Kapitel 1

Quadratur in 1D

Gegeben seien Intervallgrenzen $a, b \in \mathbb{R}$ mit $a < b$ sowie eine integrierbare Gewichtsfunktion $\omega \in C(a,b) \cap L^1(a,b)$ mit $\omega > 0$. Beispiele für solche Gewichtsfunktionen sind $\omega \equiv 1$, $\omega(x) = (1-x^2)^{-1/2}$, $\omega(x) = |\log(x)|$, $\omega(x) = \exp(x)$ etc.

Ziel ist für gegebenes $f \in C[a,b] := \{f : [a,b] \to \mathbb{R} \text{ stetig}\}$ die numerische Approximation von

$$Qf := \int_a^b f\omega \, dx \tag{1.1}$$

durch eine sogenannte **(n-Punkt) Quadraturformel**

$$Q_n f := \sum_{j=0}^{n} \omega_j f(x_j). \tag{1.2}$$

Hierbei sind $a \le x_0 < \cdots < x_n \le b$ gegebene **Stützstellen** und $\omega_0, \ldots, \omega_n \in \mathbb{R}$ gegebene **Gewichte**. Die Zahl $n \in \mathbb{N}$ bezeichnet man als **Länge** oder **Ordnung** der Quadraturformel. Aufgrund der Definition ergeben sich sofort zwei natürliche Fragen:

- Was ist eine günstige Wahl für die Stützstellen und die Gewichte?
- Unter welchen Bedingungen gilt Konvergenz $Q_n f \to Qf$ für $n \to \infty$?

Diese Fragen werden im Folgenden weitestgehend konstruktiv und elementar beantwortet.

1.1 Konstruktion

Da nach Satz von Weierstraß die Polynome

$$\mathbb{P} := \bigcup_{k=0}^{\infty} \mathbb{P}_k \quad \text{mit} \quad \mathbb{P}_k := \{p : [a,b] \to \mathbb{R} \text{ Polynom vom Grad } \le k\} \tag{1.3}$$

dicht liegen in $C[a,b]$, ist es ein natürlicher Ansatz zu fordern, dass zumindest (gewisse) Polynome p durch Q_n exakt integriert werden.

Definition 1 *Es sei Q_n eine Quadraturformel der Länge n. Existiert eine natürliche Zahl $k \in \mathbb{N}_0$ mit*

$$Q_n p = Qp \quad \textit{für alle } p \in \mathbb{P}_k, \tag{1.4}$$

*so bezeichnet man k als **Exaktheitsgrad** von Q_n.*

Bemerkung 2 *Für den Exaktheitsgrad einer Quadraturformel Q_n der Länge n gilt $k \leq 2n+1$. Um dies zu sehen, müssen wir lediglich ein Polynom $p \in \mathbb{P}_{2n+2}$ angeben, für das $Q_n p \neq Qp$ gilt. Wir definieren*

$$p(x) := \prod_{k=0}^{n} (x - x_k)^2$$

Dann gelten $p \in \mathbb{P}_{2n+2}$ und $p > 0$ fast überall. Aus der Maßtheorie wissen wir, dass $p\omega > 0$ fast überall auch $Qp > 0$ impliziert. Auf der anderen Seite gilt nach Definition aber $Q_n p = 0$, denn alle Auswertungspunkte von p sind Nullstellen von p.

Freiwillige Übung. Beweisen Sie mit Mittel der elementaren Analysis, dass für eine stetige Gewichtsfunktion $\omega \in C[a,b]$ mit $\omega = 0$ nur an endlich vielen Punkten und das konstruierte Polynom p gilt $Qp > 0$, ohne die Maßtheorie zu bemühen. ■

Bemerkung 3 *In der Regel werden Quadraturformeln auf Referenzintervallen gegeben, z.B. $[a,b] = [0,1]$ oder $[a,b] = [-1,1]$. Für allgemeine Intervalle $[a,b]$ erhält man dann Quadraturformeln mittels affiner Transformation, z.B. sind*

$$\Phi : [0,1] \to [a,b], \quad \Phi(t) = a + t(b-a) \tag{1.5}$$

oder

$$\Phi : [-1,1] \to [a,b], \quad \Phi(t) = \frac{1}{2}\big(a + b + t(b-a)\big) \tag{1.6}$$

affine Bijektionen von den Referenzintervallen auf $[a,b]$. Beispielsweise sei Q_n^{ref} eine Quadraturformel zur numerischen Approximation von $Q^{\mathrm{ref}} f := \int_{-1}^{1} f\widetilde{\omega}\, dt$. Mittels Kettenregel und Definition $\omega(x) = \widetilde{\omega}(\Phi^{-1}(x))$ erhält man nun

$$\begin{aligned} Qf = \int_a^b f\omega\, dx = \int_{-1}^{1} f(\Phi(t))\omega(\Phi(t))\Phi'(t)\, dt &= \frac{b-a}{2} \int_{-1}^{1} f(\Phi(t))\widetilde{\omega}(t)\, dt \\ &= \frac{b-a}{2}\, Q^{\mathrm{ref}}(f \circ \Phi). \end{aligned}$$

In Konsequenz erhalten wir durch

$$\frac{b-a}{2}\, Q_n^{\mathrm{ref}}(f \circ \Phi) = \frac{b-a}{2} \sum_{j=0}^{n} \widetilde{\omega}_j f(\Phi(\widetilde{x}_j)) =: \sum_{j=0}^{n} \omega_j f(x_j) = Q_n f$$

mit $x_j = \Phi(\widetilde{x}_j)$ und $\omega_j = \frac{b-a}{2}\,\widetilde{\omega}_j$ eine Quadraturformel zur Approximation von Qf. Da Φ affin ist, ist für $p \in \mathbb{P}$ die Verkettung $p \circ \Phi$ ebenfalls ein Polynom gleichen Grades. Konsequenterweise haben Q_n und Q_n^{ref} denselben Exaktheitsgrad. — Man beachte allerdings, dass die Gewichtsfunktion $\widetilde{\omega} = \omega \circ \Phi$ für Q und Q^{ref} i.a. verschieden ist (Ausnahme $\omega = 1 = \widetilde{\omega}$).

Im Folgenden bezeichnet I_n den **(Lagrange-) Interpolationsoperator**, d.h.

$$I_n : C[a,b] \to \mathbb{P}_n, \quad I_n f := \sum_{j=0}^{n} f(x_j) L_j \tag{1.7}$$

Dabei bezeichnen $x_0, \ldots, x_n$ die gegebenen Stützstellen und $L_0, \ldots, L_n$ die zugehörigen **Lagrange-Polynome**

$$L_k(x) = \prod_{\substack{j=0 \\ j \neq k}}^{n} \frac{x - x_j}{x_k - x_j}. \tag{1.8}$$

Man beachte, dass die L_j wegen der Kronecker-Eigenschaft $L_j(x_k) = \delta_{jk}$ linear unabhängig sind. Insbesondere ist $\{L_0, \dots L_n\}$ also eine Basis von $\mathbb{P}_n$. Ferner ist $I_n f \in \mathbb{P}_n$ das eindeutige Interpolationspolynom mit

$$I_n f(x_j) = f(x_j) \quad \text{für alle } j = 0, \dots, n. \tag{1.9}$$

Das folgende Lemma zeigt, dass wir de facto keine Freiheit bei der Wahl der Gewichte haben, um einen hohen Exaktheitsgrad zu erreichen.

Lemma 4 *Die folgenden beiden Aussagen sind äquivalent:*

(i) *Der Exaktheitsgrad k von Q_n erfüllt $k \geq n$.*

(ii) *Es gilt $Q_n f = Q(I_n f)$ für alle $f \in C[a,b]$.*

In diesem Fall sind die Gewichte gegeben durch $\omega_j = Q(L_j) = \int_a^b L_j \omega \, dx$.

Beweis. (i)⇒(ii): Wegen $L_i \in \mathbb{P}_n$ und Exaktheitsgrad $k \geq n$ gilt

$$\int_a^b L_i \omega \, dx = Q(L_i) = Q_n(L_i) = \sum_{j=0}^n \omega_j L_i(x_j) = \omega_i.$$

Dies zeigt insbesondere die Formel für die Gewichte ω_i. Insbesondere folgt für $I_n f = \sum_{j=0}^n f(x_j) L_j$

$$Q(I_n f) = \sum_{j=0}^n f(x_j) Q(L_j) = \sum_{j=0}^n f(x_j) \omega_j = Q_n f.$$

(ii)⇒(i): Für $f = L_j \in C[a,b]$ folgt nach Voraussetzung $Q_n(L_j) = Q(I_n L_j) = Q(L_j)$, da I_n eine Projektion ist, d.h. $I_n p = p$ für alle $p \in \mathbb{P}_n$. Insbesondere stimmen die linearen Funktionale Q_n und Q auf der Basis $\{L_0, \dots, L_n\}$ von $\mathbb{P}_n$ überein. Nach Linearer Algebra folgt deshalb $Q_n p = Q p$ für alle $p \in \mathbb{P}_n$, d.h. der Exaktheitsgrad erfüllt $k \geq n$. ■

Bisher waren die Stützstellen $x_0 < \dots < x_n$ frei und die Gewichte ω_j festgelegt durch die Forderung, dass der Exaktheitsgrad maximal sei.

Jetzt optimieren wir zusätzlich die Stützstellen, sodass der Exaktheitsgrad maximal $k = 2n+1$ wird. Dies führt auf die sogenannte Gauss-Quadratur.

Wir skizzieren im Folgenden die Konstruktion der Gauss-Quadratur:

- $(p,q)_\omega := \int_a^b pq\omega \, dx$ definiert ein Skalarprodukt auf dem Polynomraum $\mathbb{P}$.
- Man wendet Gram-Schmidt-Orthogonalisierung auf die Monombasis $\{1, x, x^2, x^3, \dots\}$ an und erhält die sogenannten **Orthogonalpolynome** $\{p_0, p_1, p_2, \dots\}$.
- Der wesentliche Beweisschritt ist nun, dass die Nullstellen $x_0, \dots, x_n$ eines Orthogonalpolynoms p_{n+1} alle einfach sind und im offenen Intervall (a,b) liegen.
- Man wählt als Stützstellen die (sortierten) Nullstellen $a < x_0 < \dots < x_n < b$ von p_{n+1} und definiert (notwendig!) $\omega_j = \int_a^b L_j \omega \, dx$. Dies liefert eine Quadraturformel Q_n, die (mindestens) exakt ist auf $\mathbb{P}_n$
- Für ein Polynom $p \in \mathbb{P}_{2n+1}$ verwenden wir Polynomdivision, um $p = \alpha p_{n+1} + \beta$ mit $\alpha, \beta \in \mathbb{P}_n$ zu schreiben. Nun beobachtet man

  $$Qp = (p,1)_\omega = (\alpha, p_{n+1})_\omega + (\beta, 1)_\omega = Q\beta$$

  aufgrund der Orthogonalität von p_{n+1} auf $\mathbb{P}_n$ sowie

  $$Q_n p = \sum_{j=0}^n \omega_j \big(\alpha(x_j) p_{n+1}(x_j) + \beta(x_j)\big) = Q_n \beta$$

  aufgrund der Wahl der x_j als Nullstellen von p_{n+1}. Wegen $Q_n \beta = Q\beta$ folgt abschließend die Exaktheit von Q_n auf $\mathbb{P}_{2n+1}$.

- Ferner erfüllen die Gewichte

$$\omega_j = Q(L_j) = Q_n(L_j) = Q_n(L_j^2) = Q(L_j^2) = \int_a^b L_j^2 \omega \, dx > 0,$$

  wobei wir $L_j(x_k) = \delta_{jk}$ sowie $L_j^2 \in \mathbb{P}_{2n}$ ausgenutzt haben.

Mit ähnlichen Tricks zeigt man, dass Q_n sogar eindeutig ist. Insgesamt gilt der folgende Satz.

Satz 5 *Es existiert eine eindeutige Quadraturformel Q_n von maximalen Exaktheitsgrad $k = 2n + 1$. Diese wird wie oben konstruiert, und es gilt für alle Gewichte insbesondere $\omega_j > 0$.* ■

1.2 Konvergenz

Satz 6 *Es sei Q_n eine Folge von Quadraturformeln $Q_n f = \sum_{j=0}^n \omega_j^{(n)} f(x_j^{(n)})$. Dann sind die folgenden Aussagen äquivalent:*

(i) *Es gilt Konvergenz $\lim_{n\to\infty} Q_n f = Qf$ für alle $f \in C[a, b]$.*

(ii) *Es gilt Konvergenz $\lim_{n\to\infty} Q_n p = Qp$ für alle Polynome $p \in \mathbb{P}_n$, und*

$$M := \sup_{n\in\mathbb{N}} \sum_{j=0}^n |\omega_j^{(n)}| < \infty. \tag{1.10}$$

Insbesondere gelten (i)–(ii), sofern Q_n die Gauss-Quadraturformel der Länge n ist.

Beweis. Wir zeigen zunächst (ii) ⇒ (i), weil dies die praktische relevante Richtung ist. Sei $\varepsilon > 0$. Es ist Folgendes zu zeigen:

$$\exists n_0 \in \mathbb{N} \forall n \geq n_0 : \quad |Qf - Q_n f| \leq \varepsilon.$$

Nach Satz von Weierstraß existiert ein Polynom $p \in \mathbb{P}_n$ mit $\|f-p\|_\infty \leq (C/2)\,\varepsilon$ mit $C := \|\omega\|_{L^1(a,b)}+M > 0$. Für dieses existiert nach (ii) ein $n_0 \in \mathbb{N}$ mit

$$\forall n \geq n_0 : \quad |Qp - Q_n p| \leq \varepsilon/2.$$

Sei $n \geq n_0$. Die Dreiecksungleichung zeigt

$$|Qf - Q_n f| \leq |(Qf - Q_n f) - Qp - Q_n p| + |Qp - Q_n p| \leq |Q(f - p)| + |Q_n(f - p)| + \varepsilon/2.$$

Für den ersten Summanden gilt nach Hölder-Ungleichung

$$|Q(f - p)| = \Big| \int_a^b (f - p)\omega \, dx \leq \|\omega\|_{L^1(a,b)} \|f - p\|_\infty.$$

Für den zweiten Summanden gilt

$$|Q_n(f - p)| \leq \sum_{j=0}^n |\omega_j^{(n)}||(f - p)(x_j)| \leq M \, \|f - p\|_\infty.$$

Insgesamt folgt also mit $|Qf - Q_n f| \leq (\|\omega\|_{L^1(a,b)} + M) \, \|f - p\|_\infty + \varepsilon/2 = \varepsilon$ die Behauptung.

Wir zeigen nun (i) ⇒ (ii), um zu sehen, dass insbesondere die Voraussetzung (1.10) nicht nur hinreichend, sondern auch notwendig ist. Dazu erinnern wir an den **Satz von Banach-Steinhaus**.

Seien X ein Banach-Raum und Y ein normierter Raum. Sei $\mathcal{A} \subseteq L(X,Y)$ eine Menge von linearen stetigen Operatoren. Dann ist gleichmäßige Beschränktheit

$$\sup_{A\in\mathcal{A}} \|A\| < \infty \quad \text{der Operatornorm } \|A\| := \sup_{x\in X\setminus\{0\}} \frac{\|Ax\|_Y}{\|X\|_X}$$

äquivalent zur punktweisen Beschränkheit

$$\forall x \in X: \quad \sup_{A\in\mathcal{A}} \|Ax\|_Y < \infty.$$

Wir wenden diesen Satz mit $X = C[a,b]$, $Y = \mathbb{R}$ und $\mathcal{A} = \{Q_n \,|\, n \in \mathbb{N}\}$ an. Nach (i) konvergiert Q_n punktweise gegen Q. Deshalb ist $\mathcal{A}$ punktweise beschränkt und somit sogar gleichmäßig beschränkt. Wir definieren

$$M := \sup_{n\in\mathbb{N}} \|Q_n\|,$$

und für die lineare Abbildung Q_n ist nur noch die Operatornorm als $\|Q_n\| = \sum_{j=0}^n |\omega_j^{(n)}|$ zu identifizieren: Für $f \in C[a,b]$ gilt

$$|Q_n f| \le \sum_{j=0}^n |\omega_j^{(n)}||f(x_j^{(n)}| \le \|f\|_\infty \sum_{j=0}^n |\omega_j^{(n)}|.$$

Dies zeigt $\|Q_n\| \le \sum_{j=0}^n |\omega_j^{(n)}||$ nach Definition der Operatornorm. Um die Gleichheit zu zeigen, müssen wir eine geeignete Funktion $s \in C[a,b]$ konstruieren:

- Sei $x_{-1}^{(n)} := a$ und $x_{n+1}^{(n)} := b$, d.h. $a = x_{-1}^{(n)} \le x_0^{(n)} < x_1^{(n)} < \cdots < x_n^{(n)} \le x_{n+1}^{(n)}$.
- Wir wählen s als stückweise affinen Streckenzug (affiner Spline) mit den Knotenwerten $s(x_{-1}^{(n)}) := s(x_0^{(n)})$, $s(x_{n+1}^{(n)}) := s(x_n^{(n)})$ und $s(x_j^{(n)}) := \operatorname{sign}(\omega_j^{(n)})$ für alle $j = 0,\dots,n$.

Dann gilt $|s(x)| \le 1$ für alle $x \in [a,b]$ und $|s(x_j^{(n)})| = 1$, d.h. $\|s\|_\infty = 1$. Ferner gilt

$$|Q_n s| = \Big|\sum_{j=0}^n \omega_j^{(n)} s(x_j^{(n)})\Big| = \sum_{j=0}^n |\omega_j^{(n)}| = \|s\|_\infty \sum_{j=0}^n |\omega_j^{(n)}|$$

Nach Definition der Operatornorm folgt also $\|Q_n\| \ge \sum_{j=0}^n |\omega_j^{(n)}|$ und damit Gleichheit. Dies zeigt (1.10), und die zweite Bedingung in (ii) ist wegen $\mathbb{P}_n \subseteq C[a,b]$ trivial.

Für den Exaktheitsgrad der Gauss-Quadratur Q_n gilt nach Konstruktion $k_n = 2n+1$, und Konvergenz $\lim_n Q_n p = Qp$ für alle $p \in \mathbb{P}_n$ ist deshalb trivial. Ferner gilt in diesem Fall $\omega_j^{(n)} > 0$, sodass

$$\sum_{j=0}^n |\omega_j^{(n)}| = \sum_{j=0}^n \omega_j^{(n)} = Q_n(1) = Q(1) = \int_a^b \omega\,dx = \|\omega\|_{L^1(a,b)},$$

und es folgt auch (1.10). ■

Bemerkung 7 *Die Konvergenzannahme $Q_n p \to Q_p$ für alle Polynome $p \in \mathbb{P}$ ist in der Praxis leicht dadurch zu erfüllen, dass man die Gewichte ω_j geeignet wählt, siehe Lemma 4. Die Bedingung* (1.10) *ist allerdings kritisch und für fast alle Quadraturformeln (mit Ausnahme der Gauss-Quadraturen) verletzt.*

Um eine Aussage über die Konvergenzgeschwindigkeit der Quadratur zu bekommen, leiten wir analog zum Beweis von Satz 6 eine Fehlerabschätzung her, die den Quadraturfehler auf einen Bestapproximationsfehler für polynomiale Approximation von f zurückführt.

Lemma 8 *Für eine Quadraturformel Q_n vom Exaktheisgrad $k \geq 0$ gilt die Fehlerabschätzung*

$$|Qf - Q_n f| \leq \Big(\|\omega\|_{L^1(a,b)} + \sum_{j=0}^{n} |\omega_j| \Big) \min_{p \in \mathbb{P}_k} \|f - p\|_\infty. \tag{1.11}$$

Beweis. Sei $p \in \mathbb{P}_k$, also $Q_n p = Qp$. Mit der Dreiecksungleichung folgt

$$\begin{aligned} |Qf - Q_n f| \leq |Qf - Qp| + |Q_n f - Q_n p| &= \Big| \int_a^b (f-p)\omega \, dx \Big| + \Big| \sum_{j=0}^{n} \omega_j \big(f(x_j) - p(x_j) \big) \Big| \\ &\leq \Big(\|\omega\|_{L^1(a,b)} + \sum_{j=0}^{n} |\omega_j| \Big) \|f - p\|_\infty, \end{aligned}$$

wobei im letzten Schritt die Hölder-Ungleichung verwendet wurde. Dies zeigt die Abschätzung (1.11) mit dem Infimum statt des Minimums auf der rechten Seite. Der Leser mag sich überlegen, dass das Infimum tatsächlich angenommen wird (siehe nachfolgende Übung). ■

Freiwillige Übung. Es sei V ein endlich-dimensionaler Teilraum eines (möglicherweise unendlich-dimensionalen) normierten Raums X. Man verwende des Satz von Bolzano-Weierstraß, um zu zeigen, dass für jedes $x \in X$ mindestens ein $v \in V$ existiert mit

$$\|x - v\|_X = \min_{w \in V} \|x - w\|_X.$$

Man wende diese Beobachtung an, um zu sehen, dass das Minimum in (1.11) tatsächlich existiert. Der *Alternantensatz von Čebyšev* zeigt übrigens, dass das Minimum eindeutig angenommen wird. ■

Bemerkung 9 *Die Konvergenzgeschwindigkeit in Satz hängt 6 von der Glattheit von f ab. Aus Lemma 8 folgt, dass der Quadraturfehler schneller fällt als der Bestapproximationsfehler für Polynome und insbesondere somit schneller als ein beliebiger Interpolationsfehler. Aus der Numerischen Mathematik wissen Sie, dass für beliebige Stützstellen $a \leq t_0 < \cdots < t_k \leq b$ und $I_k f \in \mathbb{P}_k$ das zugehörige Interpolationspolynom folgende Fehlerdarstellung gilt*

$$f(x) - I_k f(x) = \frac{f^{(k+1)}(\xi)}{(k+1)!} \prod_{j=0}^{k} (x - t_k).$$

Insbesondere folgt also auf dem Intervall $[a,b]$ die Fehlerabschätzung

$$\|f - I_k f\|_\infty \leq \frac{\|f^{(k+1)}\|_\infty}{(k+1)!} \max_{x \in [a,b]} \prod_{j=0}^{k} |x - t_j|.$$

Ebenfalls aus der Numerischen Mathematik wissen Sie, dass die eindeutige Lösung der Minimierungsaufgabe

$$\max_{x \in [a,b]} \prod_{j=0}^{k} |x - t_j| = \min_{\tau_0, \ldots, \tau_k \in [a,b]} \max_{x \in [a,b]} \prod_{j=0}^{k} |x - \tau_j|$$

die Čebyšev-Knoten sind. Dies sind die Bilder $t_j = \Phi(\zeta_j)$ der

$$\zeta_j = \cos\Big(\frac{2j+1}{2(k+1)} \pi \Big) \quad \textit{für } j = 0, \ldots, k$$

unter der affinen Abbildung $\Phi : [-1,1] \to [a,b]$, $\Phi(t) = \frac{1}{2}\big(a + b + t(b-a)\big)$ Mit dieser Wahl gilt

$$\max_{x \in [a,b]} \prod_{j=0}^{k} |x - t_j| = \Big(\frac{b-a}{2} \Big)^{k+1} 2^{-k}.$$

und somit insgesamt

$$\|f - I_k f\|_\infty \leq \frac{1}{2} \frac{(b-a)^{k+1}}{(k+1)!} \|f^{(k+1)}\|_\infty 4^{-k}.$$

Sofern die Ableitung $f^{(k+1)}$ für $k \to \infty$ nicht zu schnell aufbläst, erhalten wir für $f \in C^\infty(a,b)$ also exponentiell schnelle Konvergenz. Anderenfalls zeigt der Satz allerdings nur Konvergenz ohne Rate, d.h. die Geschwindigkeit kann beliebig langsam sein.

Ist die zu integrierende Funktion nur stückweise glatt, so bieten sich sogenannte **summierte Quadraturformeln** an. Wir betrachten den Spezialfall $\omega \equiv 1$. Es sei

$$Q_n^{(\mathrm{ref})} f := \sum_{j=0}^{n} \omega_j f(x_j) \tag{1.12}$$

eine Quadraturformel auf einem Referenzintervall, z.B. $[-1,1]$, zur Approximation von

$$Q^{(\mathrm{ref})} f := \int_{-1}^{1} f\, dt. \tag{1.13}$$

Es sei $a = y_0 < y_1 < \cdots < y_N = b$ eine Unterteilung des Intervalls $[a,b]$ in Teilintervalle $[y_{\ell-1}, y_\ell]$. Auf jedem Teilintervall $[y_{\ell-1}, y_\ell]$ sei

$$Q_n^{(\ell)} f := \sum_{j=0}^{n} \omega_j^{(\ell)} f(x_j^{(\ell)}) \tag{1.14}$$

die Quadraturformel, die durch affine Transformation aus Q_n hervorgeht, d.h. $Q_n^{(\ell)} f$ approximiert

$$Q^{(\ell)} f := \int_{y_{\ell-1}}^{y_\ell} f\, dx. \tag{1.15}$$

Dann definiert

$$Q_{nN} f := \sum_{\ell=1}^{N} Q_n^{(\ell)} f \approx \sum_{\ell=1}^{N} \int_{y_{\ell-1}}^{y_\ell} f\, dx = \int_a^b f\, dx =: Qf \tag{1.16}$$

eine Quadraturformel auf $[a,b]$.

