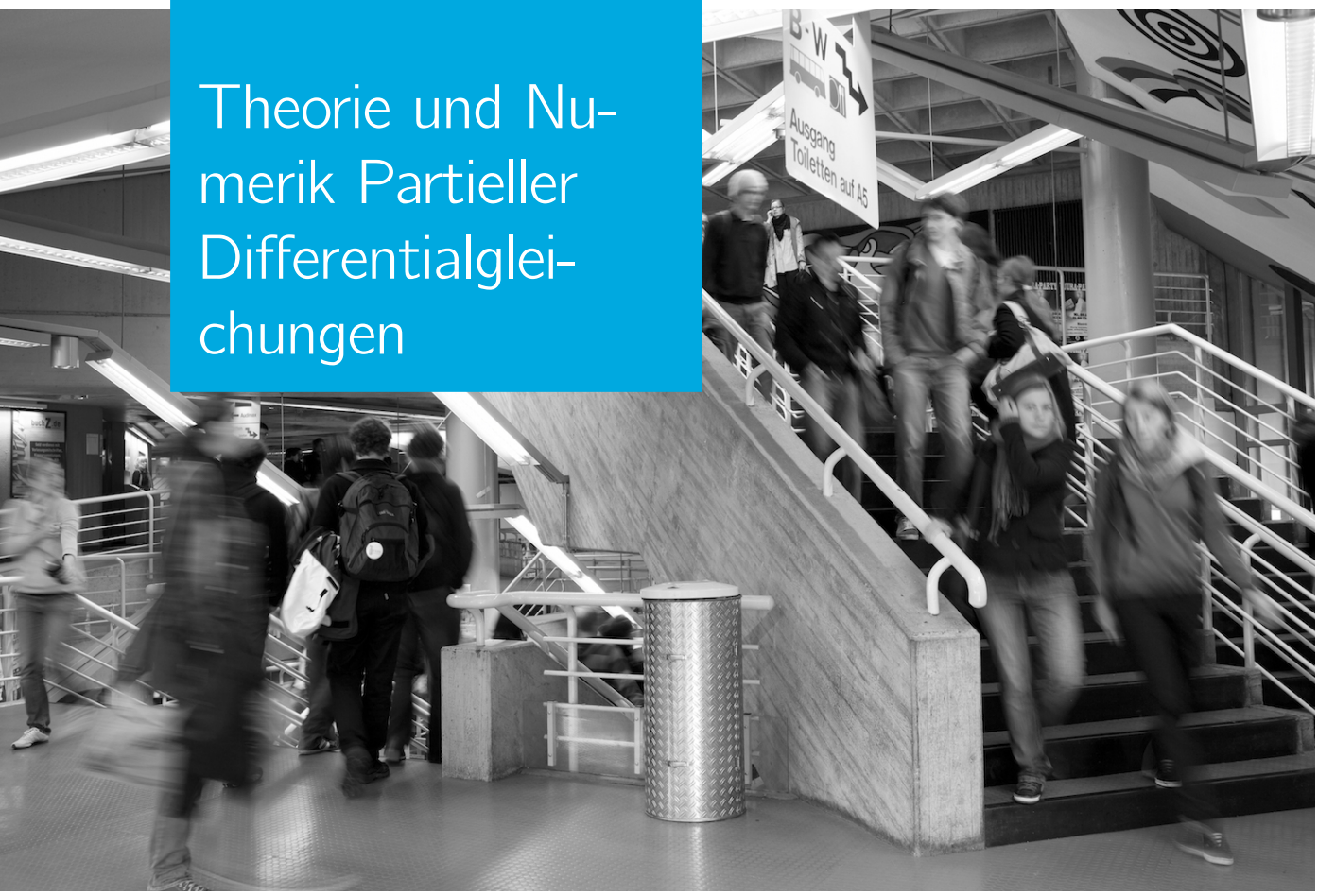
Universität Konstanz

Wintersemester 2017/18

Theorie und Numerik Partieller Differentialgleichungen



Prof. Dr. Stefan Volkwein
Numerikteil

Inhaltsverzeichnis

1 Anfangswertprobleme

In diesem Kapitel wollen wir die für die Vorlesung notwendigen Grundlagen aus dem Bereich der Anfangswertprobleme zusammenstellen. Dabei geht es im Wesentlichen um die numerische Lösung von Anfangswertproblemen für Deskriptorsysteme durch Einschrittverfahren.

1.1 Einleitung

Gesucht ist eine Funktion $u : [0, T] \to \mathbb{R}^N$, die das folgende lineare Anfangswertproblem löst:

$$M\dot{u}(t) + Su(t) = f(t) \text{ for } t \in (0, T] \quad \text{und} \quad u(0) = u_\circ. \tag{1.1}$$

In (1.1) setzen wir voraus, dass $T > 0$, $M, S \in \mathbb{R}^{N\times N}$, $f : [0, T] \to \mathbb{R}^N$ und $u_\circ \in \mathbb{R}^N$ gelten. Ferner ist die Matrix M symmetrisch und positiv definit, das heißt, es ist $M = M^\top$ und

$$\langle v, Mv\rangle_2 = v^\top Mv > 0 \quad \text{für alle } v \in \mathbb{R}^N \setminus \{0\}, \tag{1.2}$$

wobei $\langle \cdot\,, \cdot\rangle_2$ das Euklidische Skalarprodukt bezeichnet. Für (1.2) schreiben wir auch $M \succ 0$, was aus der Semidefiniten Programmierung kommt. Lineare Anfangswertprobleme, bei denen vor der Zeitableitung eine nicht notwendigerweise invertierbare Koeffizientenmatrix steht, werden in der Systemtheorie *Deskriptorsysteme* genannt. Aus der Lineren Algebra wissen wir, dass M diagonalisierbar ist. Es existieren N reelle, positive Eigenwerte $\lambda_1 \geq \ldots \geq \lambda_N > 0$ besitzt. Es gilt dann auch

$$\lambda_N |v|_2^2 \leq v^\top Mv \leq \lambda_N |v|_2^2 \quad \text{für alle } v \in \mathbb{R}^N \setminus \{0\}, \tag{1.3}$$

wobei $|\cdot|_2$ die Euklidische Norm auf $\mathbb{R}^N$ bezeichnet. Insbesondere ist die Matrix M invertierbar. Mit $A = -M^{-1}S \in \mathbb{R}^{N\times N}$ und $\tilde{f} = M^{-1}f$ können wir daher (1.1) in der Form

$$\dot{u}(t) = Au(t) + \tilde{f}(t) \text{ for } t \in (0, T] \quad \text{und} \quad u(0) = u_\circ \tag{1.4}$$

schreiben. An dieser Stelle erinnern wir an die bekannte Tatsache, dass die Berechnung der Matrix A nicht durch Invertieren von M erfolgt, sondern durch Lösen der linearen Gleichung

$$MA = -S = -\big[s_1 \,|\, \ldots \,|\, s_N\big] \in \mathbb{R}^{N\times N}$$

mit den N rechten Seiten $-s_i$, $1 \leq i \leq N$. Hier kann entweder als direktes Verfahren die *Cholesky-Zerlegung* [Lui10, Abschnitt 1.4] oder als indirekte Methode das *Conjugate-Gradient (CG) Verfahren* [Kel99, Chapter 2] eingesetzt werden.

1.2 Motivation

Die Entwicklung der Temperatur $u = u(t, \boldsymbol{x})$ an der Stelle $\boldsymbol{x}$ eines Stabes zur Zeit t ergibt sich als Lösung der parabolischen *Wärmeleitungsgleichung*

$$\frac{\partial u}{\partial t}(t, \boldsymbol{x}) = \kappa \frac{\partial^2 u}{\partial \boldsymbol{x}^2}(t, \boldsymbol{x}) \quad \text{für } t \in (0, T] \text{ und } \boldsymbol{x} \in (0, \ell).$$

Die Anfangsbedingung lautet $u(0, \boldsymbol{x}) = u_\circ(\boldsymbol{x})$ für $\boldsymbol{x} \in (0, \ell)$ mit stetiger Funktion $\mathrm{u}_\circ$ und mit $\ell > 0$. Die Randwerte sind $u(t, 0) = u(t, \ell) = 0$ (homogene Dirichlet-Randbedingungen). Mit Hilfe der *Linienmethode* kann die gesuchte Lösung $u(t, \boldsymbol{x})$ mit durch ein Systems gewöhnlicher Differentialgleichungen erster Ordnung angenähert werden. Dazu sei für $N \in \mathbb{N}$ die Orts-Schrittweite in $\boldsymbol{x}$-Richtung durch $h = \ell/(N+1)$ gegeben. Wir approximieren die zweite partielle Ableitung nach $\boldsymbol{x}$ durch den *zentralen Differenzenquotienten zweiter Ordnung*:

$$\kappa \frac{\partial^2 u}{\partial \boldsymbol{x}^2}(t, \boldsymbol{x}) \approx \kappa \frac{u(t, \boldsymbol{x}+h) - 2u(t, \boldsymbol{x}) + u(t, \boldsymbol{x}-h)}{h^2} \quad \text{für } t \in (0, T] \text{ und } \boldsymbol{x} \in [h, \ell - h]. \quad (1.5)$$

Der Diskretisierungsfehler in (1.5) lässt sich mit der Taylorentwicklung [DR11, Satz 11.27] von u abschätzen, sofern die Funktion u hinreichend glatt ist. Darauf kommen wir später zurück. Wir bezeichnen mit $\mathrm{u}_i(t)$, $0 \le i \le N+1$, die numerischen Näherungen für $u(t, \boldsymbol{x}_i)$, wobei $\boldsymbol{x}_i = ih$ für $i = 0, \ldots, N+1$ gilt. Aufgrund der Randbedingungen ist die Temperatur an $\boldsymbol{x}_0 = 0$ und $\boldsymbol{x}_{N+1} = \ell$ bekannt. Daher genügt es, Näherungen für die Temperatur u an den inneren Gitterpunkten $x_i \in (0, \ell)$, $i = 1, \ldots, N$, zu berechnen. Die Bestimmungsgleichungen ergeben sich aus dem obigen zentralen Differenzenquotienten für die zweite Ableitung nach $\boldsymbol{x}$ wie folgt:

$$\dot{\mathrm{u}}_i(t) = \kappa \frac{\mathrm{u}_{i+1}(t) - 2\mathrm{u}_i(t) + \mathrm{u}_{i-1}(t)}{h^2} \quad \text{für } t \in (0, T] \text{ und } 1 \le i \le N.$$

Setzen wir

$$\mathrm{u}(t) = \begin{pmatrix} \mathrm{u}_1(t) \\ \vdots \\ \mathrm{u}_N(t) \end{pmatrix}, \quad \mathrm{u}_\circ = \begin{pmatrix} u_\circ(\boldsymbol{x}_1) \\ \vdots \\ \mathrm{u}_\circ(\boldsymbol{x}_N) \end{pmatrix}, \quad \mathrm{A} = \frac{\kappa}{h^2} \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{N \times N},$$

so erhalten wir das (lineare) Anfangswertproblem

$$\dot{\mathrm{u}}(t) = \mathrm{A}\mathrm{u}(t) \text{ für } t \in (0, T] \quad \text{und} \quad \mathrm{u}(0) = \mathrm{u}_\circ,$$

was (1.4) entspricht, wenn wir $\tilde{\mathrm{f}} = 0$ setzen.

Der Begriff Linienmethode reflektiert hier die Bestimmung der vektorwertigen Funktion $\mathrm{u}(t)$ entlang der zur Zeitachse parallelen Linien durch die Ortsgitterpunkte. Mit Hilfe numerischer Verfahren zur Lösung des Anfangswertproblems – also über eine nachfolgende Diskretisierung der Zeit – erhalten wir auf diese Weise ein numerisches Verfahren zur Behandlung partieller Differentialgleichungen vom Typ der Wärmeleitungsgleichung (ein Beispiel einer sogenannten parabolischen Differentialgleichung).

Wir merken an dieser Stelle an, dass die Matrix A eine Tridiagonalmatrix ist, also dünn besetzt ist. Matrizen dieser Bauart nennen wir *Sparse-Matrizen*. Das Lösen von linearen Gleichungssystemen mit dünn besetzten Koeffizienten-Matrizen geht deutlich schneller als bei voll besetzen. Bei Tridiagonalmatrizen ist der Aufwand an Multiplikationen bei der Verwendung der Cholesky-Zerlegung zum Beispiel nur von der Größenordnung $\mathcal{O}(N)$ – und nicht $\mathcal{O}(N^3)$ wie im allgemeinen Fall.

1.3 Existenztheorie

Wenn $\mathrm{f} \in C([0, T]; \mathbb{R}^N)$ gilt, folgt die eindeutige Lösbarkeit von (1.4) – und damit von (1.1) – mit dem Satz von Picard-Lindelöf [DR11, Satz 16.1]. Setzen wir $\mathrm{g}(t, \mathrm{v}) = \mathrm{A}\mathrm{v} + \tilde{\mathrm{f}}(t)$ für $(t, \mathrm{v}) \in [0, T] \times \mathbb{R}^N$, so ergibt sich sofort, dass g stetig in den Argumenten t und v ist. Ferner folgt die notwendige globale Lipschitz-Stetigkeit von g bezüglich v aus

$$|\mathrm{g}(t, \mathrm{v}) - \mathrm{g}(t, \mathrm{w})|_2 = |\mathrm{A}\mathrm{v} - \mathrm{A}\mathrm{w}|_2 \le \|\mathrm{A}\|_2 |\mathrm{v} - \mathrm{w}|_2 \quad \text{für alle } t \in (0, T] \text{ und } \mathrm{v}, \mathrm{w} \in \mathbb{R}^N.$$

Hierbei bezeichnen wir mit $\|A\|_2$ die von der Euklidischen Norm $|\cdot|_2$ *induzierte Matrixnorm* [Lui10, Abschnitt 9.3]

$$\|A\|_2 = \sup\left\{|Av|_2 \,\middle|\, v \in \mathbb{R}^N \text{ mit } |v|_2 = 1\right\}$$

der Matrix A. Diese Matrixnorm wird auch *Spektralradius* oder *Spektralnorm* der Matrix A genannt, denn es gilt

$$\|A\|_2 = \max\left\{\sqrt{\lambda_{\max}} \,\middle|\, \lambda_{\max} \text{ ist der größte Eigenwert der Matrix } A^\top A\right\}.$$

Wir erinnern daran, dass $\|\cdot\|_2$ submultiplikativ und mit der Vektornorm $|\cdot|_2$ verträglich ist.

Durch Verwenden der *Variation der Konstanten* [DR11, Lemma 17.6] ist die Lösung von (1.1) sogar explizit angebbar:

$$u(t) = e^{tA}u_\circ + \int_0^t e^{(t-s)A}\tilde{f}(s)\,ds \quad \text{für } t \in [0,T],$$

wobei die Matrix-Exponentialfunktion

$$e^{tA} = \sum_{j=0}^{\infty} \frac{t^j A^j}{j!} \in \mathbb{R}^{N\times N} \quad \text{für } t \in [0,T]$$

zwar mit Hilfe der Eigenwertwertzerlegung der symmetrischen Matrix $A = M^{-1}S$ berechnet werden kann, dieses aber für großes N nicht effizient und oft auch nicht stabil ist [Lui10, Abschnitt 11.4]. Daher verwenden wir hier andere Lösungstechniken.

Da wir später allgemeinere Inhomogenitäten $f \in L^2(0,T;\mathbb{R}^N)$ betrachten werden, benötigen wir eine Verallgemeinerung des Satzes von Picard-Lindelöf. Dazu bemerken wir, dass der Funktionenraum

$$H^1(0,T;\mathbb{R}^N) = \left\{v \in L^2(0,T;\mathbb{R}^N) \,\middle|\, \dot{v} \in L^2(0,T;\mathbb{R}^N)\right\}$$

ein Hilbertraum mit dem Skalarprodukt

$$\langle v, w\rangle_{H^1(0,T;\mathbb{R}^N)} = \int_0^T v(t)^\top w(t) + \dot{v}(t)^\top \dot{w}(t)\,dt \quad \text{für } v,w \in H^1(0,T;\mathbb{R}^N)$$

ist. Das folgende Lemma ist [Dzi10, Lemma 5.27] entnommen. Es sei dabei noch an die stetige Einbettung $H^1(0,T;\mathbb{R}^N) \hookrightarrow C^0([0,T];\mathbb{R}^N) = C([0,T];\mathbb{R}^N)$ erinnert, die zum Beispiel aus dem Auswahlsatz von Rellich und Kondrachov [DR12, Satz 16.22] folgt. Insbesondere existiert eine Einbettungskonstante $c_e > 0$ mit

$$\|v\|_{C([0,T];\mathbb{R}^N)} \le c_e \|v\|_{H^1(0,T;\mathbb{R}^N)} \quad \text{für alle } v \in H^1(0,T;\mathbb{R}^N). \tag{1.6}$$

Lemma 1.1. *Sei die Matrix* S *symmetrisch und positiv semidefinit, das heißt, es gelten* $S^\top = S$ *und* $S \succeq 0$. *Dann existiert zu gegebenem* $f \in L^2(0,T;\mathbb{R}^N)$ *und* $u_\circ \in \mathbb{R}^N$ *genau eine Lösung* $u \in H^1(0,T;\mathbb{R}^N)$ *von* (1.1). *Diese Lösung* u *ist sogar Hölder-stetig mit Exponent* 1/2 *(vergeiche* [DR11, Definition 7.21]*), das heißt,*

$$u \in C^{0,1/2}([0,T];\mathbb{R}^N) = \left\{v \in C([0,T];\mathbb{R}^N) \,\middle|\, \sup\left\{\frac{|v(t)-v(\tau)|_2}{|t-\tau|^{1/2}} \,:\, t,\tau \in [0,T], t \ne \tau\right\} < \infty\right\}.$$

Beweis. Aus der Einbettung $H^1(0,T;\mathbb{R}^N) \hookrightarrow C^{0,1/2}([0,T];\mathbb{R}^N)$ (vergleiche [Dzi10, Satz 3.40]) folgt die letzte Aussage des Lemmas. Insbesondere erhalten wir die punktweise Annahme des Anfangswertes: $\lim_{t\to 0} u(t) = u(0) = u_\circ$. Um den Satz von Picard-Lindelöf anwenden zu können, wird die Funktion $f \in L^2(0,T;\mathbb{R}^N)$ durch eine Folge stetiger Funktionen approximiert. Wir zeigen, dass die Lösungen zu diesen Approximatonen von f die Lösung u von (1.1) approximieren. Nach dem

Satz von Fischer-Riesz (vergleiche [DR12, Satz 4.20]) gibt es eine Folge $\{f_m\}_{m\in\mathbb{N}} \subset C([0,T];\mathbb{R}^N)$ mit

$$\|f - f_m\|_{L^2(0,T;\mathbb{R}^N)} = \left(\int_0^T |f(t) - f_m(t)|_2^2 \, dt \right)^{1/2} \to 0 \quad \text{für } m \to \infty. \tag{1.7}$$

Da f_m in $C([0,T];\mathbb{R}^N)$ liegt für alle m, existiert nach dem Satz von Picard-Lindelöf genau eine Lösung $u_m \in C^1([0,T];\mathbb{R}^N)$ für jedes $m \in \mathbb{N}$ mit

$$M\dot{u}_m(t) + Su_m(t) = f_m(t) \text{ for } t \in (0,T] \quad \text{and} \quad u_m(0) = u_\circ. \tag{1.8}$$

Wir zeigen, dass $\{u_m\}_{m\in\mathbb{N}}$ eine Cauchy-Folge in $H^1(0,T;\mathbb{R}^N)$ ist. Seien $\mu, \nu \in \mathbb{N}$ und u_μ, $u_\nu \in H^1(0,T;\mathbb{R}^N)$ die Lösungen von (1.8) für $m = \mu$ beziehungsweise $m = \nu$. Dann gilt

$$M\big(\dot{u}_\mu(t) - \dot{u}_\nu(t)\big) + S\big(u_\mu(t) - u_\nu(t)\big) = f_\mu(t) - f_\nu(t) \quad \text{für } t \in (0,T].$$

Multiplikation mit $\dot{u}_\mu(t) - \dot{u}_\nu(t)$ und Verwenden von (1.3) ergeben

$$\begin{aligned}
&\lambda_N |\dot{u}_\mu(t) - \dot{u}_\nu(t)|_2^2 + \big(S\big(u_\mu(t) - u_\nu(t)\big)\big)^\top \big(\dot{u}_\mu(t) - \dot{u}_\nu(t)\big) \\
&\qquad \le |\dot{u}_\mu(t) - \dot{u}_\nu(t)|_M^2 + \big(S\big(u_\mu(t) - u_\nu(t)\big)\big)^\top \big(\dot{u}_\mu(t) - \dot{u}_\nu(t)\big) \\
&\qquad = \big(\big(f_\mu(t) - f_\nu(t)\big)^\top \big(\dot{u}_\mu(t) - \dot{u}_\nu(t)\big) \quad \text{für } t \in (0,T],
\end{aligned}$$

wobei wir $|v|_M^2 = v^\top M v$ für $v \in \mathbb{R}^N$ gesetzt haben. Nach Voraussetzung ist die Matrix S symmetrisch und positiv semidefinit. Daher erhalten wir mit der Seminorm $|\cdot|_S$ aus

$$\begin{aligned}
\big(S\big(u_\mu(t) - u_\nu(t)\big)\big)^\top \big(\dot{u}_\mu(t) - \dot{u}_\nu(t)\big) &= \frac{1}{2}\frac{d}{dt}\Big(\big(S\big(u_\mu(t) - u_\nu(t)\big)\big)^\top \big(u_\mu(t) - u_\nu(t)\big)\Big) \\
&= \frac{1}{2}\frac{d}{dt}|u_\mu(t) - u_\nu(t)|_S^2 \quad \text{für } t \in (0,T]
\end{aligned}$$

die Ungleichung

$$\begin{aligned}
&\lambda_N |\dot{u}_\mu(t) - \dot{u}_\nu(t)|_2^2 + \frac{1}{2}\frac{d}{dt}|u_\mu(t) - u_\nu(t)|_S^2 \\
&\qquad \le \big(\big(f_\mu(t) - f_\nu(t)\big)^\top \big(\dot{u}_\mu(t) - \dot{u}_\nu(t)\big) \quad \text{für } t \in (0,T].
\end{aligned}$$

Division durch $\lambda_N > 0$, Integration über $[0,T]$ und Verwenden von $u_\mu(0) = u_\nu(0) = u_\circ$ ergeben für beliebiges $\varepsilon > 0$ die Abschätzung:

$$\begin{aligned}
&\int_0^T |\dot{u}_\mu(t) - \dot{u}_\nu(t)|_2^2 \, dt + \frac{1}{2\lambda_N}|u_\mu(T) - u_\nu(T)|_S^2 \\
&= \int_0^T |\dot{u}_\mu(t) - \dot{u}_\nu(t)|_2^2 \, dt + \frac{1}{2\lambda_N}|u_\mu(T) - u_\nu(T)|_S^2 - \frac{1}{2\lambda_N}|u_\mu(0) - u_\nu(0)|_S^2 \\
&\le \frac{1}{\lambda_N}\int_0^T \big(\big(f_\mu(t) - f_\nu(t)\big)^\top \big(\dot{u}_\mu(t) - \dot{u}_\nu(t)\big)\, dt \\
&\le \frac{1}{\lambda_N}\int_0^T |f_\mu(t) - f_\nu(t)|_2 |\dot{u}_\mu(t) - \dot{u}_\nu(t)|_2 \, dt \\
&\le \frac{1}{\lambda_N}\int_0^T \frac{1}{2\varepsilon}|f_\mu(t) - f_\nu(t)|_2^2 + \frac{\varepsilon}{2}|\dot{u}_\mu(t) - \dot{u}_\nu(t)|_2^2 \, dt \\
&= \frac{1}{2\varepsilon\lambda_N}\|f_\mu - f_\nu\|_{L^2(0,T;\mathbb{R}^N)}^2 + \frac{\varepsilon}{2\lambda_N}\int_0^T |\dot{u}_\mu(t) - \dot{u}_\nu(t)|_2^2 \, dt,
\end{aligned} \tag{1.9}$$

wobei wir die *Cauchy-Schwarz-Ungleichung* [DR12, Satz 12.17]

$$v^\top w \le |v|_2 |w|_2 \quad \text{für } v, w \in \mathbb{R}^N$$

und die *Ungleichung von Young* [DR11, Satz 10.2]

$$ab \leq \frac{a^2}{2\varepsilon} + \frac{\varepsilon b^2}{2} \quad \text{für } a, b \geq 0 \tag{1.10}$$

mit $a = |\mathsf{f}_\mu(t) - \mathsf{f}_\nu(t)|_2$ und $b = |\dot{\mathsf{u}}_\mu(t) - \dot{\mathsf{u}}_\nu(t)|_2$ verwendet haben. Wir wählen nun $\varepsilon = \lambda_N > 0$ und erhalten mit (1.9)

$$\|\dot{\mathsf{u}}_\mu - \dot{\mathsf{u}}_\nu\|^2_{L^2(0,T;\mathbb{R}^N)}\,\mathrm{d}t \leq \frac{1}{\lambda_N^2}\,\|\mathsf{f}_\mu - \mathsf{f}_\nu\|^2_{L^2(0,T;\mathbb{R}^N)}. \tag{1.11}$$

Da $(u_\mu - u_\nu)(0) = 0$ gilt, können wir die *Poincaré-Ungleichung* [DR12, Satz 16.23] anwenden. Es existiert daher eine Poincaré-Konstante $c_\mathsf{p} > 0$ mit

$$\|\mathsf{u}_\mu - \mathsf{u}_\nu\|_{L^2(0,T;\mathbb{R}^N)} \leq c_\mathsf{p}\,\|\dot{\mathsf{u}}_\mu - \dot{\mathsf{u}}_\nu\|_{L^2(0,T;\mathbb{R}^N)}. \tag{1.12}$$

Aus (1.11) und (1.12) folgt nun

$$\begin{aligned}\|\mathsf{u}_\mu - \mathsf{u}_\nu\|^2_{H^1(0,T;\mathbb{R}^N)} &= \|\mathsf{u}_\mu - \mathsf{u}_\nu\|^2_{L^2(0,T;\mathbb{R}^N)} + \|\dot{\mathsf{u}}_\mu - \dot{\mathsf{u}}_\nu\|^2_{L^2(0,T;\mathbb{R}^N)} \\ &\leq \left(c_\mathsf{p}^2 + 1\right)\|\dot{\mathsf{u}}_\mu - \dot{\mathsf{u}}_\nu\|^2_{L^2(0,T;\mathbb{R}^N)} \leq \frac{c_\mathsf{p}^2 + 1}{\lambda_N^2}\,\|\mathsf{f}_\mu - \mathsf{f}_\nu\|^2_{L^2(0,T;\mathbb{R}^N)}.\end{aligned}$$

Wegen der Vollständigkeit des Hilbertraums $L^2(0,T;\mathbb{R}^N)$ und wegen (1.7) ist $\{\mathsf{f}_m\}_{m\in\mathbb{N}}$ eine Cauchy-Folge in $L^2(0,T;\mathbb{R}^N)$. Damit ist $\{\mathsf{u}_m\}_{m\in\mathbb{N}}$ eine Cauchy-Folge in $H^1(0,T;\mathbb{R}^N)$. Da der Hilbertraum $H^1(0,T;\mathbb{R}^N)$ vollständig ist, existiert ein Grenzelement $\mathsf{u} \in H^1(0,T;\mathbb{R}^N)$ mit $\|\mathsf{u}_m - \mathsf{u}\|_{H^1(0,T;\mathbb{R}^N)} \to 0$ für $m \to \infty$. Wir verwenden (1.8), so dass wir aus

$$\begin{aligned}\|\mathsf{M}\dot{\mathsf{u}} + \mathsf{S}\mathsf{u} - \mathsf{f}\|_{L^2(0,T;\mathbb{R}^N)} &= \|\mathsf{M}(\dot{\mathsf{u}} - \dot{\mathsf{u}}_m) + \mathsf{S}(\mathsf{u} - \mathsf{u}_m) - (\mathsf{f} - \mathsf{f}_m)\|_{L^2(0,T;\mathbb{R}^N)} \\ &\leq \|\mathsf{M}\|_2\|\dot{\mathsf{u}} - \dot{\mathsf{u}}_m\|_{L^2(0,T;\mathbb{R}^N)} + \|\mathsf{S}\|_2\|\mathsf{u} - \mathsf{u}_m\|_{L^2(0,T;\mathbb{R}^N)} + \|\mathsf{f} - \mathsf{f}_m\|_{L^2(0,T;\mathbb{R}^N)} \\ &\leq \max\left\{\|\mathsf{M}\|_2, \|\mathsf{S}\|_2, 1\right\}\left(\|\mathsf{u} - \mathsf{u}_m\|_{H^1(0,T;\mathbb{R}^N)} + \|\mathsf{f} - \mathsf{f}_m\|_{L^2(0,T;\mathbb{R}^N)}\right) \to 0 \quad \text{für } m \to \infty\end{aligned}$$

schliessen können, dass das Grenzelement u die Differentialgleichung erfüllt. Mit (1.6) ergibt sich

$$|\mathsf{u}(0) - \mathsf{u}_\circ|_2 = |\mathsf{u}(0) - \mathsf{u}_m(0)|_2 \leq \|\mathsf{u} - \mathsf{u}_m\|_{C([0,T];\mathbb{R}^N)} \leq c_\mathsf{e}\,\|\mathsf{u} - \mathsf{u}_m\|_{H^1(0,T;\mathbb{R}^N)} \to 0 \quad \text{für } m \to \infty.$$

Also ist u eine Lösung von (1.1). Die Eindeutigkeit folgt aus der Eindeutigkeit des Grenzwertes. □

1.4 Einschrittverfahren

In diesem Abschnitt wollen wir numerische Lösungsverfahren für (1.1) einführen; vergleiche [Lui10, Abschnitt 6.3]. Dazu seien $t_k = k\Delta t$, $k = 0, \ldots, M$, eine äquidistante Diskretisierung des Zeitintervalls $[0,T]$ mit Schrittweite $\Delta t = T/M$. Sei $\mathsf{u}(t) \in \mathbb{R}^N$ die (in der Regel unbekannte) Lösung von (1.1). Mit

$$\mathsf{u}^k = \begin{pmatrix} u_1^k \\ \vdots \\ u_N^k \end{pmatrix} \in \mathbb{R}^N \quad \text{für } k = 0, \ldots, M$$

bezeichnen wir eine Approximation für u am Zeitpunkt t_k für $0 \leq k \leq M$. Wir setzen voraus, dass $\mathsf{S} = \mathsf{S}^\top$ und $\mathsf{S} \succeq 0$ gelten.