Satz 10 *Die Quadraturformel $Q_n^{(\mathrm{ref})}$ sei exakt vom Grad $k \geq 0$. Ferner sei die Unterteilung von $[a,b]$ uniform, d.h. $y_\ell - y_{\ell-1} = h := (b-a)/N$ für alle $\ell = 1, \ldots, N$. Dann gilt für $f \in C[a,b]$*

$$\lim_{N \to \infty} Q_{nN} f = Qf. \tag{1.17}$$

Im Fall $f \in C^p[a,b]$ und $p > k$ gilt $|Qf - Q_{nN} f| = \mathcal{O}(N^{-(k+1)})$.

Beweis. Wie oben gezeigt, gilt

$$\omega_j^{(\ell)} = \frac{y_\ell - y_{\ell-1}}{2} \omega_j = \frac{\omega_j}{N} (b-a)$$

Ferner gilt nach Lemma 8 für $\omega \equiv 1$ auf $[y_{\ell-1}, y_\ell]$

$$\begin{aligned} |Q_n^{(\ell)} f - Q^{(\ell)} f| &\leq \Big(|y_\ell - y_{\ell-1}| + \sum_{j=0}^{n} |\omega_j^{(\ell)}| \Big) \min_{p \in \mathbb{P}_k} \|f - p\|_{\infty, [y_{\ell-1}, y_\ell]} \\ &= \frac{b-a}{N} \Big(1 + \frac{1}{2} \sum_{j=0}^{n} |\omega_j| \Big) \min_{p \in \mathbb{P}_k} \|f - p\|_{\infty, [y_{\ell-1}, y_\ell]}. \end{aligned}$$

Summation über alle Teilintervalle liefert

$$|Q_{nN}f - Qf| \le \sum_{\ell=1}^{N} |Q_n^{(\ell)}f - Q^{(\ell)}f| \le \Big(1 + \frac{1}{2}\sum_{j=0}^{n}|\omega_j|\Big) \max_{\ell=1,\dots,N} \min_{p\in\mathbb{P}_k} \|f-p\|_{\infty,[y_{\ell-1},y_\ell]}.$$

Für glattes f folgt die Aussage nun aus den Fehlerabschätzungen für den Interpolationsfehler. Für stetiges f nutzen wir die gleichmäßige Stetigkeit auf dem Kompaktum $[a,b]$, d.h.

$$\forall\varepsilon>0\exists\delta>0\forall x,y\in[a,b] \quad \big(|x-y|\le\delta \Rightarrow |f(x)-f(y)|\le\varepsilon\big).$$

Für gegebenes $\varepsilon>0$, entsprechendes $\delta>0$ und beliebiges $N\in\mathbb{N}$ mit $h=(b-a)/N\le\delta$ folgt also

$$\min_{p\in\mathbb{P}_k}\|f-p\|_{\infty,[y_{\ell-1},y_\ell]} \le \|f-f(y_\ell)\|_{\infty,[y_{\ell-1},y_\ell]} \le \varepsilon.$$

Dies beweist die Konvergenz (1.17) ■

Der folgende Sprachgebrauch hat sich in der Numerischen Mathematik eingebürgert:

- Man spricht von der **h-Methode**, wenn man für fixiertes $Q_n^{(\mathrm{ref})}$ die summierte Quadraturformel Q_{nN} nimmt und für fixes $n\in\mathbb{N}_0$ nur den Grenzübergang $N\to\infty$ betrachtet, siehe Satz 10. Dies führt i.d.R. auf eine algebraische Konvergenzrate $\mathcal{O}(N^{-\alpha})$ mit $\alpha>0$.
- Man spricht von der **p-Methode**, wenn man eine Familie von eventuell summierten Quadraturformeln Q_{nN} nimmt und für fixes $N\in\mathbb{N}$ nur den Grenzübergang $n\to\infty$ betrachtet, siehe Satz 6. Dies führt i.d.R. auf eine algebraische Konvergenzrate $\mathcal{O}(n^{-\alpha})$ mit $\alpha>0$. Für beliebig glattes f kann aber auch ein exponentielles Konvergenzverhalten $\mathcal{O}(q^n)$ mit $0<q<1$ auftreten.
- Man spricht von der **hp-Methode**, wenn man dort, wo f nicht glatt ist, das Intervall verfeinert und parallel dort, wo f glatt ist, den Quadraturgrad n erhöht. Üblicherweise führt dies für Funktionen, die außerhalb endlich vieler Singularitäten (der Ableitungen) glatt sind, auf exponentielles Konvergenzverhalten $\mathcal{O}(q^n)$ mit $0<q<1$.

1.3 Fundamentale Unsicherheit der Quadratur

Alle Konvergenzaussagen des vorausgegangenen Abschnitts sind asymptotische Aussagen. Es kann —in endlicher Rechenzeit— keine zuverlässige Aussage geben, ob der Quadraturfehler

$$|Q_n f - Qf| \le \tau \tag{1.18}$$

eine gegebene (oder notwendige) Toleranz $\tau>0$ unterschreitet. Um dies zu sehen, muss man sich nur klarmachen, dass endliche Rechenzeit bedeutet, dass der Integrand f nur endlich oft ausgewertet worden ist, und diese Auswertungen sind deterministisch aufgrund der gewählten Folge von Quadraturformeln. Sind $t_0,\dots,t_m\in[a,b]$ diese endlich vielen Auswertungspunkte, so gibt es aber überabzählbar viele nicht-negative Funktionen $f\in C[a,b]$ mit $f(t_j)=0$ für alle $j=0,\dots,m$, z.B. das Polynom

$$p(t) := \prod_{j=0}^{m}(t-t_j)^2$$

multipliziert mit jeder positiven stetigen Funktion. In diesem Fall ist das Integral $Qf>0$ und mittels linearer Skalierung sogar beliebig groß. Auf der anderen Seite liefert das numerische Verfahren (mindestens) bis zu diesem Zeitpunkt stets $Q_nf=0$, d.h. der relative Fehler ist 100%, und der absolute Fehler kann beliebig groß sein.

Überhuber [3] bezeichnet dies als *fundamentale Unsicherheit der Quadratur*. Diese ist probleminhärent und lässt sich auch durch adaptive Techniken nicht vermeiden!

1.4 Berechnung von Gauss-Quadraturen

Aufgrund der Konstruktion der Orthogonalpolynome $p_0, p_1, p_2, \dots$ mit Hilfe der Gram-Schmidt-Orthogonalisierung kann man sich leicht überlegen, dass p_j eindeutig durch eine 3-Term-Rekursion gegeben ist. Es gilt

$$p_0(x) = 1, \quad p_1(x) = x - \beta_0, \quad p_{n+1}(x) = (x - \beta_n)p_n(x) - \gamma_n^2 p_{n-1}(x) \quad \text{für } n \geq 1 \tag{1.19}$$

mit reellen Koeffizienten

$$\beta_n = \frac{(xp_n, p_n)_\omega}{(p_n, p_n)_\omega} \quad \text{und} \quad \gamma_n^2 = \frac{(p_n, p_n)_\omega}{(p_{n-1}, p_{n-1})_\omega}. \tag{1.20}$$

Ebenso ist klar, dass der Leitkoeffizient eines Orthogonalpolynoms p_j automatisch auf 1 normalisiert ist.

Im Fall von $\omega \equiv 1$ auf dem Intervall $[-1, 1]$ erhält man die (normalisierten) **Legendre-Polynome**. Man nennt die zugehörige Gauss-Quadratur deshalb auch **Gauss-Legendre-Quadratur**.

Im Fall von $\omega(x) = (1-x^2)^{-1/2}$ auf dem Intervall $[-1, 1]$ erhält man die **Čebyšev-Polynome erster Art**. Man spricht deshalb von der **Gauss-Čebyšev-Quadratur**.

In den seltensten Fällen gibt es geschlossene Formeln für die Stützstellen und Gewichte von Gauss-Quadraturen. Eine Ausnahme bildet die Gauss-Čebyšev-Quadratur, für die gilt

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\, dx \approx Q_n f = \frac{\pi}{n+1} \sum_{j=0}^{n} f\Big(\cos\Big(\frac{2j-1}{2(n+1)}\,\pi\Big)\Big), \tag{1.21}$$

siehe Hanke-Bourgeois [1, Beispiel 40.4, Seite 339].

Kennt man die Koeffizienten β_j, γ_j aus der 3-Term-Rekursion, erhält man die Stützstellen und Gewichte der Gauss-Quadratur durch Lösen eines Eigenwertproblems. Der Beweis findet sich beispielsweise in Plato [2, Theorem 6.40].

Satz 11 *Mit den Konstanten der 3-Term-Rekursion* (1.19) *sind die Eigenwerte der Matrix*

$$A = \begin{pmatrix} \beta_0 & -\gamma_1 & 0 & \cdots & 0 \\ -\gamma_1 & \beta_1 & -\gamma_2 & \ddots & \vdots \\ 0 & -\gamma_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\gamma_n \\ 0 & \cdots & 0 & -\gamma_n & \beta_n \end{pmatrix} \in \mathbb{R}_{\text{sym}}^{(n+1)\times(n+1)} \tag{1.22}$$

genau die Nullstellen $x_0, \dots, x_n$ des $(n+1)$-ten Orthogonalpolynoms p_{n+1}. Die zugehörigen Gauss-Gewichte werden gegeben durch

$$\omega_k = \Big(\int_a^b \omega\, dx\Big)\Big(\sum_{j=0}^{n} \tau_j^2 p_j(x_k)^2\Big)^{-1} \quad \textit{für } k = 0, \dots, n \tag{1.23}$$

mit $\tau_j = \prod_{\ell=1}^{j} \gamma_\ell^{-1}$. ∎

Im Spezialfall der Gauss-Legendre-Quadratur lässt sich dieser Satz weiter vereinfachen. Ein Beweis findet sich beispielsweise in Hanke-Bourgeois [1, Abschnitt 41, Seite 342].

Satz 12 (Golub, Welsch '69) *Die Eigenwerte der Matrix*

$$A = \begin{pmatrix} 0 & \alpha_1 & 0 & \cdots & 0 \\ \alpha_1 & 0 & \alpha_2 & \ddots & \vdots \\ 0 & \alpha_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \alpha_n \\ 0 & \cdots & 0 & \alpha_n & 0 \end{pmatrix} \in \mathbb{R}_{\text{sym}}^{(n+1)\times(n+1)} \quad \textit{mit} \quad \alpha_k = \frac{k}{\sqrt{4k^2-1}} \quad \textit{für } k = 1, \dots, n \tag{1.24}$$

sind genau die Stützstellen $x_0, \dots, x_n$ der Gauss-Legendre-Quadratur (d.h. $\omega \equiv 1$ auf dem Intervall $[-1,1]$). Ist $v^{(j)}$ der Eigenvektor zu x_j mit $|v^{(j)}| = 1$, so gilt

$$\omega_j = 2\,(v_1^{(j)})^2 \tag{1.25}$$

für das zu x_j gehörige Gewicht $\omega_j > 0$. ■

1.5 Integration der Ableitung

Häufig sind gewisse L^q-Normen $\|f'\|_{L^q(a,b)}$ von Ableitungen zu integrieren. Dies erfolgt zweckmäßig durch summierte Quadraturformeln

$$\|f'\|_{L^q(a,b)}^q = \int_a^b |f'|^q\,dx = \sum_{\ell=1}^N \int_{y_{\ell-1}}^{y_\ell} |f'|^q\,dx \approx \sum_{\ell=1}^N Q_n^{(\ell)}\big(|f'|^q\big) = Q_{nN}\big(|f'|^q\big). \tag{1.26}$$

Üblicherweise steht allerdings lediglich die Punktauswertung der Funktion $f(x)$, nicht aber der Ableitung $f'(x)$ zur Verfügung, sodass die bisherigen Techniken nicht direkt angewendet werden können. Auf theoretischer Seite ersetzen wir f daher durch ein stückweises Interpolationspolynom p

$$f|_{(y_{\ell-1},y_\ell)} \approx p|_{(y_{\ell-1},y_\ell)} =: p_\ell \in \mathbb{P}_m. \tag{1.27}$$

Dann gilt formal

$$\big|\|f'\|_{L^q(a,b)} - Q_{nN}\big(|p'|^q\big)^{1/q}\big| \le \|f'-p'\|_{L^q(a,b)} + \big|\|p'\|_{L^q(a,b)} - Q_{nN}\big(|p'|^q\big)^{1/q}\big|. \tag{1.28}$$

Im häufigen Fall $q=2$ ist $|p'_\ell|^2 = (p'_\ell)^2$ ein Polynom und die Quadraturformel $Q_n^{(\ell)}$ ist für geeignetes n deshalb exakt, d.h. der zweite Term auf der rechten Seite verschwindet.

Damit ergeben sich dann folgende Fragen:

- Wie wertet man $p'(x)$ aus, wenn man nur gewisse Funktionswerte $f(x)$ kennt?
- Wie kann man den Fehler $\|f'-p'\|_{L^p(a,b)}$ kontrollieren?

Ohne Beschränkung der Allgemeinheit betrachten wir von nun an nur noch ein Teilintervall $[a,b] = [y_{\ell-1}, y_\ell]$. Für fixierten Polynomgrad m seien $a \le t_0 < \cdots < t_m \le b$ die Interpolationsknoten. Mit den zugehörigen Lagrange-Polynomen L_j lässt sich das Interpolationspolynom p zu f schreiben als

$$p = \sum_{j=0}^m f(t_j) L_j. \tag{1.29}$$

Insbesondere gilt also

$$p' = \sum_{j=0}^m f(t_j) L_j'. \tag{1.30}$$

Berechnet man nun die Ableitungen L_j' der Lagrange-Polynome explizit, so kann man sie an den Quadraturknoten $x_0 < \cdots < x_n$ auswerten, d.h. lineares Postprocessing des Auswertungsvektors $\big(f(t_0), \dots, f(t_m)\big)$ liefert die benötigten Ableitungswerte $\big(p'(x_0), \dots, p'(x_n)\big)$. Verwendet man eine summierte Quadraturformel und fixiert den Interpolationsgrad m sowie die Quadraturordnung n, so muss die Berechnung von $L_j'(x_k)$ nur auf dem Referenzintervall durchgeführt werden, um die Transformationsmatrix zu bestimmen.

Der folgende Satz schätzt den Fehler $\|f^{(m)} - p^{(m)}\|_\infty$ in der L^∞-Norm ab.

Satz 13 *Seien $f \in C^{n+1}[a,b]$, $t_0 < \cdots < t_n \in [a,b]$ paarweise verschiedene Stützstellen und $p \in \mathbb{P}_n$ das eindeutige Interpolationspolynom mit $p(x_k) = f(x_k)$ für alle $k = 0, \ldots, n$. Dann gilt für jedes $m = 0, \ldots, n$ und jedes $t \in [a,b]$ die Fehlerdarstellung*

$$f^{(m)}(t) - p^{(m)}(t) = \frac{f^{(n+1)}(\xi)}{(n+1-m)!} \prod_{\ell=0}^{n-m} (t - \zeta_\ell) \tag{1.31}$$

mit geeigneten Zwischenstellen $\xi = \xi(t,m)$ und $\zeta_\ell = \zeta_\ell(m)$ in $[a,b]$. Für $m = 0$ gilt $\zeta_\ell = t_\ell$ für alle $\ell = 0, \ldots, n$. Ferner folgt aus (1.31) *die Abschätzung*

$$\|f^{(m)} - p^{(m)}\|_\infty \leq \frac{\|f^{n+1}\|_\infty}{(n+1-m)!} (b-a)^{n+1-m}. \tag{1.32}$$

Beweis. Der Interpolationsfehler $e = f - p$ hat in $[a,b]$ mindestens $n+1$ verschiedene Nullstellen. Zwischen zwei Nullstellen liegt nach Satz von Rolle eine Nullstelle der Ableitung e', d.h. e' hat mindestes n Nullstellen. Induktives Vorgehen zeigt, dass die m-te Ableitung $e^{(m)}$ mindestens $n+1-m$ paarweise verschiedene Nullstellen $\zeta_0 < \cdots < \zeta_{n-m}$. Ohne Beschränkung der Allgemeinheit gilt $t \notin \{\zeta_0, \ldots, \zeta_{m-n}\}$, denn anderenfalls ist (1.31) trivial. Wir definieren

$$F(y) := e^{(m)}(t)\omega(y) - \omega(t)e^{(m)}(y) \quad \text{mit} \quad \omega(y) := \prod_{\ell=0}^{n-m} (y - \zeta_\ell) \in \mathbb{P}_{n+1-m}.$$

Nach Konstruktion hat F insgesamt $n+2-m$ Nullstellen. Induktion mit dem Satz von Rolle zeigt, dass $F^{(n+1-m)}$ mindestens eine Nullstelle $\xi \in [a,b]$ hat, d.h.

$$\begin{aligned} 0 = F^{(n+1-m)}(\xi) &= e^{(m)}(t)\omega^{(n+1-m)}(\xi) - \omega(t)e^{(n+1)}(\xi) \\ &= e^{(m)}(t)\,(n+1-m)! - \omega(t)\,f^{(n+1)}(\xi) \end{aligned}$$

wegen $p^{(n+1)} = 0$. Durch Umformen folgt (1.31). Naive Abschätzung der rechten Seite von (1.31) zeigt

$$f^{(m)}(t) - p^{(m)}(t) \leq \frac{\|f^{n+1}\|_\infty}{(n+1-m)!} (b-a)^{n+1-m}$$

für alle $t \in [a,b]$, und es folgt (1.32). ∎

Kapitel 2

Finite Elemente Methode in 1D

2.1 Galerkin-Verfahren und Orthogonalprojektionen

2.1.1 Der funktionalanalytische Rahmen

Der folgende Satz ist das zentrale Ergebnis über Hilbert-Räume. Er charakterisiert den Dualraum H^* eines Hilbert-Raums H und wird in jeder Funktionalanalysis-Vorlesung und jedem Lehrbuch bewiesen, siehe beispielsweise WERNER [4, Theorem V.3.6].

Satz 14 (Riesz) *Es seien H ein Hilbert-Raum über $\mathbb{R}$ mit zugehörigem Skalarprodukt $\langle \cdot\,, \cdot \rangle_H$ und $F \in H^*$. Dann existiert ein eindeutiges $u \in H$ mit*

$$\langle u\,, v \rangle_H = F(v) \quad \textit{für alle } v \in H. \tag{2.1}$$

Insbesondere gilt

$$\|u\|_H = \|F\|_{H^*} := \sup_{v \in H \setminus \{0\}} \frac{F(v)}{\|v\|_H}, \tag{2.2}$$

und die Abbildung $I_H : H^* \to H$, die F auf u abbildet, ist ein isometrischer Isomorphismus, d.h. linear und bijektiv mit $\|I_H(F)\|_H = \|F\|_{H^*}$.

Beweis. Wir beweisen nur die „einfachen“ Aussagen des Satzes und verweisen für den Beweis der Existenz einer Lösung von (2.1) auf die Funktionalanalysis-Vorlesung. Um die Eindeutigkeit der Lösung zu zeigen, seien $u, \widetilde{u} \in H$ zwei Lösungen von (2.1). Dann gilt

$$\langle u - \widetilde{u}\,, v \rangle_H = 0 \quad \text{für alle } v \in H.$$

Die Wahl $v = u - \widetilde{u}$ zeigt deshalb die Eindeutigkeit $u = \widetilde{u}$.

Als Nächstes zeigen wir (2.2), um zu sehen, dass dies ebenfalls eine Konsequenz aus (2.1) ist: Zunächst gilt

$$\|u\|_H^2 = \langle u\,, u \rangle_H = F(u) \leq \|F\|_{H^*} \|u\|_H$$

und somit $\|u\|_H \leq \|F\|_{H^*}$. Andererseits gilt mit der Cauchy-Schwarz-Ungleichung

$$F(v) = \langle u\,, v \rangle_H \leq \|u\|_H \|v\|_H \quad \text{für alle } v \in H \setminus \{0\}$$

und deshalb $\|F\|_{H^*} \leq \|u\|_H$. Nach Aussage des Satzes ist die Abbildung $I_H : H^* \to H$ bijektiv. Man beachte, dass die Umkehrabbildung durch

$$I_H^{-1} : H \to H^*, u \mapsto \langle u\,, \cdot \rangle_H$$

gegeben ist. Da I_H^{-1} linear ist, ist auch I_H linear, und die Isometrie-Eigenschaft ist gerade (2.2). ■

In der Vorlesung über partielle Differentialgleichungen wird häufig das folgende *Lemma von Lax-Milgram* bewiesen. Dabei handelt es sich im Wesentlichen um den Satz von Riesz, wobei auf die Symmetrie des Skalarproduktes verzichtet werden kann.

Satz 15 (Lax-Milgram) *Es sei H ein Hilbert-Raum über $\mathbb{R}$ und $\langle\!\langle \cdot\,,\cdot \rangle\!\rangle$ eine stetige und elliptische Bilinearform auf H, d.h. es existieren Konstanten $C_1, C_2 > 0$ mit*

$$C_1 \|u\|_H^2 \leq \langle\!\langle u\,, u \rangle\!\rangle \quad \textit{und} \quad \langle\!\langle u\,, v \rangle\!\rangle \leq C_2 \|u\|_H \|v\|_H \quad \textit{für alle } u, v \in H. \tag{2.3}$$

Ferner sei $F \in H^*$. Dann existiert ein eindeutiges $u \in H$ mit

$$\langle\!\langle u\,, v \rangle\!\rangle = F(v) \quad \textit{für alle } v \in H. \tag{2.4}$$

Ferner gilt $C_1 \|u\|_H \leq \|F\|_{H^*} \leq C_2 \|u\|_H$.

Beweis. Die Eindeutigkeit von u sowie die Normabschätzungen folgen analog zu den Überlegungen zum Satz von Riesz. Wir zeigen deshalb nur die Existenz von u. Dazu betrachten wir die Funktion $\Phi(u) := u - C\, I_H(\langle\!\langle u\,, \cdot \rangle\!\rangle - F)$ mit $C := C_1/C_2^2$. Dann gelten $\Phi : H \to H$ sowie die Äquivalenz

$$\Phi(u) = u \quad \Longleftrightarrow \quad \langle\!\langle u\,, \cdot \rangle\!\rangle = F \in H^*.$$

Wir werden zeigen, Φ eine Kontraktion ist, und damit hat Φ einen eindeutigen Fixpunkt, d.h. (2.4) hat eine eindeutige Lösung. Elementare Rechnung zeigt

$$\|\Phi(u) - \Phi(v)\|_H^2 = \|u - v\|_H^2 - 2C \,\langle u - v\,, I_H\langle\!\langle u - v\,, \cdot \rangle\!\rangle \rangle_H + C^2 \|I_H\langle\!\langle u - v\,, \cdot \rangle\!\rangle\|_H^2.$$

Nach Definition von I_H gilt

$$\langle u - v\,, I_H\langle\!\langle u - v\,, \cdot \rangle\!\rangle \rangle_H = \langle\!\langle u - v\,, u - v \rangle\!\rangle \geq C_1 \|u - v\|_H^2.$$

Mit der Isometrie-Eigenschaft und der Stetigkeit von $\langle\!\langle \cdot\,,\cdot \rangle\!\rangle$ gilt

$$\|I_H\langle\!\langle u - v\,, \cdot \rangle\!\rangle\|_H^2 = \|\langle\!\langle u - v\,, \cdot \rangle\!\rangle\|_{H^*}^2 \leq C_2^2 \|u - v\|_H^2.$$

Insgesamt erhalten wir

$$\|\Phi(u) - \Phi(v)\|_H^2 \leq \|u - v\|_H^2 \left(1 - 2CC_1 + C^2C_2^2\right) = \|u - v\|_H^2 \left(1 - C_1^2/C_2^2\right).$$

Die Beobachtung $q := 1 - C_1^2/C_2^2 < 1$ beschließt den Beweis. ■

Freiwillige Übung. Es sei $\langle\!\langle \cdot\,,\cdot \rangle\!\rangle : H \times H \to \mathbb{R}$ eine Bilinearform auf einem normierten Raum H. Zeigen Sie, dass $\langle\!\langle \cdot\,,\cdot \rangle\!\rangle$ genau dann stetig ist, d.h. $\lim_n \langle\!\langle x_n\,, y_n \rangle\!\rangle = \langle\!\langle x\,, y \rangle\!\rangle$ für alle konvergenten Folgen $(x_n), (y_n)$ in H mit $x = \lim_n x_n$ und $y = \lim_n y_n$, wenn

$$\|\langle\!\langle \cdot\,,\cdot \rangle\!\rangle\| := \sup_{x,y \in H\setminus\{0\}} \frac{\langle\!\langle x\,, y \rangle\!\rangle}{\|x\|_H \|y\|_H} < \infty$$

gilt. In diesem Fall ist $\|\langle\!\langle \cdot\,,\cdot \rangle\!\rangle\|$ die kleinst mögliche Konstante $C_2 > 0$ in (2.3), und $\|\cdot\|$ definiert eine Norm auf dem Raum aller stetigen Bilinearformen auf H. ■

Freiwillige Übung. Beweisen Sie den Satz von Riesz (d.h. die Existenz einer Lösung) für $\dim H < \infty$ mit den Mitteln der Linearen Algebra. Fixieren Sie dazu eine Basis $\{\zeta_1, \ldots, \zeta_n\}$ und schreiben Sie die Lösung (sofern sie existiert) als Linearkombination $u = \sum_{j=1}^n x_j \zeta_j$. ■

2.1.2 Das abstrakte Galerkin-Verfahren

Es sei $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ eine stetige und elliptische Bilinearform auf einem Hilbert-Raum H. Es sei $X_h \leq H$ ein endlich-dimensionaler Teilraum von H mit Basis $\{\zeta_1, \ldots, \zeta_N\}$.

Wir wenden nun das Lemma von Lax-Milgram für X_h an. Dazu beachte man, dass $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ ebenfalls stetig und elliptisch ist auf X_h und dass X_h als endlich-dimensionaler Teilraum eines Hilbert-Raums ebenfalls ein Hilbert-Raum ist.

Zu $u \in H$ und mit $F := \langle\!\langle u\,, \cdot \rangle\!\rangle \in H^*$ existiert nach Lax-Milgram ein eindeutiges $u_h \in X_h$ mit

$$\langle\!\langle u_h\,, v_h \rangle\!\rangle = \langle\!\langle u\,, v_h \rangle\!\rangle \quad \text{für alle } v_h \in X_h. \tag{2.5}$$

Dieses bezeichnet man als **Galerkin-Approximation von u** (bzgl. $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ und X_h). Es ist offensichtlich eindeutig charakterisiert durch die sogenannte **Galerkin-Orthogonalität**

$$\langle\!\langle u - u_h\,, v_h \rangle\!\rangle = 0 \quad \text{für alle } v_h \in X_h. \tag{2.6}$$

Das folgende *Céa-Lemma* besagt, dass das Galerkin-Verfahren quasi-optimal ist. Das bedeutet, das die Approximation u_h von u bis auf eine Konstante so gut ist, wie es im Raum X_h überhaupt möglich ist. Die Konstante hängt dabei nur von $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ ab, sie ist also insbesondere unabhängig von X_h! In diesem Sinne ist das folgende Lemma asymptotisch interessant: Ist $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ festgelegt, so bestimmen nur mehr die Räume X_h die Konvergenzrate.

Lemma 16 (Céa) *Es sei $u_h \in X_h$ die Galerkin-Approximation von u. Mit den Konstanten $C_1, C_2 > 0$ aus* (2.3) *gilt dann*

$$\|u - u_h\|_H \leq \frac{C_2}{C_1} \min_{v_h \in X_h} \|u - v_h\|_H. \tag{2.7}$$

Insbesondere gilt $C_2/C_1 = 1$, falls $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ ein Skalarprodukt ist, das $\|\cdot\|_H$ induziert, d.h. $\|v\|_H^2 = \langle\!\langle v\,, v \rangle\!\rangle$ für alle $v \in H$.

Beweis. Für beliebiges $v_h \in X_h$ zeigt die Galerkin-Orthogonalität (2.6)

$$C_1 \|u - u_h\|_H^2 \leq \langle\!\langle u - u_h\,, u - u_h \rangle\!\rangle = \langle\!\langle u - u_h\,, u - v_h \rangle\!\rangle \leq C_2 \|u - u_h\|_H \|u - v_h\|_H.$$

Dies zeigt die Abschätzung (2.7) mit dem Infimum anstelle des Minimums auf der rechten Seite. Man überlegt sich leicht, dass aufgrund der endlichen Dimension das Infimum als Minimum angenommen wird. — In Hilbert-Räumen kann man das sogar für jeden abgeschlossenen Unterraum zeigen, und endlich-dimensionale Unterräume sind insbesondere abgeschlossen. ■

Bemerkung 17 *Ist $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ ein Skalarprodukt (also zusätzlich symmetrisch), so folgt aus der Galerkin-Orthogonalität* (2.6) *bzw. aus dem Céa-Lemma, dass die Abbildung $\mathbb{G}_h : H \to X_h, u \mapsto u_h$ die Orthogonalprojektion auf X_h bzgl. $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ ist. Um dies zu sehen, zeigt man, dass $\mathbb{G}_h$ eine lineare Projektion ist, d.h. $\mathbb{G}_h^2 = \mathbb{G}_h$. Die Lineare Algebra lehrt uns nämlich, dass die Orthogonalprojektion die einzige lineare Projektion mit* (2.6) *ist. — Insbesondere ist die Berechnung einer Orthogonaloprojektion (oder Bestapproximierenden bzgl. einer Hilbert-Norm) ein Spezialfall der Galerkin-Approximation.*

Der folgende Satz ist für uns in gewisser Weise der entscheidende. Denn er zeigt, dass man die Galerkin-Approximation durch das Lösen eines linearen Gleichungssystems berechnen kann, wenn man sich eine beliebige Basis $\{\zeta_1, \ldots, \zeta_N\}$ von X_h vorgibt.

Satz 18 *Wir definieren die Matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ und den Vektor $\mathbf{b} \in \mathbb{R}^N$ durch*

$$\mathbf{A}_{jk} = \langle\!\langle \zeta_k\,, \zeta_j \rangle\!\rangle \quad \textit{und} \quad \mathbf{b}_j = \langle\!\langle u\,, \zeta_j \rangle\!\rangle. \tag{2.8}$$

Dann ist $\mathbf{A}$ regulär, und die eindeutige Lösung $\mathbf{x} \in \mathbb{R}^N$ von $\mathbf{A}\mathbf{x} = \mathbf{b}$ erfüllt

$$u_h = \sum_{k=1}^{N} \mathbf{x}_j \zeta_j. \tag{2.9}$$

Falls $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ zusätzlich symmetrisch ist, ist $\mathbf{A}$ sogar eine SPD-Matrix.

Beweis. Es seien $\mathbf{y} \in \mathbb{R}^N$ und $v_h := \sum_{k=1}^N \mathbf{y}_k \zeta_k \in X_h$. Dann gilt

$$C_1 \|v_h\|_H^2 \leq \langle\!\langle v_h \, , v_h \rangle\!\rangle = \sum_{j,k=1}^N \mathbf{y}_j \mathbf{y}_k \langle\!\langle \zeta_k \, , \zeta_j \rangle\!\rangle = \sum_{j,k=1}^N \mathbf{y}_j \mathbf{y}_k \mathbf{A}_{jk} = \mathbf{y} \cdot \mathbf{A}\mathbf{y}.$$

Für $\mathbf{y} \neq 0$ folgt also $\mathbf{y} \cdot \mathbf{A}\mathbf{y} > 0$ und daher $\ker(\mathbf{A}) = \{0\}$. Insbesondere ist $\mathbf{A}$ also injektiv und deshalb regulär. Falls $\langle\!\langle \cdot \, , \cdot \rangle\!\rangle$ symmetrisch ist, so ist klarerweise $\mathbf{A}$ symmetrisch und deshalb SPD.

Nun sei $\mathbf{x} \in \mathbb{R}^N$ der Koeffizientenvektor der Galerkin-Lösung $u_h = \sum_{k=1}^N \mathbf{x}_k \zeta_k$. Nach Definition von u_h gilt

$$\mathbf{b}_j = \langle\!\langle u \, , \zeta_j \rangle\!\rangle = \langle\!\langle u_h \, , \zeta_j \rangle\!\rangle = \sum_{k=1}^N \mathbf{x}_k \mathbf{A}_{jk} = (\mathbf{A}\mathbf{x})_j \text{für alle } j = 1, \ldots, N,$$

d.h. $\mathbf{x}$ ist die eindeutige Lösung des linearen Gleichungssystems $\mathbf{A}\mathbf{x} = \mathbf{b}$. ■

Bemerkung 19 *Mit leichter Modifikation der Argumentation gibt der letzte Beweis auch einen Beweis für das Lemma von Lax-Milgram im Fall von* $\dim H < \infty$*: Die Matrix* $\mathbf{A}$ *ist regulär, und* $\mathbf{x}$ *sei die eindeutige Lösung von* $\mathbf{A}\mathbf{x} = \mathbf{b}$*. Dann definiert man* $u = \sum_{k=1}^n \mathbf{x}_k \zeta_k$ *und beobachtet, dass* u *die Variationsgleichung* (2.4) *löst.*

Bemerkung 20 *Obwohl im Sinne der Linearen Algebra alle Basen gleichwertig sind, muss man in der Numerik mehr aufpassen. Mögliche Ziele sind die folgenden:*

- *kleine Konditionszahl* $\operatorname{cond}_2(\mathbf{A})$*, damit die berechnete Lösung mehr als „weißes Rauschen“ ist;*
- *möglichst wenige Einträge* $\mathbf{A}_{jk}$ *ungleich Null, damit die Matrix mit möglichst wenig Aufwand aufgebaut und gespeichert werden kann;*
- *viel algebraische Struktur in der Matrix, z.B. Bandstruktur, damit das Lösen des linearen Gleichungssystems möglichst niedrigen Aufwand hat.*

Während die ersten beiden Punkte durch die Wahl der Basis fixiert werden und durch die Indizierung der Basiselemente nicht geändert werden, ist für den letzten Punkt auch die Reihenfolge der Basiselemente wesentlich.

Freiwillige Übung. Es sei H ein reeller Hilbert-Raum, $F \in H^*$ und $\mathcal{J}(v) := \frac{1}{2}\langle v \, , v \rangle_H - F(v)$. Zeigen Sie, dass die Variationsformulierung

$$\langle u \, , v \rangle_H = F(v) \quad \text{für alle } v \in H \tag{2.10}$$

äquivalent ist zum Minimierungsproblem

$$\mathcal{J}(u) = \min_{v \in H} \mathcal{J}(v), \tag{2.11}$$

wobei Sie sich für die die Implikation (2.11)⇒(2.10) nur überlegen müssen, wie die Fréchet-Ableitung von $\mathcal{J}(u)$ aussieht. ■

Bemerkung 21 *Ist* $\langle\!\langle \cdot \, , \cdot \rangle\!\rangle$ *zusätzlich symmetrisch, so zeigt die letzte Übungsaufgabe, dass*

- $\mathcal{J}(u) = \min_{v \in H} \mathcal{J}(v)$
- $\mathcal{J}(u_h) = \min_{v_h \in X_h} \mathcal{J}(v_h)$

mit $\mathcal{J}(v) := \frac{1}{2}\langle\!\langle v \, , v \rangle\!\rangle - F(v)$ *gelten. In der Physik ist* $\mathcal{J}(\cdot)$ *i.a. eine Energie, und die Natur strebt danach, energieminimierende Zustände einzunehmen. Das Galerkin-Verfahren entspricht daher gerade dem Ersetzen eines unendlich-dimensionalen Minimierungsproblems durch ein endlich-dimensionales Minimierungsproblem.*

Freiwillige Übung. Das Galerkin-Verfahren kann auch für gewisse nichtlineare Gleichungen verwendet werden. Es sei H ein Hilbert-Raum und $A : H \to H^*$ ein nicht-notwendig linearer Operator. Zu gegebenem $F \in H^*$ ist dann ein $u \in H$ mit

$$(Au)(v) = F(v) \quad \text{für alle } v \in H \tag{2.12}$$

gesucht. Analysieren Sie den Beweis vom Lemma von Lax-Milgram, um zu sehen, welche Eigenschaften A erfüllen muss, um eindeutige Lösbarkeit von (2.12) zu garantieren. Sie erhalten den sogenannten *Hauptsatz über stark monotone Operatoren*. Wie sieht in diesem Fall das Galerkin-Verfahren aus? Wie würde man z.B. das Newton-Verfahren verwenden, um die diskrete Lösung $u_h \in X_h$ zu berechnen? ■

2.1.3 L^2-Orthogonalprojektion auf $\mathcal{P}^0(\mathcal{T}_h)$

Wir betrachten $H = L^2(a,b)$ und $\langle\!\langle u\,, v \rangle\!\rangle = \int_a^b uv\,dx$. Gegeben seien Knoten

$$\mathcal{K}_h := \{a = x_0 < \cdots < x_N = b\}$$

und Elemente

$$\mathcal{T}_h = \{T_1, \ldots, T_N\} \quad \text{mit} \quad T_j := [x_{j-1}, x_j].$$

Wir betrachten den diskreten Raum

$$X_h = \mathcal{P}^0(\mathcal{T}_h) := \big\{ f : [a,b] \to \mathbb{R} \,\big|\, \forall T \in \mathcal{T}_h \quad f|_T \text{ konstant} \big\}. \tag{2.13}$$

Eine mögliche Basis $\zeta_j = \chi_{T_j}$ von X_h besteht aus den charakteristischen Funktionen zu $T_j \in \mathcal{T}_h$. In diesem Fall kann man die L^2-Orthogonalprojizierte (d.h. die Galerkin-Approximation) $u_h \in X_h$ leicht berechnen: Für $T_j \in \mathcal{T}_h$ gilt

$$\operatorname{length}(T_j)\, u_h|_{T_j} = \langle\!\langle u_h\,, \chi_{T_j} \rangle\!\rangle = \langle\!\langle u\,, \chi_{T_j} \rangle\!\rangle = \int_{T_j} u\,dx$$

Die L^2-Orthogonalprojizierte ist in diesem Fall also das $\mathcal{T}_h$-stückweise Integralmittel

$$u_h|_{T_j} = \frac{1}{\operatorname{length}(T_j)} \int_{T_j} u\,dx \quad \text{für alle } T_j \in \mathcal{T}_h.$$

In diesem Fall wäre die Galerkin-Matrix $\mathbf{A}$ also insbesondere eine Diagonalmatrix.

2.1.4 L^2-Orthogonalprojektion auf $\mathcal{S}^1(\mathcal{T}_h)$

Wir betrachten wieder $H = L^2(a,b)$ und $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle = \int_a^b uv\,dx$. Gegeben seien Knoten

$$\mathcal{K}_h := \{a = x_1 < \cdots < x_N = b\}$$

und Elemente

$$\mathcal{T}_h = \{T_1, \ldots, T_{N-1}\} \quad \text{mit} \quad T_j := [x_j, x_{j+1}].$$

Wir betrachten den diskreten Raum

$$X_h = \mathcal{S}^1(\mathcal{T}_h) := \big\{ f \in C[a,b] \,\big|\, \forall T \in \mathcal{T}_h \quad f|_T \text{ affin} \big\} \tag{2.14}$$

aller stetigen und stückweise affinen Funktionen. Offensichtlich ist eine diskrete Funktion $v_h \in X_h$ durch ihre Knotenwerte $v_h(x_j)$ für alle $j = 1, \ldots, N$ eindeutig festgelegt. Eine mögliche Basis von X_h bilden

deshalb die Hutfunktionen $\zeta_j \in X_h$, die durch die Kronecker-Eigenschaft $\zeta_j(x_k) = \delta_{jk}$ eindeutig festgelegt sind. Aufgrund der Anordnung der Knoten lässt sich ζ_j explizit als

$$\zeta_j(x) = \begin{cases} 0 & \text{für } x < x_{j-1} \\ \frac{x-x_{j-1}}{x_j-x_{j-1}} & \text{für } x \in [x_{j-1}, x_j] \\ \frac{x-x_{j+1}}{x_j-x_{j+1}} & \text{für } x \in [x_j, x_{j+1}] \\ 0 & \text{für } x > x_{j+1} \end{cases} \tag{2.15}$$

angeben. Offensichtlich gilt deshalb $\mathbf{A}_{jk} = \langle\!\langle \zeta_k \, , \zeta_j \rangle\!\rangle = 0$ für $|j-k| > 1$, d.h. wenn die zugehörigen Knoten $x_j, x_k \in \mathcal{K}_h$ nicht benachbart sind. Insbesondere hat die Galerkin-Matrix also Tridiagonalgestalt

$$\mathbf{A} = \begin{pmatrix} * & * & 0 & \cdots & 0 \\ * & * & * & \ddots & \vdots \\ 0 & * & * & * & 0 \\ \vdots & \ddots & * & * & * \\ 0 & \cdots & 0 & * & * \end{pmatrix} \tag{2.16}$$

und die direkte Lösung, z.B. mittels Gauss-Verfahren, hat lediglich Aufwand $\mathcal{O}(N)$.

Es steht noch aus, die Galerkin-Daten $\mathbf{A}_{jk}$ und $\mathbf{b}_j$ zu berechnen. Dies geschieht regelmäßig elementweise. Auf jedem Element $T_j = [x_j, x_{j+1}]$ haben genau die Hutfunktionen ζ_j und ζ_{j+1} Beiträge zu $\mathbf{A}$. Es gelten

$$\int_{x_j}^{x_{j+1}} \zeta_j \zeta_j \, dx = \int_{x_j}^{x_{j+1}} \zeta_{j+1} \zeta_{j+1} \, dx = \frac{1}{3} (x_{j+1} - x_j) \tag{2.17}$$

sowie

$$\int_{x_j}^{x_{j+1}} \zeta_j \zeta_{j+1} \, dx = \frac{1}{6} (x_{j+1} - x_j). \tag{2.18}$$

Um dies einzusehen verwendet man am besten den Transformationssatz für die affine Transformation $\Phi : [0,1] \to [x_j, x_{j+1}]$, $t \mapsto x_j + t(x_{j+1} - x_j)$. Beachten Sie, dass $\zeta_j \circ \Phi(t) = 1 - t$ und $\zeta_{j+1} \circ \Phi(t) = t$. Es folgt beispielsweise

$$\int_{x_j}^{x_{j+1}} \zeta_j \zeta_{j+1} \, dx = (x_{j+1} - x_j) \int_0^1 t(1-t) \, dt = (x_{j+1} - x_j) \left(\frac{1}{2} - \frac{1}{3} \right) = \frac{1}{6} (x_{j+1} - x_j).$$

Der Aufbau von A (sogenannte *Assemblierung*) erfolgt durch eine Schleife über alle Elemente. Die wesentliche Beobachtung ist die folgende:

$$A_{jk} = \int_a^b \zeta_j \zeta_k \, dx = \sum_{T_\ell \in \mathcal{T}_h} \int_{T_\ell} \zeta_j \zeta_k, dx,$$

wobei wir ausnutzen, dass auf $T_\ell = [x_\ell, x_{\ell+1}]$ nur die Hutfunktionen ζ_ℓ und $\zeta_{\ell+1}$ nicht verschwinden, d.h. man muss nur $j, k \in \{\ell, \ell+1\}$ berücksichtigen. Im MATLAB-ähnlichen Pseudo-Code erhalten wir damit

```
A = sparse(N,N)
b = zeros(N,1)
for j = 1:N-1
  A(j:j+1,j:j+1) = A(j:j+1,j:j+1) + [2 1;1 2]/6*(x(j+1) - x(j))
  b(j:j+1) = b(j:j+1) + ∫_{x_j}^{x_{j+1}} (u ζ_j ; u ζ_{j+1}) dx
end
```

sofern die Knoten in einem Vektor `x` der Länge $N+1$ gespeichert sind. Dabei muss die Integration zur Berechnung der Einträge von $\mathbf{b}$ natürlich noch durch eine geeignete Quadratur ersetzt werden. Beachten

Sie, dass die Schleife über die Elementindizes läuft und die elementweisen Beiträge zu **A** und **b** addiert werden.

Um den Schritt in der Implementierung von 1D FEM auf 2D FEM gering zu halten, verwenden wir den folgenden Overhead: Wir speichern im $N \times 1$-Vektor `coordinates` die Koordinaten der Knoten, d.h. `coordinates(j)` entspricht x_j. Ferner speichern wir in einer $(N-1) \times 2$-Matrix `elements` die Indizes der Knoten eines Elements $T_\ell = [x_k, x_\ell]$ in der Form `elements(j,:) = [k,ell]`. Dadurch kann auch die Sortierung der Knoten entfallen, die in 2D (oder 3D) ohnehin nicht mehr herzustellen ist. Im MATLAB-ähnlichen Pseudo-Code liest sich der Aufbau der Galerkin-Daten dann wie folgt:

```
N = size(coordinates,1)
A = sparse(N,N)
b = zeros(N,1)
for j = 1:size(elements,1)
  nodes = elements(j,:)
  hTj = abs(diff(coordinates(nodes)))
  A(nodes,nodes) = A(nodes,nodes) + [2 1;1 2]/6*hTj
  b(nodes) = b(nodes) + ∫_{T_j} (ζ_{nodes(1)}; ζ_{nodes(2)}) dx
end
```

2.2 Modellproblem

Die Finite Elemente Methode (FEM) ist ein Galerkin-Verfahren zur Lösung gewisser stationärer Differentialgleichungen, in der Regel sogenannter Diffusionsgleichungen, z.B.

$$-(\alpha u')' + \beta u' + \gamma z = f \quad \text{in } (a,b), \tag{2.19}$$

versehen mit gewissen Randbedingungen, z.B. **homogenen Dirichlet-Randbedingungen**

$$u(a) = 0 = u(b). \tag{2.20}$$

Hierbei sind die (eventuell ortsabhängigen) Daten $\alpha, \beta, \gamma, f : (a,b) \to \mathbb{R}$ sowie die Randbedingungen gegeben und die Lösung $u : [a,b] \to \mathbb{R}$ mit (2.19)–(2.20) gesucht. Die Formulierung (2.19)–(2.20) bezeichnet man als **starke Formulierung** der Differentialgleichung, und u heißt ggf. **starke Lösung**. Alternativ zu (2.20) können mit gegebenen $A, B \in \mathbb{R}$ auch (allgemeine) **Dirichlet-Randbedingungen**

$$u(a) = A, \quad u(b) = B, \tag{2.21}$$

Neumann-Randbedingungen

$$(\alpha u')(a) = A, \quad (\alpha u')(b) = B \tag{2.22}$$

und mit $A, B \geq 0$ und zusätzlich gegebenen $u_a, u_b \in \mathbb{R}$ auch **Robin-Randbedingungen**

$$A\big(u(a) - u_a\big) = +(\alpha u')(a), \quad B\big(u(b) - u_b\big) = -(\alpha u')(b) \tag{2.23}$$

vorgegeben werden. Auch gemischte Randbedingungen sind möglich, z.B. **gemischte Dirichlet-Neuman-Randbedingungen**

$$u(a) = A, \quad (\alpha u')(b) = B \quad \text{bzw.} \quad (\alpha u')(a) = A, \quad u(b) = B. \tag{2.24}$$

Um das Galerkin-Verfahren anzuwenden, wird (2.20), versehen mit Randbedingungen, in eine Variationsformulierung umgeformt. Das Vorgehen ist wie folgt:

- Man multipliziert (2.20) mit einer Testfunktion v, die am ggf. vorgegebenen Dirichlet-Rand verschwindet.
- Man integriert die Gleichung über (a,b).
- Man nutzt partielle Integration, um (mindestens) den höchsten Ableitungsterm zu reduzieren.
- Man nutzt die Randbedingungen, um u in den enstehenden Randtermen zu eliminieren.

Im Folgenden betrachten wir zunächst das homogene Dirichlet-Problem (2.19)–(2.20).

2.3 Homogenes Dirichlet-Problem

Es sei $v : [a,b] \to \mathbb{R}$ eine Testfunktion mit $v(a) = 0 = v(b)$. Mittels partieller Integration erhalten wir aus (2.19)–(2.20)

$$F(v) := \int_a^b fv\,dx = \int_a^b \big(-(\alpha u')' + \beta u' + \gamma z\big)v\,dx = \int_b^a \alpha u'v' + \beta u'v + \gamma uv\,dx =: \langle\!\langle u\,, v\rangle\!\rangle.$$

Um das Konzept des Galerkin-Verfahrens anzuwenden, muss man einen geeigneten Hilbert-Raum H definieren und zeigen, dass $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ eine stetige und elliptische Bilinearform auf H ist sowie dass $F \in H^*$ gilt.

2.3.1 Sobolev-Raum in 1D

Wir wiederholen zunächst den Begriff der absolutstetigen Funktion aus der Maßtheorie.

Definition 22 *Eine Funktion $v : [a,b] \to \mathbb{R}$ ist **absolutstetig**, falls gilt*

$$v(x) = A + \int_a^x g\,dt \quad \textit{für alle } x \in [a,b] \tag{2.25}$$

mit einer geeigneten Funktion $g \in L^1(a,b)$ und einer Zahl $A \in \mathbb{R}$.

Die absolutstetigen Funktion sind genau die Funktionen, für die der Hauptsatz der Differential- und Integralrechnung aus der elementaren Analysis gilt:

Satz 23 *Jede absolutstetige Funktion $v : [a,b] \to \mathbb{R}$ ist stetig und fast überall differenzierbar mit $v' = g$ mit der Funktion $g \in L^1(\Omega)$ aus (2.25). Ferner gilt*

$$v(x) = v(\xi) + \int_\xi^x v'\,dt \tag{2.26}$$

für alle $\xi \in [a,b]$. ■

Insbesondere ist jede absolutstetige Funktion beschränkt auf dem Kompaktum $[a,b]$ und damit in $L^p(a,b)$ für alle $1 \le p \le \infty$. Der Raum H für das Galerkin-Verfahren ist ein geeigneter Teilraum des folgenden Sobolev-Raums.

Definition 24 *Der Sobolev-Raum $H^1(a,b)$ ist in 1D durch*

$$H^1(a,b) := \big\{ v : [a,b] \to \mathbb{R} \text{ absolutstetig} \,\big|\, v' \in L^2(a,b) \big\} \tag{2.27}$$

definiert und wird mit der Norm

$$\|v\|_{H^1(a,b)} := \big(\|v\|^2_{L^2(a,b)} + \|v'\|^2_{L^2(a,b)} \big)^{1/2} \tag{2.28}$$

versehen.

Freiwillige Übung. Beweisen Sie, dass $H^1(a,b)$ ein normierter Raum ist. Verwenden Sie dazu, dass die H^1-Norm als ℓ_2-Summe von zwei Normen zusammengesetzt ist sowie die Tatsache, dass Sie bereits wissen, dass die L^2-Norm eine Norm ist. ■

Wir werden im Anschluss auch die Vollständigkeit von $H^1(a,b)$ beweisen. Dazu benötigen wir allerdings die Sobolev-Ungleichung, die im Folgenden zusammen mit zwei anderen zentralen Ungleichungen bewiesen wird.