1.4.1 Euler-Cauchy-Verfahren

Zunächst betrachten wir das *Euler-Cauchy-Verfahren*, welches oft auch *explizites Euler-Verfahren* genannt wird. Dabei wird die Zeitableitung durch den bezüglich des Zeitpunkts t_{k-1} vorwärtsgenommenen Differenzenquotienten ersetzt:

$$\mathrm{M}\,\frac{\mathrm{u}^k - \mathrm{u}^{k-1}}{\Delta t} + \mathrm{S}\mathrm{u}^{k-1} = \mathrm{f}(t_{k-1}) \text{ für } k = 1, \ldots, M \quad \text{und} \quad \mathrm{u}^0 = \mathrm{u}_\circ. \tag{1.13}$$

Damit ist $\mathrm{u}^0 \in \mathbb{R}^N$ durch den Anfangsvektor $\mathrm{u}_\circ$ gegeben. Die weiteren Vektoren $\{\mathrm{u}^k\}_{k=1}^M \subset \mathbb{R}^N$ ergeben sich als sukzessive Lösungen der folgenden linearen Gleichungssysteme

$$\mathrm{M}\mathrm{u}^k = \big(\mathrm{M} - \Delta t\mathrm{S}\big)\mathrm{u}^{k-1} + \Delta t\,\mathrm{f}(t_{k-1}) \quad \text{für } k = 1, \ldots, M$$

mit der von k unabhängigen, symmetrischen und positiv definiten Koeffizientenmatrix M. Das Euler-Cauchy-Verfahren besitzt die Konsistenzordnung eins, wenn $\mathrm{f} \in C^1([0,T];\mathbb{R}^N)$ gilt. Der Nachweis erfolgt unter Verwendung geeigneter Taylorentwicklungen von der Lösung u. Damit folgt für die Konvergenz

$$\max_{0\le k\le M} |\mathrm{u}(t_k) - \mathrm{u}^k|_2 = \mathcal{O}(\Delta t) \quad (\Delta t \to 0);$$

siehe [Lui10, Abschnitt 6.4]. Es gibt also eine von M unabhängige Konstante $c > 0$ mit

$$\max_{0\le k\le M} |\mathrm{u}(t_k) - \mathrm{u}^k|_2 \le c\Delta t \quad \text{für } \Delta t \text{ hinreichend klein.}$$

1.4.2 Implizites Euler-Verfahren

Im Gegensatz zum vorigen Abschnitt verwenden wir beim *impliziten Euler-Verfahren* nun den bezüglich des Zeitpunkts t_k rückwärtsgenommenen Differenzenquotienten zur Diskretisierung der Zeitableitung. Wir erhalten statt (1.13) die folgende Verfahrensvorschrift:

$$\mathrm{M}\,\frac{\mathrm{u}^k - \mathrm{u}^{k-1}}{\Delta t} + \mathrm{S}\mathrm{u}^{k} = \mathrm{f}(t_{k}) \text{ für } k = 1, \ldots, M \quad \text{und} \quad \mathrm{u}^0 = \mathrm{u}_\circ. \tag{1.14}$$

Wieder ist $\mathrm{u}^0 \in \mathbb{R}^N$ durch den Anfangsvektor $\mathrm{u}_\circ$ festgelegt. Allerdings berechnen sich nun die weiteren Vektoren $\{\mathrm{u}^k\}_{k=1}^M \subset \mathbb{R}^N$ sukzessive aus den folgenden linearen Gleichungssystemen

$$\big(\mathrm{M} + \Delta t\,\mathrm{S}\big)\mathrm{u}^k = \mathrm{M}\mathrm{u}^{k-1} + \Delta t\,\mathrm{f}(t_{k-1}) \quad \text{für } k = 1, \ldots, M$$

mit der symmetrischen und positiv definiten Koeffizientenmatrix $\mathrm{M} + \Delta t\,\mathrm{S}$. Insbesondere ist die Koeffizientenmatrix bei nichtäquidistanter Schrittweite von k abhängig. Das implizite Euler-Verfahren stellt sich als stabileres Verfahren heraus, wie wir später sehen werden. Für die Konvergenz gilt aber ebenso wie für das explizite Euler-Verfahren nur die Konvergenzordnung eins.

1.4.3 Crank-Nicolson-Verfahren

Wir wollen noch ein weiteres implizites Verfahren betrachten. Hier wird die Differentialgleichung an den Stellen $t_{k-1/2} = t_k - \Delta t/2$, $k = 1, \ldots, M$, betrachtet. Wir erhalten dann eine symmetrische Differenzenapproximation

$$\dot{\mathrm{u}}(t_{k-1/2}) = \frac{u(t_k) - \mathrm{u}(t_{k-1})}{\Delta t} + \mathcal{O}(\Delta t^2) \quad \text{für } k = 1, \ldots, M$$

mit Konsistenzordnung zwei, sofern $\mathrm{u} \in C^3([0,T];\mathbb{R}^N)$ – also $\mathrm{f} \in C^2([0,T];\mathbb{R}^N)$ – gilt. Wird nun der verbleibende Teil $\mathrm{f}(t) - \mathrm{S}\mathrm{u}(t)$ der Differentialgleichung auch an der Stelle $t_{k-1/2}$ ausgewertet,

sind Näherungen für u an $t_{k-1/2}$ notwendig. Das umgeht man, indem diese Näherungen durch die arithmetischen Mittel der Näherungen an t_{k-1} und t_k ersetzt wird. Für $\mathrm{u} \in C^2([0,T];\mathbb{R}^N)$ erhalten wir mit den zwei Taylorentwicklungen

$$\mathrm{u}(t_{k-1}) = \mathrm{u}(t_{k-1/2}) - \frac{1}{2}\,\dot{\mathrm{u}}(t_{k-1/2})\Delta t + \frac{1}{8}\,\ddot{\mathrm{u}}(\xi_k^-)\Delta t^2 \quad \text{für } \xi_k^- \in [t_{k-1}, t_{k-1/2}] \text{ und } k = 1,\dots,M,$$
$$\mathrm{u}(t_k) = \mathrm{u}(t_{k-1/2}) + \frac{1}{2}\,\dot{\mathrm{u}}(t_{k-1/2})\Delta t + \frac{1}{8}\,\ddot{\mathrm{u}}(\xi_k^+)\Delta t^2 \quad \text{für } \xi_k^+ \in [t_{k-1/2}, t_k] \text{ und } k = 1,\dots,M$$

die folgende Abschätzung

$$\left|\frac{\mathrm{u}(t_{k-1}) + \mathrm{u}(t_k)}{2} - \mathrm{u}(t_{k-1/2})\right| = \frac{\Delta t^2}{16}\left|\ddot{\mathrm{u}}(\xi_k^-) + \ddot{\mathrm{u}}(\xi_k^+)\right| \le \frac{\Delta t^2}{8}\max_{t\in[0,T]}|\ddot{\mathrm{u}}(t)| = \mathcal{O}(\Delta t^2).$$

Der Fehler ist also von der der Ordnung Δt^2 wie der obige symmetrische Differenzenquotient. Insgesamt erhalten wir das *Crank-Nicolson-* oder auch *Trapez-Verfahren*

$$\mathrm{M}\,\frac{\mathrm{u}^k - \mathrm{u}^{k-1}}{\Delta t} + \mathrm{S}\,\frac{\mathrm{u}^k + \mathrm{u}^{k-1}}{2} = \frac{\mathrm{f}^k + \mathrm{f}^{k-1}}{2} \text{ für } k = 1,\dots,M \quad \text{und} \quad \mathrm{u}^0 = \mathrm{u}_\circ. \tag{1.15}$$

Der Vektor $\mathrm{u}^0 \in \mathbb{R}^N$ ist durch $\mathrm{u}_\circ$ bestimmt. Die weiteren Vektoren $\{\mathrm{u}^k\}_{k=1}^M \subset \mathbb{R}^N$ erhalten wir sukzessive aus den folgenden linearen Gleichungssystemen

$$\left(\mathrm{M} + \frac{\Delta t}{2}\,\mathrm{S}\right)\mathrm{u}^k = \left(\mathrm{M} - \frac{\Delta t}{2}\,\mathrm{S}\right)\mathrm{u}^{k-1} + \frac{\Delta t}{2}\left(\mathrm{f}^{k-1} + \mathrm{f}^k\right) \quad \text{für } k = 1,\dots,M$$

mit der symmetrischen, positiv definiten Koeffizientenmatrix $\mathrm{M} + \Delta t\,\mathrm{S}/2$. Es lässt sich zeigen, dass das Crank-Nicolson-Verfahren die Konvergenzordung zwei besitzt und daher schneller als die beiden Euler-Verfahren konvergiert.

Abschließend wollen wir noch ein diskretes Schema angeben, welches die drei Verfahren (1.13)-(1.15) beinhaltet. Dazu führen wir den Parameter $\theta \in [0,1]$ ein und betrachten das sogenannte *θ-Schema*

$$\mathrm{M}\,\frac{\mathrm{u}^k - \mathrm{u}^{k-1}}{\Delta t} + \theta\mathrm{S}\mathrm{u}^k + (1-\theta)\mathrm{S}\mathrm{u}^{k-1} = \theta\mathrm{f}^k + (1-\theta)\mathrm{f}^{k-1} \text{ für } k = 1,\dots,M \quad \text{und} \quad \mathrm{u}^0 = \mathrm{u}_\circ. \tag{1.16}$$

Offenbar ergibt (1.16) mit $\theta = 0$ das Verfahren (1.13), für $\theta = 1$ die Methode (1.14) und für $\theta = 1/2$ das Verfahren (1.15). Ist u^{k-1}, $k \in \{2,\dots,M\}$, bekannt, so erhalten wir u^k durch Lösen des linearen Gleichungssystems

$$\big(\mathrm{M} + \theta\Delta t\mathrm{S}\big)\mathrm{u}^k = \big(\mathrm{M} + (\theta-1)\Delta t\mathrm{S}\big)\mathrm{u}^{k-1} + \Delta t\big(\theta\mathrm{f}^k + (1-\theta)\mathrm{f}^{k-1}\big).$$

Da für $\theta \in [0,1]$ die Matrix $\mathrm{M} + \theta\Delta t\mathrm{S}$ symmetrisch und positiv definit ist, gibt es genau eine Lösung u^k dieses Gleichungssystems.

2 Finite Differenzen für die Poissongleichung

In diesem Kapitel orientieren wir uns an [Dzi10, Kapitel 3.1.1]. Wir verweisen zur weiteren Vertiefung auch auf [AU10, Kapitel 9.1], [KA00, Kapitel 1] und [Str04, Chapter 12].

2.1 Die Poissongleichung

Für die einfachere Präsentation betrachten wir hier den Spezialfall $G = (0,1)^n \subset \mathbb{R}^n$. Für gegebene Funktionen $f : G \to \mathbb{R}$ und $g : \partial G \to \mathbb{R}$ diskretisieren wir das *Poissonproblem*

$$-\Delta u = f \text{ in } G, \quad u = g \text{ auf } \partial G \tag{2.1}$$

mit einem *Finite-Differenzenverfahren*. Wir führen die wichtigen Begriffe *Konsistenz*, *Stabilität* und *Konvergenz* ein. Dazu betrachten wir zunächst das folgende abstrakte Schema (2.2) zur Diskretisierung einer kontinuierlichen Gleichung ein, welches wir im folgenden Kapitel ebenfalls verwenden werden:

$$\begin{array}{rccc} & X_0 & & Y_0 \\ & \cup & & \cup \\ \mathcal{T}: & X & \to & Y \\ & \downarrow \mathcal{D}_h^X & & \downarrow \mathcal{D}_h^Y \\ \mathcal{T}_h: & X_h & \to & Y_h \end{array} \tag{2.2}$$

In (2.2) bezeichnen wir mit X_0, Y, X_0 und Y_0 geeignete Funktionenräume für das kontinuierliche Problem. Die Gleichung $\mathcal{T}u = b$ steht für das zu lösende, meist unendlich-dimensionale Problem. Weiter seinen X_h und Y_h geeignete diskrete Räume. Die Diskretisierungsoperatoren $\mathcal{D}_h^X : X \to X_h$ und $\mathcal{D}_h^Y : Y \to Y_h$ verknüpfen die (kontinuierlichen) Räume X beziehungsweise Y mit den (diskreten) Räumen X_h beziehungsweise Y_h. Der Subindex h steht allgemein für einen positiven Diskretisierungsparameter. Bei uns wird h für die Ortsgitterweite stehen.

Wir wollen nun die Räume und Operatoren in (2.2) für das Problem (2.1) spezifizieren. Für $\alpha \in [0,1]$ definieren wir die kontinuierlichen Räume

$$\begin{aligned} X &= \big\{ v \in C(\overline{G}) \cap C^{2,\alpha}(G) \,\big|\, \sup_{x \in G} |\Delta v(x)| < \infty \big\}, \\ Y &= \big\{ (f,g) \in C^{0,\alpha}(G) \times C(\partial G) \,\big|\, \sup_{x \in G} |f(x)| < \infty \big\}, \\ X_0 &= C(G), \quad Y_0 = \big\{ (f,g) \in C(G) \times C(\partial G) \,\big|\, \sup_{x \in G} |f(x)| < \infty \big\}. \end{aligned}$$

Offenbar gelten $X \subset X_0$ und $Y \subset Y_0$. Nun führen wir den Operator $\mathcal{T} : X \to Y$ ein:

$$\mathcal{T}u = \big(-\Delta u, u|_{\partial\Gamma} \big) \in Y \quad \text{für } u \in X$$

Offenbar ist $\mathcal{T}$ wohldefiniert und linear. Aus [DR12, Satz 22.13] erhalten wir folgendes *Maximum-* und *Minimumprinzip*.

Satz 2.1. *Sei $G = (0,1)^n \subset \mathbb{R}^n$. Gilt $\Delta u = 0$ in G für $u \in X$, so gelten*

$$\min_{x \in \overline{G}} |u(x)| = \min_{x \in \partial G} |u(x)| \quad \textit{und} \quad \max_{x \in \overline{G}} |u(x)| = \max_{x \in \partial G} |u(x)|$$

Beweis. Nach dem ersten Teil von [DR12, Satz 22.13] folgt, dass die Funktion $u \in X$ wegen $\Delta u \geq 0$ in G ihr Maximum auf dem Rand ∂G wahrnimmt. Der zweite Teil von [DR12, Satz 22.13] liefert wegen $\Delta u \leq 0$ in G, dass die Funktion u ihr Minimum am Rand von ∂G annimmt. □

Bemerkung 2.2. 1) Die Aussage von Satz 2.1 ist für unbeschränkte Gebiete im Allgemeinen falsch. Das erkennt man an dem Beispiel $G = \{\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2 \,|\, x_2 < 0\}$ und $u(\boldsymbol{x}) = x_2$ für $\boldsymbol{x} = (x_1, x_2) \in G$. Offenbar gilt $\Delta u = 0$ in G, aber u nimmt nicht sein Maximum am Rand von G an.

2) Aus Satz 2.1 folgt, dass (2.1) höchstens eine Lösung besitzt; vergleiche [DR12, Folgerung 22.15]. ◊

Satz 2.3. *Seien $G = (0,1)^n \subset \mathbb{R}^n$ und $(f, g) \in Y$. Dann existiert genau eine Lösung $u \in X$ von (2.1). Insbesondere ist der Operator $\mathcal{T}$ bijektiv und $\sup_{\boldsymbol{x} \in G} |\Delta u(\boldsymbol{x})| < \infty$.*

Beweis. Existenz einer Lösung zu (2.1) folgt aus [Dzi10, Satz 1.44]. Für $(f, g) \in Y$ erhalten wir $u = \mathcal{T}^{-1}(f, g) \in X$. Wegen $\sup_{\boldsymbol{x} \in G} |f(\boldsymbol{x})| < \infty$ folgt auch $\sup_{\boldsymbol{x} \in G} |\Delta u(\boldsymbol{x})| = \sup_{\boldsymbol{x} \in G} |f(\boldsymbol{x})| < \infty$.□

2.2 Finite Differenzen

Zur Diskretisierung führen wir nun diskrete Räume X_h und Y_h sowie Diskretisierungsoperatoren $\mathcal{D}_h^X$ und $\mathcal{D}_h^Y$ ein. Dazu wählen wir die *Schrittweite* $h = 1/N_{\boldsymbol{x}} > 0$ mit $N_{\boldsymbol{x}} \in \mathbb{N}$. Die *Menge aller Gitterpunkte* $\overline{G}_h$ definieren wir durch

$$\overline{G}_h = \overline{G} \cap h\mathbb{Z}^n.$$

In unserem Fall enthält $\overline{G}_h$ genau $(N_{\boldsymbol{x}} + 2)^n$ Gitterpunkte. Wir schreiben das in der Form $|\overline{G}_h| = (N_{\boldsymbol{x}} + 2)^n$. Als *Menge der Randgitterpunkte* ∂G_h setzen wir

$$\partial G_h = \left\{\boldsymbol{x}_h \in \overline{G}_h \,\middle|\, \operatorname{dist}(\boldsymbol{x}_h, \partial G)\right\} \subset \overline{G} \quad \text{mit } \operatorname{dist}(\boldsymbol{x}_h, \partial G) = \inf\left\{|\boldsymbol{x}_h - \boldsymbol{x}|_2 \,\middle|\, \boldsymbol{x} \in \partial G\right\}.$$

Offenbar erhalten wir $\partial G_h \subset \partial G$ für $G = (0,1)^n$ und $h = 1/N_{\boldsymbol{x}}$. Die *Menge aller inneren Gitterpunkte* G_h ergibt sich als

$$G_h = \overline{G}_h \setminus \partial G_h \subset G.$$

Wir erhalten in unserem Fall $|G_h| = N_{\boldsymbol{x}}^n$ und somit $|\partial G_h = (N_{\boldsymbol{x}} + 2)^n - N_{\boldsymbol{x}}^n$. Die diskreten Räume X_h und Y_h enthalten Funktionen, die auf den soeben eingeführten Gittermengen definiert sind:

$$X_h = \left\{v_h \,\middle|\, v_h : \overline{G}_h \to \mathbb{R}\right\} \quad \text{und} \quad Y_h = \left\{(f_h, g_h) \,\middle|\, f_h : G_h \to \mathbb{R} \text{ und } g_h : \partial G_h \to \mathbb{R}\right\}.$$

Die Diskretisierungsoperatoren ergeben sich in natürlicher Weise als Einschränkungen von stetigen Funktionen auf das Finite-Differenzengitter. Hier wählen wir

$$\mathcal{D}_h^X v = v\big|_{\overline{G}_h} \in X_h \text{ für } v \in X \quad \text{und} \quad \mathcal{D}_h^Y(f, g) = \left(f\big|_{G_h}, g\big|_{\partial G_h}\right) \in Y_h \text{ für } (f, g) \in Y.$$

Offenbar sind die Operatoren $\mathcal{D}_h^X$ und $\mathcal{D}_h^Y$ wohldefiniert und linear. Nun können wir den diskreten Operator $\mathcal{T}_h : X_h \to Y_h$ wie folgt einführen:

$$\mathcal{T}_h u_h = \left(-\Delta_h u_h, u_h\big|_{\partial G_h}\right) \in Y_h \quad \text{für } u_h \in X_h, \tag{2.3}$$

wobei wir den *diskreten Laplace-Operator* $\Delta_h : G_h \to \{v_h \,|\, v_h : G_h \to \mathbb{R}\}$ mit

$$(\Delta_h v_h)(\boldsymbol{x}_h) = \frac{1}{h^2}\left(\sum_{j=1}^{n} \left(v_h(\boldsymbol{x}_h + h\mathrm{e}_j) + v_h(\boldsymbol{x}_h - h\mathrm{e}_j)\right) - 2n v_h(\boldsymbol{x}_h)\right) \quad \text{für } \boldsymbol{x}_h \in G_h \tag{2.4}$$

verwendet haben. In (2.4) bezeichnen die Vektoren e_j, $j = 1, \ldots, n$, die j-ten kanonischen Einheitsvektoren in $\mathbb{R}^n$.

Bemerkung 2.4. Wir wissen bereits, dass Funktionen $u \in C(\overline{G}) \cap C^2(G)$ mit $\Delta u = 0$ ein Maximum- und ein Minimumprinzip erfüllen; siehe Satz 2.1. Weiter erfüllt u die *Mittelwerteigenschaft*

$$u(\boldsymbol{x}) = \frac{1}{\omega_n r^{n-1}} \int_{\partial B(\boldsymbol{x},r)} u(\boldsymbol{s})\, \mathrm{d}\boldsymbol{s} \quad \text{für } \boldsymbol{x} \in G,\ r > 0 \text{ and } \overline{B(\boldsymbol{x}, r)} \subset G$$

mit dem $(n-1)$-dimensionalen Fächeninhalt ω_n der Einheitssphäre und der n-dimensionalen offenen Kugel $B(\boldsymbol{x}, r) = \{\tilde{\boldsymbol{x}} \in \mathbb{R}^n \,|\, |\tilde{\boldsymbol{x}} - \boldsymbol{x}|_2 < r\}$; vergleiche [DR12, Bemerkung 22.16-(i)]. Gilt für $u_h \in X_h$ die Gleichung $\Delta_h u_h = 0$ in G_h, so erhalten wir mit (2.4) die Beziehung

$$u_h(\boldsymbol{x}_h) = \frac{1}{2n}\left(\sum_{j=1}^{n} \big(u_h(\boldsymbol{x}_h - h\mathrm{e}_j) + u_h(\boldsymbol{x}_h + h\mathrm{e}_j)\big)\right) \quad \text{für } \boldsymbol{x}_h \in G_h,$$

was als *diskrete Mittelwerteigenschaft* bezeichnet wird. ◊

Nun können wir das diskrete Problem formulieren: Zu gegebenem $(f_h, g_h) \in Y_h$ bestimme eine Gitterfunktion $u_h \in X_h$, so dass

$$\mathcal{T}_h u_h = (f_h, g_h) \quad \text{in } Y_h \tag{2.5}$$

erfüllt ist. Offenbar ist $\mathcal{T}_h$ wohldefiniert und linear. Bevor wir zeigen, dass $\mathcal{T}_h$ bijektiv ist, bewesen wir das folgende *diskrete Maximumprinzip*.

Lemma 2.5. *Es sei $G = (0,1)^n \subset \mathbb{R}^n$ und Δ_h wie in (2.4) definiert. Gilt $-\Delta_h u_h \le 0$ in G_h für ein $u_h \in X_h$, dann folgt*

$$\max_{\boldsymbol{x}_h \in \overline{G}_h} u_h(\boldsymbol{x}_h) = \max_{\boldsymbol{x}_h \in \partial G_h} u_h(\boldsymbol{x}_h).$$

Beweis. Sei $u_h(\bar{\boldsymbol{x}}_h) = \max_{\boldsymbol{x}_h \in \overline{G}_h} u_h(\boldsymbol{x}_h) = u_{\max}$ für ein $\bar{\boldsymbol{x}}_h \in \overline{G}_h$. Gilt $\bar{\boldsymbol{x}}_h \in \partial G_h$, so ist nichts zu zeigen. Daher betrachten wir nur den Fall, dass $\bar{\boldsymbol{x}}_h$ in G_h liegt. Aus $-\Delta_h u_h(\bar{\boldsymbol{x}}_h) \le 0$ erhalten wir

$$\sum_{j=1}^{n} \big(u_h(\bar{\boldsymbol{x}}_h) - u_h(\bar{\boldsymbol{x}}_h - h\mathrm{e}_j)\big) + \sum_{j=1}^{n} \big(u_h(\bar{\boldsymbol{x}}_h) - u_h(\bar{\boldsymbol{x}}_h + h\mathrm{e}_j)\big) \le 0. \tag{2.6}$$

Da u_h in $\bar{\boldsymbol{x}}_h$ sein Maximum $u_{\max}$ annimmt, sind alle Summanden in (2.6) nichtnegativ. Es folgt damit, dass u_h auf $\bar{\boldsymbol{x}}_h \pm h\mathrm{e}_j$, $j = 1, \ldots, n$, konstant gleich $u_h(\bar{\boldsymbol{x}}_h)$ ist. Nun existieren ein $\boldsymbol{x}_h \in \partial G_h$ und $\boldsymbol{x}_h^1, \ldots, \boldsymbol{x}_h^K \in G_h$ mit $K \in \mathbb{N}$, $\boldsymbol{x}_h^1 = \bar{\boldsymbol{x}}_h \pm h\mathrm{e}_j$, $\boldsymbol{x}_h^{i+1} = \boldsymbol{x}_h^i \pm h\mathrm{e}_j$ für $i = 1, \ldots, K-1$ und $\boldsymbol{x}_h = \boldsymbol{x}_h^K \pm h\mathrm{e}_j$ und jeweils ein $j \in \{1, \ldots, n\}$. Mit der gleichen Argumentation wie für $\bar{\boldsymbol{x}}_h$ erhalten wir, dass u_h sein Maximum $u_{\max}$ auch in $\boldsymbol{x}_h \in \partial G_h$ annimmt. Daraus folgt die Behauptung des Lemmas. □

Bemerkung 2.6. Es gilt auch ein *diskretes Minimumprinzip*: Aus $-\Delta_h u_h \ge 0$ in G_h für ein $u_h \in X_h$ folgt

$$\min_{\boldsymbol{x}_h \in \overline{G}_h} u_h(\boldsymbol{x}_h) = \min_{\boldsymbol{x}_h \in \partial G_h} u_h(\boldsymbol{x}_h).$$

Der Beweis erfolgt mit einer analogen Argumentation wie im Beweis von Lemma 2.5. ◊

Satz 2.7. *Zu $G = (0,1)^n \subset \mathbb{R}^n$ sei der Operator $\mathcal{T}_h$ wie in (2.3) definiert. Dann gibt es zu jedem $(f_h, g_h) \in Y_h$ genau eine Gitterfunktion $u_h \in X_h$, die (2.5) löst, das heißt, der Operator $\mathcal{T}_h$ ist bijektiv und $u_h = \mathcal{T}_h^{-1}(f_h, g_h)$.*

Beweis. Es gilt (2.5) genau dann, wenn

$$-\Delta_h u_h = f_h \text{ in } G_h, \quad u_h = g_h \text{ auf } \partial G_h \tag{2.7}$$

gelten. Offenbar ist (2.7) ein lineares Gleichungssystem zur Bestimmung der Werte der Gitterfunktion u_h an den Gitterpunkten in $\overline{G}_h$. Es handelt sich also um $|\overline{G}_h|$ Unbekannte, die zu berechnen

sind. Die Anzahl der Gleichiungen in (2.7) ergibt sich als $|G_h| + |\partial G_h| = |\overline{G}_h|$. Die Bijektivität von $\mathcal{T}_h$ ergibt sich also aus der Injektivität von $\mathcal{T}_h$. Wir betrachten daher das homogene Problem

$$-\Delta_h u_h = 0 \text{ in } G_h, \quad u_h = 0 \text{ auf } \partial G_h$$

und zeigen, dass $u_h = 0$ in $\overline{G}_h$ gelten muss. Aus $u_h = 0$ auf ∂G_h, Lemma 2.5 und Bemerkung 2.6 erhalten wir

$$\max_{\boldsymbol{x}_h \in \overline{G}_h} u_h(\boldsymbol{x}_h) = \max_{\boldsymbol{x}_h \in \partial G_h} u_h(\boldsymbol{x}_h) = 0 = \min_{\boldsymbol{x}_h \in \partial G_h} u_h(\boldsymbol{x}_h) = \min_{\boldsymbol{x}_h \in \overline{G}_h} u_h(\boldsymbol{x}_h).$$

Also ist $u_h = 0$ in $\overline{G}_h$, was zu zeigen war. □

Beispiel 2.8. 1) Im Fall $n = 1$ führen wir die drei Indexmengen

$$\mathbb{I}_{\partial G_h} = \{0, N_{\boldsymbol{x}} + 1\}, \quad \mathbb{I}_{G_h} = \{1, \ldots, N_{\boldsymbol{x}}\}, \quad \mathbb{I}_{\overline{G}_h} = \mathbb{I}_{\partial G_h} \cup \mathbb{I}_{G_h} = \{0, \ldots, N_{\boldsymbol{x}} + 1\}.$$

Damit setzen wir

$$G = (0,1), \quad G_h = \left\{ ih \,\middle|\, i \in \mathbb{I}_{G_h} \right\} \subset G = (0,1), \quad \partial G_h = \{0,1\} = \partial G.$$

Seien $\mathrm{u}_i = u_h(ih)$ für $i \in \mathbb{I}_{\overline{G}_h}$, $\mathrm{f}_i = f_h(ih)$ für $i \in \mathbb{I}_{G_h}$ und $\mathrm{g}_i = g_h(ih)$ für $i \in \mathbb{I}_{\partial G_h}$. Für den diskreten Laplace-Operator bekommen wir den zentralen Differenzenquotienten

$$\Delta_h \mathrm{u}_i = \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} \quad \text{für } i \in \mathbb{I}_{G_h};$$

vergleiche (1.5). Damit ergibt das diskrete Poissonproblem (2.7) die linearen Gleichungen

$$-\mathrm{u}_{i-1} + 2\mathrm{u}_i - \mathrm{u}_{i+1} = h^2 \mathrm{f}_i \text{ für } i \in \mathbb{I}_{G_h}, \quad \mathrm{u}_i = \mathrm{g}_i \text{ für } i \in \mathbb{I}_{\partial G_h}. \tag{2.8}$$

Aufgrund der Randbedingungen gelten $u_0 = g_0$ und $u_{N_x+1} = g_{N_x+1}$. Es sind also nur die $N_{\boldsymbol{x}}$ Werte $\mathrm{u}_1, \ldots, \mathrm{u}_{N_x}$ zu bestimmen. Daher definieren wir $N = N_{\boldsymbol{x}}$, die Vektoren

$$\mathrm{u} = \begin{pmatrix} \mathrm{u}_1 \\ \vdots \\ \mathrm{u}_N \end{pmatrix} \in \mathbb{R}^N, \quad \mathrm{f} = h^2 \begin{pmatrix} \mathrm{f}_1 \\ \vdots \\ \mathrm{f}_N \end{pmatrix} \in \mathbb{R}^N, \quad \mathrm{g} = \begin{pmatrix} \mathrm{g}_0 \\ 0 \\ \vdots \\ 0 \\ \mathrm{g}_{N+1} \end{pmatrix} \in \mathbb{R}^N$$

und die Tridiagonalmatrix

$$\mathrm{A} = \begin{pmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{N \times N}.$$

Dann lässt sich (2.8) in Matrixschreibweise wie folgt ausdrücken:

$$\mathrm{Au} = \mathrm{f} + \mathrm{g}.$$

Die Matrix A ist symmetrisch und positiv definit.