Lemma 25 (Fundamentale Ungleichungen für Sobolev-Funktionen in 1D)

(i) ***Sobolev-Ungleichung:*** *Die Einbettung $H^1(a,b) \subset C[a,b]$ ist stetig, d.h. es existiert eine Konstante $C_{\text{Sobolev}} > 0$ mit*

$$\|v\|_{\infty,[a,b]} \leq C_{\text{Sobolev}} \, \|v\|_{H^1(a,b)} \tag{2.29}$$

(ii) ***Friedrichs-Ungleichung:*** *Falls $v \in H^1(a,b)$ eine Nullstelle $\xi \in [a,b]$ hat, so gilt*

$$\|v\|_{L^2(a,b)} \leq |b-a| \, \|v'\|_{L^2(a,b)}. \tag{2.30}$$

(iii) ***Poincaré-Ungleichung:*** *Falls $v \in H^1(a,b)$ zusätzlich $\int_a^b v \, dx = 0$ erfüllt, so gilt*

$$\|v\|_{L^2(a,b)} \leq |b-a| \, \|v'\|_{L^2(a,b)}. \tag{2.31}$$

Beweis. Wir beweisen zunächst die Friedrichs-Ungleichung. Es gilt

$$|v(x)| = \Big| \int_\xi^x v' \, dt \Big| \leq |b-a|^{1/2} \|v'\|_{L^2(a,b)}$$

und damit

$$\|v\|^2_{L^2(a,b)} = \int_a^b |v(x)|^2 \leq |b-a|^2 \|v'\|^2_{L^2(a,b)}.$$

Als Nächstes beweisen wir die Poincaré-Ungleichung, indem wir beobachten, dass die stetige Funktion v aufgrund der Integraleigenschaft zwingend eine Nullstelle $\xi \in [a,b]$ hat. Damit ist die Poincaré-Ungleichung ein Spezialfalls der Friedrichs-Ungleichung.

Um die Sobolev-Ungleichung zu beweisen, definieren wir das Integralmittel

$$\overline{v} := \frac{1}{b-a} \int_a^b v \, dx$$

und die Funktion $w := v - \overline{v} \in H^1(a,b)$. Nach Definition gilt $\int_a^b w \, dx = 0$, und deshalb hat v eine Nullstelle $\xi \in [a,b]$. Wie beim Beweis der Friedrichs-Ungleichung erhalten wir

$$|w(x)| \leq (b-a)^{1/2} \|w'\|_{L^2(a,b)} = (b-a)^{1/2} \|v'\|_{L^2(a,b)}.$$

Ferner gilt

$$|\overline{v}| = \frac{1}{b-a} \Big| \int_a^b v \, dt \Big| \leq \frac{1}{(b-a)^{1/2}} \|v\|_{L^2(a,b)}.$$

Insgesamt erhalten wir

$$\begin{aligned} \|v\|_{\infty,[a,b]} \leq \|w\|_{\infty,[a,b]} + |\overline{v}| &\leq (b-a)^{1/2} \|v'\|_{L^2(a,b)} + \frac{1}{(b-a)^{1/2}} \|v\|_{L^2(a,b)} \\ &\leq \Big((b-a) + \frac{1}{(b-a)} \Big)^{1/2} \|v\|_{H^1(a,b)}, \end{aligned}$$

wobei wir zuletzt die Cauchy-Schwarz-Ungleichung in $\mathbb{R}^2$ verwendet haben. ■

Bemerkung 26 *Die Friedrichs- und Poincaré-Ungleichung kann man auch ganz "natürlich" motivieren: üblicherweise ist v eine physikalische Grösse, die z.B. in der Einheit V gemessen wird, und x ein Punkt im Raum, der in der Längenskala L gemessen wird. Die Einheit von $\|v\|^2_{L^2}$ wäre dann V^2L, während $\|v'\|^2_{L^2}$ in V^2L^{-1} gemessen wird. Physikalisch macht es also keinen Sinn, die Norm (2.28) zu definieren. Die Friedrichs- und Poincaré-Ungleichungen sagen nun aber, das man das darf, solange man nur bestimmte Funktionen betrachtet. Man beachte: $|b-a|$ wird wieder in L gemessen, das passt also!*

Satz 27 *Der Sobolev-Raum $H^1(a,b)$ sowie dessen Unterräume $H^1_0(a,b) := \{v \in H^1(a,b) \,|\, v(a) = 0 = v(b)\}$ und $H^1_*(a,b) := \{v \in H^1(a,b) \,|\, \int_a^b v\,dt = 0\}$ sind Hilbert-Räume.*

Beweis. Klarerweise sind sowohl $H^1_0(a,b)$ als auch $H^1_*(a,b)$ Unterräume von $H^1(a,b)$. Aufgrund der Sobolev-Ungleichung ist die Abbildungen $v \mapsto |v(a)| + |v(b)|$ eine stetige Abbildung $H^1(a,b) \to \mathbb{R}$, da Punktauswertung ein stetiges Funktional auf $C[a,b]$ ist. $H^1_0(a,b)$ ist das Urbild von $\{0\}$ unter dieser Abbildung. Da Urbilder abgeschlossener Mengen unter stetigen Funktionen abgeschlossen sind, ist $H^1_0(a,b)$ ein abgeschlossener Unterraum von $H^1(a,b)$. Dasselbe gilt für $H^1_*(a,b)$, da die Abbildung $u \mapsto \int_a^b v\,dx$ ebenfalls stetig ist. Ingesamt müssen wir also nur noch zeigen, dass $H^1(a,b)$ vollständig ist.

Sei (u_n) eine Cauchy-Folge in $H^1(a,b)$. Nach Definition der Norm sind dann (u_n) und (u_n') Cauchy-Folgen in $L^2(a,b)$. Da $L^2(a,b)$ vollständig ist, existieren die L^2-Grenzwerte $u := \lim_n u_n \in L^2(a,b)$ und $g := \lim_n u_n' \in L^2(a,b)$. Für festes $n \in \mathbb{N}$ gilt

$$u_n(x) - u_n(a) = \int_a^x u_n'\,dt = \int_a^x g\,dt + \int_a^x (u_n' - g)\,dt \quad \text{für alle } x \in (a,b).$$

Aufgrund von $g = \lim_n u_n' \in L^2(a,b)$ verschwindet der zweite Term auf der rechten Seite für $n \to \infty$.

Da (u_n) eine Cauchy-Folge in $H^1(a,b)$ ist und die Punktauswertung bei a ein stetiges lineares Funktional auf $H^1(a,b)$, ist die Folge $\big(u_n(a)\big)$ eine Cauchy-Folge in $\mathbb{R}$, und der Grenzwert $A := \lim_n u_n(a) \in \mathbb{R}$ existiert.

Ferner folgt aus der L^2-Konvergenz $u = \lim_n u_n$ für eine Teilfolge $u(x) = \lim_k u_{n_k}(x)$ für fast alle $x \in (a,b)$. Betrachten wir nun diese Teilfolge, so erhalten wir

$$u(x) - A = \int_a^x g\,dt \quad \text{für fast alle } x \in (a,b).$$

Insbesondere stimmt also der L^2-Grenzwert u fast überall —und damit im L^2-Sinn— mit der absolutstetigen Funktion $v(x) := A + \int_a^x g\,dt$ überein. Insbesondere ist $u = v$ absolutstetig mit $u' = g$, d.h. $u \in H^1(a,b)$, und die vorausgesetzten L^2-Konvergenzen zeigen H^1-Konvergenz $u = \lim_n u_n \in H^1(a,b)$. ■

2.3.2 Schwache Formulierung des homogenen Dirichlet-Problems

Zur Lösung des homogenen Dirichlet-Problems verwenden wir den Hilbert-Raum $H := H^1_0(a,b)$. Das folgende Lemma hält die offensichtlichen Eigenschaften von $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ und $F(\cdot)$ fest.

Lemma 28 *Seien $\alpha \in L^\infty(a,b)$, $\beta \in L^2(a,b)$ und $\gamma, f \in L^1(a,b)$. Dann definieren*

$$F(v) := \int_a^b f v\,dx \tag{2.32}$$

ein stetiges und lineares Funktional und

$$\langle\!\langle u\,, v \rangle\!\rangle := \int_b^a \alpha u'v' + \beta u'v + \gamma uv\,dx \tag{2.33}$$

eine stetige Bilinearform auf $H^1(a,b)$.

Beweis. Für $u, v \in H^1(a,b)$ gilt $u, v \in C[a,b] \subset L^\infty(a,b)$ und $u', v' \in L^2(a,b)$. Insbesondere sind alle Integrale wohldefiniert, F ist linear und $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ ist bilinear.

Mit Hölder- und Sobolev-Ungleichung gilt

$$|F(v)| \le \|f\|_{L^1}\|v\|_{L^\infty} \le C_{\text{Sobolev}}\|f\|_{L^1}\|v\|_{H^1} \quad \text{für alle } v \in H^1(a,b),$$

d.h. F ist stetig mit Operatornorm $\|F\| \le C_{\text{Sobolev}}\,\|f\|_{L^1(a,b)}$. Ferner gilt für die Bilinearform

$$\begin{aligned}\langle\!\langle u\,, v \rangle\!\rangle &\le \|\alpha\|_{L^\infty}\|u'\|_{L^2}\|v'\|_{L^2} + \|\beta\|_{L^2}\|u'\|_{L^2}\|v\|_{L^\infty} + \|\gamma\|_{L^1}\|u\|_{L^\infty}\|v\|_{L^\infty} \\ &\le \big(\|\alpha\|_{L^\infty} + C_{\text{Sobolev}}\,\|\beta\|_{L^2} + C^2_{\text{Sobolev}}\,\|\gamma\|_{L^1}\big)\,\|u\|_{H^1}\|v\|_{H^1}\end{aligned}$$

für alle $u, v \in H^1(a,b)$, d.h. $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ ist eine stetige Bilinearform auf $H^1(a,b)$. ■

Um das Lemma von Lax-Milgram anzuwenden, bleibt nur noch die Frage bestehen, ob $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ auch elliptisch auf $H^1_0(a,b)$ ist. Wir geben eine Beispielrechnung: Nach Friedrichs-Ungleichung gilt

$$\|v'\|_{L^2(a,b)} \le \|v\|_{H^1(a,b)} \le (1 + |b-a|^2)^{1/2}\, \|v'\|_{L^2(a,b)},$$

d.h. die Seminorm $\|v'\|_{L^2(a,b)}$ definiert eine äquivalente Norm auf $H^1_0(a,b)$. Insbesondere ist damit

$$\langle\!\langle v\,, v \rangle\!\rangle \gtrsim \|v'\|^2_{L^2(a,b)} \quad \text{für alle } v \in H^1_0(a,b)$$

zu zeigen.

Wir zeigen folgendes hinreichendes (aber i.a. nicht notwendiges) Kriterium für Elliptizität von $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$. Es kann im Einzelfall recht schwierig sein, die Elliptizität zu verifizieren.

Lemma 29 *Die Funktion β sei absolutstetig, und es gelte*

$$\alpha - \varepsilon > C_{\mathrm{ell}} > 0 \quad \textit{und} \quad \gamma - \frac{\beta'}{2} + \frac{\varepsilon}{|b-a|^2} \ge 0 \quad \textit{fast überall in } (a,b) \tag{2.34}$$

mit einer beliebigen Konstante $\varepsilon > 0$. Dann ist $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ elliptisch auf $H^1_0(a,b)$.

Beweis. Da β absolutstetig ist, zeigt partielle Integration

$$\int_a^b \beta u' u \, dx = -\int_a^b \frac{\beta'}{2}\, u^2 \, dx.$$

Damit folgt

$$\langle\!\langle v\,, v \rangle\!\rangle = \int_a^b \alpha (v')^2 \, dx + \int_a^b \Big(\gamma - \frac{\beta'}{2}\Big)\, v^2 \, dx.$$

Insbesondere sehen wir, dass nur das erste Integral die Elliptizität garantieren kann, während das zweite Integral diese de facto nur verhindert. Für $\varepsilon > 0$ zeigt die Friedrichs-Ungleichung

$$\begin{aligned}
\langle\!\langle v\,, v \rangle\!\rangle &= \int_a^b (\alpha - \varepsilon)(v')^2 \, dx + \varepsilon \int_a^b (v')^2 \, dx + \int_a^b \Big(\gamma - \frac{\beta'}{2}\Big)\, v^2 \, dx \\
&\ge \int_a^b (\alpha - \varepsilon)(v')^2 \, dx + \int_a^b \Big(\gamma - \frac{\beta'}{2} + \frac{\varepsilon}{|b-a|^2}\Big)\, v^2 \, dx
\end{aligned}$$

Unter den Voraussetzungen (2.34) erhalten wir schließlich Elliptizität

$$\langle\!\langle v\,, v \rangle\!\rangle \ge C_{\mathrm{ell}} \int_a^b (v')^2 \, dx = C_{\mathrm{ell}} \|v'\|^2_{L^2(a,b)} \quad \text{für alle } v \in H^1_0(a,b).$$

Diese beschließt den Beweis. ■

Der folgende Satz zeigt, dass das Konzept der schwachen Formulierung aus mathematischer Sicht das richtige Konzept ist, um das Modellproblem (2.19) mit homogenen Dirichlet-Randbedingungen (2.20) zu lösen.

Satz 30 (Äquivalenz starker und schwacher Formulierung) *Seien $\alpha \in L^\infty(a,b)$, $\beta \in L^2(a,b)$ und $\gamma, f \in L^1(a,b)$, und die stetige Bilinearform $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ sei elliptisch auf $H^1_0(a,b)$. Dann gelten die folgenden Aussagen:*

(i) *Löst $u \in C^1(a,b) \cap C[a,b]$ mit $\alpha u' \in C^1(a,b) \cap C[a,b]$ die starke Form*

$$-(\alpha u')' + \beta u' + \gamma u = f \ \textit{ in } (a,b) \quad \textit{mit} \quad u(a) = 0 = u(b), \tag{2.35}$$

so gilt $u \in H^1_0(a,b)$, und u löst auch die schwache Formulierung

$$\langle\!\langle u\,, v \rangle\!\rangle = F(v) \quad \textit{für alle } v \in H^1_0(a,b). \tag{2.36}$$

(ii) *Nach Lemma von Lax-Milgram gibt es eine eindeutig Lösung $u \in H_0^1(a,b)$ der schwachen Formulierung.*

(iii) *Falls die schwache Lösung $u \in H_0^1(a,b)$ zusätzliche Regularität $\alpha u' \in C^1(a,b)$ besitzt und die Daten stetig sind, so löst u auch die starke Formulierung.*

Beweis. (i) gilt nach Konstruktion der schwachen Formulierung und aufgrund der Annahmen an die Daten. (ii) folgt aus dem Lemma von Lax-Milgram aufgrund der Annahmen an die Daten. Damit ist nur noch (iii) zu zeigen. Mit Hilfe partieller Integration gilt

$$0 = \langle\!\langle u\,, v \rangle\!\rangle - F(v) = \int_a^b \alpha u'v' + \beta u'v + \gamma uv - fv\,dx = \int_a^b \big(\underbrace{-(\alpha u')' + \beta u' + \gamma u - f}_{=:g \in C(a,b)}\big) v\,dx$$

für alle $v \in H_0^1(a,b)$. Wir müssen zeigen, dass $g = 0$ gilt, und nehmen an, es gibt ein $\xi \in (a,b)$ mit $g(\xi) \neq 0$. Ohne Beschränkung der Allgemeinheit gilt $g(\xi) > 0$. Aufgrund der Stetigkeit von g findet man ein $\varepsilon > 0$, sodass $g > 0$ auf der Umgebung $U_\varepsilon(\xi)$ gilt. Es gibt aber C^1-Funktionen v mit der Eigenschaft $v > 0$ auf $U_\varepsilon(\xi)$ und $v = 0$ auf $[a,b]\backslash U_\varepsilon(\xi)$, z.B. das Polynom vom Grad 4 mit doppelten Nullstellen in $\xi \pm \varepsilon$. Dann folgt aber mit

$$0 = \int_a^b gv\,dx = \int_{\xi-\varepsilon}^{\xi+\varepsilon} gv\,dx > 0$$

ein Widerspruch. ■

Bemerkung 31 *Der vorausgegangene Satz zeigt, dass die schwache Lösung $u \in H_0^1(a,b)$ der einzige Kandidat für die Lösung der starken Formulierung ist und unter gewissen Regularitätsannahmen auch die starke Lösung liefert. Die Regularitätsannahmen in (iii) können mit dem* Hauptsatz der Variationsrechnung

$$\forall g \in L^1_{loc} \quad \Big[\Big(\forall v \in C_c^\infty(a,b) \quad \int_a^b gv\,dx = 0\Big) \Longleftrightarrow g = 0 \text{ fast überall in } (a,b)\Big]$$

abgeschwächt werden. Demnach reicht die zusätzliche Regularität, dass $\alpha u'$ absolutstetig ist, dafür, dass die schwache Lösung $u \in H_0^1(a,b)$ bereits die starke Formulierung fast überall in (a,b) erfüllt.

2.3.3 P1-FEM für homogenes Dirichlet-Problem

Definition 32 *Eine endliche Menge $\mathcal{T}_h = \{T_1, \ldots, T_N\}$ heißt **Triangulierung** von (a,b), falls die Elemente $T_j \in \mathcal{T}_h$ kompakte Intervalle positiver Länge sind mit*

$$\bigcup_{j=1}^N T_j = [a,b] \quad \text{und} \quad \#(T_j \cap T_k) \leq 1 \text{ für alle } j \neq k,$$

d.h. $\mathcal{T}_h$ überdeckt $[a,b]$, und zwei verschiedene Elemente haben maximal einen Knoten gemein. Es bezeichne $\mathcal{K}_h = \{x_1, \ldots, x_{N+1}\}$ die Menge der Knoten von $\mathcal{T}_h$, d.h. zu $T_j \in \mathcal{T}_h$ existieren genau zwei Knoten $x_k, x_\ell \in \mathcal{K}_h$ mit $T_j = [x_k, x_\ell]$.

Nach dem Hauptsatz der Differential- und Integralrechnung vil $C^1[a,b] \subset H^1(a,b)$. Das folgende Lemma besagt, dass eine $\mathcal{T}_h$-stückweise H^1-Funktion (z.B. eine stückweise C^1-Funktion) genau dann in $H^1(a,b)$ liegt, wenn sie stetig ist. Insbesondere zeigt dies, dass $\mathcal{S}^1(\mathcal{T}_h) \subset H^1(a,b)$ gilt.

Lemma 33 *Sei $\mathcal{T}_h = \{T_1, \ldots, T_N\}$ eine Triangulierung von (a,b) und $v : [a,b] \to \mathbb{R}$ mit $v \in H^1(T_j)$ für alle $T_j \in \mathcal{T}_h$. Dann gilt genau dann $v \in H^1(a,b)$, wenn $v \in C[a,b]$ gilt.*

Beweis. Jede Funktion $v \in H^1(a,b)$ ist stetig auf $[a,b]$, so dass nur die umgekehrte Inklusion zu zeigen ist, d.h. zu zeigen ist

$$\exists g \in L^2(a,b)\, \forall x \in (a,b) \quad v(x) = v(a) + \int_a^x g\, dt.$$

Sei $x \in T_j$. Ohne Beschränkung der Allgemeinheit nehmen wir an, dass die Knoten so sortiert sind, dass $T_k = [x_k, x_{k+1}]$ mit $x_k < x_{k+1}$ gilt. Mit der elementweisen Absolutstetigkeit und der Teleskopsumme gilt

$$\begin{aligned} v(x) = v(x_j) + \int_{x_j}^x v'\, dt &= v(x_1) + \sum_{k=1}^{j-1} \big(v(x_{k+1}) - v(x_k)\big) + \int_{x_j}^x v'\, dt \\ &= v(a) + \sum_{k=1}^{j-1} \int_{x_k}^{x_{k+1}} v'\, dt + \int_{x_j}^x v'\, dt \\ &= v(a) + \int_{x_1}^x v'\, dt. \end{aligned}$$

Dies zeigt die Absolutstetigkeit. ∎

Nun können wir das Galerkin-Verfahren für $X_h = \mathcal{S}^1_0(\mathcal{T}_h) := \mathcal{S}^1(\mathcal{T}_h) \cap H^1_0(\mathcal{T}_h)$ verwenden, um eine Galerkin-Approximation $u_h \in \mathcal{S}^1_0(\mathcal{T}_h)$ der schwachen Lösung zu berechnen. Eine Basis von $\mathcal{S}^1_0(\mathcal{T}_h)$ ist durch alle Hutfunktionen $\zeta_j \in \mathcal{S}^1_0(\mathcal{T}_h)$ zu inneren Knoten $x_j \in \mathcal{K}_h \backslash \{a,b\}$ gegeben, d.h. der Raum $\mathcal{S}^1_0(\mathcal{T}_h)$ hat für $\#\mathcal{T}_h = N$ Dimension $\dim \mathcal{S}^1_0(\mathcal{T}_h) = N-1$.

Beachten Sie, dass die rechte Seite $\mathbf{b} \in \mathbb{R}^{N-1}$ im Galerkin-Verfahren nun durch

$$\mathbf{b}_j := \langle\!\langle u\,, \zeta_j \rangle\!\rangle = \int_a^b f \zeta_j\, dx$$

gegeben ist, d.h. man kann die Galerkin-Lösung $u_h \in \mathcal{S}^1_0(\mathcal{T}_h)$ tatsächlich berechnen, ohne die schwache Lösung $u \in H^1_0(a,b)$ zu kennen.

2.3.4 A Priori Analysis und Konvergenz

Nach Céa-Lemma gilt für die Galerkin-Approximation $u_h \in X_h := \mathcal{S}^1_0(\mathcal{T}_h)$

$$\|u - u_h\|_{H^1(a,b)} \lesssim \min_{v_h \in X_h} \|u - v_h\|_{H^1(a,b)}. \tag{2.37}$$

Um eine konkrete a priori Fehlerabschätzung zu erhalten, muss man also nur ein geeignetes $v_h \in \mathcal{S}^1_0(\mathcal{T}_h)$ in Abhängigkeit von u konstruieren.

Lemma 34 (Approximationssatz) *Es sei $v \in H^2(a,b) := \big\{ v \in H^1(a,b) \,\big|\, v' \in H^1(a,b) \big\}$. Es sei*

$$I_h v := \sum_{j=1}^{N+1} v(z_j) \zeta_j \in \mathcal{S}^1(\mathcal{T}_h) \tag{2.38}$$

der nodale Interpolant der stetigen Funktion v. Ferner sei

$$h \in L^\infty(a,b), \quad h|_{T_j} = \operatorname{diam}(T_j) \quad \textit{für alle } T_j \in \mathcal{T}_h \tag{2.39}$$

*die **lokale Netzweite**. Dann gelten die folgenden Fehlerabschätzungen für den Interpolationsfehler:*

(i) $\|v - I_h v\|_{L^2(a,b)} \le \|h(v - I_h v)'\|_{L^2(a,b)} \le \|h v'\|_{L^2(a,b)}$,

(ii) $\|v - I_h v\|_{L^2(a,b)} \le \|h^2 v''\|_{L^2(a,b)}$,

(iii) $\|(v - I_h v)'\|_{L^2(a,b)} \le \|h v''\|_{L^2(a,b)}$.

Beweis. Der Beweis wird elementweise geführt und folgt dann durch Übergang zur ℓ_2-Summe. Wir führen dies im Fall von (i) vor: Sei $T_j \in \mathcal{T}_h$. Da die Funktion $v - I_h v$ auf T_j mindestens eine Nullstelle hat (sogar mindestens zwei), gilt nach Friedrichs-Ungleichung

$$\|v - I_h v\|_{L^2(T_j)} \leq \operatorname{diam}(T_j) \|(v - I_h v)'\|_{L^2(T_j)} = \|h(v - I_h v)'\|_{L^2(T_j)}.$$

Durch Summation der Quadrate erhalten wir

$$\|v - I_h v\|^2_{L^2(a,b)} = \sum_{j=1}^{N} \|v - I_h v\|^2_{L^2(T_j)} \leq \sum_{j=1}^{N} \|h(v - I_h v)'\|^2_{L^2(T_j)} = \|h(v - I_h v)'\|^2_{L^2(a,b)}.$$

Dies ist die erste Abschätzung in (i). Nach Definition des nodalen Interpolanten als $\mathcal{T}_h$-elementweise affine Funktion gilt

$$\mathcal{P}_0(T_j) \ni (I_h v)'|_{T_j} = \frac{v(x_{j+1}) - v(x_j)}{x_{j+1} - x_j} = \frac{1}{x_{j+1} - x_j} \int_{x_j}^{x_{j+1}} v' \, dx.$$

Wir haben bereits oben gesehen, dass Integralmittelbildung die L^2-Orthogonalprojektion Π_h auf $\mathcal{P}_0(T_j)$ ist. Es gilt also die (übrigens nur in 1D gültige) Identität

$$\Pi_h(v')|_{T_j} = (I_h v)'|_{T_j} \quad \text{für alle } T_j \in \mathcal{T}_h. \tag{2.40}$$

Damit erhalten wir

$$\|(v - I_h v)'\|_{L^2(T_j)} = \|v' - \Pi_h(v')\|_{L^2(T_j)} = \min_{\lambda \in \mathbb{R}} \|v' - \lambda\|_{L^2(T_j)} \leq \|v'\|_{L^2(T_j)}.$$

Durch h-gewichtete ℓ_2-Summation folgt die zweite Abschätzung in (i).