2) Wir wollen auch den Fall $n = 2$ betrachten und führen für die Gitterpunkte wieder drei Indexmengen in $(\mathbb{N}_0)^2$ ein:

$$\begin{aligned} \mathbb{I}_{\partial G_h} = &\left\{ (i,0) \,\middle|\, i = 0, \ldots, N_{\boldsymbol{x}} + 1 \right\} \cup \left\{ (i, N_{\boldsymbol{x}} + 1) \,\middle|\, i = 0, \ldots, N_{\boldsymbol{x}} + 1 \right\} \\ &\cup \left\{ (0,j) \,\middle|\, j = 1, \ldots, N_{\boldsymbol{x}} \right\} \cup \left\{ (N_{\boldsymbol{x}} + 1, j) \,\middle|\, j = 1, \ldots, N_{\boldsymbol{x}} \right\}, \\ \mathbb{I}_{G_h} = &\left\{ (i,j) \,\middle|\, i,j = 1, \ldots, N_{\boldsymbol{x}} \right\}, \quad \mathbb{I}_{\overline{G}_h} = \mathbb{I}_{\partial G_h} \cup \mathbb{I}_{G_h}. \end{aligned}$$

Es gelten nun

$$G_h = \left\{(ih, jh) \,\middle|\, (i,j) \in \mathbb{I}_{G_h}\right\} \subset G = (0,1)^2, \quad \partial G_h = \left\{(ih, jh) \,\middle|\, (i,j) \in \mathbb{I}_{\partial G_h}\right\} \subset \partial G.$$

Wir setzen $\mathrm{u}_{ij} = u_h(ih, jh)$ für $(i,j) \in \mathbb{I}_{\overline{G}_h}$, $\mathrm{f}_{ij} = f_h(ih, jh)$ für $(i,j) \in \mathbb{I}_{G_h}$ und $\mathrm{g}_{ij} = g_h(ih, jh)$ für $(i,j) \in \mathbb{I}_{\partial G_h}$. Wir erhalten für den diskreten Laplace-Operator

$$\Delta_h \mathrm{u}_{ij} = \frac{\mathrm{u}_{i+1,j} + \mathrm{u}_{i-1,j} + \mathrm{u}_{i,j+1} + \mathrm{u}_{i,j-1} - 4\mathrm{u}_{ij}}{h^2} \quad \text{für } (i,j) \in \mathbb{I}_{G_h}.$$

Für das diskrete Poissonproblem (2.7) bekommen wir die linearen Gleichungen

$$4\mathrm{u}_{ij} - \mathrm{u}_{i+1,j} - \mathrm{u}_{i-1,j} - \mathrm{u}_{i,j+1} - \mathrm{u}_{i,j-1} = h^2 \mathrm{f}_{ij} \text{ für } (i,j) \in \mathbb{I}_{G_h}, \quad \mathrm{u}_{ij} = \mathrm{g}_{ij} \text{ für } (i,j) \in \mathbb{I}_{\partial G_h}.$$

Offenbar sind die Werte von u_{ij} für $(i,j) \in \mathbb{I}_{\partial G_h}$ durch die Randwerte gegeben. Es ist also u_{ij} nur für alle inneren Gitterpunkte zu berechnen, wobei $|G_h| = N$ mit $N = N_x^2$ gilt. Wir verwenden die übliche lexikographische Ordnung und definieren

$$\mathrm{u} = \begin{pmatrix} \mathrm{u}_{11} \\ \vdots \\ \mathrm{u}_{N_x,1} \\ \mathrm{u}_{12} \\ \vdots \\ \mathrm{u}_{N_x,2} \\ \mathrm{u}_{13} \\ \vdots \\ \mathrm{u}_{N_x N_x} \end{pmatrix} \in \mathbb{R}^N, \quad \mathrm{f} = h^2 \begin{pmatrix} \mathrm{f}_{11} \\ \vdots \\ \mathrm{f}_{N_x,1} \\ \mathrm{f}_{12} \\ \vdots \\ \mathrm{f}_{N_x,2} \\ \mathrm{f}_{13} \\ \vdots \\ \mathrm{f}_{N_x N_x} \end{pmatrix} \in \mathbb{R}^N, \quad \mathrm{g} = \begin{pmatrix} \mathrm{g}^1 + \tilde{\mathrm{g}}^0 \\ \mathrm{g}^2 \\ \vdots \\ \mathrm{g}^{N_x - 1} \\ \mathrm{g}^{N_x} + \tilde{\mathrm{g}}^{N_x+1} \end{pmatrix} \in \mathbb{R}^N$$

mit $N = N_x^2$ und

$$\mathrm{g}^k = \begin{pmatrix} \mathrm{g}_{0k} \\ 0 \\ \vdots \\ 0 \\ \mathrm{g}_{N_x+1,k} \end{pmatrix} \in \mathbb{R}^{N_x}, \ k \in \{1, \ldots, N_x\}, \quad \tilde{\mathrm{g}}^k = \begin{pmatrix} \mathrm{g}_{1,k} \\ \vdots \\ \mathrm{g}_{N_x,k} \end{pmatrix} \in \mathbb{R}^{N_x}, \ k \in \{0, N_x + 1\}.$$

Dann lässt sich das diskrete Poissonproblem als Matrixgleichung

$$\mathrm{Au} = \mathrm{f} + \mathrm{g}$$

schreiben, wobei die Koeffisizeintenmatrix $\mathrm{A} \in \mathbb{R}^{N \times N}$ durch die *Block-Tridiagonalmatrix*

$$\mathrm{A} = \begin{pmatrix} T & -I & 0 & & 0 \\ -I & T & -I & & \\ & \ddots & \ddots & \ddots & \\ & & -I & T & -I \\ 0 & & & -I & T \end{pmatrix}, \quad T = \begin{pmatrix} 4 & -1 & 0 & & 0 \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ 0 & & & -1 & 4 \end{pmatrix} \in \mathbb{R}^{N_x \times N_x}$$

gegeben ist, wobei $I \in \mathbb{R}^{N_x \times N_x}$ die Einheitsmatrix bezeichnet. ◊

2.3 Konsistenz, Stabilität und Konvergenz

Es stellt sich die Frage, ob die Lösung u_h des diskreten Problems (2.7) eine Approximation der Lösung u des kontinuierlichen Problems (2.1) liefert. Folgt aus $h \to 0$, dass der Approximationsfehler in einer geeigneten Norm ebenfalls gegen null konvergiert?

Definition 2.9. *Wir gehen von dem Schema* (2.2) *aus. Dabei seinen* X_0, Y_0, X_h *und* Y_h *normierte Räume und* $X \subset X_0$ *und* $Y \subset Y_0$ *Teilräume. Das Schema heißt*

1) konsistent *in* $u \in X$, *falls gilt*

$$\|\mathcal{T}_h \mathcal{D}_h^X u - \mathcal{D}_h^Y \mathcal{T} u\|_{Y_h} \to 0 \quad (h \to 0),$$

2) stabil, *falls es eine von* h *unabhängige Konstante* $c_s > 0$ *gibt, so dass für alle* $v_h \in X_h$ *gilt*

$$\|v_h\|_{X_h} \leq c_s \|\mathcal{T}_h v_h\|_{Y_h},$$

3) konvergent, *falls für die Lösungen* $u \in X$ *und* $u_h \in X_h$ *von* $\mathcal{T} u = (f, g) \in Y$ *beziehungsweise* (2.5) *gilt*

$$\|u_h - \mathcal{D}_h^X u\|_{X_h} \to 0 \quad (h \to 0).$$

Satz 2.10. *Konsistenz und Stabilität implizieren Konvergenz, falls die Daten konsistent approximiert werden, also*

$$\|b_h - \mathcal{D}_h^y b\|_{Y_h} \to 0 \quad (h \to 0)$$

gilt. Insbesondere erhalten wir Konvergenz, wenn wir $b_h = \mathcal{D}_h^Y b$ *setzen.*

Beweis. Wir zeigen die Aussage analog wie im Beweis von [Dzi10, Satz 3.5]. Seien $u \in X$, $u_h \in X_h$, $b \in Y$, $b_h \in Y_h$ mit $\mathcal{T} u = b$ und $\mathcal{T}_h u_u = b_h$. Dann erhalten wir mit Hilfe der Stabilität

$$\begin{aligned}
\|u_h - \mathcal{D}_h^X u\|_{X_h} &\leq c_s \|\mathcal{T}_h(u_u - \mathcal{D}_h^X u)\|_{Y_h} = c_s \|\mathcal{T}_h u_h - \mathcal{T}_h \mathcal{D}_h^X u\|_{Y_h} \\
&= c_s \|(\mathcal{T}_h u_h - b_h) + (b_h - \mathcal{D}_h^Y b) + (\mathcal{D}_h^Y b - \mathcal{T}_h \mathcal{D}_h^X u)\|_{Y_h} \\
&\leq c_s \left(\|b_h - \mathcal{D}_h^Y b\|_{Y_h} + \|\mathcal{D}_h^Y \mathcal{T} u - \mathcal{T}_h \mathcal{D}_h^X u\|_{Y_h} \right),
\end{aligned}$$

was die Behauptung des Satzes ergibt. □

Bemerkung 2.11. 1) Für nichtlineare Abbildungen $\mathcal{T}$ ist Satz 2.10 im Allgemeinen falsch.

2) Liegt neben der Stabilität sogar Konsistenz und Datenapproximation mit der Ordnung $\beta > 0$ vor, das heißt, gelten

$$\|\mathcal{T}_h \mathcal{D}_h^X u - \mathcal{D}_h^Y \mathcal{T} u\|_{Y_h} = \mathcal{O}(h^\beta) \text{ für } h \to 0 \quad \text{und} \quad \|b_h - \mathcal{D}_h^y b\|_{Y_h} = \mathcal{O}(h^\beta) \text{ für } h \to 0,$$

so folgt Konvergenz der Ordnung β:

$$\|\mathcal{D}_h^X u - u_h\|_{X_h} = \mathcal{O}(h^\beta) \quad \text{für } h \to 0.$$

Der Beweis folgt wie im Beweis von Satz 2.10. ◊

Wir kehren nun wieder zu unserem Problem (2.1) zurück und wählen folgende Normen:

$$\begin{aligned}
\|v\|_X &= \max_{x \in \overline{G}} |v(x)| && \text{für } u \in X, \\
\|v_h\|_{X_h} &= \max_{x_h \in \overline{G}_h} |v_h(x_h)| && \text{für } u_h \in X_h, \\
\|(f, g)\|_Y &= \sup_{x \in G} |f(x)| + \sup_{x \in \partial G} |g(x)| && \text{für } (f, g) \in Y, \\
\|(f_h, g_h)\|_{Y_h} &= \max_{x_h \in \overline{G}_h} |f_h(x_h)| + \max_{x_h \in \partial G_h} |g_h(x_h)| && \text{für } (f_h, g_h) \in Y_h.
\end{aligned}$$

Wir weisen Konsistenz und Stabilität für das Finite-Differenzenverfahren nach. Dann erhalten wir Konvergenz mit Satz 2.10. Die Konsistenz erhalten wir durch Verwendung geeigneter Taylorentwicklungen. Dazu bemerken wir, dass

$$\begin{aligned}
\mathcal{T}_h \mathcal{D}_h^X u &= \left(-\Delta_h(\mathcal{D}_h^X u), (\mathcal{D}_h^X u)\big|_{\partial G_h} \right) = \left(-\Delta_h\big(u|_{\overline{G}_h}\big), u\big|_{\partial G_h} \right), \\
\mathcal{D}_h^Y \mathcal{T} u &= \mathcal{D}_h^Y \left(-\Delta u, u\big|_{\partial G} \right) = \left(-(\Delta u)\big|_{G_h}, u\big|_{\partial G_h} \right).
\end{aligned}$$

gelten. Also erhalten wir

$$\mathcal{T}_h\mathcal{D}_h^X u - \mathcal{D}_h^Y\mathcal{T}u = \Big(-\Delta_h\big(u|_{\overline{G}_h}\big) + (\Delta u)\big|_{G_h}, 0\Big)$$

und damit dann

$$\|\mathcal{T}_h\mathcal{D}_h^X u - \mathcal{D}_h^Y\mathcal{T}u\|_{Y_h} = \max_{x_h\in G_h}\big|\Delta_h\big(u|_{\overline{G}_h}\big)(x_h) - (\Delta u)(x_h)\big|.$$

Unter Verwendung der Taylorentwcklung (vergleiche [DR11, Satz 11.27]) erhalten wir für $u \in C^2(\overline{G})$ an einem Gitterpunkt $x_h \in G_h$ die beiden Gleichungen

$$u(x_h \pm he_j) = u(x_h) \pm u_{x_j}(x_h)h + \frac{1}{2}u_{x_jx_j}(x_h)h^2 + o(h^2) \quad (h\to 0) \quad \text{für } j = 1,\dots,n,$$

die wir kompakt als eine Gleichung geschrieben haben. Addition beider Gleichungen ergibt nach Summation über $j = 1,\dots,n$

$$\Delta u(x_h) - \Delta_h\big(u|_{\overline{G}_h}\big)(x_h) = \Delta u(x_h) + \frac{1}{h^2}\left(2nu(x_h) - \sum_{j=1}^{n}\big(u(x_h+he_j)+u(x_h-he_j)\big)\right) \to 0 \quad (h\to 0).$$

Wir erhalten damit das folgende Lemma.

Lemma 2.12. *Seien $G = (0,1)^n$ und $u \in C^2(\overline{G})$. Dann gilt für den in (2.4) definierten diskreten Laplace-Operator Δ_h*

$$\max_{x_h\in G_h}\Big|\Delta u(x_h) - \Delta_h\big(u|_{\overline{G}_h}\big)(x_h)\Big| \to 0 \quad (h\to 0).$$

Damit ist das Finite-Differenzenverfahren konsistent. Gilt sogar $u \in C^4(\overline{G})$, so erhalten wir Konsistenz der Ordnung zwei:

$$\max_{x_h\in G_h}\Big|\Delta u(x_h) - \Delta_h\big(u|_{\overline{G}_h}\big)(x_h)\Big| = \mathcal{O}(h^2) \quad (h\to 0).$$

Beweis. Den ersten Teil des Satzes haben wir bereits bewiesen. Der zweite folgt, wenn wir die obige Taylorentwicklung fortsetzen:

$$u(x_h \pm he_j) = u(x_h) \pm u_{x_j}(x_h)h + \frac{1}{2}u_{x_jx_j}(x_h)h^2 \pm \frac{1}{6}u_{x_jx_jx_j}(x_h)h^3 + \frac{1}{24}u_{x_jx_jx_jx_j}(\xi_{j,h}^{\pm})h^4$$

mit Zwischenstellen $\xi_{j,h}^{\pm}$ auf den Verbindungslinien x_h und $x_h \pm he_j$ für $j = 1,\dots,n$. Nun folgt die Konsistenzordnung wieder durch Addition beider Gleichungen und Summation über j. Ferner ist das Maximum von $u_{x_jx_jx_jx_j}$ auf $\overline{G}$ beschränkt nach Voraussetzung. □

Wir zeigen nun die Stabilität des Differenzenverfahrens (vergleiche [Dzi10, Lemma 3.6]).

Lemma 2.13. *Sei $G = (0,1)^n$. Dann ist das Finite-Differenzenverfahren stabil, das heißt, es gibt eine Konstante $c_s > 0$, so dass für alle $h > 0$ und alle $v_h \in X_h$ die Abschätzung*

$$\|v_h\|_{X_h} \le c_s\|\mathcal{T}_h v_h\|_{Y_h} = c_s\left(\max_{x_h\in G_h}|\Delta_h v(x_h)| + \max_{x_h\in\partial G_h}|v_h(x_h)|\right)$$

erfüllt ist.

Beweis. Wir wählen ein beliebiges $v_h \in X_h$ und definieren die Gitterfunktion $f_h : G_h \to \mathbb{R}$ durch $f_h(x_h) = -\Delta_h v_h(x_h)$ für $x_h \in G_h$. Offenbar folgt

$$-\Delta_h v_h(x_h) = f_h(x_h) \le c_h \quad \text{für alle } x_h \in G_h \text{ mit } c_h = \max\left\{0, \max_{\tilde{x}_h\in G_h} f_h(\tilde{x}_h)\right\} \ge 0. \tag{2.9}$$

Für die quadratische Funktion $w(\boldsymbol{x}) = x_1^2$, $\boldsymbol{x} \in (x_1, \ldots, x_n) \in \overline{G}$, erhalten wir

$$\begin{aligned} -\Delta_h w(\boldsymbol{x}) &= \frac{1}{h^2}\left(2nw(\boldsymbol{x}) - \sum_{j=1}^{n}\left(w(\boldsymbol{x}+h\mathrm{e}_j) + w(\boldsymbol{x}-h\mathrm{e}_j)\right)\right) \\ &= \frac{1}{h^2}\left(2nx_1^2 - (x_1+h)^2 - (n-1)x_1^2 - (x_1-h)^2 - (n-1)x_1^2\right) \\ &= \frac{1}{h^2}\left(2x_1^2 - (x_1^2 + 2x_1h + h^2) - (x_1 - 2x_1h + h^2)\right) = -2 \end{aligned} \tag{2.10}$$

für $\boldsymbol{x} \in [h, 1-h]^n \subsetneq G$. Wir bemerken hier, dass damit $\Delta w(\boldsymbol{x}) = \Delta_h w(\boldsymbol{x})$ gilt. Mit der in (2.9) definierten Konstante c_h sei die Gitterfunktion $u_h \in X_h$ gegeben durch $u_h(\boldsymbol{x}_h) = v_h(\boldsymbol{x}_h) + c_h w(\boldsymbol{x}_h)/2$ für $\boldsymbol{x}_h \in \overline{G}_h$. Dann erhalten wir mit (2.10) die Ungleichung

$$-\Delta_h u_h(\boldsymbol{x}_h) = -\Delta_h v_h(\boldsymbol{x}_h) + \frac{c_h}{2}(-\Delta_h w)(\boldsymbol{x}_h) \le c_h + \frac{c_h}{2}(-2) = 0 \quad \text{für } \boldsymbol{x}_h \in G_h.$$

Definieren wir die Gitterfunktion $g_h(\boldsymbol{x}_h) = \max_{\tilde{\boldsymbol{x}}_h \in \partial G_h} v_h(\tilde{\boldsymbol{x}}_h)$ für $\boldsymbol{x}_h \in \partial G_h$, so können wir wegen $-\Delta u_h \le 0$ in G_h das diskrete Maximumprinzip (Lemma 2.5) anwenden:

$$\begin{aligned} \max_{\boldsymbol{x}_h \in \overline{G}_h} u_h(\boldsymbol{x}_h) &= \max_{\boldsymbol{x}_h \in \partial G_h} u_h(\boldsymbol{x}_h) = \max_{\boldsymbol{x}_h \in \partial G_h}\left(v_h(\boldsymbol{x}_h) + \frac{c_h}{2}x_{h,1}^2\right) \le \max_{\boldsymbol{x}_h \in \partial G_h} g_h(\boldsymbol{x}_h) + \frac{c_h}{2}1^2 \\ &\le \max_{\boldsymbol{x}_h \in \partial G_h} g_h(\boldsymbol{x}_h) + \frac{1}{2}\max\left\{0, \max_{\tilde{\boldsymbol{x}}_h \in G_h} f_h(\tilde{\boldsymbol{x}}_h)\right\} \\ &\le \max_{\boldsymbol{x}_h \in \partial G_h} |g_h(\boldsymbol{x}_h)| + \frac{1}{2}\max_{\tilde{\boldsymbol{x}}_h \in G_h} |f_h(\tilde{\boldsymbol{x}}_h)|, \end{aligned} \tag{2.11}$$

wobei wir $c_h \ge 0$ und die Notation $\boldsymbol{x}_h = (x_{h,1}, \ldots, x_{h,n}) \in \partial G_h$ verwendet haben. Statt (2.9) gilt auch

$$-\Delta_h v_h(\boldsymbol{x}_h) = f_h(\boldsymbol{x}_h) \ge \tilde{c}_h \quad \text{für alle } \boldsymbol{x}_h \in G_h \text{ mit } \tilde{c}_h = \min\left\{0, \min_{\tilde{\boldsymbol{x}}_h \in G_h} f_h(\tilde{\boldsymbol{x}}_h)\right\} \le 0.$$

Wir führen nun $\tilde{u}_h \in X_h$ durch $\tilde{u}_h(\boldsymbol{x}_h) = v_h(\boldsymbol{x}_h) + \tilde{c}_h w(\boldsymbol{x}_h)/2$ für $\boldsymbol{x}_h \in \overline{G}_h$ ein. Dann gilt

$$-\Delta_h \tilde{u}_h(\boldsymbol{x}_h) = -\Delta_h v_h(\boldsymbol{x}_h) + \frac{\tilde{c}_h}{2}(-\Delta_h w)(\boldsymbol{x}_h) \ge \tilde{c}_h + \frac{\tilde{c}_h}{2}(-2) = 0 \quad \text{für } \boldsymbol{x}_h \in G_h.$$

Jetzt lässt sich wegen $-\Delta_h \tilde{u}_h \ge 0$ in G_h das diskrete Minimumprinzip (vergleiche Bemerkung 2.6) anwenden. Es folgt daher wegen $\tilde{c}_h \le 0$

$$\begin{aligned} \min_{\boldsymbol{x}_h \in \overline{G}_h} u_h(\boldsymbol{x}_h) &= \min_{\boldsymbol{x}_h \in \partial G_h} u_h(\boldsymbol{x}_h) = \min_{\boldsymbol{x}_h \in \partial G_h}\left(v_h(\boldsymbol{x}_h) + \frac{\tilde{c}_h}{2}x_{h,1}^2\right) \ge \min_{\boldsymbol{x}_h \in \partial G_h} g_h(\boldsymbol{x}_h) + \frac{\tilde{c}_h}{2}1^2 \\ &\ge \min_{\boldsymbol{x}_h \in \partial G_h} g_h(\boldsymbol{x}_h) + \frac{1}{2}\min\left\{0, \min_{\tilde{\boldsymbol{x}}_h \in G_h} f_h(\tilde{\boldsymbol{x}}_h)\right\} \\ &\ge -\min_{\boldsymbol{x}_h \in \partial G_h} |g_h(\boldsymbol{x}_h)| - \frac{1}{2}\min_{\tilde{\boldsymbol{x}}_h \in G_h} |f_h(\tilde{\boldsymbol{x}}_h)|. \end{aligned} \tag{2.12}$$

Aus (2.11) und (2.12) schließen wir

$$\max_{\boldsymbol{x} \in \overline{G}_h} |u_h(\boldsymbol{x}_h)| = \left\{\max_{\boldsymbol{x} \in \overline{G}_h} u_h(\boldsymbol{x}_h), -\min_{\boldsymbol{x}_h \in \overline{G}_h} u_h(\boldsymbol{x}_h)\right\} \le \max_{\boldsymbol{x}_h \in \partial G_h} |g_h(\boldsymbol{x}_h)| + \frac{1}{2}\max_{\boldsymbol{x}_h \in G_h} |f_h(\boldsymbol{x}_h)|.$$

Damit und mit $v_h(\boldsymbol{x}_h) = u_h(\boldsymbol{x}_h) - c_h w(\boldsymbol{x}_h)/2$ für $\boldsymbol{x}_h \in \overline{G}_h$ folgt

$$\begin{aligned} \|v_h\|_{X_h} &= \max_{\boldsymbol{x}_h \in \overline{G}_h} |v_h(\boldsymbol{x}_h)| = \max_{\boldsymbol{x}_h \in \overline{G}_h}\left|u_h(\boldsymbol{x}) - \frac{c_h}{2}w_h(\boldsymbol{x}_h)\right| \le \max_{\boldsymbol{x}_h \in \overline{G}_h} |u_h(\boldsymbol{x}_h)| + \frac{|c_h|}{2} \\ &\le \max_{\boldsymbol{x}_h \in \partial G_h} |g_h(\boldsymbol{x}_h)| + \max_{\boldsymbol{x}_h \in G_h} |f_h(\boldsymbol{x})| \le \max_{\boldsymbol{x}_h \in \partial G_h} |v_h(\boldsymbol{x}_h)| + \max_{\boldsymbol{x}_h \in G_h} |\Delta_h v_h(\boldsymbol{x})| = c_{\mathsf{s}} \|\mathcal{T}_h v_h\|_{Y_h} \end{aligned}$$

mit der Konstante $c_{\mathsf{s}} = 1$. □

Wir erhalten nun das folgende Konvergenzresultat (vergleiche [Dzi10, Satz 3.7]).

Satz 2.14. *Seien $G = (0,1)^n \subset \mathbb{R}^n$ und das Finite-Differenzengitter $\overline{G}_h = G_h \cup \partial G_h$ wie oben definiert. Zu gegebenem $(f, g) \in Y$ sei $u \in X$ die eindeutige Lösung von (2.1). Wir setzen $(f_h, g_h) = (\mathcal{D}_h^X f, \mathcal{D}_h^Y g) \in Y_h$. Dann existiert genau eine Lösung $u_h \in X_h$ von (2.7). Gilt $u \in C^2(\overline{G}) \subset X$, so ist das Finite-Differenzenverfahren konvergent, das heißt, es gilt*

$$\|u_h - \mathcal{D}_h^X u\|_{X_h} \to 0 \quad \text{für } h \to 0.$$

Haben wir sogar $u \in C^4(\overline{G})$, so erhalten wir Konvergenzordnung zwei:

$$\|u_h - \mathcal{D}_h^X u\|_{X_h} = \mathcal{O}(h^2) \quad (h \to 0).$$

Beweis. Der Beweis der Aussage folgt direkt aus Satz 2.3, Lemma 2.12 und Lemma 2.13. □

3 Galerkin-Verfahren für das Poissonproblem

In diesem Abschnitt werden wir uns mit Diskretisierungsmethoden für die schwache Formulierung des Poissonproblems. Für mehr Details verweisen wir auf [BS08, Kapitel 5].

3.1 Schwache Formulierung des Poissonproblems

Sei G ein beschränktes *Gebiet* in $\mathbb{R}^n$, das heißt, $G \subset \mathbb{R}^n$ ist offen, beschränkt und zusammenhängend. Punkte in G bezeichnen wir beispelsweise mit $\boldsymbol{x} = (x_1, \dots, x_n)$. Wir setzen $H = L^2(\Omega)$. Dann ist H ein Hilbertraum mit dem Skalarprodukt

$$\langle \varphi, \phi \rangle_H = \int_G \varphi \phi \, \mathrm{d}\boldsymbol{x} \quad \text{für } \varphi, \phi \in H.$$

Ferner definieren wir den Raum $V = H_0^1(G)$ und setzen

$$\langle \varphi, \phi \rangle_V = \int_G \nabla\varphi \cdot \nabla\phi \, \mathrm{d}\boldsymbol{x} = \sum_{i=1}^{n} \int_G \frac{\partial \varphi}{\partial x_i} \frac{\partial \phi}{\partial x_i} \, \mathrm{d}\boldsymbol{x} \quad \text{für } \varphi, \phi \in V.$$

Wir erinnern an die erste *Poincaré-Ungleichung* [DR12, Satz 16.23]: Es gibt eine Poincaré-Konstante $c_\mathsf{p} > 0$ mit

$$\|\varphi\|_H \leq c_\mathsf{p} \|\varphi\|_V \quad \text{für alle } \varphi \in V. \tag{3.1}$$

Die Ungleichung (3.1) garantiert, dass $\langle \cdot, \cdot \rangle_V$ ein Skalarprodukt auf V ist. Ferner gilt $V \hookrightarrow H$. Zu gegebenem $f \in V' = H^{-1}(G)$ betrachten wir das Variationsproblem

$$\int_G \nabla u \cdot \nabla\varphi \, \mathrm{d}\boldsymbol{x} = \langle f, \varphi \rangle_{V',V} \quad \text{für alle } \varphi \in V. \tag{3.2}$$

Wir schreiben (3.2) auch in der Form

$$-\Delta u = f \quad \text{in } V'$$

und verstehen den Operator $-\Delta : V \to V'$ als

$$\langle -\Delta\varphi, \phi \rangle_{V',V} = \int_G \nabla\varphi \cdot \nabla\phi \, \mathrm{d}\boldsymbol{x} \quad \text{für } \varphi, \phi \in V.$$

Satz 3.1. *Zu jedem $f \in V'$ existiert genau ein $u \in V$, welches* (3.2) *löst. Ferner gilt die Abschätzung*

$$\|u\|_V \leq \|f\|_{V'}.$$

Beweis. Offenbar lässt sich (3.2) in der Form

$$\langle u, \varphi \rangle_V = \langle f, \varphi \rangle_{V',V} \quad \text{für alle } \varphi \in V.$$

schreiben. Damit folg die Existenz einer eindeutigen Lösung direkt aus dem Darstellungssatz von Riesz [DR12, Satz 12.24]. Zum Nachweis der Ungleichung wählen wir $\varphi = u$ in (3.2). Dann erhalten wir

$$\|u\|_V^2 = \langle f, u \rangle_{V',V} \leq \|f\|_{V'} \|u\|_V.$$

Division durch $\|u\|_V$ ergibt im Fall $u \neq 0$ die Behauptung. Für $u = 0$ ist die Abschätzung ohnehin klar. □

Bemerkung 3.2. 1) Die eindeutige Lösung der Minimierungsaufgabe

$$\min_{\varphi\in V} J(\varphi) \quad \text{mit } J : V \to \mathbb{R},\ J(\varphi) = \frac{1}{2}\|\varphi\|_V^2 - \langle f, \varphi\rangle_{V',V} \text{ für } \varphi \in V$$

ist charakterisiert durch die Gleichung (3.2). Für den Beweis zeigt man, dass die Zielfunktion J in V strikt konvex ist und J in u Gâteaux-differenzierbar ist. Dabei ist J Gâteaux-differenzierbar in u, wenn in u die *Richtungsableitung* [DR11, Definition 11.12]

$$J'(u;\varphi) = \lim_{t\to 0} \frac{J(u+t\varphi) - J(u)}{t}$$

von J für alle $\varphi \in V$ existieren und es eine lineare, beschränkte Abbildung $\mathcal{A}_u : V \to \mathbb{R}$ gibt mit

$$\mathcal{A}_u\varphi = J'(u;\varphi) \quad \text{für alle } \varphi \in V.$$

Wir setzen $J'(u) = \mathcal{A}_u$ und bezeichnen $J'(u)$ als *Gâteaux-Ableitung*. Die Gleichung (3.2) entspricht dann der hinreichenden Optimalitätsbedingung erster Ordnung

$$\langle J'(u), \varphi\rangle_{V',V} = 0 \quad \text{für alle } \varphi \in V,$$

wobei $J'(u) \in V'$ die Gâteaux-Ableitung von J an u bezeichnet.

2) Gilt sogar $f \in H \hookrightarrow V'$, so können wir die rechte Seite in (3.2) auch in der Form

$$\langle f, \varphi\rangle_{V',V} = \langle f, \varphi\rangle_H = \int_G f\varphi \,\mathrm{d}\boldsymbol{x} \quad \text{für } \varphi \in H \tag{3.3}$$

schreiben.

3) Angenommen, $u \in C^2(G)\cap C(\overline{G})$ ist eine *starke Lösung des Poissonproblems* mit homogenen Dirichlet-Randbedingungen, das heißt, es gilt

$$-\Delta u = f \text{ in } G, \quad u = 0 \text{ auf } \partial G \tag{3.4}$$

für ein $f \in C(\overline{G})$. Multiplikation (3.4) mit $\varphi \in C_0^1(G) \cap C(\overline{G})$ und Integration über G ergibt

$$-\int_G \Delta u\varphi \,\mathrm{d}\boldsymbol{x} = \int_G f\varphi \,\mathrm{d}\boldsymbol{x}. \tag{3.5}$$

Ist der Rand ∂G von G glatt gemäß [DR11, Definition 13.64], so gilt mit der *Greenschen Formel* [DR11, Folgerung 13.70-(ii)]

$$-\int_G \Delta u\varphi \,\mathrm{d}\boldsymbol{x} = \int_G \nabla u\cdot\nabla\varphi \,\mathrm{d}\boldsymbol{x} - \int_{\partial G} \langle\varphi\nabla u, \boldsymbol{\nu}\rangle_2 \,\varphi \,\mathrm{d}A = \int_G \nabla u\cdot\nabla\varphi \,\mathrm{d}\boldsymbol{x},$$

wobei wir $\varphi = 0$ auf ∂G genutzt haben und $\boldsymbol{\nu}$ die äußere Normale bezeichnet. Für die Definition des Oberflächenintegrals

$$\begin{aligned}\int_{\partial G} \langle\varphi\nabla u, \boldsymbol{\nu}\rangle_2 \,\mathrm{d}A &= \int_{\partial G} \langle\varphi(\boldsymbol{x})\nabla u(\boldsymbol{x}), \boldsymbol{\nu}(\boldsymbol{x})\rangle_2 \,\mathrm{d}A(\boldsymbol{x}) = \int_{\partial G} \varphi(\boldsymbol{x})\,\langle\nabla u(\boldsymbol{x}), \boldsymbol{\nu}(\boldsymbol{x})\rangle_2 \,\mathrm{d}A(\boldsymbol{x}) \\ &= \int_{\partial G} \varphi(\boldsymbol{x})\,\nabla u(\boldsymbol{x})\cdot\boldsymbol{\nu}(\boldsymbol{x}) \,\mathrm{d}A(\boldsymbol{x}) = \int_{\partial G} \varphi(\boldsymbol{x})\,\frac{\partial u}{\partial\boldsymbol{\nu}}(\boldsymbol{x}) \,\mathrm{d}A(\boldsymbol{x})\end{aligned}$$

mittels geeigneter Parametrisierung der $(n-1)$-dimensionalen Fläche ∂G verweisen wir auf [DR11, Folgerung 13.57]. Wegen $f \in C(\overline{G}) \hookrightarrow H$ können wir (3.3) anwenden. Wir erhalten daher für (3.5)

$$\int_G \nabla u\cdot\nabla\varphi \,\mathrm{d}\boldsymbol{x} = \langle f, \varphi\rangle_{V',V}.$$

Da $\varphi \in C^1(G) \cap C(\overline{G})$ beliebig gewählt ist und $C^1(G) \cap C(\overline{G})$ eine dichte Teilmenge in V ist, genügt eine starke Lösung des Poissonproblems der Variationsgleichung (3.2). ◊

Bemerkung 3.2-3) motiviert die folgende Definition.

Definition 3.3. *Die eindeutige Lösung des Variationsproblems* (3.2) *wird* schwache Lösung des Poisonproblem *genannt.*

3.2 Das Galerkin-Verfahren für das Poissonproblem

Wir wählen N linear unabhänige Funktionen $\{\varphi_i\}_{i=1}^N \subset V$ und definieren den endlich-dimensionalen Teilraum

$$V_h = \operatorname{Span}\{\varphi_1, \dots, \varphi_N\} \subset V. \tag{3.6}$$

Die Idee des *Galerkin-Verfahrens* ist es nun, ein $u_h \in V_h$ zu bestimmen, so dass

$$\int_G \nabla u_h \cdot \nabla \varphi_h \, \mathrm{d}\boldsymbol{x} = \langle f, \varphi_h \rangle_{V',V} \quad \text{für alle } \varphi \in V_h \tag{3.7}$$

erfüllt ist.

Satz 3.4. *Seien $\varphi_1, \dots, \varphi_N$ linear unabhängige Funktionen in V und V_h wie in* (3.6) *definiert. Dann gibt es zu jedem $f \in V'$ genau ein $u_h \in V_h$, welches* (3.2) *löst. Ferner gilt die Abschätzung*

$$\|u_h\|_V \leq \|f_h\|_{(V_h)'} \quad \textit{mit } f_h = f|_{V_h}.$$

Beweis. Wir können (3.7) in der Form

$$\langle u_h, \varphi_h \rangle_V = \langle f, \varphi_h \rangle_{V',V} \quad \text{für alle } \varphi \in V_h \tag{3.8}$$

schreiben, Offenbar ist V_h als Teilraum von V wieder ein Hilbertraum mit der gleichen Topologie wie in V. Ferner ist $f_h = f|_{V_h}$ ein Element in $(V_h)'$. Nach dem Darstellungssatz von Riesz [DR12, Satz 12.24] existiert damit eine eindeutige Lösung $u_h \in V_h$ von (3.8), die damit auch (3.7) löst. Zum Nachweis der Ungleichung wählen wir $\varphi_h = u_h$ in (3.8) und bekommen

$$\|u_h\|_V^2 = \langle u_h, u_h \rangle_V = \langle f, u_h \rangle_{V',V} = \langle f_h, u_h \rangle_{(V_h)',V_h} \leq \|f_h\|_{(V_h)'} \|u_h\|_V,$$

woraus die Abschätzung folgt. □

Wir wollen nun die Lösung u_h von (3.7) berechnen. Da jedes $\varphi_h \in V_h$ als Linearkombination der N linear unabhängigen Funktionen $\{\varphi_i\}_{i=1}^N$ geschrieben werden kann, ist (3.7) äquivalent mit den N linearen Gleichungen

$$\int_G \nabla u_h \cdot \nabla \varphi_i \, \mathrm{d}\boldsymbol{x} = \langle f, \varphi_i \rangle_{V',V} \quad \text{für } i = 1, \dots, N \tag{3.9}$$

zur Bestimmung von $u_h \in V_h$. Wegen (3.6) hat u_h die Darstellung

$$u_h(\boldsymbol{x}) = \sum_{j=1}^N u_j \varphi_j(\boldsymbol{x}) \quad \text{für } \boldsymbol{x} \in \overline{G} \tag{3.10}$$

mit einem zu bestimmenden Koeffizientenvektor $\mathrm{u} = (\mathrm{u}_1, \dots, \mathrm{u}_N)^\top \in \mathbb{R}^N$ Einsetzen von (3.10) in (3.9) ergibt

$$\sum_{j=1}^N \mathrm{u}_j \int_G \nabla \varphi_j \cdot \nabla \varphi_i \, \mathrm{d}\boldsymbol{x} = \langle f, \varphi_i \rangle_{V',V} \quad \text{für } i = 1, \dots, N. \tag{3.11}$$

Wir definieren die Steifigkeitsmatrix

$$\mathrm{S} = \big((\mathrm{S}_{ij})\big) \in \mathbb{R}^{N \times N} \quad \text{mit } \mathrm{S}_{ij} = \int_G \nabla \varphi_j \cdot \nabla \varphi_i \, \mathrm{d}\boldsymbol{x} = \langle \varphi_j, \varphi_i \rangle_V \text{ für } 1 \leq i, j \leq N \tag{3.12}$$

und den Vektor

$$\mathrm{f} = \big(\mathrm{f}_i\big) \in \mathbb{R}^N \quad \text{mit } \mathrm{f}_i = \langle f, \varphi_i \rangle_{V',V} \text{ für } 1 \leq i \leq N$$

für die rechte Seite. Dann lässt sich (3.11) als Lineares Gleichungssystem in Matrixform schreiben:

$$\mathrm{S}\mathrm{u} = \mathrm{f}.$$

Lemma 3.5. *Seinen $\{\varphi_i\}_{i=1}^N \subset V$ linear unabhängig in V. Dann ist die Steifigkeitsmatrix* S *symmetrisch und positiv definit.*

Beweis. Die Symmetrie von S ist offensichtlich. Sei $\mathsf{v} = (\mathsf{v}_i) \in \mathbb{R}^N$ beliebig gewählt. Wir setzen $v_h = \sum_{j=1}^N \mathsf{v}_j \varphi_j \in V_h$ und erhalten mit (3.1) die Abschätzung

$$\mathsf{v}^\top \mathsf{S}\mathsf{v} = \sum_{i=1}^N \sum_{j=1}^N \mathsf{v}_i \mathsf{S}_{ij} \mathsf{v}_j = \sum_{i=1}^N \sum_{j=1}^N \mathsf{v}_i \mathsf{v}_j \langle \varphi_j, \varphi_i \rangle_V = \left\langle \sum_{j=1}^N \mathsf{v}_j \varphi_j, \sum_{i=1}^N \mathsf{v}_i \varphi_i \right\rangle_V = \|v_h\|_V^2 \geq \frac{1}{c_{\mathsf{p}}^2} \|v_h\|_H^2 \geq 0.$$

Es gilt $\mathsf{v}^\top \mathsf{S}\mathsf{v} = 0$ genau dann, wenn $v_h = 0$ ist. Das bedeutet aber, dass $\mathsf{v} = 0$ gilt. □

Bemerkung 3.6. Wir hätten im Beweis von Lemma 3.5 auch so argumentieren können: Aus $\mathsf{v}^\top \mathsf{S}\mathsf{v} = 0$ folgt, dass $v_h \in V_h \subset V$ konstant sein muss. Das geht nur, wenn $v_h = 0$ gilt. ◊

Wir können nun den Fehler zwischen der kontinuierlichen und diskreten Lösung wie folgt abschätzen.

Satz 3.7 (Céa Lemma)**.** *Seinen $G \subset \mathbb{R}^n$ ein beschränktes Gebiet, $f \in V'$ und V_h wie in* (3.6) *mit linear unabhängigen Funktionen $\{\varphi_i\}_{i=1}^N$. Dann gilt für die Lösung $u \in V$ von* (3.2) *und die Lösung $u_h \in V_h$ von* (3.9) *die Gleichung*

$$\|u - u_h\|_V = \inf \left\{ \|u - \varphi_h\|_V \,\middle|\, \varphi_h \in V_h \right\}. \tag{3.13}$$

Beweis. Da wir $\varphi_h = u_h \in V_h$ wählen können, ist

$$\|u - u_h\|_V \geq \inf \left\{ \|u - \varphi_h\|_V \,\middle|\, \varphi_h \in V_h \right\} \tag{3.14}$$

klar. Um Satz 3.7 zu zeigen, orientieren wir uns an dem Beweis von [Dzi10, Satz 3.10]. Wegen $V_h \subset V$ können wir in (3.2) auch $\varphi = \varphi_h \in V_h$ wählen und bekommen

$$\int_G \nabla u \cdot \nabla \varphi_h \, \mathrm{d}\boldsymbol{x} = \langle f, \varphi_h \rangle_{V',V} \quad \text{für alle } \varphi_h \in V_h. \tag{3.15}$$

Subtraktion der Gleichungen (3.15) und (3.9) ergibt die *Galerkin-Orthogonalität des Fehlers*

$$\int_G \nabla(u - u_h) \cdot \nabla \varphi_h \, \mathrm{d}\boldsymbol{x} = \langle u - u_h, \varphi_h \rangle_V = 0 \quad \text{für alle } \varphi_h \in V_h. \tag{3.16}$$

Für jedes $\varphi_h \in V_h$ gilt daher

$$\begin{aligned}
\|u - u_h\|_V^2 &= \int_G \nabla(u - u_h) \cdot \nabla u \, \mathrm{d}\boldsymbol{x} - \underbrace{\int_G \nabla(u - u_h) \cdot \nabla u_h \, \mathrm{d}\boldsymbol{x}}_{=0 \text{ wegen (3.16) und } u_h \in V_h} \\
&= \int_G \nabla(u - u_h) \cdot \nabla u \, \mathrm{d}\boldsymbol{x} \\
&= \int_G \nabla(u - u_h) \cdot \nabla u \, \mathrm{d}\boldsymbol{x} - \underbrace{\int_G \nabla(u - u_h) \cdot \nabla \varphi_h \, \mathrm{d}\boldsymbol{x}}_{=0 \text{ wegen (3.16) und } \varphi_h \in V_h} \\
&= \int_G \nabla(u - u_h) \cdot \nabla(u - \varphi_h) \, \mathrm{d}\boldsymbol{x} \leq \|u - u_h\|_V \|u - \varphi_h\|_V.
\end{aligned}$$

Da $\varphi_h \in V_h$ beliebig gewählt ist, erhalten wir

$$\|u - u_h\|_V \leq \inf \left\{ \|u - \varphi_h\|_V \,\middle|\, \varphi_h \in V_h \right\}.$$

Daraus und aus (3.14) folgt (3.13). □

Wir wollen nun das beschriebene Galerkin-Verfahren in der Form des Schemas (2.2) beschreiben, Dazu wählen wir $X = V$ und $Y = X' = V'$. Ferner seinen $X_0 = H \supset V$ und $Y_0 = Y$. Wir definieren den linearen Operator $\mathcal{T} : X \to X'$ durch

$$\mathcal{T} = -\Delta, \quad \langle \mathcal{T}u, \varphi \rangle_{X',X} = \int_G \nabla u \cdot \nabla \varphi \, \mathrm{d}\boldsymbol{x} \text{ für } u, \varphi \in X.$$

Die Menge X_h ist der in (3.6) eingeführte endlich-dimensionale Teilraum V_h. Dabei verwenden wir auf X_h das gleiche Skalarprodukt und die gleiche Norm wie in X. Ferner setzen wir $Y_h = (X_h)' = (V_h)'$. Wegen $X_h \subset X$ können wir den Operator $\mathcal{T}_h : X_h \to Y_h$ als Restriktion definieren:

$$\mathcal{T}_h = \mathcal{T}\big|_{X_h}, \quad \langle \mathcal{T}_h u_h, \varphi_h \rangle_{(X_h)',X_h} = \int_G \nabla u_h \cdot \nabla \varphi_h \, \mathrm{d}\boldsymbol{x} \text{ für } u_h, \varphi_h \in X_h.$$

Für die Diskretisierungsoperator $\mathcal{D}_h^Y : Y \to Y_h$ wählen wir den linearen Restriktionsoperator

$$\mathcal{D}_h^Y f = f\big|_{X_h} \quad \text{für } f \in Y.$$

Aus

$$\begin{aligned}
\|\mathcal{D}_h^Y f\|_{Y_h} &= \sup\left\{ \langle \mathcal{D}_h^Y f, \varphi_h \rangle_{(X_h)',X_h} \,\middle|\, \varphi_h \in X_h \text{ mit } \|\varphi_h\|_{X_h} = 1 \right\} \\
&= \sup\left\{ \langle f, \varphi_h \rangle_{X',X} \,\middle|\, \varphi_h \in X_h \text{ mit } \|\varphi_h\|_X = 1 \right\} \\
&\le \sup\left\{ \langle f, \varphi \rangle_{X',X} \,\middle|\, \varphi \in X \text{ mit } \|\varphi\|_X = 1 \right\} = \|f\|_{X'} = \|f\|_Y
\end{aligned}$$

schließen wir, dass $\mathcal{D}_h^Y$ auch stetig ist. Die Wahl des Diskretisierungsoperators $\mathcal{D}_h^X : X \to X_h$ werden wir später genauer untersuchen. Hier an dieser Stelle wählen wir zunächst die lineare *Ritz-Projektion*: Zu $u \in X$ ist $u_h = \mathcal{D}_h^X u \in X_h$ die Lösung des Variationsproblems

$$\int_G \nabla u_h \cdot \nabla \varphi_h \, \mathrm{d}\boldsymbol{x} = \int_G \nabla u \cdot \nabla \varphi_h \, \mathrm{d}\boldsymbol{x} \quad \text{für alle } \varphi_h \in X_h.$$

Da $f : X \to \mathbb{R}$ mit

$$\langle f, \varphi \rangle_{X',X} = \int_G \nabla u \cdot \nabla \varphi \, \mathrm{d}\boldsymbol{x} \quad \text{für } \varphi_h \in X_h$$

ein stetiges, lineares Funktional auf X ist, folgt aus Satz 3.4, dass der Operator $\mathcal{D}_h^X$ wohldefiniert ist.

Nach Definition 2.9-1) erhalten wir für die Konsistenz

$$\begin{aligned}
\|\mathcal{T}_h \mathcal{D}_h^X u - \mathcal{D}_h^Y \mathcal{T} u\|_{Y_h} &= \sup\left\{ \langle \mathcal{T}_h \mathcal{D}_h^X u - \mathcal{D}_h^Y \mathcal{T} u, \varphi_h \rangle_{(X_h)',X_h} \,\middle|\, \varphi_h \in X_h \text{ mit } \|\varphi_h\|_{X_h} = 1 \right\} \\
&= \sup\left\{ \int_G \nabla\big(\mathcal{D}_h^X u - u\big) \cdot \nabla \varphi_h \, \mathrm{d}\boldsymbol{x} \,\middle|\, \varphi_h \in X_h \text{ mit } \|\varphi_h\|_{X_h} = 1 \right\} \\
&\le \sup\left\{ \|\mathcal{D}_h^X u - u\|_X \|\varphi_h\|_X \,\middle|\, \varphi_h \in X_h \text{ mit } \|\varphi_h\|_X = 1 \right\} = \|\mathcal{D}_h^X u - u\|_X.
\end{aligned}$$

Damit impliziert

$$\|\mathcal{D}_h^X u - u\|_X \to 0 \quad \text{für } h \to 0 \tag{3.17}$$

die Konsistenz des Galerkin-Verfahrens. Wir werden später (3.17) durch Konstruktion oder geeignete Voraussetzungen erfüllen. Aus Satz 3.7 erhalten wir

$$\|\mathcal{D}_h^X u - u\|_X = \inf\left\{ \|u - \varphi_h\|_X \,\middle|\, \varphi_h \in X_h \right\}.$$

Wir werden also X_h so konstruieren, dass

$$\inf\left\{ \|u - \varphi_h\|_X \,\middle|\, \varphi_h \in X_h \right\} \to 0 \quad \text{für } h \to 0$$

für $u \in X$ gilt. Der Nachweis der Stabilität des Galerkin-Verfahrens folgt direkt aus der Ungleichung (3.1)

$$\begin{aligned}
\|u_h\|_{X_h} = \frac{\|u_h\|_X^2}{\|u_h\|_X} &= \frac{\langle \mathcal{T}_h u_h, u_h \rangle_{(X_h)',X_h}}{\|u_h\|_X} = \left\langle \mathcal{T}_h u_h, \frac{u_h}{\|u_h\|_X} \right\rangle_{(X_h)',X_h} \\
&\le \sup\left\{ \langle \mathcal{T}_h u_h, \varphi_h \rangle_{(X_h)',X_h} \,\middle|\, \|\varphi_h\|_{X_h} = 1 \right\} = \|\mathcal{T}_h u_h\|_{Y_h}
\end{aligned}$$

Nun folgt die Konvergenz des Galerkin-Verfahrens aus Satz 2.10, sofern (3.17) erfüllt ist.

3.3 Finite Elemente

In diesem Abschnitt führen wir das *Finite-Elemente-Verfahren* ein. Als weiterführende Literatur sei zum Beispiel auf das Buch [Bra07] verwiesen.

3.3.1 Simplexe

Wir wollen nun geeignete endlich-dimensionale Teilräume X_h konstruieren und orientieren uns dabei an [Dzi10, Abschnitt 3.2.1]. Grundlage ist die Zerlegung des Gebietes G in Simplexe. Im Fall $n = 2$ führt das auf ein Dreiecksgitter, im Fall $n = 3$ auf ein Tetraedergitter.

Definition 3.8 ([Dzi10, Definition 3.11])**.** 1) *Für $s \in \{1, \dots, n\}$ seien die Vektoren $\mathrm{a}_0, \dots, \mathrm{a}_s \in \mathbb{R}^n$ derart, dass die Vektoren $\{\mathrm{a}_j - \mathrm{a}_0\}_{j=1}^s$ linear unabhängig sind. Dann heißt*

$$\mathcal{T} = \left\{ \boldsymbol{x} \in \mathbb{R}^n \,\middle|\, x = \sum_{j=0}^{s} \lambda_j \mathrm{a}_j,\ 0 \le \lambda_j \le 1 \text{ und } \sum_{j=0}^{s} \lambda_j = 1 \right\}$$

ein (nicht-degeneriertes) s-dimensionales Simplex im $\mathbb{R}^n$*. Die Punkte $\mathrm{a}_0, \dots, \mathrm{a}_s$ heißen* Ecken des Simplex*. Sind $\tilde{\mathrm{a}}_0, \dots, \tilde{\mathrm{a}}_r \in \{\mathrm{a}_0, \dots, \mathrm{a}_s\}$ mit $r \in \{0, \dots, s\}$, so nennen wir*

$$\tilde{\mathcal{T}} = \left\{ \boldsymbol{x} \in \mathbb{R}^n \,\middle|\, x = \sum_{j=0}^{r} \lambda_j \tilde{\mathrm{a}}_j,\ 0 \le \lambda_j \le 1 \text{ und } \sum_{j=0}^{r} \lambda_j = 1 \right\}$$

ein r-dimensionales Seitensimplex von $\mathcal{T}$*. Die eindimensionalen Seitensimplexe werden* Kanten*, die nulldimensionalen* Punkte *genannt.*

2) *Ein Simplex $\mathcal{T}_\circ$ zu $\mathrm{a}_0 = 0$ und $\mathrm{a}_j = \mathrm{e}_j$ für $j = 1, \dots, n$ mit den kanonischen Einheitsvektoren $\{\mathrm{e}_j\}_{j=1}^n \subset \mathbb{R}^n$ heißt n*-dimensionales Einheitssimplex*.*

3) *Die Zahl*

$$h(\mathcal{T}) = \max\left\{ |\mathrm{a}_i - \mathrm{a}_j|_2 \,\middle|\, i, j = 0, \dots, s \right\}$$

wird Durchmesser des s*-dimensionalen Simplex $\mathcal{T}$ genannt. Wir nennen*

$$\varrho(\mathcal{T}) = 2 \sup\left\{ r \,\middle|\, B(x_\circ, r) \subset \mathcal{T} \text{ für } x_\circ \in \mathcal{T} \right\}$$

den Inkugeldurchmesser des Simplex $\mathcal{T}$*. Wir definieren den Quotienten*

$$\sigma(\mathcal{T}) = \frac{h(\mathcal{T})}{\varrho(\mathcal{T})}$$

für ein Simplex $\mathcal{T}$.

4) *Als* Schwerpunkt eines Simplex $\mathcal{T}$ *bezeichnen wir den Punkt $\boldsymbol{x}_\mathcal{T} = \sum_{j=0}^s \mathrm{a}_j/(s+1) \in \mathcal{T}$.*

Im Fall $n = 2$ ist das zweidimensionale Simplex ein nicht-degeneriertes Dreieck mit den drei Ecken a_0, a_1, a_2 und den drei Kanten $\mathrm{a}_1 - \mathrm{a}_0$, $\mathrm{a}_2 - \mathrm{a}_0$, $\mathrm{a}_2 - \mathrm{a}_1$. Im Fall $n = 3$ ist das dreidimensionale Simplex ein Tetraeder mit den vier Ecken a_0, a_1, a_2, a_3, mit sechs Kanten und mit vier zweidimensionalen Seitensimplexen.

Lemma 3.9. *Seien $\mathcal{T}$ ein nicht-degeneriertes s-dimensionales Simplex in $\mathbb{R}^n$ mit den Ecken $\{\mathrm{a}_j\}_{j=0}^s$ und $\boldsymbol{x} = (x_1, \dots, x_n)^\top \in \mathcal{T}$. Dann existiert eine eindeutige Lösung $\boldsymbol{\lambda} = (\lambda_0, \dots, \lambda_s)^\top \in \mathbb{R}^{s+1}$ des linearen Gleichungssystems*

$$\sum_{j=0}^{s} \lambda_j \mathrm{a}_j = \boldsymbol{x}, \quad \sum_{j=0}^{s} \lambda_j = 1. \tag{3.18}$$

Beweis. Die Lösbarkeit folgt aus der Voraussetzung $\boldsymbol{x} \in \mathcal{T}$ und Definition 3.8-1). Wir haben daher nur die Eindeutigkeit zu zeigen. Das Gleichungssystem (3.18) lässt wie folgt in Matrixschreibweise ausdrücken:

$$\left(\begin{array}{c|c|c|c} \mathrm{a}_0 & \mathrm{a}_1 & \dots & \mathrm{a}_s \\ 1 & 1 & \dots & 1 \end{array}\right) \begin{pmatrix} \lambda_0 \\ \vdots \\ \lambda_s \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Für den Rang der $(n+1) \times (s+1)$ Koeffizientenmatrix gilt

$$\begin{aligned} \operatorname{Rang} \left(\begin{array}{c|c|c|c} \mathrm{a}_0 & \mathrm{a}_1 & \dots & \mathrm{a}_s \\ 1 & 1 & \dots & 1 \end{array}\right) &= \operatorname{Rang} \left(\begin{array}{c|c|c|c} \mathrm{a}_0 & \mathrm{a}_1 - \mathrm{a}_0 & \dots & \mathrm{a}_s - \mathrm{a}_0 \\ 1 & 0 & \dots & 0 \end{array}\right) \\ &= 1 + \operatorname{Rang} \left(\begin{array}{c|c|c} \mathrm{a}_1 - \mathrm{a}_0 & \dots & \mathrm{a}_s - \mathrm{a}_0 \end{array}\right) = s + 1, \end{aligned}$$

da $\{\mathrm{a}_j - \mathrm{a}_0\}_{j=1}^s$ linear unabhängig sind und sich der Rang der Koeffizientenmatrix nicht ändert, wenn wir von der j-ten Spalte, $j = 2, \dots, s$, die erste Spalte abziehen. Daraus folgt die Eindeutigkeit der Lösung $\boldsymbol{\lambda} \in \mathbb{R}^{s+1}$. □

Lemma 3.9 motiviert die folgende Definition.

Definition 3.10. *Als* baryzentrische Koordinaten $\boldsymbol{\lambda} = (\lambda_0, \dots, \lambda_s)^\top \in \mathbb{R}^{s+1}$ *eines Punktes* $x \in \mathcal{T}$ *des* s*-dimensionalen Simplex* $\mathcal{T}$ *bezeichnen wir die nach Lemma* 3.9 *eindeutige Lösung des linearen Gleichungssystems* (3.18).

Wir zitieren hier das folgende Lemma, in dem es um die Transformation des Einheitssimplex $\mathcal{T}_\circ$ auf ein gegebenes Simplex $\mathcal{T}$ geht. Für einen Beweis sei auf [Dzi10, Hilfssatz 3.13] verwiesen.

Lemma 3.11. *Jedes* s*-dimensionale Simplex* $\mathcal{T}$ *im* $\mathbb{R}^s$ *is affin-äquivalent zum* s*-dimensionalen Einheitssimplex* $\mathcal{T}_\circ$*. Es gibt genau eine affin-lineare Abbildung*

$$F : \mathcal{T}_\circ \to \mathcal{T}, \quad F(\tilde{\boldsymbol{x}}) = \mathrm{A}\tilde{\boldsymbol{x}} + \mathrm{b}$$

mit einer regulären $s \times s$ *Matrix* A *und einem Vektor* $\mathrm{b} \in \mathbb{R}^s$*, so dass* $F(\mathrm{e}_j) = \mathrm{a}_j$ *für* $j = 0, \dots, s$ *gilt. Außerdem sind folgende Abschätzungen erfüllt:*

$$|\mathrm{A}|_2 \le \frac{h(\mathcal{T})}{\varrho(\mathcal{T}_\circ)}, \quad |\mathrm{A}^{-1}|_2 \le \frac{h(\mathcal{T}_\circ)}{\varrho(\mathcal{T})}, \quad |\det \mathrm{A}| = \frac{|\mathcal{T}|}{|\mathcal{T}_\circ|}, \quad \hat{c}_1 \varrho(\mathcal{T})^s \le |\det \mathrm{A}| \le \hat{c}_2 h(\mathcal{T})^s,$$

wobei $|\mathcal{T}|$ *das Maß des Simplex* $\mathcal{T}$ *bezeichnet und die positiven Konstanten* $\hat{c}_1$*,* $\hat{c}_2$ *nur von* s *abhängen.*

Bemerkung 3.12. Die Matrix A ist durch den Basiswechsel von e_j auf $\mathrm{a}_j - \mathrm{a}_0$, $j = 1, \dots, s$, gegeben: Wir bekommen mit $\mathrm{e}_0 = 0 \in \mathbb{R}^s$

$$\mathrm{a}_0 = F(\mathrm{e}_0) = \mathrm{A}\mathrm{e}_0 + \mathrm{b} = \mathrm{b}$$

und daher

$$\mathrm{a}_j = F(\mathrm{e}_j) = \mathrm{A}\mathrm{e}_j + \mathrm{b} = \mathrm{A}_{\cdot j} + \mathrm{a}_0 \text{ für } j = 1, \dots, s \Rightarrow \mathrm{A} = \left(\mathrm{a}_1 - \mathrm{a}_0 \,\middle|\, \dots \,\middle|\, \mathrm{a}_s - \mathrm{a}_0\right),$$

wobei $\mathrm{A}_{\cdot j} \in \mathbb{R}^s$ die j-te Spalte von A bezeichnet. ◊

Wir führen in der folgenden Definition eine Triangulierung des beschränkten Gebietes G ein. Dabei betrachten wir für die einfachere Präsentation nur Gebiete G, die polygonal berandet sind.