Als nächstes zeigen wir (iii). Da $v - I_h v$ auf T_j mindestens zwei verschiedene Nullstellen hat, hat die Ableitung nach Mittelwertsatz mindestens eine Nullstelle in T_j. Mit der Friedrichs-Ungleichung erhalten wir wieder

$$\|(v - I_h v)'\|_{L^2(T_j)} \leq \operatorname{diam}(T_j) \, \|(v - I_h v)''\|_{L^2(T_j)} = \operatorname{diam}(T_j) \, \|v''\|_{L^2(T_j)} = \|h v''\|_{L^2(T_j)}.$$

Durch ℓ_2-Summation folgt (iii).

Abschließend zeigen wir (ii). Die Kombination der bewiesenen lokalen Abschätzungen zeigt

$$\|v - I_h v\|_{L^2(T_j)} \leq \operatorname{diam}(T_j) \, \|(v - I_h v)'\|_{L^2(T_j)} \leq \operatorname{diam}(T_j)^2 \, \|v''\|_{L^2(T_j)} = \|h^2 v''\|_{L^2(T_j)},$$

und ℓ_2-Summation beschließt den Beweis. ■

Bemerkung 35 *Die Aussage von Lemma 34 (i) gilt nur in 1D, wohingegen die Aussagen (ii)-(iii) mit anderem Beweis auch für 2D und 3D korrekt sind.*

Freiwillige Übung. Zeigen Sie, dass in Lemma 34 (i) sogar $\|v - I_h v\|_{L^2(a,b)} \leq \frac{1}{\sqrt{2}} \|h(v - I_h)'\|_{L^2(a,b)}$ gilt, indem Sie den Beweis kritisch analysieren. ■

Freiwillige Übung. Zeigen Sie, dass $W^{1,p}(a,b) := \big\{ v : [a,b] \to \mathbb{R} \text{ absolutstetig} \,\big|\, v' \in L^p(a,b) \big\}$ versehen mit $\|v\|_{W^{1,p}(a,b)} := \big(\|v\|^p_{L^p(a,b)} + \|v'\|^p_{L^p(a,b)} \big)^{1/p}$ ein Banach-Raum ist. Zeigen Sie, dass die Aussage von Lemma 34 (ii)–(iii) auch in L^p-Normen gilt. ■

Freiwillige Übung. Nutzen Sie Ihre Kenntnis über den 1D Interpolationsfehler aus Satz 13, um Lemma 34 (ii) für $p \geq 1$ und $\mathcal{S}^p(\mathcal{T}_h) := \big\{ v_h \in C[a,b] \,\big|\, \forall T \in \mathcal{T}_h \quad v_h|_T \in \mathbb{P}_p \big\}$ zu beweisen. Welche Konvergenzrate $\mathcal{O}(h^\alpha)$ erwarten Sie für $u \in H^1_0(a,b) \cap H^q(a,b)$ in Abhängigkeit von $q \in \mathbb{N}$? ■

Der folgende Satz ist das Hauptresultat in diesem Abschnitt. Er zeigt, dass P^1-FEM für das Modellproblem im Allgemeinen bestenfalls mit linearer Ordnung $\mathcal{O}(h)$ konvergiert. Ferner zeigt er Konvergenz unabhängig von zusätzlicher Regularität der schwachen Lösung. Es ist eine Besonderheit des Galerkin-Verfahrens, dass man auch ohne zusätzliche Annahmen immer Konvergenz erhält, sofern die lokale Netzweite gegen Null geht.

Satz 36 (A Priori Fehlerabschätzung) *Es sei $u \in H^1_0(a,b)$ die schwache Lösung des homogenen Dirichlet-Problems und $u_h \in \mathcal{S}^1_0(\mathcal{T}_h)$ die Galerkin-Lösung zu einer Triangulierung $\mathcal{T}_h$. Dann gelten die folgenden Aussagen*

(i) *Für $\|h\|_{L^\infty(a,b)} \to 0$ folgt Konvergenz $u_h \to u$ in $H^1(a,b)$.*

(ii) *Falls u zusätzliche Regularität $u \in H^1_0(a,b) \cap H^2(a,b)$ erfüllt, so gilt $\|u-u_h\|_{H^1(a,b)} \leq C_1 \|hu''\|_{L^2(a,b)}$.*

(iii) *Es gibt schwache Lösungen $u \in H^1_0(a,b) \cap H^2(a,b)$ mit $\|hu''\|_{L^2(a,b)} \leq C_2 \|u-u_h\|_{H^1(a,b)}$, sodass $u'' \neq 0$ fast überall ist, d.h.* (i) *ist nicht verbesserbar.*

Die Konstante $C_1 > 0$ hängt nur von $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ ab, nicht aber von u oder $\mathcal{T}_h$. Die Konstante $C_2 > 0$ hängt nur von u ab.

Beweis von Satz 36 (ii). Man beachte, dass wegen $I_h u \in \mathcal{S}^1_0(\mathcal{T}_h)$ und Céa-Lemma

$$\|u-u_h\|_{H^1(a,b)} \lesssim \min_{v_h \in X_h} \|u-v_h\|_{H^1(a,b)} \leq \|u - I_h u\|_{H^1(a,b)} \leq (1+\max\{1,(b-a)\}^2)^{1/2} \|hu''\|_{L^2(a,b)}$$

gilt, wobei wir (ii) und (iii) aus Lemma 34 verwendet und für den L^2-Anteil die lokale Netzweite mit $h \leq \max\{1,(b-a)\}$ abgeschätzt haben. ∎

Die anderen beiden Behauptungen in Satz 36 brauchen zwei weitere Lemmata.

Lemma 37 (Inverse Abschätzung) *Für jedes $\mathcal{T}_h$-stückweises Polynom $v_h \in \mathcal{P}^p(\mathcal{T}_h)$ gilt*

$$\|hv_h'\|_{L^2(a,b)} \leq C \|v_h\|_{L^2(a,b)}, \tag{2.41}$$

wobei die Konstante $C > 0$ nur vom Polynomgrad $p \in \mathbb{N}_0$ abhängt.

Beweis. Im ersten Schritt betrachten wir einen endlich-dimensionalen Raum X, auf dem eine Norm $\|\cdot\|$ und eine Halbnorm $|\cdot|$ gegeben seien. Wir betrachten den Kern der Halbnorm $Y := \{x \in X \,|\, |x| = 0\}$ und den zugehörigen Faktorraum X/Y. Auf dem ebenfalls endlich-dimensionalen Raum X/Y definieren

$$\|x+Y\|_{X/Y} := \inf_{y \in Y} \|x+y\| \quad \text{und} \quad |x+Y|_{X/Y} := \inf_{y \in Y} |x+y| = |x|$$

zwei Normen, wie man ggf. leicht nachrechnet. Da alle Normen auf einem endlich-dimensionalen Raum äquivalent sind, existiert also eine Konstante $C > 0$ mit

$$|x| = |x+Y|_{X/Y} \leq C \|x+Y\|_{X/Y} \leq C \|x\| \quad \text{für alle } x \in X.$$

Im zweiten Schritt, wenden wir dies Ergebnis auf $X = \mathbb{P}_p(0,1)$, $\|\cdot\| = \|\cdot\|_{L^2(0,1)}$ und $|\cdot| = \|(\cdot)'\|_{L^2(0,1)}$ an. Es folgt

$$\|w_h'\|_{L^2(0,1)} \leq C \|w_h\|_{L^2(0,1)} \quad \text{für alle } w_h \in \mathbb{P}_p(0,1),$$

wobei $C > 0$ nur von p und vom Intervall $[0,1]$ abhängt.

Der dritte Beweisschritt folgt mittels eines sogenannten *Skalierungsarguments*: Sei $T = [z_j, z_k] \in \mathcal{T}_h$. Wir nutzen die affine Bijektion

$$\Phi_T : [0,1] \to T, \quad \Phi_T(t) = z_j + t(z_k - z_j).$$

Für $v_h \in \mathcal{P}^p$ gelten dann $v_h \circ \Phi_T \in \mathbb{P}_p(0,1)$ und mit Transformationssatz

$$\|v_h\|_{L^2(T)}^2 = \int_a^b v_h^2 \, dx = \operatorname{diam}(T) \int_0^1 (v_h \circ \Phi_T)^2 \, dt = \operatorname{diam}(T) \, \|v_h \circ \Phi_T\|_{L^2(0,1)}^2.$$

Nach Kettenregel gilt $(v_h \circ \Phi_T)'(t) = v_h'(\Phi_T(t))\Phi_T'(t) = \operatorname{diam}(T)\, v_h'(\Phi_T(t))$ und damit

$$\|(v_h \circ \Phi_T)'\|_{L^2(0,1)}^2 = \operatorname{diam}(T)^2 \, \|v_h' \circ \Phi_T\|_{L^2(0,1)}^2 = \operatorname{diam}(T) \, \|v_h'\|_{L^2(T)}^2.$$

Mit Hilfe des zweiten Beweisschritts und $w_h := v_h \circ \Phi_T$ gilt deshalb

$$\operatorname{diam}(T) \, \|v_h'\|_{L^2(T)} = \operatorname{diam}(T)^{1/2} \, \|(v_h \circ \Phi_T)'\|_{L^2(0,1)} \le C \operatorname{diam}(T)^{1/2} \, \|v_h \circ \Phi_T\|_{L^2(0,1)} = C \, \|v_h\|_{L^2(T)},$$

wobei die Konstante $C > 0$ nur von $p \in \mathbb{N}_0$ und dem Referenzintervall $[0,1]$ abhängt.

Insgesamt ist damit

$$\|h v_h'\|_{L^2(T)} = \operatorname{diam}(T) \, \|v_h'\|_{L^2(T)} \le C \, \|v_h\|_{L^2(T)} \quad \text{für alle } T \in \mathcal{T}_h \text{ und } v_h \in \mathcal{P}^p(\mathcal{T})$$

gezeigt, und ℓ_2-Summation beschließt den Beweis. ■

Bemerkung 38 *Das Skalierungsargument, d.h. der Übergang zu einem Referenzintervall, ist nötig, um zu sehen, dass die Konstante $C > 0$ nicht von der Triangularierung $\mathcal{T}_h$ abhängt. Anderenfalls wüsste man nicht, dass C für $h \to 0$ beschränkt bleibt.*

Freiwillige Übung. Wie sieht die inverse Abschätzung (2.41) aus, wenn man $\|p'\|_{L^q(a,b)}$ durch $\|p\|_{L^r(a,b)}$ abschätzt, d.h. wie hängt die h-Potenz an $1 \le p, q \le \infty$? ■

Freiwillige Übung. Zeigen Sie, dass $\|x + Y\|_{X/Y}$ und $|x + Y|_{X/Y}$ im Beweis von Lemma 37 Normen auf X/Y definieren. ■

Freiwillige Übung. Man kann auch Skalierungsargumente zwischen Halbnormen machen. Beweisen Sie das folgende Statement: Es seinen $|\cdot|_1$ und $|\cdot|_2$ Halbnormen auf einem endlich-dimensionalen Raum X. Dann sind die folgenden beiden Aussagen äquivalent:

(i) Es gibt eine Konstante $C > 0$ mit $|x|_1 \le C \, |x|_2$ für alle $x \in X$.

(ii) Es gilt die Inklusion $\big\{x \in X \,\big|\, |x|_2 = 0\big\} \subseteq \big\{x \in X \,\big|\, |x|_1 = 0\big\}$.

Hinweis. Wiederholen Sie das Faktorraum-Argument aus dem Beweis von Lemma 37, um die Aussage auf das dortige Argument zurückzuführen (wenn $|\cdot|_2$ eine Norm wäre). ■

Beweis von Satz 36 (iii). Es sei $u \in \mathcal{P}_2(a,b) \cap H_0^1(a,b)$ eine Blase mit Nullstellen a und b, d.h. $u' \neq 0$ fast überall. Dann gilt

$$\|h u''\|_{L^2(a,b)} = \inf_{v_h \in \mathcal{S}_0^1(\mathcal{T}_h)} \|h(u - v_h)''\|_{L^2(a,b)} \lesssim \inf_{v_h \in \mathcal{S}_0^1(\mathcal{T}_h)} \|(u - v_h)'\|_{L^2(a,b)} \le \|u - u_h\|_{H^1(a,b)},$$

wobei wir die inverse Abschätzung für $(u - v_h)' \in \mathcal{P}^1(\mathcal{T}_h)$ verwendet haben. ■

Das folgende Resultat zeigt man mit Hilfe sogenannter Glättungstechniken, die Üblicherweise in der Vorlesung zu partiellen Differentialgleichungen oder in der Analysis 3 vorgeführt werden. Obwohl in 1D nicht schwierig, ist der Beweis dennoch technisch und länglich. Wir bemerken nur, dass die Inklusionen $C^\infty[a,b] \subset H^1(a,b)$ und $C_c^\infty(a,b) \subset H_0^1(a,b)$ offensichtlich sind.

Lemma 39 (Satz von Meyers-Serrin '64) (i) *$C^\infty[a,b]$ ist ein dichter Teilraum in $H^1(a,b)$.*

(ii) $C_c^\infty(a,b) := \big\{v \in C^\infty[a,b] \,\big|\, \operatorname{supp}(v) \subset (a,b)\big\}$ *ist ein dichter Teilraum von* $H_0^1(a,b)$. ■

Beweis von Satz 36 (i). Wir müssen folgende Quantorenkette verifizieren

$$\exists C > 0 \forall \varepsilon > 0 \exists h_0 > 0 \forall \mathcal{T}_h \text{ mit } \|h\|_{L^\infty(a,b)} \le h_0 \quad \|u - u_h\|_{H^1(a,b)} \le C\,\varepsilon.$$

Sei $\varepsilon > 0$. Nach Satz von Meyers-Serrin existiert ein $v \in C_c^\infty(a,b)$ mit $\|u - v\|_{H^1(a,b)} \le \varepsilon$. Nach Approximationssatz gilt $\|v - I_h v\|_{H^1(a,b)} \lesssim \|hv''\|_{L^2(a,b)}$. Insbesondere existiert also ein $h_0 > 0$ mit $\|v - I_h v\|_{H^1(a,b)} \le \varepsilon$ für alle $\mathcal{T}_h$ mit $\|h\|_{L^\infty(a,b)} \le \varepsilon$. Mit Céa-Lemma und Dreiecksungleichung folgt

$$\|u - u_h\|_{H^1(a,b)} \lesssim \|u - I_h v\|_{H^1(a,b)} \le \|u - v\|_{H^1(a,b)} + \|v - I_h v\|_{H^1(a,b)} \le 2\,\varepsilon,$$

was den Beweis beschließt. ■

Bemerkung 40 *In 1D hat man oft zusätzliche Regularität für die schwache Lösung. Falls*

- α *absolutstetig mit* $0 < \alpha_0 \le \alpha$ *fast überall mit* $\alpha_0 \in \mathbb{R}$,
- $\beta \in L^\infty(a,b)$,
- $\gamma, f \in L^2(a,b)$,

so erfüllt die schwache Lösung $u \in H_0^1(a,b) \cap H^2(a,b)$. *Zum Beweis muss man zeigen, dass* $\delta := -\alpha u'$ *absolutstetig ist. Aufgrund der starken Formulierung folgt dann* $\delta' = f - \beta u' - \gamma u \in L^2(a,b)$ *und damit* $\delta \in H^1(a,b)$. *Mit der Quotientenregel gilt* $u' = -\delta/\alpha \in H^1(a,b)$ *und damit* $u \in H^2(a,b)$.

2.3.5 FEM und Quadratur

In der Praxis müssen die auftretenden Integrale mittels Quadratur berechnet werden. Diesen zusätzlichen Approximationsfehler haben wir bisher nicht berücksichtigt. Dass Galerkin-Verfahren bezüglich Datenstörungen stabil sind, zeigt das folgende Strang-Lemma.

Satz 41 (Erstes Strang-Lemma) *Es sei* $\langle\!\langle \cdot\,,\cdot \rangle\!\rangle$ *eine stetige und elliptische Bilinearform auf einem Hilbert-Raum* H, $F \in H^*$ *und* $u \in H$ *die eindeutige Lösung der Variationsformulierung*

$$\langle\!\langle u\,, v \rangle\!\rangle = F(v) \quad \textit{für alle } v \in H. \tag{2.42}$$

Für $h > 0$ *sei* X_h *ein endlich-dimensionaler Teilraum von* H, F_h *sei ein lineares Funktional auf* X_h *und* $\langle\!\langle \cdot\,,\cdot \rangle\!\rangle_h$ *sei eine Bilinearform auf* X_h. *Es gelte*

$$\lim_{h\to 0} C_h = 0 \quad \textit{mit} \quad C_h := \sup_{v_h, w_h \in X_h \setminus \{0\}} \frac{|\langle\!\langle v_h\,, w_h \rangle\!\rangle - \langle\!\langle v_h\,, w_h \rangle\!\rangle_h|}{\|v_h\|_H \|w_h\|_H}. \tag{2.43}$$

Dann existieren eine (h-unabhängige) Konstante $C_{\mathrm{ell}} > 0$ *und* $h_0 > 0$, *sodass für alle* $h < h_0$ *die diskrete Bilinearform gleichmäßig elliptisch ist auf* X_h, *d.h.*

$$C_{\mathrm{ell}} \|v_h\|_H^2 \le \langle\!\langle v_h\,, v_h \rangle\!\rangle_h \quad \textit{für alle } v_h \in X_h. \tag{2.44}$$

Für $h < h_0$ *garantiert das Lemma von Lax-Milgram insbesondere eine eindeutige gestörte Galerkin-Lösung* $u_h \in X_h$ *von*

$$\langle\!\langle u_h\,, v_h \rangle\!\rangle_h = F_h(v_h) \quad \textit{für alle } v_h \in X_h. \tag{2.45}$$

Ferner gilt die Fehlerabschätzung

$$C^{-1}\,\|u - u_h\|_H \le \inf_{v_h \in X_h} \big(\|u - v_h\|_H + \|\langle\!\langle v_h\,, \cdot \rangle\!\rangle - \langle\!\langle v_h\,, \cdot \rangle\!\rangle_h\|_{X_h^*}\big) + \|F - F_h\|_{X_h^*}, \tag{2.46}$$

wobei die Konstante $C > 0$ *nur von der diskreten Elliptizitätskonstante* $C_{\mathrm{ell}} > 0$ *und der Stetigkeitskonstanten von* $\langle\!\langle \cdot\,,\cdot \rangle\!\rangle$ *abhängt. Insbesondere gilt*

$$C^{-1}\|u - u_h\|_H \le \min_{v_h \in X_h} \|u - v_h\|_H + C_h\,\|u\|_H + \|F - F_h\|_{X_h^*} \tag{2.47}$$

mit der Konstanten $C_h > 0$ *aus* (2.43).

Beweis. Es sei $0 < \varepsilon < C$ mit $C > 0$ der Elliptizitätskonstanten von $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$. Nach Voraussetzung (2.43) existiert dann ein $h_0 > 0$, sodass $C_h \leq \varepsilon$ für alle $h < h_0$ gilt. Es folgt

$$C \, \|v_h\|_H^2 \leq \langle\!\langle v_h \,, v_h \rangle\!\rangle \leq \langle\!\langle v_h \,, v_h \rangle\!\rangle_h + |\langle\!\langle v_h \,, v_h \rangle\!\rangle - \langle\!\langle v_h \,, v_h \rangle\!\rangle_h| \leq \langle\!\langle v_h \,, v_h \rangle\!\rangle_h + C_h \, \|v_h\|_H^2$$

Umsortieren zeigt die Elliptizität

$$(C - \varepsilon) \, \|v_h\|_H^2 \leq \langle\!\langle v_h \,, v_h \rangle\!\rangle_h \quad \text{für alle } v_h \in X_h \text{ und } h < h_0.$$

Dies beweist (2.44) mit $C_{\mathrm{ell}} = C - \varepsilon > 0$. Da (bi-)lineare Abbildungen auf endlich-dimensionalen Räumen *immer* stetig sind, sind die Voraussetzungen des Lemmas von Lax-Milgram erfüllt, und es folgt die eindeutige Existenz von $u_h \in X_h$ mit (2.45).

Sei $v_h \in X_h$. Nach Dreiecksungleichung gilt

$$\|u - u_h\|_H \leq \|u - v_h\|_H + \|v_h - u_h\|_H.$$

Da der erste Summand auf der rechten Seite von (2.46) auftritt, muss nur noch der zweite Summand durch die rechte Seite von (2.46) abgeschätzt werden. Aufgrund der diskreten Elliptizität gilt

$$\|v_h - u_h\|_H^2 \lesssim \langle\!\langle v_h - u_h \,, v_h - u_h \rangle\!\rangle_h = \langle\!\langle v_h \,, v_h - u_h \rangle\!\rangle_h - F_h(v_h - u_h).$$

Wir addieren nun auf der rechten Seite eine Null mit Hilfe der Identität

$$\langle\!\langle u - v_h \,, v_h - u_h \rangle\!\rangle = F(v_h - u_h) - \langle\!\langle v_h \,, v_h - u_h \rangle\!\rangle.$$

Jetzt erhalten wir

$$\begin{aligned}
\|v_h - u_h\|_H^2 &\lesssim \big(\langle\!\langle v_h \,, v_h - u_h \rangle\!\rangle_h - \langle\!\langle v_h \,, v_h - u_h \rangle\!\rangle\big) + \big(F(v_h - u_h) - F_h(v_h - u_h)\big) - \langle\!\langle u - v_h \,, v_h - u_h \rangle\!\rangle \\
&\leq \big(\|\langle\!\langle v_h \,, \cdot \rangle\!\rangle - \langle\!\langle v_h \,, \cdot \rangle\!\rangle_h\|_{X_h^*} + \|F - F_h\|_{X_h^*}\big) \, \|v_h - u_h\|_H - \langle\!\langle u - v_h \,, v_h - u_h \rangle\!\rangle \\
&\lesssim \big(\|\langle\!\langle v_h \,, \cdot \rangle\!\rangle - \langle\!\langle v_h \,, \cdot \rangle\!\rangle_h\|_{X_h^*} + \|F - F_h\|_{X_h^*} + \|u - v_h\|_H\big) \, \|v_h - u_h\|_H,
\end{aligned}$$

wobei wir bei der zweiten Abschätzung die Definition der dualen Norm $\|\cdot\|_{X_h^*}$ und in der dritten die Stetigkeit von $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ verwendet haben. Division durch $\|v_h - u_h\|_H$ zeigt

$$\|v_h - u_h\|_H \lesssim \|\langle\!\langle v_h \,, \cdot \rangle\!\rangle - \langle\!\langle v_h \,, \cdot \rangle\!\rangle_h\|_{X_h^*} + \|F - F_h\|_{X_h^*} + \|u - v_h\|_H$$

Kombination mit der zuerst genannten Dreiecksungleichung und Übergang zum Infimum beschließen den Beweis von (2.46).

Es sei $w_h \in X_h$ mit

$$\|u - w_h\|_H = \inf_{v_h \in X_h} \|u - v_h\|_H.$$

Aus (2.46) und der Definition von C_h folgt insbesondere

$$\begin{aligned}
\|u - u_h\|_H &\lesssim \|u - w_h\|_H + \big(\|\langle\!\langle w_h \,, \cdot \rangle\!\rangle - \langle\!\langle w_h \,, \cdot \rangle\!\rangle_h\|_{X_h^*}\big) + \|F - F_h\|_{X_h^*} \\
&\leq \|u - w_h\|_H + C_h \, \|w_h\|_H + \|F - F_h\|_{X_h^*}.
\end{aligned}$$

Da w_h die Orthogonalprojizierte von u bezüglich des H-Skalarproduktes ist, gilt $\|w_h\|_H \leq \|u\|_H$, was den Beweis beschließt. ■

Freiwillige Übung. Zeigen Sie, dass das Infimum in (2.46) als Minimum angenommen wird. ■

Freiwillige Übung. Folgern Sie, dass die Konvergenzaussage (i) von Satz 36 auch für das gestörte Galerkin-Verfahren gilt, sofern die Datenfehler verschwinden, d.h. $C_h + \|F - F_h\|_{X_h^*} \to 0$ für $h \to 0$. ■

Eine Konsequenz aus dem Strang-Lemma ist, dass man die Integrale der FEM mittels Quadratur berechnen darf, solange die Quadraturfehler

$$\|F - F_h\|_{X_h^*} + \sup_{v_h \in X_h \setminus \{0\}} \frac{\|\langle\!\langle v_h\,, \cdot \rangle\!\rangle - \langle\!\langle v_h\,, \cdot \rangle\!\rangle_h\|_{X_h^*}}{\|v_h\|_H}$$

mindestens mit der Ordnung des Bestapproximationsfehlers

$$\min_{v_h \in X_h} \|u - v_h\|_H$$

fallen. Dies reduziert dann nicht einmal die Konvergenzordnung. Da die P1-FEM nach Satz 36 bestenfalls mit Ordnung $\mathcal{O}(h)$ fällt, werden wir im Folgenden sehen, dass Mittelpunktsquadratur auf uniformen Gittern bereits hinreichend ist.