Definition 3.13. *Sei* $G \subset \mathbb{R}^n$ *ein beschränktes Gebiet. Es gelten*

$$\overline{G} = \bigcup_{j=1}^{m} \mathcal{T}_j, \quad \partial G = \bigcup_{j=1}^{\tilde{m}} \tilde{\mathcal{T}}_j, \quad m, \tilde{m} \in \mathbb{N}, \tag{3.19}$$

mit n-dimensionalen Simplexen $\mathcal{T}_j$ ($j = 1, \dots, m$) und $(n-k)$-dimensionalen Simplexen $\tilde{\mathcal{T}}_j$ ($k \in \{1, \dots, n\}$), die Seitensimplexe der $\mathcal{T}_j$ sind. Wir nennen

$$\Delta(G) = \left\{ \mathcal{T}_j \,\middle|\, j = 1, \dots, m \right\}$$

eine Triangulierung von G. *Die Triangulierung heißt* zulässig, *wenn für je zwei Simplexe $\mathcal{T}_1, \mathcal{T}_2 \in \Delta(G)$ gilt, dass $\mathcal{T}_1 \cap \mathcal{T}_2 = \mathcal{S}$ mit $\mathcal{S} = \emptyset$ oder $\mathcal{S}$ ist ein gemeinsames $(n-k)$-dimensionales Seitensimplex von $\mathcal{T}_1$ und $\mathcal{T}_2$ ist ($k \in \{1, \dots, n\}$). Für eine zulässige Triangulierung f definieren wir*

$$h = \max\left\{ h(\mathcal{T}) \,\middle|\, \mathcal{T} \in \Delta(G) \right\}, \quad \varrho = \min\left\{ \varrho(\mathcal{T}) \,\middle|\, \mathcal{T} \in \Delta(G) \right\}, \quad \sigma = \max\left\{ \sigma(\mathcal{T}) \,\middle|\, \mathcal{T} \in \Delta(G) \right\}.$$

Die Zahl h nennen wir globale Gitterweite *oder* Feinheit *von* $\Delta(G)$.

Das folgende Resultat ermöglicht die Zusammensetzung von stückweise polynomialen Funktionen auf einer gegebenen Triangulierung, um auf diese Weise endlich-dimensionale Teilräume $X_h \subset X$ zu erhalten.

Satz 3.14. *Seien G zulässig trianguliert und $v \in C(\overline{G})$. Gilt $v|_{\mathcal{T}} \in C^1(\mathcal{T})$ für jedes Simplex $\mathcal{T} \in \Delta(G)$, so folgt $v \in H^1(G)$.*

.

Beweis. Wir müssen zeigen, dass v eine schwache Ableitung besitzt. Dazu verwenden wir die Eigenschaft (3.19) und die *partielle Integration* [DR11, Folgerung 13.70-(iii)]. Für $i \in \{1, \dots, n\}$ und $\boldsymbol{x} = (x_1, \dots, x_n) \in \overline{G}$ erhalten wir

$$\int_G v\varphi_{x_i} \,\mathrm{d}\boldsymbol{x} = \sum_{j=1}^{m} \int_{\mathcal{T}_j} v\varphi_{x_i} \,\mathrm{d}\boldsymbol{x} = \sum_{j=1}^{m} \left(\int_{\partial G} v\varphi\nu_i \,\mathrm{d}A - \int_{\mathcal{T}_j} v_{x_i}\varphi \,\mathrm{d}\boldsymbol{x} \right) = -\int_G v_{x_i}\varphi \,\mathrm{d}\boldsymbol{x},$$

für $\varphi \in C_0^\infty(G)$, wobei wir $v \in C(\overline{G})$ auf den Simplexrändern und $v|_{\mathcal{T}_j} \in C^1(\mathcal{T}_j)$, $j = 1, \dots, m$, bei der Anwendung der partiellen Integration verwendet haben. Die Randintegrale verschwinden wegen $\varphi = 0$ auf ∂G und wegen der Tatsache, dass sich die Randintegrale benachbarter Simplexe aufgrund der gegensätzlichen Orientierung der Orientierung der Normale weghheben. Damit setzt sich die schwache Ableitung stückweise aus den partiellen Ableitungen von v zusammen. □

3.3.2 Lagrange-Elemente

Die in Definition 3.10 eingeführten baryzentichen Koordinaten werden nun dazu verwendet, Polynome auf Simplexen einzuführen. Der Raum aller Polynome vom Grad kleiner oder gleich k (mit $k \in \mathbb{N}_0$) wird mit

$$\mathbb{P}_k = \left\{ p : \mathbb{R}^n \to \mathbb{R} \,\middle|\, p(\boldsymbol{x}) = \sum_{|\alpha|=0}^{k} c_\alpha \boldsymbol{x}^\alpha \text{ mit } c_\alpha \in \mathbb{R} \right\}$$

bezeichnet. Hierbei ist $\alpha = (\alpha_1, \dots, \alpha_n) \in (\mathbb{N}_0)^n$ ein Multiindex und

$$|\alpha| = \sum_{i=1}^{n} \alpha_i, \quad x^\alpha = \prod_{i=1}^{n} x_i^{\alpha_i} \text{ für } \boldsymbol{x} = (x_1, \dots, x_n) \in \mathbb{R}^n.$$

Wir setzen

$$\mathbb{P}_k(\mathcal{M}) = \left\{ p|_{\mathcal{M}} \,\middle|\, p \in \mathbb{P}_k \right\} \quad \text{für eine Menge } \mathcal{M} \subset \mathbb{R}^n.$$

Sei $\mathcal{T}$ ein n-dimensionales Simplex und $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathcal{T}$. Aus (3.18) schließen wir, dass die i-te Komponente x_i, $i \in \{1, \dots, n\}$, von $\boldsymbol{x}$ durch

$$x_i = x_i(\lambda) = \sum_{j=0}^{n} a_{ji}\lambda_j \quad \text{mit } \lambda = (\lambda_0, \dots, \lambda_n) \in \mathbb{R}^{n+1}$$

gegeben ist, wobei a_{ji} die i-te Komponente des Vektors $a_j \in \mathbb{R}^n$ ist. Daher erhalten wir für ein Polynom $p \in \mathbb{P}_k(\mathcal{T})$

$$p(\boldsymbol{x}) = \sum_{|\alpha|=0}^{k} c_\alpha \boldsymbol{x}^\alpha = \sum_{|\alpha|=0}^{k} c_\alpha \prod_{i=1}^{n} x_i^{\alpha_i} = \sum_{|\alpha|=0}^{k} c_\alpha \prod_{i=1}^{n} \Big(\sum_{j=0}^{n} a_{ji}\lambda_j \Big)^{\alpha_i}$$
$$= c_{(0,\dots,0)} + \sum_{|\alpha|=1}^{k} c_\alpha \prod_{i=1}^{n} \Big(\sum_{j=0}^{n} a_{ji}\lambda_j \Big)^{\alpha_i} = \bar{p}(\boldsymbol{\lambda}).$$

Wegen der zweiten Gleichung in (3.18) gilt auch

$$c_{(0,\dots,0)} = c_{(0,\dots,0)} \sum_{j=0}^{n} \lambda_j.$$

Damit lässt sich $\bar{p}$ als ein Polynom vom Grad k in $\boldsymbol{\lambda}$ ohne konstanten Term schreiben. Nach Umnummerierung erhalten wir die Darstellung

$$\bar{p}(\boldsymbol{\lambda}) = \sum_{|\beta|=1}^{n} \tilde{c}_\beta \boldsymbol{\lambda}^\beta \quad \text{für } \boldsymbol{\lambda} = (\lambda_0, \dots, \lambda_n)^\top \in \mathbb{R}^{n+1}$$

mit einem Multiindex $\beta = (\beta_0, \dots, \beta_n) \in (\mathbb{N}_0)^{n+1}$. Wir zeigen das für den Spezialfall $n = 2$ und $k = 1$:

$$\begin{aligned} p(\boldsymbol{\lambda}) &= \sum_{|\alpha|=0}^{1} c_\alpha \prod_{i=1}^{2} \Big(\sum_{j=0}^{2} a_{ji}\lambda_j \Big)^{\alpha_i} = c_{(0,0)} + c_{(1,0)} \sum_{j=0}^{2} a_{j1}\lambda_j + c_{(0,1)} \sum_{j=0}^{2} a_{j2}\lambda_j \\ &= c_{(0,0)}\big(\lambda_0 + \lambda_1 + \lambda_2\big) + \big(c_{(1,0)}a_{01} + c_{(0,1)}a_{02}\big)\lambda_0 + \big(c_{(1,0)}a_{11} + c_{(0,1)}a_{12}\big)\lambda_1 \\ &\quad + \big(c_{(1,0)}a_{21} + c_{(0,1)}a_{22}\big)\lambda_2 \\ &= \big(c_{(0,0)} + c_{(1,0)}a_{01} + c_{(0,1)}a_{02}\big)\lambda_0 + \big(c_{(0,0)} + c_{(1,0)}a_{11} + c_{(0,1)}a_{12}\big)\lambda_1 \\ &\quad + \big(c_{(0,0)} + c_{(1,0)}a_{21} + c_{(0,1)}a_{22}\big)\lambda_2. \end{aligned}$$

Nun setzen wir

$$\begin{aligned} \tilde{c}_{(1,0,0)} &= c_{(0,0)} + c_{(1,0)}a_{01} + c_{(0,1)}a_{02}, \\ \tilde{c}_{(0,1,0)} &= c_{(0,0)} + c_{(1,0)}a_{11} + c_{(0,1)}a_{12}, \\ \tilde{c}_{(0,0,1)} &= c_{(0,0)} + c_{(1,0)}a_{21} + c_{(0,1)}a_{22} \end{aligned}$$

und erhalten

$$p(\boldsymbol{\lambda}) = \sum_{|\beta|=1} \tilde{c}_\beta \boldsymbol{\lambda}^\beta \quad \text{für } \boldsymbol{\lambda} = (\lambda_0, \lambda_1, \lambda_2)^\top \in \mathbb{R}^3.$$

Satz 3.15 (Lineares Element)**.** 1) *Sei $\mathcal{T}$ ein n-dimensionales Simplex. Dann ist durch die Vorgabe von Stützwerten $\{p_j\}_{j=0}^n \subset \mathbb{R}$ für die Stützstellen $\{\mathrm{a}_j\}_{j=0}^n \subset \mathcal{T}$ ein Polynom $p \in \mathbb{P}_1(\mathcal{T})$ eindeutig bestimmt. Für jedes Polynom $p \in \mathbb{P}_1(\mathcal{T})$ gilt an $\boldsymbol{x} \in \mathcal{T}$ die Darstellung*

$$p(\boldsymbol{x}) = p(\boldsymbol{x}(\boldsymbol{\lambda})) = \bar{p}(\boldsymbol{\lambda}) = \sum_{j=0}^{n} p(\mathrm{a}_j)\lambda_j \quad \textit{für } \boldsymbol{\lambda} = (\lambda_0, \dots, \lambda_n)^\top \in \mathbb{R}^{n+1} \tag{3.20}$$

und wir haben $\dim \mathbb{P}_1(\mathcal{T}) = n + 1$.

2) *Ist das beschränkte Gebiet $G \subset \mathbb{R}^n$ zulässig trianguliert und sind $\{\bar{\mathrm{a}}_j\}_{j=1}^{\bar{m}} \subset \overline{G}$ die Ecken der Triangulierung $\Delta(G)$, so ist durch Vorgabe von $\{u_h(\bar{\mathrm{a}}_j)\}_{j=1}^{\bar{m}}$ eindeutig eine Funktion*

$$u_h \in X_h = \big\{ \varphi_h \in C(\overline{G}) \,\big|\, \varphi_h|_{\mathcal{T}} \in \mathbb{P}_1(\mathcal{T}) \textit{ für } \mathcal{T} \in \Delta(G) \big\} \subset H^1(G)$$

bestimmt.

3) *Eine Basis von X_h ist durch die Funktionen*

$$\phi_i \in X_h, \quad \phi_i(\bar{a}_j) = \delta_{ij} \text{ für } i,j = 1,\ldots,\bar{m}$$

gegeben. Diese Basis nennen wir Knotenbasis.

Beweis.

1) Gegeben seien die Stützwerte $\{p_j\}_{j=0}^n$ und Stützstellen $\{a_j\}_{j=0}^n$. Zur Bestimmung der $n+1$ unbekannten Koeffizienten $c_0,\ldots,c_n$ des linearen Polynoms

$$p \in \mathbb{P}_1(\mathcal{T}), \quad p(\boldsymbol{x}) = c_0 + \sum_{i=1}^n c_i x_i \text{ für } \boldsymbol{x} = (x_1,\ldots,x_n) \in \mathcal{T}$$

sind dann die $n+1$ Gleichungen

$$p(a_j) = c_0 + \sum_{i=1}^n c_i a_{ji} = p_j \quad \text{für } j = 0,\ldots,n \tag{3.21}$$

zu lösen, wobei wieder a_{ji} die i-te Komponente des Vektor a_j bezeichnet. Wir geben eine Lösung an, die dann auch eindeutig ist. Sei $\{e_j\}_{j=0}^n \subset \mathbb{R}^{n+1}$ die kanonische Basis des $\mathbb{R}^{n+1}$. Mit der Notation

$$p(\boldsymbol{x}(\boldsymbol{\lambda})) = \bar{p}(\boldsymbol{\lambda}) = \sum_{|\beta|=1} \tilde{c}_\beta \boldsymbol{\lambda}^\beta = \sum_{j=0}^n \tilde{c}_{e_j} \boldsymbol{\lambda}^{e_j} = \sum_{j=0}^n \tilde{c}_{e_j} \lambda_j \quad \text{für } \boldsymbol{\lambda} = (\lambda_0,\ldots,\lambda_n)^\top \in \mathbb{R}^{n+1}$$

und wegen

$$\boldsymbol{x}(e_i) = \sum_{j=0}^n a_j \delta_{ij} = a_i \quad \text{für } i = 0,\ldots,n$$

bekommen wir

$$p(a_i) = p(\boldsymbol{x}(e_i)) = \bar{p}(e_i) = \sum_{j=0}^n \tilde{c}_{e_j} (e_i)^{e_j} = \sum_{j=0}^n \tilde{c}_{e_j} \delta_{ij} = \tilde{c}_{e_i} \quad \text{für } i = 0,\ldots,n.$$

Damit löst

$$p(\boldsymbol{x}(\boldsymbol{\lambda})) = \bar{p}(\boldsymbol{\lambda}) = \sum_{j=0}^n p(a_j)\lambda_j$$

das Gleichungssystem (3.21). Offenbar gilt dann auch $\mathbb{P}_1(\mathcal{T}) = n+1$.

2) Wir zeigen, dass u_h stetig auf $\overline{G}$ ist. Nach Satz 3.14 ist dann u_h auch ein Element in $H^1(G)$. Sind $\mathcal{T}_1$ und $\mathcal{T}_2$ zwei Simplexe der Triangulierung $\Delta(G)$ und ist $\mathcal{T}_1 \cap \mathcal{T}_2 = \mathcal{S}$ mit einem gemeinsamen $(n-k)$-dimensionalen Seitensimplex $\mathcal{S}$, so ist $u_h|_{\mathcal{S}} \in \mathbb{P}_1(\mathcal{S})$ nach Teil 1) des Beweises bereits durch die Werte an den Ecken von $\mathcal{S}$ eindeutig bestimmt.

3) Offenbar ist die Dimension von X_h gleich $\bar{m}$. Ferner liegen die Funktionen $\phi_1,\ldots,\phi_{\bar{m}}$ in X_h aufgrund von Satz 3.14. Um die lineare Unabhängigkeit zu zeigen, betrachten wir die Gleichung

$$\sum_{i=1}^{\bar{m}} \gamma_i \phi_i = 0 \quad \text{in } \overline{G} \tag{3.22}$$

mit reellen Koeffizienten $\gamma_1,\ldots,\gamma_{\bar{m}}$. Wir müssen beweisen, dass alle γ_i's gleich null sein müssen. Aus (3.22), $\{\bar{a}_j\}_{j=1}^{\bar{m}} \subset \overline{G}$ und $\phi_i(\bar{a}_j) = \delta_{ij}$ für $i,j = 1,\ldots,\bar{m}$ folgt bereits

$$0 = \sum_{i=1}^{\bar{m}} \gamma_i \phi_i(a_j) = \sum_{i=1}^{\bar{m}} \gamma_i \delta_{ij} = \gamma_j \quad \text{für } j = 1,\ldots,\bar{m},$$

was zu zeigen war. □

Bemerkung 3.16. Wir weisen an dieser Stelle darauf hin, dass der in Satz 3.15 definierte Finite-Elemente-Raum X_h kein Teilraum von $H^2(G)$ ist. Es ist um einiges schwerer, Finite-Elemente-Räume von $H^2(G)$ zu kontruieren. ◊

Analog wie in Satz 3.15 kann man auch eine Finite-Elemente-Raum mit quadratischen Ansatzfunktionen einführen. Einen Beweis des folgenden Satzes findet man in [Dzi10, Element 3.17].

Satz 3.17 (Quadratisches Element)**.** 1) *Sei $\mathcal{T}$ ein n-dimensionales Simplex mit den Kantenmittelpunkten $\mathrm{a}_{ij} = (\mathrm{a}_i + \mathrm{a}_j)/2$ für $i, j = 1, \ldots, n$ und $i < j$. Dann ist durch die Vorgabe von Stützwerten $p_j \in \mathbb{R}$ $(j = 0, \ldots, n)$ und $p_{ij} \in \mathbb{R}$ $(i, j = 1, \ldots, n$ und $i < j)$ für die Stützstellen $\mathrm{a}_j \in \mathcal{T}$ $(j = 0, \ldots, n)$ beziehungsweise $\mathrm{a}_{ij} \in \mathcal{T}$ $(i, j = 1, \ldots, n$ und $i < j)$ ein Polynom $p \in \mathbb{P}_2(\mathcal{T})$ eindeutig bestimmt. Für jedes Polynom $p \in \mathbb{P}_2(\mathcal{T})$ gilt an $\boldsymbol{x} \in \mathcal{T}$ die Darstellung*

$$p(\boldsymbol{x}) = p(\boldsymbol{x}(\boldsymbol{\lambda})) = \bar{p}(\boldsymbol{\lambda}) = \sum_{j=0}^{n} p_j \lambda_j (2\lambda_j - 1) + 4 \sum_{j=0}^{n} \sum_{i=0}^{j-1} p_{ij} \lambda_i \lambda_j \tag{3.23}$$

für $\boldsymbol{\lambda} = (\lambda_0, \ldots, \lambda_n)^\top \in \mathbb{R}^{n+1}$ und wir haben $\dim \mathbb{P}_2(\mathcal{T}) = (n+1)(n+2)/2$.

2) *Ist das beschränkte Gebiet $G \subset \mathbb{R}^n$ zulässig trianguliert und sind $\{\bar{\mathrm{a}}_j\}_{j=1}^{\bar{m}} \subset \overline{G}$ die Ecken und Kantenmittelpunkte der Triangulierung $\Delta(G)$, so ist durch Vorgabe von $\{u_h(\bar{\mathrm{a}}_j)\}_{j=1}^{\bar{m}}$ eindeutig eine Funktion*

$$u_h \in X_h = \left\{ \varphi_h \in C(\overline{G}) \,\middle|\, \varphi_h|_{\mathcal{T}} \in \mathbb{P}_2(\mathcal{T}) \text{ für } \mathcal{T} \in \Delta(G) \right\} \subset H^1(G)$$

bestimmt.

3) *Eine Basis von X_h ist durch die Funktionen*

$$\phi_i \in X_h, \quad \phi_i(\bar{\mathrm{a}}_j) = \delta_{ij} \text{ für } i, j = 1, \ldots, \bar{m}$$

gegeben. Diese Basis nennen wir Knotenbasis.

Beispiel 3.18. Wir betrachten den Fall $n = 1$ und das Einheitssimplex $\mathcal{T}_\circ = [0, 1]$ mit den Ecken $a_0 = 0$ und $a_1 = 1$.

1) Wir betrachten lineare Finite Elemente. Die baryzentrischen Koordinaten sind

$$x = \sum_{j=0}^{1} \lambda_j a_j \;\Rightarrow\; \lambda_1 = x, \quad \sum_{j=0}^{1} \lambda_j = 1 \;\Rightarrow\; \lambda_0 = 1 - x.$$

Für die Basisfunktionen erhalten wir daher mit (3.20)

$$\bar{\phi}_0(\lambda) = \sum_{j=0}^{n} \phi_0(a_j)\lambda_j = \lambda_0 = 1 - x = \phi_0(x), \quad \bar{\phi}_1(\lambda) = \sum_{j=0}^{n} \phi_1(a_j)\lambda_j = \lambda_1 = x = \phi_1(x).$$

2) Für quadratische Finite Elemente müssen wir zusätzlich zu den Ecken den Mittelpunkt $a_{01} = 1/2$ berücksichtigen. Mit (3.23) bekommen wir

$$\bar{\phi}_0(\lambda) = \sum_{j=0}^{1} \phi_0(a_j)\lambda_j(2\lambda_j - 1) + 4\phi_0(a_{01})\lambda_0\lambda_1 = \lambda_0(2\lambda_0 - 1) = 2x^2 - 3x + 1 = \phi_0(x),$$

$$\bar{\phi}_1(\lambda) = \sum_{j=0}^{1} \phi_1(a_j)\lambda_j(2\lambda_j - 1) + 4\phi_1(a_{01})\lambda_0\lambda_1 = \lambda_1(2\lambda_1 - 1) = 2x^2 - x = \phi_1(x),$$

$$\bar{\phi}_{01}(\lambda) = \sum_{j=0}^{1} \phi_{01}(a_j)\lambda_j(2\lambda_j - 1) + 4\phi_{01}(a_{01})\lambda_0\lambda_1 = 4\lambda_0\lambda_1 = 4(x - x^2) = \phi_{01}(x)$$

als Knotenbasis. ◊

3.4 Interpolation

Seien $G \subset \mathbb{R}^n$ ein beschränktes Gebiet und $\Delta(G)$ eine zulässige Triangulierung von G. In Satz 3.7 haben wir die Gleichung

$$\|u - u_h\|_V = \inf\left\{\|u - \varphi_h\|_V \,\middle|\, \varphi_h \in V_h\right\}.$$

bewiesen, wobei $u \in V$ die Lösung von (3.2) und $u_h \in V_h$ die von (3.9) bezeichnet. Unser Ziel in diesem Abschnitt ist es nun eine Abschätzung der Form

$$\|u - u_h\|_V \le c h^r, \quad h = \max\left\{h(\mathcal{T}) \,\middle|\, \mathcal{T} \in \Delta(G)\right\},$$

mit einer möglichst nur von den Daten abhäbgenden Konstante $c > 0$ und einem möglichst großen Exponenten $r > 0$ herzuleiten.

Wir erinnern an die folgende Schreibweise für höhere partielle Ableitungen: Für einen Multiindex $\alpha = (\alpha_1, \ldots, \alpha_n) \in (\mathbb{N}_0)^n$ setzen wir

$$\partial^\alpha = \partial_1^{\alpha_1} \ldots \partial_n^{\alpha_n} = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}};$$

vergleiche [DR11, Seite 154]. Wir benötigen die zweite *Poincaré-Ungleichung*, die in [DR12, Satz 16.24] bewiesen wird.

Satz 3.19. *Seien $G \subset \mathbb{R}^n$ ein beschränktes Gebiet, welches* (3.19) *erfüllt. Dann existiert eine Poincaré-Konstante $\tilde{c}_\mathsf{p} > 0$ mit*

$$\|u\|_H \le \tilde{c}_\mathsf{p}\left(\|u\|_V + \left|\int_G u \,\mathrm{d}\boldsymbol{x}\right|\right)$$

Bemerkung 3.20. Die Eigenschaft (3.19) garantiert, dass G eine sogenante Segmenteigenschaft [DR12, Definition 16.19-(ii)] erfüllt, die für die Gültigkeit der zweiten Poincaréschen Ungleichung notwendig ist. □

Das folgende Resultat ist ein Spezialfall von [Dzi10, Satz 3.24].

Folgerung 3.21. *Seien $G \subset \mathbb{R}^n$ ein beschränktes Gebiet, $\varphi \in H^\ell(G)$ mit $\ell \in \mathbb{N}$ und*

$$\int_G \partial^\alpha \varphi \,\mathrm{d}\boldsymbol{x} = 0 \quad \text{für } \alpha \in (\mathbb{N}_0)^n \text{ und } |\alpha| = 0, \ldots, \ell - 1. \tag{3.24}$$

Dann folgt die Abschätzung

$$\|\varphi\|_{H^\ell(G)} \le c\,|\varphi|_{H^\ell(G)} \quad \text{mit der Halbnorm } |\varphi|_{H^\ell(G)} = \left(\sum_{|\alpha|=\ell} \|\partial^\alpha \varphi\|_H^2\right)^{1/2}$$

mit einer nur von G und ℓ abhängigen Konstante $c > 0$.

Beweis. Da für $|\alpha| = 0, \ldots, \ell - 1$ die partielle Ableitung $\partial^\alpha \varphi$ in $H^{\ell-|\alpha|}(G)$ liegt, folgt die Behauptung aus mehrfacher Anwendung von Satz 3.19. □

Lemma 3.22. *Zu $u \in H^{k+1}(G)$ gibt es genau ein Polynom $p \in \mathbb{P}_k(G)$ mit*

$$\int_G \partial^\alpha (u - p) \,\mathrm{d}\boldsymbol{x} = 0 \quad \text{für } \alpha \in (\mathbb{N}_0)^n \text{ und } |\alpha| = 0, \ldots, k. \tag{3.25}$$

Beweis. Ein $p \in \mathbb{P}_k(G)$ besitzt die Darstellung

$$p(\boldsymbol{x}) = \sum_{|\beta|=0}^{k} c_\beta \boldsymbol{x}^\beta \quad \text{für } \boldsymbol{x} \in \overline{G}.$$

mit einem Koeffizientenvektor $c_\beta = (c_\beta)$. Damit ist (3.25) äquivalent mit dem linearen Gleichungssystem

$$\sum_{|\beta|=0}^{k} \left(\int_G \partial^\alpha x^\beta \, \mathrm{d}x \right) c_\beta = \int_G \partial^\alpha u \, \mathrm{d}x \quad \text{für } \alpha \in (\mathbb{N}_0)^n \text{ und } |\alpha| = 0, \ldots, k,$$

was auch in der Form

$$\sum_{|\beta|=0}^{k} A_{\alpha\beta} c_\beta = b_\alpha \quad \text{für } \alpha \in (\mathbb{N}_0)^n \text{ und } |\alpha| = 0, \ldots, k \tag{3.26}$$

geschrieben werden kann, wobei wir

$$A = \big((A_{\alpha\beta})\big) \text{ mit } A_{\alpha\beta} = \int_G \partial^\alpha x^\beta \, \mathrm{d}x, \quad b = \big(b_\alpha\big) \text{ mit } b_\alpha = \int_G \partial^\alpha u \, \mathrm{d}x$$

gesetzt haben. Das lineare Gleichungssystem enthält genau so viele Gleichungen wie unbekannte Koeffizienten. Es reicht also, die Eindeutigkeit einer Lösung zu zeigen. Dazu betrachten wir

$$\sum_{|\beta|=0}^{k} A_{\alpha\beta} c_\beta = 0 \quad \text{für } \alpha \in (\mathbb{N}_0)^n \text{ und } |\alpha| = 0, \ldots, k,$$

was äquivalent mit

$$\int_G \partial^\alpha p \, \mathrm{d}x = 0 \quad \text{für } \alpha \in (\mathbb{N}_0)^n \text{ und } |\alpha| = 0, \ldots, k,$$

ist. Da $p \in \mathbb{P}_k(G)$ ist, muss p das Nullpolynom sein, was zu zeigen war. □

Seien $G \subset \mathbb{R}^n$ ein beschränktes Gebiet und $k \in \mathbb{N}_0$. Dann ist $\mathbb{P}_k(G)$ ein Untervektorraum des Vektorraumes $H^{k+1}(G)$. Der Quotientenvektorraum $H^{k+1}(G)/\mathbb{P}_k(G)$ ist die Menge

$$H^{k+1}(G)/\mathbb{P}_k(G) = \big\{ [\varphi] \,\big|\, \varphi \in H^{k+1}(G) \big\}$$

aller Äquivalenzklassen

$$[\varphi] = \varphi + \mathbb{P}_k(G) = \big\{ \varphi + p \,\big|\, p \in \mathbb{P}_k(G) \big\}.$$

Die Vektorraumoperationen auf $H^{k+1}(G)/\mathbb{P}_k(G)$ sind wie folgt definiert:

$$[\varphi_1] + [\varphi_2] = [\varphi_1 + \varphi_2], \quad \lambda \cdot [\varphi_1] = [\lambda \varphi_1]$$

für $[\varphi_1], [\varphi_2] \in H^{k+1}(G)/\mathbb{P}_k(G)$ und $\lambda \in \mathbb{R}$. Im Folgenden unterscheiden wir in der Notation nicht zwischen φ und $[\varphi]$ für $\varphi \in H^{k+1}(G)$. Die folgende Aussage ist nun ein Spezialfall von [Dzi10, Satz 3.26].

Lemma 3.23. *Seien $G \subset \mathbb{R}^n$ ein beschränktes Gebiet und $k \in \mathbb{N}_0$. Dann existiert eine Konstante $c > 0$, die von G und k abhängt, so dass für alle $u \in H^{k+1}(G)/\mathbb{P}_k(G)$ die Abschätzung*

$$\|u\|_{H^{k+1}(G)/\mathbb{P}_k(G)} \leq c \, |u|_{H^{k+1}(G)}$$

gilt, wobei

$$\|u\|_{H^{k+1}(G)/\mathbb{P}_k(G)} = \inf \big\{ \|u + q\|_{H^{k+1}(G)} \,\big|\, q \in \mathbb{P}_k(G) \big\} \tag{3.27}$$

Beweis. Wir wählen zu $u \in H^{k+1}(G)$ das Polynom $q = -p$, wobei $p \in \mathbb{P}_k(G)$ das nach Lemma 3.21 eindeutig bestimmte Polynom ist, und erhalten

$$\|u\|_{H^{k+1}(G)/\mathbb{P}_k(G)} = \inf \big\{ \|u + q\|_{H^{k+1}(G)} \,\big|\, q \in \mathbb{P}_k(G) \big\} \leq \|u - p\|_{H^{k+1}(G)}.$$

Da $u - p \in H^{k+1}(G)$ aufgrund der Wahl von p die Eigenschaft (3.25) erfüllt, können wir Folgerung 3.21 für $\varphi = u - p$ und $\ell = k + 1$ anwenden. Wir bekommen

$$\|u\|_{H^{k+1}(G)/\mathbb{P}_k(G)} \le c\,|u-p|_{H^{k+1}(G)} = c\left(\sum_{|\alpha|=k+1} \|\partial^\alpha (u-p)\|_H^2\right)^{1/2} = c\,|u|_{H^{k+1}(G)},$$

da $\partial^\alpha p = 0$ gilt wegen $|\alpha| = k + 1$ und $p \in \mathbb{P}_k(G)$. □

Das folgende Resultat – vergleiche [Dzi10, Folgerung 3.28] – ergibt sich aus Lemma 3.23. Hier wird der Interpolationsfehler statt durch die H^{k+1}-Norm nur durch eine entsprechende Halbnorm abgeschätzt. In dieser Vorlesung werden wir allerdings die Folgerung später nicht verwenden.

Folgerung 3.24. *Seien $G \subset \mathbb{R}^n$ ein beschränktes Gebiet, $k, \ell \in \mathbb{N}_0$ mit $k + 1 \ge \ell$ und $\mathcal{E} : H^{k+1}(G) \to H^\ell(G)$ ein linearer, stetiger Einbettungsoperator mit $\mathcal{E}p = p$ für alle $p \in \mathbb{P}_k(G)$. Ferner bezeichne $\mathcal{I} : H^{k+1}(G) \to H^\ell(G)$ einen linearen und stetigen Interpolationsoperator, der auf $\mathbb{P}_k(G)$ invariant ist. Dann gibt es eine Konstante $c > 0$, die nur von ℓ, m und den Operatornormen $\|\mathcal{E}\|$, $\|\mathcal{I}\|$ abhängt, so dass*

$$\|\mathcal{E}u - \mathcal{I}u\|_{H^{k+1}(G)} \le c\,|u|_{H^\ell(G)} \quad \text{für alle } u \in H^{k+1}(G) \tag{3.28}$$

gilt.

Beweis. Sei $u \in H^{k+1}(G)$. Da $\mathcal{E} : H^{k+1}(G) \to H^\ell(G)$ und $\mathcal{I} : H^{k+1}(G) \to H^\ell(G)$ auf $\mathbb{P}_k(G)$ invariant sind, erhalten wir für ein Polynom $p \in \mathbb{P}_k(G)$

$$\begin{aligned}\|\mathcal{E}u - \mathcal{I}u\|_{H^\ell(G)} &= \|\mathcal{E}u + p - p - \mathcal{I}u\|_{H^\ell(G)} = \|\mathcal{E}(u+p) - \mathcal{I}(u+p)\|_{H^\ell(G)} \\ &\le \big(\|\mathcal{E}\| + \|\mathcal{I}\|\big)\|u+p\|_{H^\ell(G)}.\end{aligned}$$

Da p beliebig gewählt ist, folgt mit der Konstanten $c = \|\mathcal{E}\| + \|\mathcal{I}\|$

$$\|\mathcal{E}u - \mathcal{I}u\|_{H^\ell(G)} \le c \inf\left\{\|u+p\|_{H^{k+1}(G)} \,\middle|\, p \in \mathbb{P}_k(G)\right\} = c\,\|u\|_{H^{k+1}(G)/\mathbb{P}_k(G)}.$$

Nun folgt die Abschätzung (3.28) mit Lemma 3.23. □

Satz 3.25. *Seien G_1, G_2 zwei beschränkte Gebiete in $\mathbb{R}^n$, die affin-äquivalent sind, das heißt es gibt eine invertierbare Abbildung*

$$F : G_1 \to F(G_1) = G_2, \quad \boldsymbol{x} = F(\tilde{\boldsymbol{x}}) = \mathrm{A}\tilde{\boldsymbol{x}} + \mathrm{b} \text{ für } \tilde{\boldsymbol{x}} \in G_1, \boldsymbol{x} \in G_2.$$

Für $\ell \in \mathbb{N}_0$ seien $u \in H^\ell(G_2)$ und $\tilde{u} = u \circ F \in H^\ell(G_1)$. Dann gelten

$$|\tilde{u}|_{H^\ell(G_1)} \le c_1\,|\mathrm{A}|_2^\ell\,|\det \mathrm{A}|^{-1/2}|u|_{H^\ell(G_2)}, \quad |u|_{H^\ell(G_2)} \le c_2\,|\mathrm{A}^{-1}|_2^\ell\,|\det \mathrm{A}|^{1/2}|\tilde{v}|_{H^\ell(G_1)}$$

mit positiven Konstanten c_1, c_2, die nur von ℓ und n abhängen.

Beweis. Die Aussage wird in [Dzi10, Satz 3.29] unter Verwendung der Transformationsformel bewiesen. Offenbar erhalten wir die zweite Abschätzung aus der ersten, wenn wir A durch A^{-1} ersetzen. □

Wir geben noch eine direkte Folgerung aus Satz 3.25 und Lemma 3.11 an, die wir in der Vorlesung aber nicht weiter verwenden werden. An dieser Stelle sei auch auf [Dzi10, Folgerung 3.30] verwiesen.

Folgerung 3.26. *Im Kontext von Satz 3.25 seien $G_1 = \mathcal{T}_\circ$ das n-dimensionale Einheitssimplex, $G_2 = \mathcal{T}$ ein n-dimensionales Simplex und $F : \mathcal{T}_\circ \to \mathcal{T}$ eine invertierbare Abbildung mit $\mathcal{T} = F(\mathcal{T}_\circ)$. Für $\ell \in \mathbb{N}_0$ seien $u \in H^\ell(\mathcal{T})$ und $\tilde{u} = u \circ F \in H^\ell(\mathcal{T}_\circ)$ gelten dann die Abschätzungen*

$$|\tilde{u}|_{H^\ell(\mathcal{T}_\circ)} \le \tilde{c}_1\,\frac{h(\mathcal{T})^\ell \varrho(\mathcal{T})^{-n/2}}{\varrho(\mathcal{T}_\circ)^\ell}\,|u|_{H^\ell(\mathcal{T})}, \quad |u|_{H^\ell(\mathcal{T})} \le \tilde{c}_2\,\frac{h(\mathcal{T}_\circ)^\ell h(\mathcal{T})^{n/2}}{\varrho(\mathcal{T})^\ell}\,|\tilde{u}|_{H^\ell(\mathcal{T}_\circ)}$$

mit positiven Konstanten $\tilde{c}_1, \tilde{c}_2$, die nur von ℓ und n abhängen.

Beweis. Die Aussage ergibt sich aus Lemma 3.11 und Satz 3.25: Wegen $|\det \mathsf{A}|^{-1} \leq \varrho(\mathcal{T})^{-n}/\hat{c}_1$

$$|\tilde{u}|_{H^\ell(\mathcal{T}_\circ)} \leq c_1 \, |\mathsf{A}|_2^\ell |\det \mathsf{A}|^{-1/2} |u|_{H^\ell(\mathcal{T})} \leq \frac{c_1}{\hat{c}_1^{1/2}} \left(\frac{h(\mathcal{T})}{\varrho(\mathcal{T}_\circ)} \right)^\ell \varrho(\mathcal{T})^{-n/2} |u|_{H^\ell(\mathcal{T})},$$

was die erste Abschätzung ergibt mit $\tilde{c}_1 = c_1/\hat{c}_1^{1/2}$. Die zweite Abschätzung folgt wegen $|\det \mathsf{A}| \leq \hat{c}_2 h(\mathcal{T})^n$ aus

$$|u|_{H^\ell(\mathcal{T})} \leq c_2 \, |\mathsf{A}^{-1}|_2^\ell |\det \mathsf{A}|^{1/2} |\tilde{u}|_{H^\ell(\mathcal{T}_\circ)} \leq c_2 \hat{c}_2^{1/2} \left(\frac{h(\mathcal{T}_\circ)}{\varrho(\mathcal{T})} \right)^\ell h(\mathcal{T})^{n/2} |\tilde{u}|_{H^\ell(\mathcal{T}_\circ)}$$

mit der Konstanten $\tilde{c}_2 = c_2 \hat{c}_2^{1/2}$. □

Nun kommen wir zu dem zentralen Satz; vergleiche [Dzi10, Satz 3.31].

Satz 3.27. *Seien $\mathcal{T} \subset \mathbb{R}^n$ ein n-dimensionales Simplex, $\mathcal{T}_\circ \subset \mathbb{R}^n$ das n-dimensionales Einheitssimplex und $F : \mathcal{T}_\circ \to \mathcal{T}$ die affin-lineare Abbildung aus Lemma 3.11. Seien $k, \ell \in \mathbb{N}_0$ mit $k+1 \geq \ell$. Insbesondere gilt damit $H^{k+1}(\mathcal{T}_\circ) \hookrightarrow H^\ell(\mathcal{T}_\circ)$. Ferner sei $\mathcal{I}_\circ : H^{k+1}(\mathcal{T}_\circ) \to H^\ell(\mathcal{T}_\circ)$ ein linearer, stetiger Interpolationsoperator, der auf $\mathbb{P}_k(\mathcal{T}_\circ)$ invariant ist. Dann gilt für den linearen, stetigen Interpolationsoperator $\mathcal{I} : H^{k+1}(\mathcal{T}) \to H^\ell(\mathcal{T})$, der durch*

$$(\mathcal{I}u)(F(\tilde{\boldsymbol{x}})) = (\mathcal{I}_\circ \tilde{u})(\tilde{\boldsymbol{x}}) \quad \text{für } \tilde{\boldsymbol{x}} \in \mathcal{T}_\circ,\ u \in H^{k+1}(\mathcal{T}) \text{ und } \tilde{u} = u \circ F \in H^{k+1}(\mathcal{T}_\circ)$$

beziehungsweise

$$(\mathcal{I}u)(\boldsymbol{x}) = (\mathcal{I}_\circ \tilde{u})(F^{-1}(\boldsymbol{x})) \quad \text{für } \boldsymbol{x} \in \mathcal{T},\ u \in H^{k+1}(\mathcal{T}) \text{ und } \tilde{u} = u \circ F \in H^{k+1}(\mathcal{T}_\circ)$$

definiert ist, dass jedes $u \in H^{k+1}(\mathcal{T})$ die Abschätzung

$$|u - \mathcal{I}u|_{H^\ell(\mathcal{T})} \leq \hat{c} \sigma(\mathcal{T})^\ell h(\mathcal{T})^{k+1-\ell} \, |u|_{H^{k+1}(\mathcal{T})}$$

erfüllt mit einer von n, ℓ, k und $\|\mathcal{I}_\circ\|$ abhängenden Konstanten $c > 0$.

Beweis. Aus der Stetigkeit von $\mathcal{I}_\circ : H^{k+1}(\mathcal{T}_\circ) \to H^\ell(\mathcal{T}_\circ)$ ergibt sich, dass auch $\mathcal{I} : H^{k+1}(\mathcal{T}) \to H^\ell(\mathcal{T})$ stetig ist. Wegen $k+1 \geq \ell$ gilt $u \in H^\ell(\mathcal{T})$ für $u \in H^{k+1}(\mathcal{T})$. Mit Satz 3.25 erhalten wir

$$|u - \mathcal{I}u|_{H^\ell(\mathcal{T})} \leq c_2 \, |\mathsf{A}^{-1}|_2^\ell |\det \mathsf{A}|^{1/2} |(u - \mathcal{I}u) \circ F|_{H^\ell(\mathcal{T}_\circ)} = c_2 \, |\mathsf{A}^{-1}|_2^\ell |\det \mathsf{A}|^{1/2} |\tilde{u} - \mathcal{I}_\circ \tilde{u}|_{H^\ell(\mathcal{T}_\circ)}.$$

Für ein beliebig gewähltes $p_\circ \in \mathbb{P}_k(\mathcal{T}_\circ)$ folgt aus $\mathcal{I}_\circ p_\circ = p_\circ$ somit

$$\begin{aligned} |u - \mathcal{I}u|_{H^\ell(\mathcal{T})} &\leq c_2 \, |\mathsf{A}^{-1}|_2^\ell |\det \mathsf{A}|^{1/2} |\tilde{u} - p_\circ - \mathcal{I}_\circ(\tilde{u} - p_\circ)|_{H^\ell(\mathcal{T}_\circ)} \\ &\leq c_2 \, |\mathsf{A}^{-1}|_2^\ell |\det \mathsf{A}|^{1/2} \Big(|\tilde{u} - p_\circ|_{H^\ell(\mathcal{T}_\circ)} + |\mathcal{I}_\circ(\tilde{u} - p_\circ)|_{H^\ell(\mathcal{T}_\circ)} \Big). \end{aligned} \tag{3.29}$$

Da $H^{k+1}(\mathcal{T}_\circ) \hookrightarrow H^\ell(\mathcal{T}_\circ)$ gilt, existiert eine Einbettungskonstante $\tilde{c}_1 > 0$ mit

$$|\tilde{u} - p_\circ|_{H^\ell(\mathcal{T}_\circ)} \leq \tilde{c}_1 \, \|\tilde{u} - p_\circ\|_{H^{k+1}(\mathcal{T}_\circ)}. \tag{3.30}$$

Aufgrund der Stetigkeit von $\mathcal{I}_\circ$ ergibt sich aus (3.29) und (3.30)

$$\begin{aligned} |u - \mathcal{I}u|_{H^\ell(\mathcal{T})} &\leq c_2 \, |\mathsf{A}^{-1}|_2^\ell |\det \mathsf{A}|^{1/2} \big(\tilde{c}_1 + \|\mathcal{I}_\circ\| \big) \|\tilde{u} - p_\circ\|_{H^{k+1}(\mathcal{T}_\circ)} \\ &\leq \tilde{c}_2 \, |\mathsf{A}^{-1}|_2^\ell |\det \mathsf{A}|^{1/2} \|\tilde{u} - p_\circ\|_{H^{k+1}(\mathcal{T}_\circ)} \end{aligned}$$

mit der Konstante $\tilde{c}_2 = c_2(\tilde{c}_1 + \|\mathcal{I}_\circ\|) > 0$. Da wir das Polynom $p_\circ \in \mathbb{P}_k(\mathcal{T}_\circ)$ beliebig gewählt haben, erhalten wir mit (3.27) die Abschätzung

$$\begin{aligned} |u - \mathcal{I}u|_{H^\ell(\mathcal{T})} &\leq \tilde{c}_2 \, |\mathsf{A}^{-1}|_2^\ell |\det \mathsf{A}|^{1/2} \inf \big\{ \|\tilde{u} - p_\circ\|_{H^{k+1}(\mathcal{T}_\circ)} \,\big|\, p_\circ \in \mathbb{P}_k(\mathcal{T}_\circ) \big\} \\ &= \tilde{c}_2 \, |\mathsf{A}^{-1}|_2^\ell |\det \mathsf{A}|^{1/2} \|\tilde{u}\|_{H^{k+1}(\mathcal{T}_\circ)/\mathbb{P}_k(\mathcal{T}_\circ)}. \end{aligned}$$

Aus Lemma 3.23 und Satz 3.25 bekommen wir

$$\begin{aligned}|u-\mathcal{I}u|_{H^\ell(\mathcal{T})} &\le c\tilde{c}_2\,|\mathsf{A}^{-1}|_2^\ell|\det\mathsf{A}|^{1/2}|\tilde{u}|_{H^{k+1}(\mathcal{T}_\circ)/\mathbb{P}_k(\mathcal{T}_\circ)}\\ &\le \tilde{c}_3\,|\mathsf{A}^{-1}|_2^\ell|\det\mathsf{A}|^{1/2}|\mathsf{A}|_2^{k+1}|\det\mathsf{A}|^{-1/2}|u|_{H^{k+1}(\mathcal{T})} = \tilde{c}_3\,|\mathsf{A}^{-1}|_2^\ell|\mathsf{A}|_2^{k+1}|u|_{H^{k+1}(\mathcal{T})}\end{aligned}$$

mit der Konstante $\tilde{c}_3 = cc_1\tilde{c}_2 > 0$. Jetzt lässt sich Lemma 3.11 anwenden:

$$\begin{aligned}|u-\mathcal{I}u|_{H^\ell(\mathcal{T})} &\le \tilde{c}_3\left(\frac{h(\mathcal{T}_\circ)}{\varrho(\mathcal{T})}\right)^\ell\left(\frac{h(\mathcal{T})}{\varrho(\mathcal{T}_\circ)}\right)^{k+1}|u|_{H^{k+1}(\mathcal{T})}\\ &= \tilde{c}_3 h(\mathcal{T}_\circ)^\ell\varrho(\mathcal{T}_\circ)^{-k-1}\sigma(\mathcal{T})^\ell h(\mathcal{T})^{k+1-\ell}\,|u|_{H^{k+1}(\mathcal{T})}\\ &= \hat{c}\,\sigma(\mathcal{T})^\ell h(\mathcal{T})^{k+1-\ell}\,|u|_{H^{k+1}(\mathcal{T})},\end{aligned}$$

wobei wir $\hat{c} = \tilde{c}_3 h(\mathcal{T}_\circ)^\ell\varrho(\mathcal{T}_\circ)^{-k-1} > 0$ gesetzt haben. □

3.5 Konvergenz des Galerkin-Verfahren

Sei $\mathcal{T}_\circ$ das zweidimensionale Einheitssimplex in $\mathbb{R}^n$, $n\in\{1,2,3\}$, mit den Ecken $\{\mathsf{e}_j\}_{j=0}^n$. Für $\tilde{\boldsymbol{x}}\in\mathcal{T}_\circ$ erhalten wir mit den baryzentrischen Koordinaten $\boldsymbol{\lambda} = (\lambda_0,\dots,\lambda_n)^\top\in\mathbb{R}^{n+1}$

$$\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{x}}(\boldsymbol{\lambda}) = \sum_{j=0}^n\lambda_j\mathsf{e}_j = \begin{pmatrix}\lambda_1\\ \vdots\\ \lambda_n\end{pmatrix}.$$

Die nodale Basis $\{\tilde{\phi}_i\}_{i=0}^n$ auf $\mathcal{T}_\circ$ ist gemäß Satz 3.15 gegeben durch

$$\tilde{\phi}_i(\tilde{\boldsymbol{x}}) = \tilde{\phi}_i(\tilde{\boldsymbol{x}}(\boldsymbol{\lambda})) = \sum_{j=0}^n\tilde{\phi}_i(\mathsf{e}_j)\lambda_j = \sum_{j=0}^n\delta_{ij}\lambda_j = \lambda_i = \lambda_i(\tilde{\boldsymbol{x}})\quad\text{für }\tilde{\boldsymbol{x}}\in\mathcal{T}_\circ\text{ und }0\le i\le n. \tag{3.31}$$

Wir führen nun den Interpolationsoperator $\mathcal{I}_\circ$ aus Satz 3.27 ein und wählen dazu $k=1$ und $\ell\in\{0,1\}$. Aus dem Auswahlsatz von Rellich und Kondrachov [DR12, Satz 16.22-(iii)] folgt wegen $n\le 3$, dass $H^{k+1}(\mathcal{T}_\circ) = H^2(\mathcal{T}_\circ)\hookrightarrow C(\mathcal{T}_\circ)$ gilt. Nun definieren wir $\mathcal{I}_\circ : H^2(\mathcal{T}_\circ)\to H^\ell(\mathcal{T}_\circ)$ wie folgt:

$$\mathcal{I}_\circ\tilde{\varphi}\in\mathbb{P}_1(\mathcal{T}_\circ)\hookrightarrow H^\ell(\mathcal{T}_\circ),\quad (\mathcal{I}_\circ\tilde{\varphi})(\mathsf{e}_j) = \tilde{\varphi}(\mathsf{e}_j)\text{ für }\tilde{\varphi}\in H^2(\mathcal{T}_\circ)\hookrightarrow C(\mathcal{T}_\circ)\text{ und }j=0,1\dots,n.$$

Wir interploieren also Funktionen in $H^2(\mathcal{T}_\circ)$ durch stückweise lineare Finite Elemente. Es ist klar, dass $\mathcal{I}_\circ$ linear ist. Wir wollen zeigen, dass $\mathcal{I}_\circ$ auch ein beschränkter Operator ist. Offenbar gilt wegen Satz 3.15-1) und (3.31)

$$(\mathcal{I}_\circ\tilde{\varphi})(\tilde{\boldsymbol{x}}) = \sum_{i=0}^n\tilde{\varphi}(\mathsf{e}_i)\lambda_i = \sum_{i=0}^n\tilde{\varphi}(\mathsf{e}_i)\tilde{\phi}_i(\tilde{\boldsymbol{x}})\quad\text{für }\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{x}}(\boldsymbol{\lambda})\in\mathcal{T}_\circ.$$

Wegen $H^2(\mathcal{T}_\circ)\hookrightarrow C(\mathcal{T}_\circ)$ existiert eine Einbettungskonstante $c_1>0$ mit

$$\|\tilde{\varphi}\|_{C(\mathcal{T}_\circ)}\le c_1\,\|\tilde{\varphi}\|_{H^2(\mathcal{T}_\circ)}\quad\text{für alle }\tilde{\varphi}\in H^2(\mathcal{T}_\circ).$$

Wegen $h(\mathcal{T}_\circ) = \sqrt{2}$ hängt die Konstante

$$c_2 = \sum_{i=0}^n\|\tilde{\phi}_i\|_{H^\ell(\mathcal{T}_\circ)} > 0$$

nur von $n\in\{1,2,3\}$ und $\ell\in\{0,1\}$ ab. Also bekommen wir für $\tilde{\varphi}\in H^2(\mathcal{T}_\circ)$ die Abschätzung

$$\|\mathcal{I}_\circ\tilde{\varphi}\|_{H^\ell(\mathcal{T}_\circ)}\le\sum_{i=0}^n|\tilde{\varphi}(\mathsf{e}_i)|\|\tilde{\phi}_i\|_{H^\ell(\mathcal{T}_\circ)}\le\|\tilde{\varphi}\|_{C(\mathcal{T}_\circ)}\sum_{i=0}^n\|\tilde{\phi}_i\|_{H^\ell(\mathcal{T}_\circ)}\le c_3\,\|\tilde{\varphi}\|_{H^2(\mathcal{T}_\circ)}$$

mit einer nur von $n \in \{1,2,3\}$ und $\ell \in \{0,1\}$ abhängigen Konstante $c_3 = c_1 c_2 > 0$. Damit haben wir bewiesen, dass der lineare Operator $\mathcal{I}_\circ$ stetig ist.

Seien nun $\varphi \in H^2(\mathcal{T})$ und $F : \mathcal{T}_\circ \to \mathcal{T}$ die affin-lineare Abbildung aus Lemma 3.11. Wir setzen $\tilde{\varphi} = \varphi \circ F \in H^2(\mathcal{T}_\circ)$. Die nodale Basis $\{\phi_i\}_{i=0}^n$ auf $\mathcal{T}$ ist gegeben durch $\phi_i = \tilde{\phi}_i \circ F^{-1} : \mathcal{T} \to \mathbb{R}$ für $i = 0, \dots, n$. Insbesondere gilt

$$\phi_i(\mathrm{a}_j) = \tilde{\phi}_i(F^{-1}\mathrm{a}_j) = \tilde{\phi}_i(\mathrm{e}_j) = \delta_{ij} \quad \text{für } i,j = 0,\dots,n.$$

Nun ist der lineare Interpolationsoperator $\mathcal{I} : H^2(G) \to H^\ell(\mathcal{T})$ gegeben durch

$$(\mathcal{I}\varphi)(F(\tilde{\boldsymbol{x}})) = (\mathcal{I}_\circ\tilde{\varphi})(\tilde{\boldsymbol{x}}) = \sum_{i=0}^{2} \tilde{\varphi}(\mathrm{e}_i)\tilde{\phi}_i(\tilde{\boldsymbol{x}}) \quad \text{für alle } \tilde{\boldsymbol{x}} \in \mathcal{T}_\circ \tag{3.32}$$

oder

$$(\mathcal{I}\varphi)(\boldsymbol{x}) = (\mathcal{I}_\circ\tilde{\varphi})(F^{-1}\boldsymbol{x}) = \sum_{i=0}^{n} \tilde{\varphi}(\mathrm{e}_i)\tilde{\phi}_i(F^{-1}\boldsymbol{x}) = \sum_{i=0}^{n} \varphi(\mathrm{a}_i)\phi_i(\boldsymbol{x}) \quad \text{für alle } \boldsymbol{x} \in \mathcal{T}.$$

Wir erhalten dann aus Satz 3.25 die Abschätzung

$$|\varphi - \mathcal{I}\varphi|_{H^\ell(\mathcal{T})} \le c\sigma(\mathcal{T})^\ell h(\mathcal{T})^{2-\ell} |\varphi|_{H^2(\mathcal{T})} \quad \text{für alle } \varphi \in H^2(\mathcal{T}) \text{ und } \ell = 0,1.$$

mit einer nur von $n \in \{1,2,3\}$, von $\ell \in \{0,1\}$ und von $\|\mathcal{I}_\circ\|$ abhängenden Konstante $c > 0$. Insbesondere gelten im Fall $\ell = 0$

$$\|\varphi - \mathcal{I}\varphi\|_{L^2(\mathcal{T})} = |\varphi - \mathcal{I}\varphi|_{H^0(\mathcal{T})} \le c h(\mathcal{T})^2 |\varphi|_{H^2(\mathcal{T})} \quad \text{für alle } \varphi \in H^2(\mathcal{T}) \tag{3.33}$$

und im Fall $\ell = 1$

$$\|\varphi - \mathcal{I}\varphi\|_{H^1_0(\mathcal{T})} = |\varphi - \mathcal{I}\varphi|_{H^1(\mathcal{T})} \le c\sigma(\mathcal{T}) h(\mathcal{T}) |\varphi|_{H^2(\mathcal{T})} \quad \text{für alle } \varphi \in H^2(\mathcal{T}). \tag{3.34}$$

Wir kehren nun zu unserem ursprünglichen Poissonproblem (3.2) zurück. Sei $G \subset \mathbb{R}^n$, $n \in \{1,2,3\}$, ein beschränktes Gebiet, dass die Eigenschaft (3.19) erfüllt mit einer zulässigen Triangulierung $\Delta(G) = \{\mathcal{T}_j \,|\, j = 1,\dots,m\}$. Aus dem Auswahlsatz von Rellich und Kondrachov [DR12, Satz 16.22-(iii)] folgt wiederum wegen $n \le 3$, dass $H^2(G) \hookrightarrow C(\overline{G})$ gilt. Sei $u \in V$ die Lösung des kontinuierlichen Variationsproblems (3.2). Wir nehmen zusätzlich an, dass $u \in H^2(G)$ erfüllt ist. Dann bekommen wir im Fall $\ell = 0$ mit (3.33)

$$\begin{aligned}\|u - \mathcal{I}u\|_H^2 = |u - \mathcal{I}u|_{H^0(G)}^2 &= \sum_{j=1}^{m} |u - \mathcal{I}u|_{H^0(\mathcal{T}_j)}^2 \le \sum_{j=1}^{m} c^2 h(\mathcal{T}_j)^4 |u|_{H^2(\mathcal{T}_j)}^2 \\ &\le c^2 h^4 \sum_{j=1}^{m} |u|_{H^2(\mathcal{T}_j)}^2 = c^2 h^4 |u|_{H^2(G)}^2.\end{aligned}$$

und für den Fall $\ell = 1$ mit (3.34)

$$\begin{aligned}\|u - \mathcal{I}u\|_V^2 = |u - \mathcal{I}u|_{H^1(G)}^2 &= \sum_{j=1}^{m} |u - \mathcal{I}u|_{H^1(\mathcal{T}_j)}^2 \le \sum_{j=1}^{m} c^2 \sigma(\mathcal{T}_j)^2 h(\mathcal{T}_j)^2 |u|_{H^2(\mathcal{T}_j)}^2 \\ &\le c^2\sigma^2 h^2 \sum_{j=1}^{m} |u|_{H^2(\mathcal{T}_j)}^2 = c^2\sigma^2 h^2 |u|_{H^2(G)}^2.\end{aligned}$$

Damit gelten

$$\begin{aligned}\|u - \mathcal{I}u\|_H &\le c h^2 |u|_{H^2(G)} = \mathcal{O}(h) \quad (h \to 0), \\ \|u - \mathcal{I}u\|_V &\le c\sigma h |u|_{H^2(G)} = \mathcal{O}(h) \quad (h \to 0),\end{aligned} \tag{3.35}$$

sofern σ nach oben beschränkt ist. Sei u_h die Lösung des diskreten Variationsproblems (3.7). Dann folgt mit (3.13) und (3.35) die Konvergenz des Galerkin-Verfahrens:

$$\|u - u_h\|_V = \inf\left\{\|u - \varphi_h\|_V \,\middle|\, \varphi_h \in V_h\right\} \leq \|u - \mathcal{I}u\|_V \leq c\sigma h\,|u|_{H^2(G)} = \mathcal{O}(h) \quad (h \to 0),$$

wenn σ nach oben beschränkt ist. In dem folgenden Satz ist das erhaltene Resultat noch einmal zusammengefasst.