Lemma 42 *Es sei $f \in H^1(a,b)$ und $F(v) := \int_a^b fv\,dx$. Es sei $\mathcal{T}_h = \{T_1, \dots, T_N\}$ eine uniforme Triangulierung von (a,b) mit $h = \operatorname{diam}(T_j)$ für alle $T_j \in \mathcal{T}_h$. Mit $X_h := \mathcal{S}^1(\mathcal{T}_h) \supset \mathcal{S}^1_0(\mathcal{T}_h)$ und*

$$F_h(v_h) := \sum_{j=1}^N \operatorname{diam}(T_j) f(m_j) v_h(m_j) \quad \text{für } T_j = [x_j, x_{j+1}] \text{ und } m_j = \frac{x_j + x_{j+1}}{2} \tag{2.48}$$

gilt dann

$$\|F - F_h\|_{X_h^*} \le \|hf'\|_{L^2(\Omega)} = \mathcal{O}(h).$$

Beweis. Es gilt

$$F(v_h) = \sum_{j=1}^N \int_{x_j}^{x_{j+1}} f v_h\,dx \quad \text{und} \quad \operatorname{diam}(T_j) v_h(m_j) = \int_{x_j}^{x_{j+1}} v_h\,dx \quad \text{für alle } v_h \in X_h.$$

Mit der Friedrichs-Ungleichung folgt

$$\begin{aligned}\Big| \int_{x_j}^{x_{j+1}} f v_h\,dx - \operatorname{diam}(T_j) f(m_j) v_h(m_j) \Big| = \Big| \int_{x_j}^{x_{j+1}} \big(f - f(m_j)\big) v_h\,dx \Big| &\le \|f - f(m_j)\|_{L^2(T_j)} \|v_h\|_{L^2(T_j)} \\ &\le \operatorname{diam}(T_j) \|f'\|_{L^2(T_j)} \|v_h\|_{L^2(T_j)}.\end{aligned}$$

Summation über alle Elemente und Cauchy-Schwarz-Ungleichung im $\mathbb{R}^N$ zeigen nun

$$\begin{aligned}|F(v_h) - F_h(v_h)| &\le \sum_{j=1}^N \Big| \int_{x_j}^{x_{j+1}} f v_h\,dx - \operatorname{diam}(T_j) f(m_j) v_h(m_j) \Big| \le \sum_{j=1}^N \|hf'\|_{L^2(T_j)} \|v_h\|_{L^2(T_j)} \\ &\le \Big(\sum_{j=1}^N \|hf'\|^2_{L^2(T_j)} \Big)^{1/2} \Big(\sum_{j=1}^N \|v_h\|^2_{L^2(T_j)} \Big)^{1/2} = \|hf'\|_{L^2(a,b)} \|v_h\|_{L^2(a,b)}.\end{aligned}$$

Wegen $\|v_h\|_{L^2(a,b)} \le \|v_h\|_{H^1(a,b)}$ erhalten wir also insgesamt $\|F - F_h\|_{X_h^*} \le \|hf'\|_{L^2(a,b)}$. ∎

Freiwillige Übung. Beweisen Sie, dass —unter gewissen Regularitätsannahmen an α, β und γ— das Integral der Bilinearform $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ ebenfalls durch $\mathcal{T}_h$-elementweise Mittelpunktsquadratur ersetzt werden darf. Betrachten Sie dazu alle Bestandteile des Integrals, z.B. $\int_a^b \alpha u_h' v_h'\,dx$, getrennt. ∎

Bemerkung 43 *Man kann übrigens zeigen, dass für $f \in H^2(a,b)$ die $\mathcal{T}_h$-elementweise Mittelpunktsquadratur bereits auf*

$$\|F - F_h\|_{X_h^*} = \mathcal{O}(h^2).$$

führt, d.h. verglichen mit dem Verfahrensfehler von höherer Ordnung ist. Um dies zu sehen, definieren wir $p := I_h f \in \mathcal{S}^1(\mathcal{T}_h)$*. Wie oben gilt elementweise*

$$\int_{x_j}^{x_{j+1}} f v_h \, dx - \operatorname{diam}(T_j) f(m_j) v_h(m_j) =: \langle f - f(m_j) \, , v_h \rangle_{T_j}$$
$$= \langle f - p \, , v_h \rangle_{T_j} + \langle f(m_j) - p(m_j) \, , v_h \rangle_{T_j} + \langle p - p(m_j) \, , v_h \rangle_{T_j}.$$

Wir betrachten alle drei Terme auf der rechten Seite getrennt. Nach Approximationssatz gilt

$$\langle f - p \, , v_h \rangle_{T_j} \leq \|f - p\|_{L^2(T_j)} \|v_h\|_{L^2(T_j)} \leq \|h^2 f''\|_{L^2(T_j)} \|v_h\|_{L^2(T_j)}.$$

Für den zweiten Summanden gilt analog

$$\langle f(m_j) - p(m_j) \, , v_h \rangle_{T_j} \leq \big| f(m_j) - p(m_j) \big| \|v_h\|_{L^1(T_j)} = \Big| \int_{x_j}^{m_j} f' - p' \, dx \Big| \|v_h\|_{L^1(T_j)}$$
$$\leq \operatorname{diam}(T_j) \|f' - p'\|_{L^2(T_j)} \|v_h\|_{L^2(T_j)} \leq \|h^2 f''\|_{L^2(T_j)} \|v_h\|_{L^2(T_j)}.$$

Für den dritten Summanden benutzen wir, dass das Integralmittel

$$\overline{p} := \frac{1}{\operatorname{diam}(T_j)} \int_{x_j}^{x_{j+1}} p \, dx = p(m_j)$$

die $\mathcal{P}_0(T_j)$*-Bestapproximation in* $L^2(T_j)$ *ist. Insbesondere ist also* $p - \overline{p} = p - p(m_j)$ *orthogonal auf* $\mathcal{P}_0(T_j)$*. Mit dieser Beobachtung und dem Integralmittel* $\overline{v}_h$ *von* v_h *gilt*

$$\langle p - p(m_j) \, , v_h \rangle_{T_j} = \langle p - \overline{p} \, , v_h - \overline{v}_h \rangle_{T_j} \leq \|p - \overline{p}\|_{L^2(T_j)} \|v_h - \overline{v}_h\|_{L^2(T_j)}$$
$$\leq \operatorname{diam}(T_j)^2 \|p'\|_{L^2(T_j)} \|v_h'\|_{L^2(T_j)} \leq \operatorname{diam}(T_j)^2 \|f'\|_{L^2(T_j)} \|v_h'\|_{L^2(T_j)} = \|h^2 f'\|_{L^2(T_j)} \|v_h'\|_{L^2(T_j)},$$

wobei wir die Poincaré-Ungleichung sowie die Abschätzung $\|p'\|_{L^2(T_j)} \leq \|f'\|_{L^2(T_j)}$ *verwendet haben. Letzteres folgt, weil* $p' = (I_h f)' = \Pi_h(f')$ *die Orthogonalprojizierte von* f' *auf* $\mathcal{P}^0(T_j)$ *in* $L^2(T_j)$ *ist. Durch Summation über alle Elemente erhält man wieder die Behauptung.*

2.3.6 A Posteriori Analysis und lokale Netzverfeinerung

Bei der a priori Analysis interessiert man sich für Fehlerschranken für den Galerkin-Fehler $\|u - u_h\|_H$ in Abhängigkeit von der Regularität von u und *ohne Kenntnis* einer diskreten Lösung u_h. Diese Fehlerschranken sind in der Regel in der Praxis nicht berechenbar, aber implizieren Aussagen über die Konvergenz des Verfahrens. Bei der a posteriori Analysis interessiert man sich für Fehlerschranken für den Galerkin-Fehler $\|u - u_h\|_H$ für eine *konkret berechnete* Galerkin-Lösung u_h. In der Regel interessiert man sich dabei nur für Fehlerschranken, die mit Hilfe von u_h numerisch berechnet werden können. Lokalisiert man diese Fehlerschranken, so erhält man eine heuristische Information, wo man das Netz lokal verfeinern sollte, um den Fehler wesentlich zu verringern.

Das Hauptaugenmerk dieses Abschnitts liegt darin, eine geeignete Folge von Triangulierungen $(\mathcal{T}_\ell)_{\ell \in \mathbb{N}}$ zu generieren, wobei $\mathcal{T}_{\ell+1}$ jeweils aus $\mathcal{T}_\ell$ durch geeignete lokale Verfeinerung entsteht. Aus diesem Grund ändern wir in diesem Abschnitt die Notation. Diskrete Größen haben nun stets den Index ℓ anstelle von h, z.B. bezeichnet $u_\ell \in X_\ell$ eine Galerkin-Lösung im diskreten Raum $X_\ell = \mathcal{S}^1_0(\mathcal{T}_\ell)$.

Wir formulieren zunächst die Idee eines *adaptiven Algorithmus*: Es sei $\mathcal{T}_\ell = \{T_1, \ldots, T_N\}$ eine Triangulierung von (a, b) und

$$\rho_\ell := \Big(\sum_{j=1}^{N} \rho_\ell(T)^2 \Big)^{1/2} \tag{2.49}$$

ein **a posteriori Fehlerschätzer** mit berechenbaren lokalen **Verfeinerungsindikatoren** $\rho_\ell(T)$, die (zumindest heuristisch)

$$\rho_\ell(T) \approx \|u - u_\ell\|_{H^1(T)} \quad \text{für alle } T \in \mathcal{T}_\ell \tag{2.50}$$

erfüllen. In diesem Fall ist dann ρ_ℓ eine Approximation von $\|u - u_h\|_{H^1(a,b)}$.

Sei nun $0 < \theta \leq 1$ ein gegebener Adaptivitätsparameter, $\mathcal{T}_0$ eine gegebene (grobe) Anfangstriangulierung von (a, b) und $\ell := 0$. Der übliche **adaptive Algorithmus** liest sich dann wie folgt:

(i) Berechne Galerkin-Lösung $u_\ell \in X_\ell := \mathcal{S}^1_0(\mathcal{T}_\ell)$.

(ii) Berechne Verfeinerungsindikatoren $\rho_\ell(T)$ für alle $T \in \mathcal{T}_\ell$.

(iii) Stop, falls ρ_ℓ hinreichend klein.

(iv) Anderenfalls, bestimme eine minimale Menge $\mathcal{M}_\ell \subseteq \mathcal{T}_\ell$ mit

$$\theta\, \rho_\ell^2 = \theta \sum_{T\in\mathcal{T}_\ell} \rho_\ell(T)^2 \le \sum_{T\in\mathcal{M}_\ell} \rho_\ell(T)^2. \tag{2.51}$$

(v) Erzeuge eine neue Triangulierung $\mathcal{T}_{\ell+1}$, indem alle markierten Elemente $T \in \mathcal{M}_\ell$ halbiert werden.

(vi) Erhöhe Zähler $\ell \mapsto \ell + 1$ und Sprung nach (i).

Das klassische Beispiel dieser Idee ist der sogennante **$(h - h/2)$-Fehlerschätzer**, der häufig bei gewöhnlichen Differentialgleichungen verwendet wird: Es sei $\mathcal{T}_\ell$ eine gegebene Triangulierung und $\widehat{\mathcal{T}}_\ell$ deren uniforme Verfeinerung, d.h. alle Elemente von $\mathcal{T}_\ell$ werden halbiert. Es seien $u_\ell \in X_\ell = \mathcal{S}^1_0(\mathcal{T}_\ell)$ und $\widehat{u}_\ell \in \widehat{X}_\ell := \mathcal{S}^1_0(\widehat{\mathcal{T}}_\ell)$ die zugehörigen Galerkin-Lösungen. Heuristisch ist $\widehat{u}_\ell$ dann eine wesentlich bessere Approximation an u als u_ℓ und deshalb nimmt man

$$\rho_\ell(T) := \|\widehat{u}_\ell - u_\ell\|_{H^1(T)} \approx \|u - u_\ell\|_{H^1(T)} \quad \text{für alle } T \in \mathcal{T}_\ell \tag{2.52}$$

an.

Lemma 44 *Der $(h - h/2)$-Fehlerschätzer ρ_ℓ ist immer **effizient**, d.h. es gilt*

$$\rho_\ell \le C_{\text{eff}}\, \|u - u_\ell\|_{H^1(a,b)}. \tag{2.53}$$

*Die Konstante $C_{\text{eff}} > 0$ hängt dabei nur von der zugrundeliegenden Bilinearform $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ ab. Unter der sogennanten **Saturationsannahme***

$$\|u - \widehat{u}_\ell\|_{H^1(a,b)} \le q\, \|u - u_\ell\|_{H^1(a,b)} \quad \textit{mit einer } \ell\textit{-unabhängigen Konstante } 0 < q < 1 \tag{2.54}$$

*ist ρ_ℓ auch **zuverlässig**, d.h. es gilt*

$$\|u - u_\ell\|_{H^1(a,b)} \le C_{\text{rel}}\, \rho_\ell \quad \textit{mit } C_{\text{rel}} = \frac{1}{1-q}. \tag{2.55}$$

Ist $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ symmetrisch, so ist die Zuverlässigkeit (2.55) *äquivalent zur Saturationsannahme in der induzierten Energienorm $|\!|\!| \cdot |\!|\!|$*

$$|\!|\!| u - \widehat{u}_\ell |\!|\!| \le q\, |\!|\!| u - u_\ell |\!|\!| \quad \textit{mit einer } \ell\textit{-unabhängigen Konstante } 0 < q < 1. \tag{2.56}$$

In diesem Fall impliziert also die Saturationsannahme (2.54) *in der H^1-Norm die Saturationsannahme* (2.56) *in der Energienorm. Ferner gilt für den äquivalenten $(h-h/2)$-Fehlerschätzer $\varrho_\ell = |\!|\!| \widehat{u}_\ell - u_\ell |\!|\!| \simeq \rho_\ell$ die Effizienz auch in der konstante-freien Form*

$$\varrho_\ell \le |\!|\!| u - u_\ell |\!|\!|. \tag{2.57}$$

Beweis. Mit Dreiecksungleichung und Céa-Lemma folgt aufgrund der Schachtelung $X_\ell \subset \widehat{X}_\ell$

$$\rho_\ell = \|\widehat{u}_\ell - u_\ell\|_{H^1(a,b)} \le \|u - \widehat{u}_\ell\|_{H^1(a,b)} + \|u - u_\ell\|_{H^1(a,b)} \le (C_{\text{Cea}} + 1)\, \|u - u_\ell\|_{H^1(a,b)}.$$

Mit der Saturationsannahme (2.54) gilt sogar

$$\|\widehat{u}_\ell - u_\ell\|_{H^1(a,b)} \le (q+1)\, \|u - u_\ell\|_{H^1(a,b)},$$

und die Zuverlässigkeit (2.55) folgt aus

$$\|u - u_\ell\|_{H^1(a,b)} \le \|\widehat{u}_\ell - u_\ell\|_{H^1(a,b)} + \|u - \widehat{u}_\ell\|_{H^1(a,b)} \le \rho_\ell + q\, \|u - u_\ell\|_{H^1(a,b)}$$

mit $C_{\rm rel} = 1/(1-q)$. Ist $\langle\!\langle \cdot\,,\cdot \rangle\!\rangle$ symmetrisch, so erfüllt die induzierte Norm aufgrund der Galerkin-Orthogonalität den Satz von Pythagoras

$$|\!|\!| u - u_\ell |\!|\!|^2 + |\!|\!| v_\ell |\!|\!|^2 = |\!|\!| u - (u_\ell \pm v_\ell) |\!|\!|^2 \quad \text{für alle } v_\ell \in X_\ell.$$

Wendet man dieses auf $X_\ell \subset \widehat{X}_\ell$ an, so sehen wir

$$\varrho_\ell^2 \le |\!|\!| u - \widehat{u}_\ell |\!|\!|^2 + \varrho_\ell^2 = |\!|\!| u - u_\ell |\!|\!|^2.$$

Da $|\!|\!| \cdot |\!|\!|$ eine äquivalente Norm auf $H^1(a,b)$ ist, ist (2.55) äquivalent zu

$$|\!|\!| u - u_\ell |\!|\!| \le C\, \varrho_\ell \tag{2.58}$$

mit einer Konstanten $C \ge 1$, und wir erhalten die Saturationsannahme (2.56) aus

$$|\!|\!| u - \widehat{u}_\ell |\!|\!|^2 = |\!|\!| u - u_\ell |\!|\!|^2 - \varrho_\ell^2 \le (1 - C^{-2})\, |\!|\!| u - u_\ell |\!|\!|^2$$

mit $q = 1 - C^{-2} \in [0,1)$. Auf der anderen Seite impliziert die Saturationsannahme (2.56) wegen

$$|\!|\!| u - u_\ell |\!|\!|^2 = |\!|\!| u - \widehat{u}_\ell |\!|\!|^2 + \varrho_\ell^2 \le q^2\, |\!|\!| u - u_\ell |\!|\!|^2 + \varrho_\ell^2,$$

die Zuverlässigkeit (2.58) mit $C = 1/(1-q^2)^{1/2}$. ■

Bemerkung 45 *Formal folgt die Saturationsannahme* (2.54) *aus dem asymptotischen Konvergenzverhalten des numerischen Verfahrens: Mit dem formalen Ansatz* $\|u - u_\ell\|_{H^1(a,b)} = CN^{-\alpha}$ *mit* $C, \alpha > 0$ *und* $N = \#\mathcal{T}_\ell$ *folgt wegen* $\|u - \widehat{u}_\ell\|_{H^1(a,b)} = C(\#\widehat{\mathcal{T}}_\ell)^{-\alpha} = 2^{-\alpha}\, \|u - u_\ell\|_{H^1(a,b)}$ *die Saturationsannahme mit* $q = 2^{-\alpha}$. *Allerdings lässt sich im Allgemeinen nicht vorhersagen, wann das asymptotische Konvergenzverhalten einsetzt.*

Freiwillige Übung. Es seien $X_\ell \subset \widehat{X}_\ell$ endlich-dimensionale Teilräume eines Hilbert-Raums H und $\langle\!\langle \cdot\,,\cdot \rangle\!\rangle$ eine stetige und elliptische Bilinearform auf H. Schließlich seien $u_\ell \in X_\ell$ und $\widehat{u}_\ell \in \widehat{X}_\ell$ Galerkin-Approximationen einer Funktion $u \in H$. Zeigen Sie, dass u_ℓ auch eine Galerkin-Approximation von $\widehat{u}_\ell$ ist. ■

Ein Nachteil des Schätzers ρ_ℓ ist, dass sowohl die Galerkin-Lösung u_ℓ als auch die genauere Galerkin-Lösung $\widehat{u}_\ell$ berechnet werden müssen. Jede vernünftige Implementierung wird allerdings nur $\widehat{u}_\ell$ zurückgeben. Das folgende Lemma zeigt, dass auf die Berechnung von u_ℓ tatsächlich auch verzichtet werden kann. Es wird ein zu ρ_ℓ äquivalenter Schätzer $\widetilde{\rho}_\ell$ eingeführt.

Lemma 46 *Es gibt Konstanten* $C_1, C_2 > 0$, *die weder von* $\mathcal{T}_\ell$ *noch von einer der Galerkin-Lösungen abhängen, sodass gilt*

$$\widetilde{\rho}_\ell := \|\widehat{u}_\ell - I_\ell \widehat{u}_\ell\|_{H^1(a,b)} \le C_1\, \rho_\ell \le C_2\, \widetilde{\rho}_\ell, \tag{2.59}$$

wobei $I_\ell : C[a,b] \to \mathcal{S}^1(\mathcal{T}_\ell)$ *die nodale Interpolation bezeichne.*

Beweis. Wir nutzen wieder die Identität $(I_\ell v)' = \Pi_\ell(v')$ mit $\Pi_\ell : L^2(\Omega) \to \mathcal{P}^0(\mathcal{T}_\ell)$ der L^2-Orthogonalprojektion. Mit der Friedrichs-Ungleichung gilt Normäquivalenz $\|v\|_{H^1(a,b)} \eqsim \|v'\|_{L^2(a,b)}$ für alle $v \in H^1_0(a,b)$. Es folgt

$$\widetilde{\rho}_\ell \eqsim \|(\widehat{u}_\ell - I_\ell \widehat{u}_\ell)'\|_{L^2(a,b)} = \min_{g_\ell \in \mathcal{P}^0(\mathcal{T}_\ell)} \|\widehat{u}_\ell' - g_\ell\|_{L^2(a,b)} \le \|(\widehat{u}_\ell - u_\ell)'\|_{L^2(a,b)} \eqsim \rho_\ell.$$

Auf der anderen Seite gilt nach Céa Lemma

$$\rho_\ell = \|\widehat{u}_\ell - u_\ell\|_{H^1(a,b)} \lesssim \min_{v_\ell \in S^1_0(\mathcal{T}_\ell)} \|\widehat{u}_\ell - v_\ell\|_{H^1(a,b)} \le \widetilde{\rho}_\ell,$$

wobei wir $X_\ell \subset \widehat{X}_\ell$ und $I_\ell \widehat{u}_\ell \in X_\ell$ ausgenutzt haben. ■

Bemerkung 47 *Das Konzept der $(h-h/2)$-Fehlerschätzer ρ_ℓ und $\widetilde{\rho}_\ell$ funktioniert mit jeder äquivalenten Norm auf dem Energieraum. Für das homogene Dirichlet-Problem kann deshalb anstatt der vollen H^1-Norm $\|\cdot\|_{H^1(a,b)}$ auch die Seminorm $\|(\cdot)'\|_{L^2(a,b)}$ verwendet werden. Für symmetrische Bilinearformen ist es in der Regel sinnvoller, bei der Definition von ρ_ℓ bzw. $\widetilde{\rho}_\ell$ die Energienorm $|||\cdot|||$ zu betrachten.*

Freiwillige Übung. Nutzen Sie die Beweistechnik der inversen Abschätzung, um zu zeigen, dass auch ohne Nullrandbedingung (und somit ohne Friedrichsungleichung) die Äquivalenz $\widetilde{\rho}_\ell \simeq \rho_\ell$ gilt. Zeigen Sie dazu, dass für diskrete Funktionen $\widehat{v}_\ell \in \mathcal{S}^1(\widehat{\mathcal{T}}_\ell)$ die nodale Interpolation I_ℓ elementweise quasi-optimal bezüglich der $H^1(T_j)$-Norm ist. ■

Da der $(h - h/2)$-Schätzer nur unter der Voraussetzung der Saturationsannahme zuverlässig ist, betrachten wir noch einen weiteren Schätzer, der immer zuverlässig ist.

Lemma 48 (Gewichteter Residualschätzer) *Wir setzen voraus, dass die gegebenen Daten $\alpha \in H^1(a,b)$, $\beta \in L^\infty(a,b)$ sowie $f, \gamma \in L^2(a,b)$ erfüllen. Der gewichtete Residualschätzer ρ_ℓ mit lokalen Beiträgen*

$$\rho_\ell(T) := \|h_\ell(f + (\alpha u_\ell')' - \beta u_\ell' - \gamma u_\ell)\|_{L^2(T)} \quad \text{für } T \in \mathcal{T}_\ell \tag{2.60}$$

ist immer zuverlässig, d.h. es gilt

$$C_{\mathrm{rel}}^{-1} \|u - u_\ell\|_{H^1(a,b)} \le \rho_\ell := \Big(\sum_{T \in \mathcal{T}_\ell} \rho_\ell(T)^2 \Big)^{1/2} \tag{2.61}$$

mit einer Konstanten $C_{\mathrm{rel}} > 0$, die nur von der Elliptizität von $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ abhängt.

Beweis. Aufgrund der Elliptizität und der Galerkin-Orthogonalität gilt

$$\|u - u_\ell\|_{H^1(a,b)}^2 \lesssim \langle\!\langle u - u_\ell \,, u - u_\ell \rangle\!\rangle = \langle\!\langle u - u_\ell \,, u - u_\ell - v_\ell \rangle\!\rangle \quad \text{für alle } v_\ell \in X_\ell = \mathcal{S}_0^1(\mathcal{T}_\ell).$$

Im Folgenden wählen wir $v_\ell := I_\ell(u - u_\ell) \in \mathcal{S}_0^1(\mathcal{T}_\ell)$ und definieren

$$w := u - u_\ell - v_\ell = (1 - I_\ell)(u - u_\ell) \in H_0^1(a,b).$$

Insbesondere gilt $w(x_k) = 0$ für alle Knoten $x_k \in \mathcal{K}_\ell$. Nach Definition von u erhalten wir mit partieller Integration

$$\begin{aligned}
\|u - u_\ell\|_{H^1(a,b)}^2 \lesssim \langle\!\langle u - u_\ell \,, w \rangle\!\rangle &= F(w) - \langle\!\langle u_\ell \,, w \rangle\!\rangle \\
&= \sum_{j=1}^N \int_{x_j}^{x_{j+1}} (-\alpha u_\ell' w')\, dx + \int_{x_j}^{x_{j+1}} (f - \beta u_\ell' - \gamma u_\ell) w \, dx \\
&= \sum_{j=1}^N \int_{x_j}^{x_{j+1}} (\alpha u_\ell')' w \, dx + \int_{x_j}^{x_{j+1}} (f - \beta u_\ell' - \gamma u_\ell) w \, dx \\
&= \sum_{j=1}^N \int_{x_j}^{x_{j+1}} \big(f + (\alpha u_\ell')' - \beta u_\ell' - \gamma u_\ell\big) w \, dx \\
&\le \sum_{j=1}^N \|f + (\alpha u_\ell')' - \beta u_\ell' - \gamma u_\ell\|_{L^2(T_j)} \|w\|_{L^2(T_j)}
\end{aligned}$$

Nach Wahl von w gilt $\|w\|_{L^2(T_j)} \le \operatorname{diam}(T_j) \|(u - u_\ell)'\|_{L^2(T_j)}$, und wir erhalten

$$\|u - u_\ell\|_{H^1(a,b)}^2 \lesssim \sum_{j=1}^N \operatorname{diam}(T_j) \|f + (\alpha u_\ell')' - \beta u_\ell' - \gamma u_\ell\|_{L^2(T_j)} \|(u - u_\ell)'\|_{L^2(T_j)} \le \rho_\ell \, \|u - u_\ell\|_{H^1(a,b)},$$

wobei wir abschließend die Cauchy-Schwarz-Ungleichung in $\mathbb{R}^N$ sowie $\|(\cdot)'\|_{L^2(a,b)} \le \|\cdot\|_{H^1(a,b)}$ verwendet haben. Division durch $\|u - u_\ell\|_{H^1(a,b)}$ beschließt den Beweis. ■

Bemerkung 49 *In 1D ist der residuale Fehlerschätzer wenig sinnvoll, da unter den gegebenen Voraussetzungen bereits $u \in H^1_0(a,b) \cap H^2(a,b)$ gilt und uniforme Netzverfeinerung deshalb (asymptotisch) optimal ist. Dies ist anders in 2D, wo die Geomtrie H^2-Regularität im Allgemeinen verhindert. Da H^1-Funktionen in 2D im Allgemeinen allerdings* nicht *stetig sind, ist die Verwendung der nodalen Interpolation I_ℓ nicht erlaubt. Dies führt auf zusätzliche Randintegrale, die in 1D wegen $w(x_k) = 0$ für alle $x_k \in \mathcal{K}_\ell$ wegfallen.*

2.3.7 Konvergenz adaptiver Verfahren

Uniforme Netzverfeinerung garantiert $\|h_\ell\|_{L^\infty(a,b)} \to 0$ und nach Satz 36 deshalb Konvergenz $u_\ell \to u$ in $H^1_0(a,b)$ für $\ell \to \infty$. Bei adaptiver Netzverfeinerung gilt im Allgemeinen $\|h_\ell\|_{L^\infty(a,b)} \not\to 0$, und deshalb ist die Konvergenz von u_ℓ gegen u nicht gesichert. In diesem Abschnitt wollen wir einige neuere Forschungsergebnisse vorstellen, die sich mit dieser Frage beschäftigen.

Historisch war Dörfler (1996) der erste, der sich mit dieser Frage beschäftigt hat. Er betrachtete in 2D das Modellproblem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \Gamma \end{aligned}$$

mit dem zugehörigen Residualschätzer. Dies entspricht in 1D im Wesentlichen dem Fall $\alpha \in \mathbb{R}_{>0}$ und $\beta = 0 = \gamma$. Er führte die Markierungsstrategie (2.51) ein und bewies, dass adaptive FEM für $\|h_\ell f\|_{L^2(\Omega)} \leq \varepsilon$ bis auf einen Fehler $\limsup_\ell |\!|\!| u - u_\ell |\!|\!| \simeq \varepsilon$ konvergiert. Dieses Ergebnis wurde von Morin, Nochetto und Siebert (2000) darin verbessert, dass sie zeigten, dass die Konvergenz der Oszillationen $\|h_\ell(f - f_\ell)\|_{L^2(\Omega)} \to 0$ mit $f_\ell \in \mathcal{P}^0(\mathcal{T}_\ell)$ dem $\mathcal{T}_\ell$-stückweisen Integralmittel auf Konvergenz $u_\ell \to u$ in $H^1_0(\Omega)$ führt. Binev, Dahmen und Devore (2004) erweiterten das Verfahren von Morin, Nochetto und Siebert um eine adaptive Netzvergröberung und zeigten, dass dieses adaptive Verfahren bezüglich der Elementanzahl (asymptotisch) auf die optimale Konvergenzrate führt. Stevenson (2007) verbesserte dieses Resultat und zeigte, dass auf die Vergröberung verzichtet werden kann und man trotzdem (asymptotisch) die optimale Konvergenzrate erhält. Ein Nachteil des Verfahrens von Morin, Nochetto und Siebert war, dass markierte Elemente vergleichsweise stark verfeinert wurden. Ihre sogenannte bisec_5-Regel entspricht in 1D etwa der Verfeinerung eines markierten Intervalls in 4 Teile anstatt in 2. Diese Einschränkung an die Netzverfeinerung wurde von Cascon, Kreuzer, Nochetto und Siebert 2008 eliminiert.

Die genannten Ergebnisse sind alle allerdings auf symmetrische Modellprobleme beschränkt. In 2009 hat unsere Arbeitsgruppe an der TU Wien das Konzept der *Schätzerreduktion* entwickelt. Dieses erlaubt einen elementaren Konvergenzbeweis, den wir im Folgenden vorstellen wollen. Im Kern stehen dabei die beiden folgenden Beobachtungen.

Lemma 50 (A priori Konvergenz adaptiver Verfahren) *Es seien $\langle\!\langle \cdot\,, \cdot \rangle\!\rangle$ eine elliptische Bilinearform auf einem Hilbert-Raum H und $u \in H$. Für eine Folge $X_\ell \subseteq X_{\ell+1}$ endlich-dimensionaler Teilräume von H sei $u_\ell \in X_\ell$ die Galerkin-Approximation von u. Dann existiert der Limes*

$$\lim_{\ell\to\infty} u_\ell =: u_\infty \in \overline{\bigcup_{\ell=0}^\infty X_\ell}, \tag{2.62}$$

stimmt aber nicht zwingend mit u überein.

Beweis. Nach Definition ist $X_\infty := \overline{\bigcup_{\ell=0}^\infty X_\ell}$ ein abgeschlossener Unterraum von H und deshalb selbst ein Hilbert-Raum. Nach Lemma von Lax-Milgram existiert also eine eindeutige Lösung $u_\infty \in X_\infty$ von

$$\langle\!\langle u_\infty\,, v_\infty \rangle\!\rangle = \langle\!\langle u\,, v_\infty \rangle\!\rangle \quad \text{für alle } v_\infty \in X_\infty.$$

Aufgrund der Inklusion $X_\ell \subseteq X_\infty$, ist $u_\ell \in X_\ell$ auch die Galerkin-Approximation von u_∞, und es gilt das Céa Lemma

$$\|u_\infty - u_\ell\|_H \lesssim \min_{v_\ell \in X_\ell} \|u_\infty - v_\ell\|_H.$$

Sei nun $\varepsilon > 0$. Nach Definition von X_∞, existiert ein Index ℓ_0 und ein Element $v_{\ell_0} \in X_{\ell_0}$ mit

$$\|u_\infty - v_{\ell_0}\|_H \leq \varepsilon.$$

Für alle $\ell \geq \ell_0$ folgt wegen $X_{\ell_0} \leq X_\ell$ deshalb

$$\|u_\infty - u_\ell\|_H \lesssim \min_{v_\ell \in X_\ell} \|u_\infty - v_\ell\|_H \leq \|u_\infty - v_{\ell_0}\|_H \leq \varepsilon.$$

Gemäß Definition zeigt dies $\lim_{\ell\to\infty} u_\ell = u_\infty$. ■

Lemma 51 (Prinzip der Schätzerreduktion) *Die Folge $\rho_\ell \geq 0$ der Schätzer erfülle*

$$\rho_{\ell+1} \leq q\,\rho_\ell + \alpha_\ell \quad \textit{für alle } \ell \in \mathbb{N}_0 \tag{2.63}$$

mit einer Konstante $0 < q < 1$ und einer Nullfolge $\alpha_\ell \geq 0$ mit $\alpha_\ell \to 0$. Dann folgt $\lim_{\ell\to\infty} \rho_\ell = 0$.

Beweis. Induktiv folgt aus der Kontraktionsbedingung (2.63)

$$\rho_{\ell+1} \leq q^{\ell+1}\,\rho_0 + \sum_{k=0}^{\ell} q^{\ell-k}\alpha_k \quad \text{für alle } \ell \geq 0.$$

Da jede konvergente Folge beschränkt ist, folgt $\|(\alpha_\ell)\|_\infty < \infty$. Wir erhalten also Beschränktheit

$$\rho_\ell \leq q^\ell \rho_0 + \|(\alpha_\ell)\|_\infty \sum_{k=0}^{\ell-1} q^k \leq \rho_0 + \frac{\|(\alpha_\ell)\|_\infty}{1-q} < \infty \quad \text{für alle } \ell \in \mathbb{N}.$$

Insbesondere existiert $M := \limsup_\ell \rho_\ell$. Mit der Voraussetzung erhalten wir

$$M = \limsup_{\ell\to\infty} \rho_{\ell+1} \leq q \limsup_{\ell\to\infty} \rho_\ell + \limsup_{\ell\to\infty} \alpha_\ell = q\,M$$

Insbesondere folgt $0 = \liminf_\ell \rho_\ell \leq \limsup_\ell \rho_\ell = M = 0$ und deshalb die behauptete Konvergenz. ■

Satz 52 *Es sei $\overline{\rho}_\ell := \|(\widehat{u}_\ell - I_\ell \widehat{u}_\ell)'\|_{L^2(a,b)}$ ein modifizierter $(h-h/2)$-Schätzer. Für $0 < \theta < 1$ garantiert der $\overline{\rho}_\ell$-basierte adaptive Algorithmus*

$$\overline{\rho}_{\ell+1} \leq (1-\theta)^{1/2}\,\overline{\rho}_\ell + \|\widehat{u}_{\ell+1} - \widehat{u}_\ell\|_{H^1(a,b)}. \tag{2.64}$$

Insbesondere folgt deshalb $\lim_\ell \overline{\rho}_\ell = 0$ nach dem Prinzip der Schätzerreduktion.

Beweis. Mit der Dreiecksungleichung gilt

$$\overline{\rho}_{\ell+1} = \|(\widehat{u}_{\ell+1} - I_{\ell+1}\widehat{u}_{\ell+1})'\|_{L^2(a,b)} \leq \|(\widehat{u}_\ell - I_{\ell+1}\widehat{u}_\ell)'\|_{L^2(a,b)} + \|[(1-I_{\ell+1})(\widehat{u}_{\ell+1} - \widehat{u}_\ell)]'\|_{L^2(a,b)}.$$

Mit dem Approximationssatz lässt sich der zweite Term durch

$$\|[(1-I_{\ell+1})(\widehat{u}_{\ell+1} - \widehat{u}_\ell)]'\|_{L^2(a,b)} \leq \|(\widehat{u}_{\ell+1} - \widehat{u}_\ell)'\|_{L^2(a,b)} \leq \|\widehat{u}_{\ell+1} - \widehat{u}_\ell\|_{H^1(a,b)}$$

abschätzen. Der erste Term wird elementweise abgeschätzt. Aufgrund der Netzverfeinerung gilt

$$\|(\widehat{u}_\ell - I_{\ell+1}\widehat{u}_\ell)'\|_{L^2(T)} = 0 \quad \text{für } T \in \mathcal{M}_\ell.$$

Für nicht-markierte (und damit nicht-verfeinerte) Elemente $T \in \mathcal{T}_\ell \backslash \mathcal{M}_\ell$ gilt ferner

$$\|(\widehat{u}_\ell - I_{\ell+1}\widehat{u}_\ell)'\|_{L^2(T)} = \|(1-\Pi_{\ell+1})\widehat{u}_\ell'\|_{L^2(T)} \leq \|(1-\Pi_\ell)\widehat{u}_\ell'\|_{L^2(T)} = \|(\widehat{u}_\ell - I_\ell\widehat{u}_\ell)'\|_{L^2(T)} = \overline{\rho}_\ell(T),$$

wobei $\Pi_\ell : L^2(a,b) \to \mathcal{P}^0(\mathcal{T}_\ell)$ die L^2-Orthogonalprojektion bezeichne. Nun folgt mit der Dörfler-Markierung (2.51)

$$\|(\widehat{u}_\ell - I_{\ell+1}\widehat{u}_\ell)'\|^2_{L^2(a,b)} = \sum_{T\in\mathcal{T}_\ell\backslash\mathcal{M}_\ell} \|(\widehat{u}_\ell - I_{\ell+1}\widehat{u}_\ell)'\|^2_{L^2(T)} \leq \sum_{T\in\mathcal{T}_\ell} \overline{\rho}_\ell(T)^2 - \sum_{T\in\mathcal{M}_\ell} \overline{\rho}_\ell(T)^2 \leq (1-\theta)\,\overline{\rho}_\ell^2,$$

was den Beweis beschließt. ■

Korollar 53 *Es sei* $\widetilde{\rho}_\ell := \|\widehat{u}_\ell - I_\ell \widehat{u}_\ell\|_{H^1(a,b)}$ *ein weiterer modifizierter* $(h - h/2)$*-Schätzer. Dann garantiert der* $\widetilde{\rho}_\ell$*-basierte adaptive Algorithmus ebenfalls* $\lim_\ell \widetilde{\rho}_\ell = 0$.

Beweis. Wir nehmen an, dass $\mathcal{M}_\ell$ die Menge der markierten Elemente im ℓ-ten Schritt des $\widetilde{\rho}_\ell$-basierte Algorithmus sei. Aufgrund des Approximationssatzes gilt

$$\widetilde{\rho}_\ell(T)^2 = \|\widehat{u}_\ell - I_\ell \widehat{u}_\ell\|^2_{H^1(a,b)} \leq (\|h_\ell\|^2_{L^\infty} + 1)\, \|(\widehat{u}_\ell - I_\ell \widehat{u}_\ell)'\|^2_{L^2(a,b)} \leq C\, \overline{\rho}_\ell(T)^2 \quad \text{für alle } T \in \mathcal{T}_\ell$$

mit $C := 1 + (b-a)^2 > 1$. Es folgt

$$\theta\, \overline{\rho}_\ell^2 \leq \theta\, \widetilde{\rho}_\ell^2 \leq \sum_{T \in \mathcal{M}_\ell} \widetilde{\rho}_\ell(T)^2 \leq C \sum_{T \in \mathcal{M}_\ell} \overline{\rho}_\ell(T)^2,$$

d.h. $\mathcal{M}_\ell$ erfüllt das Dörfler-Kriterium (2.51) für $\overline{\rho}_\ell$ mit $\overline{\theta} = \theta/C \in (0,1)$. Nach Satz 52 folgt $\lim_\ell \overline{\rho}_\ell = 0$ und insbesondere auch $\lim_\ell \widetilde{\rho}_\ell = 0$ wegen $\widetilde{\rho}_\ell \simeq \overline{\rho}_\ell$. ■

Bemerkung 54 *Die Konvergenz des* ρ_ℓ*-basierten adaptiven Algorithmus mit dem kanonischen* $(h - h/2)$*-Fehlerschätzer* $\rho_\ell = \|\widehat{u}_\ell - u_\ell\|_{H^1(a,b)}$ *ist komplizierter. Mit* $M_\ell := \bigcup_{T \in \mathcal{M}_\ell} T$ *zeigt die Dörfler Markierung* (2.51) *und die a priori Konvergenz* $\widehat{u}_\ell \to \widehat{u}_\infty$

$$\begin{aligned}\sqrt{\theta}\, \|\widehat{u}_\infty - u_\ell\|_{H^1(a,b)} &\leq \|\widehat{u}_\infty - u_\ell\|_{H^1(M_\ell)} + (1 + \sqrt{\theta})\, \|\widehat{u}_\infty - \widehat{u}_\ell\|_{H^1(a,b)} \\ &\leq \|\widehat{u}_\infty - u_\ell\|_{H^1(M_\ell)} + 2\, \|\widehat{u}_\infty - \widehat{u}_\ell\|_{H^1(a,b)}.\end{aligned}$$

Falls $\widehat{u}_\infty \not/ u_\infty = \lim_\ell u_\ell$*, so existieren* $\ell_0 \in \mathbb{N}$ *mit* $\sqrt{\theta}/4\, \|\widehat{u}_\infty - u_\ell\|_{H^1(a,b)} \geq \|\widehat{u}_\infty - \widehat{u}_\ell\|_{H^1(a,b)}$ *für alle* $\ell \geq \ell_0$*. Wir erhalten also*

$$\sqrt{\theta}/2\, \|\widehat{u}_\infty - u_\ell\|_{H^1(a,b)} \leq \|\widehat{u}_\infty - u_\ell\|_{H^1(M_\ell)} \quad \textit{für alle } \ell \geq \ell_0,$$

d.h. mit der „exakten Lösung“ $\widehat{u}_\infty$ *und ihrer Galerkin-Approximation* u_ℓ *ist ab Schritt* ℓ_0 *immer das Markierungskriterium mit dem Galerkin-Fehler erfüllt. Insbesondere werden dann immer die Elemente mit dem größten lokalen Fehler verfeinert. Dadurch kann man einen Widerspruch erzeugen zu* $u_\ell \not\to \widehat{u}_\infty$.

Freiwillige Übung. Es sei ρ_ℓ der gewichtete Residualschaetzer (2.60) Zeigen Sie, dass der ρ_ℓ-basierte adaptive Algorithmus die Schätzerreduktion

$$\rho_{\ell+1} \leq (1 - \theta/2)^{1/2}\, \rho_\ell + C\, \|u_{\ell+1} - u_\ell\|_{H^1(a,b)} \quad \text{für alle } \ell \in \mathbb{N}_0 \tag{2.65}$$

mit einer ℓ-unabhängigen Konstante $C > 0$ erfüllt. Insbesondere folgt deshalb

$$\lim_{\ell \to \infty} \rho_\ell = 0 = \lim_{\ell \to \infty} \|u - u_\ell\|_{H^1(a,b)}, \tag{2.66}$$

d.h. die Folge der Galerkin-Lösungen konvergiert gegen die kontinuierliche Lösung. ■

Theoretisch könnte die Konvergenz $\rho_\ell \to 0$ beliebig langsam sein. Für symmetrische Probleme ist die Konvergenz aber zumindest linear, wie der folgende Dazu zeigt.

Satz 55 (Cascon, Kreuzer, Nochetto, Siebert '08) *Die Bilinearform* $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ *sei zusätzlich symmetrisch, d.h. ein äquivalentes Skalarprodukt auf* $H^1_0(a,b)$*. Es sei* ρ_ℓ *der gewichtete Residualschätzer* (2.60)*. Dann existieren* ℓ*-unabhängige Konstanten* $0 < q, \kappa < 1$*, sodass*

$$\Delta_{\ell+1} \leq q\, \Delta_\ell \quad \textit{mit} \quad \Delta_\ell := |\!|\!| u - u_\ell |\!|\!|^2 + \kappa\, \rho_\ell^2, \tag{2.67}$$

wobei $|\!|\!| \cdot |\!|\!|$ *die induzierte Energienorm bezeichne.*

Beweis. Wir beginnen mit einer Variante der **Young-Ungleichung**: Für $a, b \geq 0$ gilt

$$2ab \leq a^2 + b^2$$

und deshalb für $\varepsilon > 0$

$$2ab = 2(a\varepsilon)\,\frac{b}{\varepsilon} \leq a^2\varepsilon^2 + \frac{b^2}{\varepsilon^2}.$$

Insbesondere erhalten wir nun für $\delta = \varepsilon^2$

$$(a+b)^2 \leq (1+\delta)a^2 + (1+\delta^{-1})b^2 \quad \text{für alle } a, b \geq 0 \text{ und } \delta > 0.$$

Wir wenden dies auf die Schätzerreduktion (2.65) an und erhalten

$$\rho_{\ell+1}^2 \leq (1+\delta)(1-\theta/2)\rho_\ell^2 + (1+\delta^{-1})C^2\,\|\!|u_{\ell+1} - u_\ell|\!\|^2$$

Laut Galerkin-Orthogonalität und $X_\ell \subset X_{\ell+1}$ gilt der Satz von Pythagoras

$$|\!|\!|u - u_{\ell+1}|\!|\!|^2 + |\!|\!|u_{\ell+1} - u_\ell|\!|\!|^2 = |\!|\!|u - u_\ell|\!|\!|^2. \tag{2.68}$$

Wir erhalten zunächst

$$\Delta_{\ell+1} = |\!|\!|u - u_{\ell+1}|\!|\!|^2 + \kappa\,\rho_{\ell+1}^2 \leq |\!|\!|u - u_\ell|\!|\!|^2 + \kappa(1+\delta)(1-\theta/2)\,\rho_\ell^2 + \big(\kappa(1+\delta^{-1})C^2 - 1\big)\,|\!|\!|u_{\ell+1} - u_\ell|\!|\!|^2.$$

Als Erstes fixieren wir also $\delta > 0$ klein genug, sodass $(1+\delta)(1-\theta/2) < 1$. Im nächsten Schritt fixieren wir nun $\kappa > 0$ klein genug, sodass $\kappa(1+\delta^{-1})C^2 - 1 \leq 0$. Dann gilt

$$\Delta_{\ell+1} \leq |\!|\!|u - u_\ell|\!|\!|^2 + \kappa(1+\delta)(1-\theta/2)\,\rho_\ell^2.$$

Schließlich nutzen wir noch die Zuverlässigkeit des Residualschätzers in der Form

$$|\!|\!|u - u_\ell|\!|\!| \leq C_{\mathrm{rel}}\,\rho_\ell.$$

Für $\varepsilon > 0$ gilt

$$|\!|\!|u - u_\ell|\!|\!|^2 = (1-\kappa\varepsilon)|\!|\!|u - u_\ell|\!|\!|^2 + \kappa\varepsilon|\!|\!|u - u_\ell|\!|\!|^2 \leq (1-\kappa\varepsilon)|\!|\!|u - u_\ell|\!|\!|^2 + \kappa\varepsilon C_{\mathrm{rel}}^2\,\rho_\ell^2,$$

sodass

$$\Delta_{\ell+1} \leq (1-\kappa\varepsilon)\,|\!|\!|u - u_\ell|\!|\!|^2 + \kappa\big((1+\delta)(1-\theta/2) + \varepsilon C_{\mathrm{rel}}^2\big)\,\rho_\ell^2.$$

Wegen $(1+\delta)(1-\theta/2) < 1$ können wir schließlich $\varepsilon > 0$ klein genug wählen, sodass $(1+\delta)(1-\theta/2) + \varepsilon C_{\mathrm{rel}}^2 < 1$. Damit ergibt sich abschließend

$$\Delta_{\ell+1} \leq (1-\varepsilon)\,|\!|\!|u - u_\ell|\!|\!|^2 + \kappa\big((1+\delta)(1-\theta/2) - \varepsilon C_{\mathrm{rel}}^2\big)\,\rho_\ell^2 \leq q\,\big(|\!|\!|u - u_\ell|\!|\!|^2 + \kappa\,\rho_\ell^2\big) = q\,\Delta_\ell$$

mit $q := \max\{1 - \kappa\varepsilon\,,\,(1+\delta)(1-\theta/2) - \varepsilon C_{\mathrm{rel}}^2\} < 1$. ∎

Bemerkung 56 *Der letzte Beweis nutzt in* (2.68) *die Symmetrie der Bilinearform. Für unsymmetrische Probleme (d.h. $\beta \neq 0$ und möglicherweise $c < 0$) ist lineare Konvergenz erst jüngst gezeigt worden: Feischl, Führer und Praetorius (2014) zeigen, dass es in diesem Fall Konstanten $C > 0$ und $0 < q < 1$ gibt mit*

$$\rho_{\ell+k} \leq C\,q^k\,\rho_\ell. \tag{2.69}$$

Überlegen Sie sich, dass für symmetrische Problem (2.69) *direkt aus* (2.67) *folgt. Der direkte Beweis von* (2.69) *im Fall unsymmetrischer Probleme ist allerdings weniger elementar als der von Satz 55. Wir verweisen auf die Spezialvorlesung* Adaptive Finite Elemente Methode *im Sommersemester 2015, in der auch bewiesen wird, dass Adaptivität auf optimale Konvergenzraten führt.*

2.4 Inhomogenes Dirichlet-Problem

Wir betrachten nun das Modellproblem

$$-(\alpha u')' + \beta u' + \gamma u = f \quad \text{in } (a,b), \tag{2.70}$$

mit **inhomogenen Dirichlet-Randbedingungen**

$$u(a) = A, \quad u(b) = B \tag{2.71}$$

für gegebene $A, B \in \mathbb{R}$. Bisher haben wir nur den Fall $A = 0 = B$ betrachtet. Die schwache Form lautet nun wie folgt: Finde $u \in H^1(a,b)$ mit (2.71) und

$$\langle\!\langle u\,, v \rangle\!\rangle = F(v) \quad \text{für alle } v \in H^1_0(a,b). \tag{2.72}$$

Dabei sind wie bisher

$$\langle\!\langle u\,, v \rangle\!\rangle := \int_a^b \alpha u'v' + \beta u'v + \gamma uv \, dx \quad \text{und} \quad F(v) := \int_a^b fv \, dx. \tag{2.73}$$

Man beachte, dass in diesem Fall der Ansatzraum $H^1(a,b)$, in dem wir die schwache Lösung u suchen, und der Testraum $H^1_0(a,b)$ für die Testfunktionen v verschieden sind.