Satz 3.28. *Sei $G \subset \mathbb{R}^n$, $n \in \{1, 2, 3\}$ ein beschränktes Gebiet, dass die Eigenschaft* (3.19) *erfüllt mit einer zulässigen Triangulierung $\Delta(G) = \{\mathfrak{T}_j \,|\, j = 1, \ldots, m\}$. Es gelte*

$$0 < \sigma = \max_{1 \leq j \leq m} \sigma(\mathfrak{T}_j) = \max_{1 \leq j \leq m} \frac{h(\mathfrak{T}_j)}{\varrho(\mathfrak{T}_j)} \leq \sigma_\circ$$

für eine von $\Delta(G)$ unabhängige Konstante $\sigma_\circ > 0$. Wir nehmen an, dass die eindeutige Lösung $u \in V$ des kontinuierlichen Variationsproblems (3.2) *im Raum $H^2(G)$ liegt. Als endlichdimensionalen Raum $V_h \subset V$ wählen wir den Raum der stückweise linearen Finite Elemente, das heißt, wir setzen*

$$V_h = \left\{\varphi \in C(\overline{G}) \,\middle|\, \varphi|_{\mathfrak{T}_j} \in \mathbb{P}_1(\mathfrak{T}_j) \text{ für } j = 1, \ldots, m\right\}.$$

Sei $u_h \in V_h$ die Lösung des diskreten Variationsproblems (3.7)*. Dann existiert eine Konstante $\bar{c} > 0$, so dass*

$$\|u - \mathcal{I}u\|_V \leq \bar{c}h\,|u|_{H^2(G)} = \mathcal{O}(h) \quad (h \to 0) \tag{3.36}$$

gilt.

4 Galerkin-Verfahren für die Wärmeleitungsgleichung

In diesem Kapitel setzen wir voraus, dass G ein beschränktes Gebiet in $\mathbb{R}^n$ ist. Ferner sind $H = L^2(G)$ und $V = H_0^1(G)$. Zur Theorie parabolischer Probleme verweisen wir auf [DR12, Kapitel 23]. Weitere Literatur zur numerischen Behandlung parabolischer Probleme finden Sie in [Dzi10, Tho97].

4.1 Ritz-Projektion

Definition 4.1. *Sei $V_h = \operatorname{Span}\{\varphi_1, \dots, \varphi_N\}$ ein endlich-dimensionaler Teilraum von V. Dann heißt die Abbildung $\mathcal{P}_h : V \to V_h$, die gegeben ist durch*

$$\langle \mathcal{P}_h\varphi, \varphi_h\rangle_V = \langle \varphi, \varphi_h\rangle_V \quad \text{für alle } \varphi_h \in V_h \text{ und für } \varphi \in V, \tag{4.1}$$

die Ritz-Projektion. *Das Element $\mathcal{P}_h\varphi \in V_h$ wird* Ritz-Projektion der Funktion $\varphi \in V$ *genannt.*

Lemma 4.2. *Sei $V_h = \operatorname{Span}\{\varphi_1, \dots, \varphi_N\}$ ein endlich-dimensionaler Teilraum von V. Die Ritz-Projektion $\mathcal{P}_h : V \to V_h$ ist wohldefiniert, linear, und es gilt*

$$\|\mathcal{P}_h\varphi - \varphi\|_V = \inf\left\{\|\varphi - \varphi_h\|_V \,\middle|\, \varphi_h \in V_h\right\} \quad \text{für } \varphi \in V.$$

Beweis. Zu gegebenem $\varphi \in V$ ist die in in Definition 4.1 eingeführte Ritz-Projektion $\mathcal{P}_h\varphi \in V_h$ von φ genau der Riesz-Repräsentant in V_h des Funktionals $g_h : V_h \to \mathbb{R}$, $g_h = g|_{V_h}$ mit $g : V \to \mathbb{R}$, $g = \langle \varphi, \cdot\rangle_V$. Daher ist $\mathcal{P}_h$ nach dem Darstellungssatz von Riesz [DR12, Satz 12.24] wohldefiniert. Die Linearität von $\mathcal{P}_h$ ist klar. Wir orientieren wir uns nun an dem Beweis von Satz 3.7; vergeiche [Dzi10, Satz 3.10]. Es ist wieder

$$\|\mathcal{P}_h\varphi - \varphi\|_V \le \inf\left\{\|\varphi - \varphi_h\|_V \,\middle|\, \varphi_h \in V_h\right\} \quad \text{für } \varphi \in V$$

zu zeigen, da '$\ge$' offenbar ist erfüllt ist. Aus (4.1) erhalten wir wieder die *Galerkin-Orthogonalität*

$$\int_G \nabla(\varphi - \mathcal{P}_h\varphi) \cdot \nabla\varphi_h \,\mathrm{d}\boldsymbol{x} = \langle \varphi - \mathcal{P}_h\varphi, \varphi_h\rangle_V = 0 \quad \text{für alle } \varphi_h \in V_h. \tag{4.2}$$

Für jedes $\varphi_h \in V_h$ gilt daher

$$\begin{aligned}
\|\varphi - \mathcal{P}_h\varphi\|_V^2 &= \int_G \nabla(\varphi - \mathcal{P}_h\varphi) \cdot \nabla\varphi \,\mathrm{d}\boldsymbol{x} - \underbrace{\int_G \nabla(\varphi - \mathcal{P}_h\varphi) \cdot \nabla(\mathcal{P}_h\varphi) \,\mathrm{d}\boldsymbol{x}}_{=0 \text{ wegen (4.2) und } \mathcal{P}_h\varphi \in V_h} \\
&= \int_G \nabla(\varphi - \mathcal{P}_h\varphi) \cdot \nabla\varphi \,\mathrm{d}\boldsymbol{x} \\
&= \int_G \nabla(\varphi - \mathcal{P}_h\varphi) \cdot \nabla\varphi \,\mathrm{d}\boldsymbol{x} - \underbrace{\int_G \nabla(\varphi - \mathcal{P}_h\varphi) \cdot \nabla\varphi_h \,\mathrm{d}\boldsymbol{x}}_{=0 \text{ wegen (4.2) und } \varphi_h \in V_h} \\
&= \int_G \nabla(\varphi - \mathcal{P}_h\varphi) \cdot \nabla(\varphi - \varphi_h) \,\mathrm{d}\boldsymbol{x} \le \|\varphi - \mathcal{P}_h\varphi\|_V \|\varphi - \varphi_h\|_V.
\end{aligned}$$

Da $\varphi_h \in V_h$ beliebig gewählt ist, erhalten wir

$$\|\varphi - \mathcal{P}_h\varphi\|_V \le \inf\left\{\|\varphi - \varphi_h\|_V \,\middle|\, \varphi_h \in V_h\right\}$$

was zu zeigen war. □

Bemerkung 4.3. Sei $u \in V$ die Lösung von (3.2). Dann ist $u_h = \mathcal{P}_h u$ genau die Lösung des diskreten Poissonproblems (3.7). ◊

Wir werden die folgende Annahme benötigen.

Voraussetzung 1. *Für ein $\varphi \in H^2(G) \cap V$ gelte die Abschätzung*

$$\|\varphi - \mathcal{P}_h\varphi\|_H + h\,\|\varphi - \mathcal{P}_h\varphi\|_V \le C_1 h^2\,\|\varphi\|_{H^2(G)}$$

mit einer Konstanten $C_1 > 0$, die nicht von φ abhängt.

Bemerkung 4.4. Das beschränkte Gebiet G erfülle (3.19) mit einer zulässigen Triangulierung $\Delta(G) = \{\mathfrak{T}_j \,|\, j = 1, \dots, m\}$. Ferner sei $n \in \{1, 2, 3\}$. Aus (3.36) folgt, dass

$$h\,\|\mathcal{P}_h\varphi - \varphi\|_V = h \inf\left\{\|\varphi - \varphi_h\|_V \,\middle|\, \varphi_h \in V_h\right\} \le \tilde{c}_1 h^2\,\|\varphi\|_{H^2(G)} \quad \text{für } \varphi \in H^2(G) \tag{4.3}$$

gilt, wobei $\tilde{c}_1 > 0$ nicht von φ abhängt und der endlich-dimensionale Raum V_h durch stückweise lineare Finite Elemente aufgespannt wird. ◊

Wir können nun Voraussetzung 1 unter der folgenden Regularitätsannahme beweisen, die wir auch bereits zur Herleitung von (3.36) verwendet haben.

Voraussetzung 2. *Für jedes $f \in H \hookrightarrow V'$ ist die Lösung $u \in V$ von $-\Delta u = f$ in H aus dem Raum $H^2(G)$ und erfüllt die Abschätzung*

$$\|u\|_{H^2(G)} \le C_2\,\|f\|_H$$

mit einer Konstanten $C_2 > 0$, die nicht von u und f abhängt.

Lemma 4.5. *Das Gebiet G erfülle (3.19), $\Delta(G)$ sei eine zulässige Triangulierung von G und $\sigma = \max_{1\le j\le m} \sigma(\mathfrak{T}_j)$ sei durch $\sigma_\circ > 0$ nach oben beschränkt. Ferner gelte $n \in \{1, 2, 3\}$. Der Raum V_h sei aufgespannt von den modalen Basisfunktionen $\{\phi_i\}_{i=1}^{\bar{m}} \subset C(\overline{G})$, die stückweise linear sind, und es gelte Vorraussetzung 2. Dann gilt für $v \in V \cap H^2(G)$*

$$\|\mathcal{P}_h v - v\|_H \le \tilde{c}_2 h^2\,\|v\|_{H^2(G)}$$

mit einer von v unabhängigen Konstante $\tilde{c}_2$.

Beweis. Wir wenden den *Aubin-Nitsche-Trick* an. Offenbar gilt $\mathcal{P}_h v - v \in H$. Sei $w \in V$ die Lösung von

$$\langle w, \varphi\rangle_V = \langle \mathcal{P}_h v - v, \varphi\rangle_V \quad \text{für alle } \varphi \in V.$$

Aus Voraussetzung 2 schließen wir, dass

$$\|w\|_{H^2(G)} \le C_2\,\|\mathcal{P}_h v - v\|_H \tag{4.4}$$

erfüllt ist. Nun wenden wir (4.2), (4.3) und (4.4) an. Es ergibt sich dann für ein beliebiges $\varphi \in V_h$

$$\begin{aligned}\|\mathcal{P}_h v - v\|_H^2 &= \langle \mathcal{P}_h v - v, \mathcal{P}_h v - v\rangle_H = \langle w, \mathcal{P}_h v - v\rangle_V \\ &= \langle w - \varphi_h, \mathcal{P}_h v - v\rangle_V \le \|w - \varphi_h\|_V \|\mathcal{P}_h v - v\|_V \le \|w - \varphi_h\|_V c_1 h\,\|v\|_{H^2(G)} \\ &= c_1 h\,\|v\|_{H^2(G)} \|w - \varphi_h\|_V.\end{aligned}$$

Nun wählen wir $\varphi_h = \mathcal{I} w$, wobei $\mathcal{I}$ der in (3.32) eingeführte Interpolationsoperator ist. Dann bekommen wir

$$\begin{aligned}\|\mathcal{P}_h v - v\|_H^2 &\le c_1 h\,\|v\|_{H^2(G)} \|w - \mathcal{I} w\|_V \le c_1 h\,\|v\|_{H^2(G)} c\sigma_\circ h |w|_{H^2(G)} \\ &\le c c_1 \sigma_\circ h^2\,\|v\|_{H^2(G)} \|w\|_{H^2(G)} \le c c_1 \sigma_\circ h^2\,\|v\|_{H^2(G)} C_2\,\|\mathcal{P}_h v - v\|_H \\ &= \tilde{c}_2 h^2\,\|v\|_{H^2(G)} \|\mathcal{P}_h v - v\|_H\end{aligned}$$

mit $\tilde{c}_2 = c c_1 \sigma_\circ C_2 > 0$, was zu zeigen war. □

4.2 Anfangsrandwertproblem für die Wärmeleitungsgleichung

Für $T > 0$ definieren wir $Q = (0,T) \times G$ und $\Sigma = (0,T) \times \partial G$. Seien $u_\circ \in H$ und $f \in L^2(0,T;H)$ gegeben. Dann betrachten wir das folgende Anfangsrandwertproblem für die Wärmeleitungsgleichung

$$\begin{aligned} u_t(t,\boldsymbol{x}) - \Delta u(t,\boldsymbol{x}) &= f(t,\boldsymbol{x}) && \text{für fast alle } (t,\boldsymbol{x}) \in Q, \\ u(t,\boldsymbol{x}) &= 0 && \text{für fast alle } (t,\boldsymbol{x}) \in \Sigma, \\ u(0,\boldsymbol{x}) &= u_\circ(\boldsymbol{x}) && \text{für fast alle } \boldsymbol{x} \in G. \end{aligned} \tag{4.5}$$

Definition 4.6. *Eine schwache Lösung u von (4.5) ist eine Funktion, die $u(0,\cdot) = u_\circ$ in H, $u(t,\cdot) \in V$ für fast alle $t \in [0,T]$, $u \in H^1(Q) \simeq H^1(0,T;H) \cap L^2(0,T;V)$ und*

$$\int_G u_t(t,\cdot)\varphi + \nabla u(t,\cdot) \cdot \nabla\varphi \,\mathrm{d}\boldsymbol{x} = \int_G f(t,\cdot)\varphi \,\mathrm{d}\boldsymbol{x} \quad \textit{für alle } \varphi \in V \textit{ und } t \in (0,T]$$

beziehungsweise

$$\langle u_t(t,\cdot),\varphi\rangle_H + \langle u(t,\cdot),\varphi\rangle_V = \langle f(t,\cdot),\varphi\rangle_H \quad \textit{für alle } \varphi \in V \textit{ und } t \in (0,T] \tag{4.6}$$

erfüllt.

4.3 Ortsdiskretisierung

Nun werden wir das Galerkin-Verfahren verwenden, um (4.6) numerisch zu lösen. Wir wählen nun wieder wie in Kapitel 3 linear unabhänige Funktionen $\{\varphi_i\}_{i=1}^N \subset V$ und definieren den endlichdimensionalen Teilraum V_h wie in (3.6). Für eine approximative Lösung u_h verlangen wir $u_h(t,\cdot) \in V_h$ für fast alle $t \in [0,T]$. Es gilt also

$$u_h(t,\boldsymbol{x}) = \sum_{j=1}^N u_j(t)\varphi_j(\boldsymbol{x}) \quad \text{für } (t,\boldsymbol{x}) \in \overline{Q} \tag{4.7}$$

gelten. Ferner soll

$$\langle u_{ht}(t,\cdot),\varphi_h\rangle_H + \langle u_h(t,\cdot),\varphi\rangle_V = \langle f(t,\cdot),\varphi_h\rangle_H \quad \text{für alle } \varphi_h \in V_h \text{ und } t \in (0,T] \tag{4.8}$$

gelten, wobei wir u_{ht} statt $(u_h)_t$ für die Zeitableitung von u_h schreiben. Wir setzen (4.7) in (4.8) ein und wählen $\varphi_h = \varphi_i$ für $i = 1,\dots,N$:

$$\sum_{j=1}^N \dot{u}_j(t)\langle\varphi_j,\varphi_i\rangle_H + \langle\varphi_j,\varphi_i\rangle_V = \langle f(t,\cdot),\varphi_i\rangle_H \quad \text{für } i = 1,\dots,N \text{ und } t \in (0,T]. \tag{4.9}$$

Mit der Steifigkeitsmatrix, die wir in (3.12) eingeführt haben, mit der Massematrix

$$\mathrm{M} = \big((\mathrm{M}_{ij})\big) \in \mathbb{R}^{N\times N} \quad \text{mit } \mathrm{M}_{ij} = \int_G \varphi_j\varphi_i \,\mathrm{d}\boldsymbol{x} = \langle\varphi_j,\varphi_i\rangle_H \text{ für } 1 \le i,j \le N$$

und mit den zeitabhängigen Vektoren

$$\begin{aligned} \mathrm{u}(t) &= \big(\mathrm{u}_i(t)\big) \in \mathbb{R}^N && \text{mit} && \mathrm{u}_i(t) = u_i(t) && \text{für } 1 \le i \le N \text{ und } t \in [0,T], \\ \mathrm{f}(t) &= \big(\mathrm{f}_i(t)\big) \in \mathbb{R}^N && \text{mit} && \mathrm{f}_i(t) = \langle f(t,\cdot),\varphi_i\rangle_H && \text{für } 1 \le i \le N \text{ und } t \in [0,T] \end{aligned}$$

läßt sich (4.9) als Differentialgleichungssystem in $\mathbb{R}^N$ schreiben:

$$\mathrm{M}\ddot{\mathrm{u}}(t) + \mathrm{S}\mathrm{u}(t) = \mathrm{f}(t) \text{ für } t \in (0,T]; \tag{4.10a}$$

vergleiche (1.1). Offenbar gilt $\mathrm{f} \in L^2(0,T;\mathbb{R}^N)$. Es folgt aus Lemma 1.1, dass (4.10) eine eindeutige Lösung $\mathrm{u} \in H^1(0,T;\mathbb{R}^N)$ besitzt, die Hölder-stetig in $[0,T]$ ist. Damit ist die durch (4.7) gegebene Funktion u_h die eindeutige Lösung des Variationsproblems (4.8) zusammen mit der Anfangsbedingung $u_h(0,\cdot) = u_{\circ,h}$ in V_h. Für die Anfangsbedingung ersetzen wir $u_\circ$ durch eine passende Funktion

$$u_{\circ,h} = \sum_{j=1}^{N} u_{\circ,i}\varphi_j \in V^h,$$

die wir später spezifizieren. Mit

$$\mathrm{u}_\circ = \left(\mathrm{u}_{\circ,i}\right) \in \mathbb{R}^N \quad \text{mit } \mathrm{u}_{\circ,i} = u_{\circ,i} \text{ für } 1 \le i \le N$$

bekommen wir die Anfangsbedingung

$$\mathrm{u}(0) = \mathrm{u}_\circ. \tag{4.10b}$$

Wir wollen nun den Fehler zwischen der schwachen Lösung u von (4.5) und u_h abschätzen. Aus (4.6) und (4.8) erhalten wir

$$\langle (u_t - u_{ht})(s), \varphi_h\rangle_H + \langle (u - u_h)(s), \varphi_h\rangle_V = 0 \quad \text{für alle } \varphi_h \in V_h \text{ und } s \in (0,T], \tag{4.11}$$

wobei wir beispielsweise $u(s)$ für die auf G definierte Funktion $u(s,\cdot)$ schreiben. Wir verwenden die in Definition 4.1 eingeführte Ritz-Projektion $\mathcal{P}_h$. In der folgenden Herleitung nehmen wir jeweils an, dass die kontinuierliche Lösung u hinreichend regulär ist. Aus (4.11) bekommen wir

$$\begin{aligned}&\langle (\mathcal{P}_h u_t - u_{ht})(s), \varphi_h\rangle_H + \langle (\mathcal{P}_h u - u_h)(s), \varphi_h\rangle_V \\ &\quad = \langle (\mathcal{P}_h u_t - u_t)(s), \varphi_h\rangle_H + \langle (\mathcal{P}_h u - u)(s), \varphi_h\rangle_V \quad \text{für alle } \varphi_h \in V_h \text{ und } s \in (0,T].\end{aligned} \tag{4.12}$$

In (4.12) stehen für $s \in (0,T]$ auf der linken Seite Ausdrücke des diskreten Fehlers $(P_h u - u_h)(s) \in V_h$, während auf der rechten Seite der Ausdrücke des Approximationsfehlers $(P_h u - u)(s) \in V$ stehen. Die Ritz-Projektion hat die Orthogonalitäts-Eigenschaft, dass

$$\langle (\mathcal{P}_h u - u)(s), \varphi_h\rangle_V = 0 \quad \text{für alle } \varphi_h \in V_h \text{ und } s \in (0,T]$$

gilt, das heißt, wir haben $(\mathcal{P}_h u - u)(s) \in V_h^\top$ für $t \in (0,T]$. Daher erhalten wir aus (4.12) die Gleichung

$$\begin{aligned}&\langle (\mathcal{P}_h u_t - u_{ht})(s), \varphi_h\rangle_H + \langle (\mathcal{P}_h u - u_h)(s), \varphi_h\rangle_V \\ &\quad = \langle (\mathcal{P}_h u_t - u_t)(s), \varphi_h\rangle_H \quad \text{für alle } \varphi_h \in V_h \text{ und } s \in (0,T].\end{aligned} \tag{4.13}$$

In (4.13) wählen wir $\varphi_h = (\mathcal{P}_h u - u_h)(s) \in V_h$ für $s \in (0,T]$. Dann gilt

$$\begin{aligned}&\langle (\mathcal{P}_h u_t - u_{ht})(s), (\mathcal{P}_h u - u_h)(s)\rangle_H + \|(\mathcal{P}_h u - u_h)(s)\|_V^2 \\ &\quad = \langle (\mathcal{P}_h u_t - u_t)(s), (\mathcal{P}_h u - u_h)(s)\rangle_H \quad \text{für } s \in (0,T].\end{aligned} \tag{4.14}$$

Mit

$$\langle (\mathcal{P}_h u_t - u_{ht})(s), (\mathcal{P}_h u - u_h)(s)\rangle_H = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}s}\|(\mathcal{P}_h u - u_h)(s)\|_H^2 \quad \text{für } s \in (0,T]$$

und mit der *Cauchy-Schwarz-Ungleichung* [DR12, Satz 12.17] ergibt sich für (4.14)

$$\begin{aligned}&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}s}\|(\mathcal{P}_h u - u_h)(s)\|_H^2 + \|(\mathcal{P}_h u - u_h)(s)\|_V^2 \\ &\quad \le \|(\mathcal{P}_h u_t - u_t)(s)\|_H \|(\mathcal{P}_h u - u_h)(s)\|_H \quad \text{für } s \in (0,T].\end{aligned} \tag{4.15}$$

Wir betrachten die rechte Seite der Ungleichung (4.15). Unter Verwendung von Voraussetzung 1, von der ersten Poincaré-Ungleichung (3.1) und von der Ungleichung von Young (1.10) mit $\varepsilon = 1$, $a = c_\mathrm{p} C_1 h^2 \|u_t(s)\|_{H^2(G)}$, $b = \|(\mathcal{P}_h u - u_h)(s)\|_V$ erhalten wir

$$\begin{aligned}\|(\mathcal{P}_h u_t - u_t)(s)\|_H \|(\mathcal{P}_h u - u_h)(s)\|_H &\le C_1 h^2 \|u_t(s)\|_{H^2(G)} \|(\mathcal{P}_h u - u_h)(s)\|_H \\ &\le C_1 h^2 \|u_t(s)\|_{H^2(G)} c_\mathrm{p} \|(\mathcal{P}_h u - u_h)(s)\|_V \\ &\le \frac{c_1}{2} h^4 \|u_t(s)\|_{H^2(G)}^2 + \frac{1}{2}\|(\mathcal{P}_h u - u_h)(s)\|_V^2,\end{aligned}$$

für $s \in (0,T]$, wobei wir $c_1 = c_\mathsf{p}^2 C_1^2 > 0$ gesetzt haben. Wir schließen daher aus (4.15) die Ungleichung

$$\frac{\mathrm{d}}{\mathrm{d}s} \|(\mathcal{P}_h u - u_h)(s)\|_H^2 + \|(\mathcal{P}_h u - u_h)(s)\|_V^2 \leq c_1 h^4 \|u_t(s)\|_{H^2(G)}^2 \quad \text{für } s \in (0,T].$$

Integration über das Intervall $(0,t)$ mit $t \in (0,T]$ ergibt

$$\begin{aligned}
\|(\mathcal{P}_h u - u_h)(t)\|_H^2 + \int_0^t \|(\mathcal{P}_h u - u_h)(s)\|_V^2 \,\mathrm{d}s &\leq c_1 h^4 \int_0^t \|u_t(s)\|_{H^2(G)}^2 \,\mathrm{d}s + \|(\mathcal{P}_h u - u_h)(0)\|_H^2 \\
&\leq c_1 h^4 \int_0^T \|u_t(s)\|_{H^2(G)}^2 \,\mathrm{d}s + \|\mathcal{P}_h u_\circ - u_{\circ,h}\|_H^2 \\
&\leq c_1 h^4 \|u_t\|_{L^2(0,T;H^2(G))}^2 + \|\mathcal{P}_h u_\circ - u_{\circ,h}\|_H^2.
\end{aligned}$$

Wegen

$$\sqrt{a^2+b^2} \leq \sqrt{a^2+2ab+b^2} = \sqrt{(a+b)^2} = a+b \quad \text{für } a,b \geq 0$$

bekommen wir mit

$$a = \sqrt{c_1} h^2 \|u_t\|_{L^2(0,T;H^2(G))}, \quad b = \|\mathcal{P}_h u_\circ - u_{\circ,h}\|_H$$

die beiden Abschätzungen

$$\begin{aligned}
\|\mathcal{P}_h u - u_h\|_{L^\infty(0,T;H)} &= \operatorname{ess\,sup} \left\{ \|(\mathcal{P}_h u - u_h)(t)\|_H \,\middle|\, t \in [0,T] \right\} \\
&\leq \sqrt{c_1} \|u_t\|_{L^2(0,T;H^2(G))} h^2 + \|\mathcal{P}_h u_\circ - u_{\circ,h}\|_H, \qquad (4.16a) \\
\|\mathcal{P}_h u - u_h\|_{L^2(0,T;V)} &= \left(\int_0^T \|(\mathcal{P}_h u - u_h)(t)\|_V^2 \,\mathrm{d}t \right)^{1/2} \\
&\leq \sqrt{c_1} \|u_t\|_{L^2(0,T;H^2(G))} h^2 + \|\mathcal{P}_h u_\circ - u_{\circ,h}\|_H. \qquad (4.16b)
\end{aligned}$$

Es bleibt, den Approximationsfehler $(\mathcal{P}_h u - u)(s) \in V$ abzuschätzen. Hier können wir Voraussetzung 1 verwenden:

$$\|\mathcal{P}_h u - u\|_{L^\infty(0,T;H)} = \operatorname{ess\,sup} \left\{ \|(\mathcal{P}_h u - u)(t)\|_H \,\middle|\, t \in [0,T] \right\} \leq C_1 h^2 \|u\|_{L^\infty(0,T;H^2(G))}, \quad (4.17a)$$

$$\|\mathcal{P}_h u - u\|_{L^2(0,T;V)} = \left(\int_0^T \|(\mathcal{P}_h u - u)(t)\|_V^2 \,\mathrm{d}t \right)^{1/2} \leq C_1 h \|u\|_{L^2(0,T;H^2(G))}. \quad (4.17b)$$

Insgesamt erhalten wir den folgenden Konvergenzsatz; vergleiche [Dzi10, Satz 5.29].

Satz 4.7 (Konvergenz)**.** *Seien $G \subset \mathbb{R}^n$, $n \in \{1,2,3\}$, ein beschränktes Gebiet, $\Delta(G)$ eine zulässige Triangulierung von G, $\sigma = \max_{1 \leq j \leq m} \sigma(\mathfrak{T}_j)$ durch $\sigma_\circ > 0$ nach oben beschränkt, $u_\circ \in H$, $f \in L^2(0,T;H)$ und $u \in H^1(Q)$ die Lösung von* (4.5) *mit*

$$\int_0^T \|u_t(t)\|_{H^2(G)}^2 \,\mathrm{d}t < \infty, \quad \|u\|_{L^\infty(0,T;H^2(G))} < \infty.$$

Weiter sei $V_h = \operatorname{Span}\{\varphi_1, \ldots, \varphi_N\}$ ein endlich-dimensionaler Teilraum von V, der Voraussetzung 1 erfüllt. Sei $u_h \in H^1(0,T;V_h)$ die Lösung des diskreten Problems (4.8) *mit dem Anfangswert $u_{\circ,h} \in V_h$. Dann gelten*

$$\|u - u_h\|_{L^\infty(0,T;H)} \leq \hat{c}_1 h^2 + \|\mathcal{P}_h u_\circ - u_{\circ,h}\|_H, \qquad (4.18a)$$

$$\|u - u_h\|_{L^2(0,T;V)} \leq \hat{c}_2 h + \hat{c}_3 h^2 + \|\mathcal{P}_h u_\circ - u_{\circ,h}\|_H \qquad (4.18b)$$

mit den positiven Konstanten

$$\begin{aligned}
\hat{c}_1 &= C_1 \|u\|_{L^\infty(0,T;H^2(G))} + \sqrt{c_1} \|u_t\|_{L^2(0,T;H^2(G))}, \\
\hat{c}_2 &= C_1 \|u\|_{L^2(0,T;H^2(G))}, \quad \hat{c}_3 = \sqrt{c_1} \|u_t\|_{L^2(0,T;H^2(G))}.
\end{aligned}$$

Beweis. Unter Verwendung der Dreiecksungleichung, von (4.16a) und von (4.17a) erhalten wir

$$\begin{aligned}\|u-u_h\|_{L^\infty(0,T;H)} &\le \|u-\mathcal{P}_h u\|_{L^\infty(0,T;H)} + \|\mathcal{P}_h u - u_h\|_{L^\infty(0,T;H)} \\ &\le C_1 \|u\|_{L^\infty(0,T;H^2(G))} h^2 + \sqrt{c_1}\,\|u_t\|_{L^2(0,T;H^2(G))} h^2 + \|\mathcal{P}_h u_\circ - u_{\circ,h}\|_H,\end{aligned}$$

woraus sich sofort (4.18a) ergibt. Um (4.18b) zu zeigen, müssen wir ebenfalls die Dreiecksungleichung anwenden. Wir bekommen dann mit (4.16b) und (4.17b)

$$\begin{aligned}\|u-u_h\|_{L^2(0,T;V)} &\le \|u-\mathcal{P}_h u\|_{L^2(0,T;V)} + \|\mathcal{P}_h u - u_h\|_{L^2(0,T;V)} \\ &\le C_1 \|u\|_{L^2(0,T;H^2(G))} h + \sqrt{c_1}\,\|u_t\|_{L^2(0,T;H^2(G))} h^2 + \|\mathcal{P}_h u_\circ - u_{\circ,h}\|_H,\end{aligned}$$

was (4.18b) ergibt. □

Bemerkung 4.8. Gilt $u_\circ \in V$, so ist $u_{\circ,h} = \mathcal{P}_h u_\circ$ eine gute Wahl für den Anfangswert des diskreten Problems (4.8). Dazu muss das Problem

$$\langle u_{\circ,h}, \varphi_h\rangle_V = \langle u_\circ, \varphi_h\rangle_V \quad \text{für alle } \varphi_h \in V_h \tag{4.19}$$

gelöst werden. Die Lösung von (4.19) erhalten wir durch die Berechnung der eindeutigen Lösung des linearen Gleichungssystems

$$\mathrm{S}\mathrm{u}_\circ = \mathrm{b} \quad \text{mit } \mathrm{b} = \left(\langle u_\circ, \varphi_i\rangle_V\right) \in \mathbb{R}^N$$

mit der symmetrischen und positiv-definiten Steifigkeitsmatrix S; vergleiche (3.12). ◊

4.4 Zeitdiskretisierung

Wir wollen nun eine Zeitdiskretisierung einführen für das Anfangswertproblem (4.10) einführen. Dabei können wir auf Kapitel 1 zurückgreifen, da (4.10) mit (1.1) übereinstimmt. Seien wieder $t_k = k\Delta t$, $k = 0, \ldots, M$, eine äquidistante Diskretisierung des Zeitintervalls $[0, T]$ mit Schrittweite $\Delta t = T/M$. Sei $\mathrm{u}(t) \in \mathbb{R}^N$ die (in der Regel unbekannte) Lösung von (4.10). Mit

$$\mathrm{u}^k = \begin{pmatrix} u_1^k \\ \vdots \\ u_N^k \end{pmatrix} \in \mathbb{R}^N \quad \text{für } k = 0, \ldots, M$$

bezeichnen wir eine Approximation für u am Zeitpunkt t_k für $0 \le k \le M$. In Abschnitt 1.4 haben wir verschiedene Einschrittverfahren bereits vorgestellt. Insbesondere hatten wir in (1.16) für $\theta \in [0, 1]$ das sogenannte θ-Schema eingeführt, welches hier

$$\mathrm{M}\,\frac{\mathrm{u}^k - \mathrm{u}^{k-1}}{\Delta t} + \theta \mathrm{S}\mathrm{u}^k + (1-\theta)\mathrm{S}\mathrm{u}^{k-1} = \theta \mathrm{f}^k + (1-\theta)\mathrm{f}^{k-1} \text{ für } k = 1, \ldots, M \quad \text{und} \quad \mathrm{u}^0 = \mathrm{u}_\circ \tag{4.20}$$

lautet. Für $\theta = 0$ ergibt (4.20) das explizite Euler-Verfahren für (4.10), für $\theta = 1/2$ das Crank-Nicolson-Verfahren und für $\theta = 1$ das imlizite Euler-Verfahren. Wenn wir (4.20) wieder als Variationsformulierung formulieren, so erhalten wir für die diskrete Lösung

$$u_h^k(\boldsymbol{x}) = \sum_{j=1}^N u_j^k \varphi_j(\boldsymbol{x}) \quad \text{für } \boldsymbol{x} \in \overline{G} \text{ und } k = 0, \ldots, M$$

das Problem

$$\begin{aligned}&\frac{1}{\Delta t}\langle u_h^k - u_h^{k-1}, \varphi_h\rangle_H + \langle \theta u_h^k + (1-\theta) u_h^{k-1}, \varphi_h\rangle_V \\ &\qquad = \langle \theta f(t_k) + (1-\theta) f(t_{k-1}), \varphi_h\rangle_H \quad \text{für alle } \varphi_h \in V_h \text{ und } k = 1, \ldots, M.\end{aligned} \tag{4.21}$$

Wir wollen $\theta \in [1/2, 1]$ betrachten. Ferner beschränken wir uns auf den Fall $f = 0$. Wir betrachten also statt (4.21) das vereinfachte Problem

$$\frac{1}{\Delta t} \langle u_h^k - u_h^{k-1}, \varphi_h \rangle_H + \langle \theta u_h^k + (1-\theta) u_h^{k-1}, \varphi_h \rangle_V = 0 \text{ für alle } \varphi_h \in V_h \text{ und } k = 1, \dots, M. \quad (4.22)$$

Nun wählen wir die Testfunktion $\varphi_h = \theta u_h^k + (1-\theta) u_h^{k-1} \in V_h$ und erhalten für den ersten Term in (4.22)

$$\begin{aligned} \langle u_h^k - u_h^{k-1}, \theta u_h^k + (1-\theta) u_h^{k-1} \rangle_H &= \theta \|u_h^k - u_h^{k-1}\|_H^2 + \langle u_h^k - u_h^{k-1}, u_h^{k-1} \rangle_H \\ &= \theta \|u_h^k - u_h^{k-1}\|_H^2 + \langle u_h^k, u_h^{k-1} \rangle_H - \|u_h^{k-1}\|_H^2. \end{aligned}$$

Mit

$$\langle u_h^k, u_h^{k-1} \rangle_H = \frac{1}{2} \|u_h^k\|_H^2 + \frac{1}{2} \|u_h^{k-1}\|_H^2 - \frac{1}{2} \|u_h^k - u_h^{k-1}\|_H^2$$

bekommen wir

$$\langle u_h^k - u_h^{k-1}, \theta u_h^k + (1-\theta) u_h^{k-1} \rangle_H = \left(\theta - \frac{1}{2}\right) \|u_h^k - u_h^{k-1}\|_H^2 + \frac{1}{2} \|u_h^k\|_H^2 - \frac{1}{2} \|u_h^{k-1}\|_H^2.$$

Es ergibt sich also

$$\frac{1}{2\Delta t} \left(\|u_h^k\|_H^2 - \|u_h^{k-1}\|_H^2 \right) + \frac{1}{\Delta t} \left(\theta - \frac{1}{2}\right) \|u_h^k - u_h^{k-1}\|_H^2 + \|\theta u_h^k + (1-\theta) u_h^{k-1}\|_V^2 = 0.$$

Nun summieren wir über $k = 1, \dots, \ell$ mit $\ell \leq M$. Dann folgt unter Ausnutzung der Teleskopeigenschaft und nach Multiplikation mit Δt

$$\frac{1}{2} \left(\|u_h^\ell\|_H^2 - \|u_h^0\|_H^2 \right) + \left(\theta - \frac{1}{2}\right) \sum_{k=1}^{\ell} \|u_h^k - u_h^{k-1}\|_H^2 + \Delta t \sum_{k=1}^{\ell} \|\theta u_h^k + (1-\theta) u_h^{k-1}\|_V^2 = 0,$$

woraus wir wegen $\theta \geq 0$ und $u_h^0 = u_{\circ,h}$ sofort das folgende Resultat bekommen; vergleiche [Dzi10, Satz 5.31].

Lemma 4.9 (Stabilität)**.** *Seien $0 < T < \infty$ und $G \subset \mathbb{R}^n$ ein beschränktes Gebiet, welches* (3.19) *erfüllt für eine zulässige Triangulierung $\Delta(G)$ und $f = 0$. Die Anfangsbedingung $u_{\circ,h} \in V_h$ erfülle*

$$u_{\circ,h} = \sum_{j=1}^{N} u_{\circ,i} \varphi_j \in V^h.$$

Dann erhalten wir für die Folge $\{u_h^k\}_{k=0}^M \subset V_h$, die (4.21) *und $u_h^0 = u_{h,\circ}$ für $\theta \in [1/2, 1]$ erfüllt, die Abschätzung*

$$\max_{0 \leq k \leq M} \|u_h^k\|_H \leq \|u_{\circ,h}\|_H.$$

Bemerkung 4.10. 1) An dieser Stelle erinnern wir an [DR12, Satz23.13-1)], in dem ein exponentieller Abfall der Energie von $t \mapsto \|u(t)\|_H^2$ für die kontinuierliche Lösung der Wärmeleitungsgleichung gezeigt wird.

2) Für $\theta \in [0, 1/2)$ muss zusätzlich die Bedingung

$$(1-\theta) \Delta t \lambda_h^2 \leq 1 \quad (4.23)$$

gelten, wobei die Konstante λ_h durch

$$\lambda_h = \sup \left\{ \frac{\|v_h\|_V}{\|v_h\|_H} \,\middle|\, v_h \in V_h \setminus \{0\} \right\} \quad (4.24)$$

gegeben ist. Aus (4.24) folgt die *inverse Ungleichung*

$$\|v_h\|_V \leq \lambda_h \|v_h\|_H.$$

Für einen beliebigen Vektor $\mathsf{v} = (\mathsf{v}_j) \in \mathbb{R}^N$ setzen wir

$$v_h(\boldsymbol{x}) = \sum_{j=1}^{N} \mathsf{v}_j \varphi_j(\boldsymbol{x}) \quad \text{für } \boldsymbol{x} \in \overline{G}.$$

Dann erhalten wir

$$\mathsf{v}^\top \mathsf{S}\mathsf{v} = \sum_{i=1}^{N}\sum_{j=1}^{N} \mathsf{v}_i \mathsf{v}_j \langle \varphi_i, \varphi_j \rangle_V = \|v_h\|_V^2 \leq \lambda_h^2 \|v_h\|_H^2 = \lambda_h^2 \sum_{i=1}^{N}\sum_{j=1}^{N} \mathsf{v}_i \mathsf{v}_j \langle \varphi_i, \varphi_j \rangle_H = \lambda_h^2 \mathsf{v}^\top \mathsf{M}\mathsf{v}.$$

Nun setzen wir $\mathsf{w} = \mathsf{M}^{1/2}\mathsf{v} \in \mathbb{R}^N$. Dann bekommen wir $\mathsf{v} = \mathsf{M}^{-1/2}\mathsf{w}$ und somit wegen $\mathsf{M} = \mathsf{M}^{1/2}\mathsf{M}^{1/2}$ die Ungleichung

$$\mathsf{w}^\top \big(\mathsf{M}^{-1/2}\mathsf{S}\mathsf{M}^{-1/2}\big)\mathsf{w} = \mathsf{v}^\top \mathsf{S}\mathsf{v} \leq \lambda_h^2 \mathsf{v}^\top \mathsf{M}\mathsf{v} = \lambda_h^2 |\mathsf{w}|_2^2$$

Daher ist die Zahl λ_h^2 der größte Eigenwert der symmetrischen und positive definiten Matrix $\mathsf{M}^{-1/2}\mathsf{S}\mathsf{M}^{-1/2}$.

3) Es lässt sich zeigen [Dzi10, Satz 5.31], dass die in (4.24) eingeführte Konstante die Ungleichung

$$\lambda_h \leq \frac{c}{h_{\mathsf{min}}} \quad \text{mit } h_{\mathsf{min}} = \min_{1\leq j\leq m} h(\mathcal{T}_j)$$

für eine Konstante $c > 0$ erfüllt, sofern $\sigma = \max_{1\leq j\leq m}(h(\mathcal{T}_j)/\varrho(\mathcal{T}_j))$ durch eine von h unabhängige Konstante $\sigma_\circ$ nach oben beschränkt ist. Damit lässt sich (4.23) garantieren, sofern

$$c^2(1-\theta)\Delta t \leq h_{\mathsf{min}}^2$$

gilt. Insbesondere für das explizite Euler-Verfahren ($\theta = 0$) erhalten wir die Forderung $c^2\Delta t \leq h_{\mathsf{min}}^2$ für die Zeit- und Ortsschrittweite. ◊

Nun kommen wir zum Nachweis der Konvergenz, wobei wir uns an dem Beweis von [Dzi10, Satz 5.35] orientieren. Dazu verwenden wir wider die in Definition 4.1 eingeführte Ritz-Projektion und setzen

$$e_h^k = u_h^k - \mathcal{P}_h u(t_k) = \big(u_h^k - u(t_k)\big) + \big(u(t_k) - \mathcal{P}_h u(t_k)\big) \quad \text{für } k = 0, \ldots, M.$$

Der Term $u(t_k) - \mathcal{P}_h u(t_k)$, $k = 0, \ldots, M$, lässt sich mit Hilfe von Voraussetzung 1 abschätzen. Aus (4.8), (4.1) und (4.6) erhalten wir für $\theta \in [0,1]$ und für alle $\varphi_h \in V_h$

$$\begin{aligned}
&\frac{1}{\Delta t}\langle e_h^k - e_h^{k-1}, \varphi_h\rangle_H + \langle \theta e_h^k + (1-\theta)e_h^{k-1}, \varphi_h\rangle_V \\
&\quad = \frac{1}{\Delta t}\langle u_h^k - u_h^{k-1}, \varphi_h\rangle_H - \frac{1}{\Delta t}\langle \mathcal{P}_h u(t_k) - \mathcal{P}_h u(t_{k-1}), \varphi_h\rangle_H + \langle \theta u_h^k + (1-\theta)u_h^{k-1}, \varphi_h\rangle_V \\
&\qquad - \langle \theta \mathcal{P}_h u(t_k) + (1-\theta)\mathcal{P}_h u(t_{k-1}), \varphi_h\rangle_V \\
&\quad = \langle \theta f(t_k) + (1-\theta) f(t_{k-1}), \varphi_h\rangle_H - \frac{1}{\Delta t}\langle \mathcal{P}_h u(t_k) - \mathcal{P}_h u(t_{k-1}), \varphi_h\rangle_H \\
&\qquad - \langle \mathcal{P}_h(\theta u(t_k) + (1-\theta)u(t_{k-1})), \varphi_h\rangle_V \\
&\quad = \theta\big(\langle f(t_k), \varphi_h\rangle_H - \langle u(t_k), \varphi_h\rangle_V\big) + (1-\theta)\big(\langle f(t_{k-1}), \varphi_h\rangle_H - \langle u(t_{k-1}), \varphi_h\rangle_V\big) \\
&\qquad - \frac{1}{\Delta t}\langle \mathcal{P}_h u(t_k) - \mathcal{P}_h u(t_{k-1}), \varphi_h\rangle_H \\
&\quad = \theta\langle u_t(t_k), \varphi_h\rangle_H + (1-\theta)\langle u_t(t_{k-1}), \varphi_h\rangle_H - \frac{1}{\Delta t}\langle \mathcal{P}_h u(t_k) - \mathcal{P}_h u(t_{k-1}), \varphi_h\rangle_H.
\end{aligned}$$

Auf der rechten Seite stehen nur Terme mit der kontinuierlichen Lösung u. Wir setzen

$$G_h^k = \theta u_t(t_k) + (1-\theta)u_t(t_{k-1}) - \frac{u(t_k) - u(t_{k-1})}{\Delta t} \in H,$$
$$F_h^k = \frac{1}{\Delta t}\left(\big(u(t_k) - \mathcal{P}_h u(t_k)\big) - \big(u(t_{k-1}) - \mathcal{P}_h u(t_{k-1})\big)\right) \in H.$$

Dann erhalten wir

$$\frac{1}{\Delta t}\langle e_h^k - e_h^{k-1}, \varphi_h\rangle_H + \langle \theta e_h^k + (1-\theta)e_h^{k-1}, \varphi_h\rangle_V = \langle G_h^k + F_h^k, \varphi_h\rangle_H \quad \text{für alle } \varphi_h \in V_h. \tag{4.25}$$

Aus der Taylorformel mit dem integralen Restglied [DR11, Abschnitt 9.3] erhalten wir für eine Funktion $v \in C^3([0,T])$ die Darstellung

$$v(s) = v(t) + (s-t)\dot{v}(t) + \frac{1}{2}(s-t)^2\ddot{v}(t) + \frac{1}{2}\int_t^s (s-\tau)^2 v^{(3)}(\tau)\,\mathrm{d}\tau \quad \text{für alle } s,t \in [0,T].$$

Für $s = t_k$, $t = t_{k-1}$, $s - t = \Delta t$ beziehungsweise für $s = t_{k-1}$, $t = t_k$, $s - t = -\Delta t$ bekomen wir

$$v(t_k) = v(t_{k-1}) + \dot{v}(t_{k-1})\Delta t + \ddot{v}(t_{k-1})\frac{\Delta t^2}{2} + \frac{1}{2}\int_{t_{k-1}}^{t_k} (t_k - \tau)^2 v^{(3)}(\tau)\,\mathrm{d}\tau \tag{4.26a}$$

$$v(t_{k-1}) = v(t_k) - \dot{v}(t_k)\Delta t + \ddot{v}(t_k)\frac{\Delta t^2}{2} - \frac{1}{2}\int_{t_{k-1}}^{t_k} (t_{k-1} - \tau)^2 v^{(3)}(\tau)\,\mathrm{d}\tau \tag{4.26b}$$

für $k \in \{1,\dots,M\}$. Aus (4.26a) folgen

$$\frac{v(t_k) - v(t_{k-1})}{\Delta t} = \dot{v}(t_{k-1}) + \ddot{v}(t_{k-1})\frac{\Delta t}{2} + \frac{1}{2\Delta t}\int_{t_{k-1}}^{t_k} (t_k - \tau)^2 v^{(3)}(\tau)\,\mathrm{d}\tau$$

und daher

$$\dot{v}(t_{k-1}) - \frac{v(t_k) - v(t_{k-1})}{\Delta t} = -\ddot{v}(t_{k-1})\frac{\Delta t}{2} - \frac{1}{2\Delta t}\int_{t_{k-1}}^{t_k} (t_k - \tau)^2 v^{(3)}(\tau)\,\mathrm{d}\tau \tag{4.27}$$

für $k \in \{1,\dots,M\}$. Mit (4.26b) schließen wir

$$\frac{v(t_k) - v(t_{k-1})}{\Delta t} = \dot{v}(t_k) - \ddot{v}(t_k)\frac{\Delta t}{2} + \frac{1}{2\Delta t}\int_{t_{k-1}}^{t_k} (t_{k-1} - \tau)^2 v^{(3)}(\tau)\,\mathrm{d}\tau$$

und deshalb

$$\dot{v}(t_k) - \frac{v(t_k) - v(t_{k-1})}{\Delta t} = \ddot{v}(t_k)\frac{\Delta t}{2} - \frac{1}{2\Delta t}\int_{t_{k-1}}^{t_k} (t_{k-1} - \tau)^2 v^{(3)}(\tau)\,\mathrm{d}\tau \tag{4.28}$$

für $k \in \{1,\dots,M\}$. Wir multiplizieren (4.27) mit $1-\theta$ und (4.28) mit θ. Dann ergibt sich

$$\begin{aligned}(1-\theta)\dot{v}(t_{k-1}) + \theta\dot{v}(t_k) - \frac{v(t_k) - v(t_{k-1})}{\Delta t}\\ = -\frac{\Delta t}{2}\big((1-\theta)\ddot{v}(t_{k-1}) - \theta\ddot{v}(t_k)\big) - \frac{1}{2\Delta t}\int_{t_{k-1}}^{t_k}\big((1-\theta)(t_k-\tau)^2 + \theta(t_{k-1}-\tau)^2\big)v^{(3)}(\tau)\,\mathrm{d}\tau\end{aligned}$$

für $k \in \{1,\dots,M\}$. Da $v \in C^3([0,T])$ erfüllt ist, erhalten wir

$$\ddot{v}(t_{k-1}) = \ddot{v}(t_k) - \big(\ddot{v}(t_k) - \ddot{v}(t_{k-1})\big) = \ddot{v}(t_k) - \int_{t_{k-1}}^{t_k} v^{(3)}(\tau)\,\mathrm{d}\tau$$

und daher

$$-\frac{\Delta t}{2}(1-\theta)\ddot{v}(t_{k-1}) = -\frac{\Delta t}{2}(1-\theta)\ddot{v}(t_k) + \frac{\Delta t}{2}(1-\theta)\int_{t_{k-1}}^{t_k} v^{(3)}(\tau)\,\mathrm{d}\tau$$

für $k \in \{1, \dots, M\}$. Somit haben wir

$$\begin{aligned}
&(1-\theta)\dot{v}(t_{k-1}) + \theta\dot{v}(t_k) - \frac{v(t_k) - v(t_{k-1})}{\Delta t} \\
&\quad = -\frac{\Delta t}{2}(1-2\theta)\,\ddot{v}(t_k) + \frac{\Delta t}{2}(1-\theta)\int_{t_{k-1}}^{t_k} v^{(3)}(\tau)\,\mathrm{d}\tau \\
&\qquad - \frac{1}{2\Delta t}\int_{t_{k-1}}^{t_k} \big((1-\theta)(t_k-\tau)^2 + \theta(t_{k-1}-\tau)^2\big) v^{(3)}(\tau)\,\mathrm{d}\tau
\end{aligned}$$

für $k \in \{1, \dots, M\}$. Für $\tau \in [t_{k-1}, t_k]$ gilt

$$\big|(1-\theta)(t_k-\tau)^2 + \theta(t_{k-1}-\tau)^2\big| \leq (1-\theta)\Delta t^2 + \theta\Delta t^2 = \Delta t^2.$$

Mit der Wahl $v(t) = u(t,\cdot)$ erhalten wir daher mit der Cauchy-Schwarz-Ungleichung [DR12, Satz 12.17]

$$\begin{aligned}
\int_{t_{k-1}}^{t_k} \|u_{ttt}(\tau)\|_H\,\mathrm{d}\tau &= \int_{t_{k-1}}^{t_k} 1 \cdot \|u_{ttt}(\tau)\|_H\,\mathrm{d}\tau = \langle 1, \|u_{ttt}(\cdot)\|_H\rangle_{L^2(t_{k-1},t_k)} \\
&\leq \|1\|_{L^2(t_{k-1},t_k)} \big\|\|u_{ttt}(\cdot)\|_H\big\|_{L^2(t_{k-1},t_k)} \\
&= \left(\int_{t_{k-1}}^{t_k} 1\,\mathrm{d}\tau\right)^{1/2} \left(\int_{t_{k-1}}^{t_k} \|u_{ttt}(\tau)\|_H^2\,\mathrm{d}\tau\right)^{1/2} = \Delta t^{1/2}\,\|u_{ttt}\|_{L^2(t_{k-1},t_k;H)}
\end{aligned}$$

die Abschätzung

$$\begin{aligned}
\|G_h^k\|_H &= \left\|(1-\theta)u_t(t_{k-1}) - \theta u_t(t_k) - \frac{u(t_k)-u(t_{k-1})}{\Delta t}\right\|_H \\
&\leq \frac{\Delta t}{2}|1-2\theta|\,\|u_{tt}(t_k)\|_H + \frac{\Delta t}{2}(1-\theta)\int_{t_{k-1}}^{t_k} \|u_{ttt}(\tau)\|_H\,\mathrm{d}\tau + \frac{\Delta t}{2}\int_{t_{k-1}}^{t_k} \|u_{ttt}(\tau)\|_H\,\mathrm{d}\tau \\
&= \frac{\Delta t}{2}|1-2\theta|\,\|u_{tt}(t_k)\|_H + \Delta t\int_{t_{k-1}}^{t_k} \|u_{ttt}(\tau)\|_H\,\mathrm{d}\tau \\
&\leq \frac{\Delta t}{2}|1-2\theta|\,\|u_{tt}(t_k)\|_H + \Delta t^{3/2}\,\|u_{ttt}\|_{L^2(t_{k-1},t_k;H)}
\end{aligned}$$

für $k \in \{1, \dots, M\}$. Für F_h^k, $k = 1, \dots, M$, erhalten wir mit Voraussetzung 1 die Abschätzung

$$\begin{aligned}
\|F_h^k\|_H^2 &= \left\|\frac{1}{\Delta t}\Big(\big(u(t_k) - \mathcal{P}_h u(t_k)\big) - \big(u(t_{k-1}) - \mathcal{P}_h u(t_{k-1})\big)\Big)\right\|_H \\
&= \frac{1}{\Delta t}\,\|(u(t_k) - u(t_{k-1}) - \mathcal{P}_h(u(t_k) - u(t_{k-1}))\|_H \leq \frac{C_1 h^2}{\Delta t}\,\|u(t_k) - u(t_{k-1})\|_{H^2(G)} \\
&= \frac{C_1 h^2}{\Delta t}\int_{t_{k-1}}^{t_k} \frac{\mathrm{d}}{\mathrm{d}t}\,\|u(\tau)\|_{H^2(G)}\,\mathrm{d}\tau = \frac{C_1 h^2}{\Delta t}\int_{t_{k-1}}^{t_k} \|u_t(\tau)\|_{H^2(G)}\,\mathrm{d}\tau \\
&\leq \frac{C_1 h^2}{\Delta t^{1/2}}\left(\int_{t_{k-1}}^{t_k} \|u_t(\tau)\|_{H^2(G)}^2\,\mathrm{d}\tau\right)^{1/2} = \frac{C_1 h^2}{\Delta t^{1/2}}\,\|u_t\|_{L^2(t_{k-1},t_k;H^2(G))}
\end{aligned}$$

für $k \in \{1, \dots, M\}$. Wir wählen nun in (4.25) als Testfunktion $\varphi_h = \theta e_h^k + (1-\theta)e_h^{k-1}$. Dann folgt mit der Cauchy-Schwarz-Ungleichung [DR12, Satz 12.17]

$$\begin{aligned}
&\frac{1}{\Delta t}\,\langle e_h^k - e_h^{k-1}, \theta e_h^k + (1-\theta)e_h^{k-1}\rangle_H + \|\theta e_h^k + (1-\theta)e_h^{k-1}\|_V^2 \\
&\quad \leq \big(\|F_h^k\|_H + \|G_h^k\|_H\big)\|\theta e_h^k + (1-\theta)e_h^{k-1}\|_H \\
&\quad \leq \left(\frac{C_1 h^2}{\Delta t^{1/2}}\,\|u_t\|_{L^2(t_{k-1},t_k;H^2(G))} + \frac{\Delta t}{2}|1-2\theta|\,\|u_{tt}(t_k)\|_H + \Delta t^{3/2}\,\|u_{ttt}\|_{L^2(t_{k-1},t_k;H)}\right) \\
&\qquad \cdot \|\theta e_h^k + (1-\theta)e_h^{k-1}\|_H
\end{aligned}$$

für $k \in \{1, \dots, M\}$. Nun folgt das folgende Resultat unter Verwendung der Poincaré- [DR12, Satz 16.23] und der Young-Ungleichung (1.10) sowie einer Aufsummierung über k. Für die Details verweisen wir auf [Dzi10, Satz 3.35].

Satz 4.11 (Konvergenz)**.** *Seien $0 < T < \infty$ und $G \subset \mathbb{R}^n$ ein beschränktes Gebiet, welches* (3.19) *erfüllt für eine zulässige Triangulierung $\Delta(G)$ und $V_H = \operatorname{Span}\{\varphi_1, \dots, \varphi_N\}$ mit linear unabhängigen Funktionen $\{\varphi_i\}_{i=1}^N \subset V$, so dass Voraussetzung 1 gilt. Ferner setzen wir $u_\circ \in V$, $f \in L^2(0,T;H)$ und $u_{\circ,h} \in V_h$ voraus. Für $\theta \in [0,1]$ seien $u \in H^1(Q)$ die Lösung von* (4.6) *und $\{u_h^k\}_{k=0}^M$ die von* (4.8)*. Im Fall $\theta < 1/2$ setzen wir zusätzlich noch* (4.23) *voraus. Dann existiert ein $c_0 > 0$ mit*

$$\max_{1\le k\le M} \|u(t_k) - u_h^k\|_H \le c_0\big(|2\theta-1|c_1\Delta t + c_2\Delta t^2 + c_3 h^2\big) + \|\mathcal{P}_h u_\circ - u_{\circ,h}\|_H,$$

$$\left(\Delta t \sum_{k=1}^{M} \|u(t_k) - u_h^k\|_V^2\right)^{1/2} \le c_0\big(|2\theta-1|c_1\Delta t + c_2\Delta t^2 + c_3 h^2 + c_4 h\big) + \|\mathcal{P}_h u_\circ - u_{\circ,h}\|_H,$$

sofern die Konstanten

$$c_1 = \sup\big\{\|u_{tt}(t)\|_H \,\big|\, t \in [0,T]\big\}, \qquad c_2 = \|u_{ttt}\|_{L^2(0,T;H)},$$
$$c_3 = \|u_t\|_{L^2(0,T;H^2(G))}, \qquad c_4 = \sup\big\{\|u_t(t)\|_{H^2(G)} \,\big|\, t \in [0,T]\big\}$$

beschränkt sind.

Literaturverzeichnis

[AU10] W. Arendt und K. Urban. *Partielle Differenzialgleichungen*. Spektrum Akademischer Verlag, Heidelberg, 2010.

[Bra07] D. Braess. *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer-Lehrbuch Masterclass, 2007.

[BS08] S.C. Brenner und L.R. Scott. *The Mathematical Theory of Finite Element Methods*. Texts in Applied Mathematics, 3rd edition, Springer, 2008.

[DR11] R. Denk und R. Racke. *Kompendium der Analysis. Band 1: Differential- und Integralrechnung, Gewöhnliche Differentialgleichungen*. Vieweg+Teubner Verlag, Springer Fachmedien Wiesbaden GmbH, 2011.

[DR12] R. Denk und R. Racke. *Kompendium der Analysis. Band 2: Maß- und Integrationstheorie, Funktionentheorie, Funktionalanalysis, Partielle Differentialgleichungen*. Vieweg+Teubner Verlag, Springer Fachmedien Wiesbaden GmbH, 2012.

[Dzi10] G. Dziuk. *Theorie und Numerik partieller Differentialgleichungen*. De Walter de Gruyter GmbH & Co. KG, Berlin, 2010.

[Kel99] C.T. Kelley. *Iterative Methods for Optimization*. Frontiers in Applied Mathematics, SIAM, Philadelphia, PA, 1999.

[KA00] P. Knabner und L. Angermann. *Numerik partieller Differentialgleichungen*. Springer-Lehrbuch, 2000.

[Lui10] E. Luik. Numerik I. Skript zur Vorlesung *Numerik I*, Wintersemester 2010/11, 2010.

[Str04] J.C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations*. Second edition, SIAM, Philadelphia, PA, 2004.

[Tho97] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems*. Springer Series in Computational Mathematics, Springer, 1997.

Stichwortverzeichnis

Zentraler Differenzenquotient 2. Ordnung, 4