2.4.1 Eindeutige Lösbarkeit

Um zu zeigen, dass die schwache Form (2.71)–(2.72) eine eindeutige Lösung hat, gehen wir wie folgt vor: Wie wählen eine beliebige Fortsetzung u_D der Dirichlet-Daten, d.h. $u_D \in H^1(a,b)$ mit $u_D(a) = A$ und $u_D(b) = B$. Eine mögliche Wahl ist die affine Funktion

$$u_D(x) = A + \frac{x-a}{b-a}\,(B - A) \in \mathcal{S}^1(\mathcal{T}_h).$$

Alternativ kann man zu gegebener Triangulierung $\mathcal{T}_h$ von (a,b) auch die $\mathcal{T}_h$-stückweise Funktion

$$u_D \in \mathcal{S}^1(\mathcal{T}_h) \quad \text{mit} \quad u_D(a) = A, \quad u_D(b) = B, \quad u_D(x_k) = 0 \text{ für alle } x_k \in \mathcal{K}_h \backslash \{a,b\}$$

wählen. Nun definieren wir $u_0 := u - u_D$. Falls die schwache Lösung $u \in H^1(a,b)$ existiert, gilt

$$u_0 \in H^1_0(a,b) \quad \text{mit} \quad \langle\!\langle u_0\,, v \rangle\!\rangle = F(v) - \langle\!\langle u_D\,, v \rangle\!\rangle \quad \text{für alle } v \in H^1_0(a,b). \tag{2.74}$$

Unter denselben Voraussetzungen wie für das homogene Dirichlet-Problem existiert also ein eindeutiges $u_0 \in H^1_0(a,b)$ mit (2.74) und deshalb ein eindeutiges $u \in H^1(a,b)$ mit (2.71)–(2.72).

Satz 30 über die Äquivalenz von starker und schwacher Formulierung gilt analog.

2.4.2 FEM für inhomogenes Dirichlet-Problem

Zu gegebener Triangulierung $\mathcal{T}_h$ sei $u_D \in \mathcal{S}^1(\mathcal{T}_h)$ eine Fortsetzung der gegebenen Dirichlet-Daten. Wir berechnen dann die eindeutige Galerkin-Approximation $u_{0h} \in \mathcal{S}^1_0(\mathcal{T}_h)$ von $u_0 \in H^1_0(a,b)$ als Lösung von

$$\langle\!\langle u_{0h}\,, v_h \rangle\!\rangle = F(v_h) - \langle\!\langle u_D\,, v_h \rangle\!\rangle \quad \text{für alle } v_h \in \mathcal{S}^1_0(\mathcal{T}_h). \tag{2.75}$$

Abschließend definiert $u_h := u_{0h} + u_D \in \mathcal{S}^1(\mathcal{T}_h)$ eine Approximation der schwachen Lösung $u \in H^1(a,b)$. Man beachte, dass sich wegen

$$\|u - u_h\|_{H^1(a,b)} = \|u_0 - u_{0h}\|_{H^1(a,b)}$$

alle Ergebnisse für das homogene Dirichlet-Problem auch auf das inhomogene Dirichlet-Problem übertragen.

2.5 Gemischtes Randwertproblem

Als Abschluss dieses Abschnitts betrachten wir nun das Modellproblem

$$-(\alpha u')' + \beta u' + \gamma u = f \quad \text{in } (a,b), \tag{2.76}$$

mit **gemischten Dirichlet-Neumann-Randbedingungen**

$$u(a) = A, \quad (\alpha u')(b) = B \tag{2.77}$$

für gegebene $A, B \in \mathbb{R}$.

Multiplikation der starken Form (2.76) mit einer Testfunktion

$$v \in H^1_D(a,b) := \big\{ v \in H^1(a,b) \,\big|\, v(a) = 0 \big\} \tag{2.78}$$

und partielle Integration liefern

$$-\int_a^b (\alpha u')' v \, dx = \int_a^b \alpha u' v' \, dx - (\alpha u' v)(b) + (\alpha u' v)(a) = \int_a^b \alpha u' v' \, dx - B\, v(b). \tag{2.79}$$

Die schwache Form lautet also wie folgt: Finde $u \in H^1(a,b)$ mit

$$u(a) = A \tag{2.80}$$

und

$$\langle\!\langle u \,, v \rangle\!\rangle = F(v) \quad \text{für alle } v \in H^1_D(a,b). \tag{2.81}$$

Hierbei ist

$$\langle\!\langle u \,, v \rangle\!\rangle := \int_a^b \alpha u' v' + \beta u' v + \gamma u v \, dx \quad \text{und} \quad F(v) := \int_a^b f v \, dx + B\, v(b). \tag{2.82}$$

Man beachte, dass die Neumann-Randbedingung ein Teil der Variationsformulierung ist. Deshalb bezeichnet man die Neumann-Randbedingungen auch als *natürliche* Randbedingungen.

2.5.1 Eindeutige Lösbarkeit

Um zu zeigen, dass die schwache Form (2.77) & (2.81) eine eindeutige Lösung hat, gehen wir wieder wie folgt vor: Wir wählen eine beliebige Fortsetzung u_D der Dirichlet-Daten, d.h. $u_D \in H^1(a,b)$ mit $u_D(a) = A$. Eine mögliche Wahl ist die konstante Funktion

$$u_D(x) = A \in \mathcal{S}^1(\mathcal{T}_h).$$

Alternativ kann man zu gegebener Triangulierung $\mathcal{T}_h$ von (a,b) auch die $\mathcal{T}_h$-stückweise Funktion

$$u_D \in \mathcal{S}^1(\mathcal{T}_h) \quad \text{mit} \quad u_D(a) = A, \quad u_D(x_k) = 0 \text{ für alle } x_k \in \mathcal{K}_h \backslash \{a\}$$

wählen. Nun definieren wir $u_0 := u - u_D$. Falls die schwache Lösung $u \in H^1(a,b)$ existiert, gilt

$$u_0 \in H^1_D(a,b) \quad \text{mit} \quad \langle\!\langle u_0 \,, v \rangle\!\rangle = F(v) - \langle\!\langle u_D \,, v \rangle\!\rangle \quad \text{für alle } v \in H^1_D(a,b). \tag{2.83}$$

Offensichtlich ist $H^1_D(a,b)$ ein abgeschlossener Unterraum von $H^1(a,b)$ und deshalb ein Hilbert-Raum. Unter analogen Voraussetzungen wie für das homogene Dirichlet-Problem existiert also ein eindeutiges $u_0 \in H^1_D(a,b)$ mit (2.83) und deshalb ein eindeutiges $u \in H^1(a,b)$ mit (2.77)–(2.81).

Satz 30 über die Äquivalenz von starker und schwacher Formulierung gilt analog.

2.5.2 FEM für gemischtes Randwertproblem

Zu gegebener Triangulierung $\mathcal{T}_h$ sei $u_D \in \mathcal{S}^1(\mathcal{T}_h)$ eine Fortsetzung der gegebenen Dirichlet-Daten. Wir berechnen dann die eindeutige Galerkin-Approximation $u_{0h} \in \mathcal{S}^1_D(\mathcal{T}_h) := \mathcal{S}^1(\mathcal{T}_h) \cap H^1_D(a,b)$ von $u_0 \in H^1_D(a,b)$ als Lösung von

$$\langle\!\langle u_{0h}\,, v_h \rangle\!\rangle = F(v_h) - \langle\!\langle u_D\,, v_h \rangle\!\rangle \quad \text{für alle } v_h \in \mathcal{S}^1_D(\mathcal{T}_h). \tag{2.84}$$

Abschließend definiert $u_h := u_{0h} + u_D \in \mathcal{S}^1(\mathcal{T}_h)$ eine Approximation der schwachen Lösung $u \in H^1(a,b)$. Wieder gilt

$$\|u - u_h\|_{H^1(a,b)} = \|u_0 - u_{0h}\|_{H^1(a,b)},$$

und alle Ergebnisse für das homogene Dirichlet-Problem lassen sich auch auf das gemischte Randwertproblem übertragen.

Kapitel 3

Finite Elemente Methode in 2D

Kapitel 4

Randelementmethode in 2D

Die Randelementmethode ist ein numerisches Verfahren zur Lösung elliptischer Differentialgleichungen, sofern die Fundamentallösung des Differentialoperators bekannt ist. Als Modellproblem betrachten wir in diesem Abschnitt wieder die Laplace-Gleichung

$$-\Delta u = f \quad \text{in } \Omega, \tag{4.1}$$

versehen mit Dirichlet- und/oder Neumann-Randbedingungen. Die zentrale Beobachtung ist, dass man die Lösung $u \in H^1(\Omega)$ dieser Gleichung in Form von drei Integraloperatoren

$$u = \widetilde{N} f + \widetilde{V}(\partial_n u) + \widetilde{K}(u|_\Gamma) \quad \text{punktweise fast überall in } \Omega \tag{4.2}$$

schreiben kann, sofern die Spur $u|_\Gamma$ und die Normalenableitung $\partial_n u$ auf Γ bekannt sind. Da die Spur $u|_\Gamma$ üblicherweise nur auf dem Dirichlet-Rand Γ_D bekannt ist und die Normalenableitung $\partial_n u$ nur auf dem Neumann-Rand, ist das Ziel der Randelementmethode (*engl.* boundary element method, BEM), die fehlenden Daten zu approximieren.

Um Gleichungen für die unbekannten Anteile der Randbedingungen zu erhalten, bildet man für die Identität in (4.2) die Spur und die Normalenableitung. Dies führt dann auf gewisse Integralgleichungen am Rand.

Ein Vorteil der BEM gegenüber der FEM ist, dass nicht Ω, sondern lediglich Γ diskretisiert werden muss, d.h. für ein 2D Problem ist lediglich eine 1D Mannigfaltigkeit zu diskretisieren. Ferner werden wir sehen, dass im Fall niedrigster Ordnung, d.h. Approximation der Normalenableitung durch stückweise konstante Funktionen und Approximation der Spur durch stückweise affine und global stetige Funktionen (d.h. wie in der P1-FEM), die BEM bereits mit Ordnung $\mathcal{O}(h^{3/2})$ konvergiert, wohingegen die P1-FEM nur auf Konvergenzordnung $\mathcal{O}(h)$ führen kann.

Ein Nachteil der BEM gegenüber der FEM ist allerdings, dass der Anwendungsbereich der BEM wesentlich eingeschränkter ist als der der FEM, d.h. ein besseres numerisches Verfahren erhält man nur für eine kleinere Klasse von Differentialgleichungen.

4.1 Darstellungsformel

In diesem Abschnitt werden wir die Darstellungsformel (4.2) (oder: Dritte Greensche Formel) herleiten. Dazu definieren wir zunächst den **Newton-Kern** (oder: **Fundamentallösung** von $-\Delta$) als

$$G(z) := \begin{cases} -\frac{1}{2\pi} \log|z| & \text{für } d = 2, \\ +\frac{1}{4\pi} \frac{1}{|z|} & \text{für } d = 3. \end{cases} \tag{4.3}$$

Dabei bezeichnet d die Raumdimension für $\Omega \subset \mathbb{R}^d$. Laut Mittelschule gilt für das Maß $|S_2^d|$ der Sphäre S_2^d im $\mathbb{R}^d$ die Formel $|S_2^2| = 2\pi$ (Umfang des Einheitskreises) bzw. $|S_2^3| = 4\pi$ (Oberfläche der Einheitskugel).

Lemma 1 (i) *Es gilt $G \in C^\infty(\mathbb{R}^d \backslash \{0\})$ mit ersten und zweiten Ableitungen*

$$\partial_j G(z) = -\frac{1}{|S_2^d|} \frac{z_j}{|z|^d} \quad \textit{and} \quad \partial_{jk} G(z) = -\frac{1}{|S_2^d|} \frac{\delta_{jk}|z|^2 - d z_j z_k}{|z|^{d+2}}. \tag{4.4}$$

(ii) *Es gilt* $-\Delta G(z) = 0$ *für* $z \neq 0$.
(iii) $G \in L^p_{\ell oc}(\mathbb{R}^d)$ *für* $d < 2p/(p-1)$ *und insbesondere* $G \in L^2_{\ell oc}(\mathbb{R}^d)$.
(iv) $\partial_j G \in L^p_{\ell oc}(\mathbb{R}^d)$ *für* $d < p/(p-1)$ *uns insbesondere* $\partial_j G \in L^1_{\ell oc}(\mathbb{R}^d)$.

Beweis. (i) und (ii) folgen durch elementares Differenzieren. Um (iii) und (iv) zu zeigen, muss gezeigt werden, dass die entsprechenden L^p-Normen auf allen Kugeln $U_R(0)$ endlich sind. Da $|G(z)|^p$ und $|\partial_j G(z)|^p$ nur von der Euklidischen Länge $|z|$ abhängt, ist es sinnvoll zu Polarkoordinaten überzugehen. Dies reduziert die Aussagen auf die Existenz von Integralen der Gestalt $\int_0^R s^\alpha \, ds$ mit gewissen $\alpha \in \mathbb{R}$. ■

Satz 2 (Darstellungsformel) *Es sei* Ω *eine beschränkte offene Menge in* $\mathbb{R}^d$ *mit hinreichend glattem Rand* $\Gamma := \partial\Omega$. *Für* $u \in C^2(\overline{\Omega})$ *mit* $f := -\Delta u \in C(\overline{\Omega})$ *gilt*

$$u(x) = \int_\Omega G(x-y) f(y)\, dy + \int_\Gamma G(x-y) \partial_{n(y)} u(y)\, ds_y - \int_\Gamma \partial_{n(y)} G(x-y)\, u(y)\, ds_y \tag{4.5}$$

für alle $x \in \Omega$. *Hierbei bezeichnet* $n(y)$ *den äußeren Normalenvektor bei* $y \in \Gamma$ *und* $\partial_{n(y)}$ *ist die Normalenableitung in diesem Punkt.* ■

Wir führen nun folgende Notation ein:

- **Newton-Potential** von $f : \Omega \to \mathbb{R}$

$$\widetilde{N} f(x) := \int_\Omega G(x-y) f(y)\, dy \tag{4.6}$$

- **Einfachschichtpotential** von $\phi : \Gamma \to \mathbb{R}$

$$\widetilde{V} \phi(x) := \int_\Gamma G(x-y) \phi(y)\, ds_y \tag{4.7}$$

- **Doppelschichtpotential** von $v : \Gamma \to \mathbb{R}$

$$\widetilde{K} v(x) := \int_\Gamma \partial_{n(y)} G(x-y)\, v(y)\, ds_y \tag{4.8}$$

Mit dieser Notation haben wir zumindest die Darstellungsformel (4.2) für $u \in C^2(\overline{\Omega})$ und punktweise in Ω bewiesen. Aus der Maßtheorie sollte bekannt sein, dass für Funktionen $f \in L^p(\mathbb{R}^d)$ und $g \in L^{p'}(\mathbb{R}^d)$ die Faltung $f * g \in L^\infty(\mathbb{R}^d)$ gleichmäßig stetig ist. Hieraus kann man mit ein wenig mathematischer Technik ableiten, dass das Newton-Potential $\widetilde{N} f = G * f$ für $f \in L^2(\Omega)$ zumindest stetig ist, d.h. $\widetilde{N} f \in C(\mathbb{R}^d)$.

Ein ähnliches Argument kann man auch für das Einfachschichtpotential anwenden, was gerade einer Faltung mit dem Newton-Kerns $G \in L^1(\Gamma)$ auf der $(d-1)$-dimensionalen Mannigfaltigkeit Γ entspricht. Man kann dann zeigen, dass für $\phi \in L^\infty(\Gamma)$, das Einfachschichtpotential $\widetilde{V}\phi \in C(\mathbb{R}^d)$ stetig ist.

Anders verhält es sich mit dem Doppelschichtpotential $\widetilde{K} v$. Man beachte, dass die Ableitung $\partial_{n(y)} G(x-y)$ des Newton-Kerns eine hat Singularität der Ordnung $1/r^{d-1}$ auf der $(d-1)$-dimensionalen Mannigfaltigkeit Γ hat, d.h. $\partial_n G(x - \cdot)$ ist nicht integrierbar über Γ. Man kann sich sogar leicht überlegen, dass $\widetilde{K}$ auf Γ springt:

Bemerkung 3 *Es gilt*

$$\widetilde{K} 1(x) = \int_\Gamma \partial_{n(y)} G(x-y)\, ds_y = \begin{cases} -1 & \text{für } x \in \Omega, \\ 0 & \text{für } x \in \mathbb{R}^d \backslash \overline{\Omega}. \end{cases} \tag{4.9}$$

Der Beweis folgt aus der Darstellungsformel (4.5) *aus Satz 2. Für die konstante Funktion* $v \equiv 1$ *gilt*

$$1 = -\widetilde{K} v(x) \quad \text{für alle } x \in \Omega$$

wegen $-\Delta v = 0 = \partial_n v$. *Nun seien* $x \in \mathbb{R}^d \backslash \overline{\Omega}$ *und* $\varepsilon > 0$ *mit* $\overline{U_\varepsilon(x)} \subset \mathbb{R}^d \backslash \overline{\Omega}$. *Wir betrachten das neue Gebiet* $\Omega_\varepsilon := \Omega \cup U_\varepsilon(x)$ *und wieder* $v \equiv 1$. *Wenden wir die letzte Beobachtung sowohl auf* Ω_ε *als auch auf* $U_\varepsilon(x)$ *an, so sehen wir*

$$-1 = \widetilde{K}_{\Omega_\varepsilon}(x) = \widetilde{K}_{\Omega}(x) + \widetilde{K}_{U_\varepsilon(x)}(x) = \widetilde{K}_{\Omega}(x) - 1,$$

wobei die Indizierung das betrachtete Gebiet bezeichnet und $\partial\Omega_\varepsilon = \Gamma \cup \partial U_\varepsilon(x)$ *und* $\Gamma \cap \partial U_\varepsilon(x) = \emptyset$ *ausgenutzt wurde. Es folgt* $\widetilde{K}_\Omega(x) = 0$, *wie oben behauptet.* ■

Durchweg verwenden wir folgende **Notationskonvention**: Betrachten wir einen Integraloperator und werten ihn für $x \in \Omega$ aus, so schreiben wir $\widetilde{N}$, $\widetilde{V}$ sowie $\widetilde{K}$. Werten wir denselben Integraloperator für $x \in \Gamma$ aus, so schreiben wir lediglich N, V bzw. K: Wir betrachten nun auf Γ die formale Funktion

$$Kv(x) = \int_\Gamma \partial_{n(y)} G(x-y)\, v(y)\, ds_y \quad \text{für } x \in \Gamma. \tag{4.10}$$

Man kann zeigen, dass für fast alle Punkte auf Γ

$$(K - 1/2)v : \Gamma \to \mathbb{R} \tag{4.11}$$

die stetige Fortsetung von $\widetilde{K}v : \Omega \to \mathbb{R}$ von Ω auf Γ ist.

Bemerkung 4 *Man bezeichnet* G *als Fundamentallösung von* $-\Delta$, *da auf* $\mathbb{R}^d$ *und für* $f \in C_c^\infty(\Omega) =: \mathcal{D}(\Omega)$ *die Faltung* $u = \widetilde{N}f = G * f$ *die Lösung von* $-\Delta u = f$ *gibt, was sofort aus der Darstellungsformel* (4.5) *folgt.*

4.2 Idee der Randelementmethode

Wir betrachten zunächst als Modellproblem

$$-\Delta u = f \quad \text{in } \Omega, \tag{4.12}$$
$$u = g \quad \text{auf } \Gamma. \tag{4.13}$$

Nach Darstellungsformel gilt

$$u = \widetilde{V}\phi - \widetilde{K}g \quad \text{in } \Omega \tag{4.14}$$

mit der *unbekannten* Normalenableitung $\phi = \partial_n u$ auf Γ. Wir nutzen nun aus, dass das Einfachschichtpotential $\widetilde{V}$ auf stetige Funktionen führt, wohingegen das Doppelschichtpotential $\widetilde{K}$ auf Γ springt. Spurbildung der letzten Gleichung liefert dann

$$g = V\phi - (K - 1/2)g \quad \text{fast überall auf } \Gamma. \tag{4.15}$$

Umstellung dieser Gleichung nach der Unbekannten zeigt

$$V\phi = (K + 1/2)g \quad \text{fast überall auf } \Gamma. \tag{4.16}$$

Durch Skalierung des Gebietes Ω auf $\operatorname{diam}(\Omega) < 1$ (nur in 2D nötig) kann man erreichen, dass V ein Isomorphismus zwischen gewissen Sobolev-Räumen ist. Damit ist die Randintegralgleichung (4.16) dann eine äquivalente Formulierung der Differentialgleichung (4.12).

4.2.1 Diskrete Räume für die Randelementmethode

Wir verwenden nun ein Galerkin-Verfahren, um die Lösung ϕ von (4.16) zu approximieren. Wir beschränken uns auf den 2D Fall: Es sei $\mathcal{E}_h = \{E_1, \dots, E_N\}$ eine Partition des Randes in affine Randstücke

$$E_j := [a_j, b_j] := \operatorname{conv}\{a_j, b_j\} \quad \text{mit} \quad a_j, b_j \in \mathbb{R}^2. \tag{4.17}$$

Zu jedem Randstück E_j fixieren wir eine Parametrisieriung

$$\gamma_j : [-1,1] \to \mathbb{R}, \quad \gamma_j(s) = \frac{1}{2}\big(a_j + b_j + s(b_j - a_j)\big). \tag{4.18}$$

Wir betonen, dass im Allgemeinen die folgenden diskreten Räume von den gewählten Parametrisierungen abhängen. Wir definieren den Raum der $\mathcal{E}_h$-stückweisen Polynome auf Γ durch

$$\mathcal{P}^p(\mathcal{E}_h) := \big\{ v : \Gamma \to \mathbb{R} \,\big|\, \forall E_j \in \mathcal{E}_h \quad v \circ \gamma_j \in \mathcal{P}^p[-1,1] \big\} \tag{4.19}$$

und $\mathcal{S}^p(\mathcal{E}_h) := \mathcal{P}^p(\mathcal{E}_h) \cap C(\Gamma)$.

Bemerkung 5 *In der Analysis wird gezeigt, dass die Definition von Randintegralen*

$$\int_{E_j} v\, ds := \int_{-1}^{1} v \circ \gamma_j(s) |\gamma_j'(s)|\, ds = \frac{\mathrm{length}(E_j)}{2} \int_{-1}^{1} v \circ \gamma_j(s)\, ds$$

unabhängig von der gewählten Parametrisierung γ_j von E_j sind.

4.2.2 BEM-Diskretisierung

In der Galerkin-Formulierung von (4.16) suchen wir eine Approximation $\phi_h \in \mathcal{P}^0(\mathcal{E}_h)$ von ϕ bei bekannter Approximation $g_h \in \mathcal{S}^1(\mathcal{E}_h)$ der gegebenen Dirichlet-Daten g. Die Variationsformulierung lautet dann

$$\langle V\phi_h \,, \psi_h \rangle_\Gamma = \langle (K + 1/2) g_h \,, \psi_h \rangle \quad \text{für alle } \psi_h \in \mathcal{P}^0(\mathcal{E}_h). \tag{4.20}$$

Hierbei bezeichnet

$$\langle v \,, w \rangle_\Gamma := \int_\Gamma vw\, ds \tag{4.21}$$

das $L^2(\Gamma)$-Skalarprodukt. Zur Implementierung von (4.20) müssen noch Basen von $\mathcal{P}^0(\mathcal{E}_h)$ und $\mathcal{S}^1(\mathcal{E}_h)$ gewählt werden:

- Für $\mathcal{P}^0(\mathcal{E}_h)$ nehmen wir die Basis $\{\chi_1, \dots, \chi_N\}$ der charakteristischen Funktionen χ_j zu $E_j \in \mathcal{E}_h$.
- Für $\mathcal{S}^1(\mathcal{E}_h)$ nehmen wir die Basis $\{\zeta_1, \dots, \zeta_N\}$ der Hutfunktionen, d.h. $\zeta_j \in \mathcal{S}^1(\mathcal{E}_h)$ erfüllt $\zeta_j(z_k) = \delta_{jk}$ für alle Knoten $z_k \in \mathcal{K}_h := \{z_1, \dots, z_N\}$.

Man beachte, dass es aufgrund des geschlossenen Randes genauso viele Knoten wie Randstücke gibt, d.h. es gilt $N := \#\mathcal{E}_h = \#\mathcal{K}_h$.

Nach Fixierung der Basis geht (4.20) in das lineare Gleichungssystem

$$\mathbf{V}\mathbf{x} = (\mathbf{K} + \tfrac{1}{2}\mathbf{M})\mathbf{g} \tag{4.22}$$

über. Dabei sind die Matrizen $\mathbf{V}, \mathbf{K}, \mathbf{M} \in \mathbb{R}^{N \times N}$ gegeben durch

$$\mathbf{V}_{jk} = \langle V\chi_k \,, \chi_j \rangle_\Gamma, \quad \mathbf{K}_{jk} = \langle K\zeta_k \,, \chi_j \rangle_\Gamma, \quad \mathbf{M}_{jk} = \langle \zeta_k \,, \chi_j \rangle_\Gamma. \tag{4.23}$$

Der Vektor $\mathbf{g} \in \mathbb{R}^N$ ist bei nodaler Interpolation der Knotenvektor der gegebenen Dirichlet-Daten

$$\mathbf{g}_j = g(z_j) \quad \text{und} \quad g_h = \sum_{j=1}^{N} \mathbf{g}_j \zeta_j. \tag{4.24}$$

Der Vektor $\mathbf{x} \in \mathbb{R}^N$ ist der gesuchte Koeffizienten-Vektor der approximativen Normalenableitung

$$\phi_h = \sum_{j=1}^{N} \mathbf{x}_j \chi_j. \tag{4.25}$$

Man kann zeigen, dass unter der Voraussetzung $\mathrm{diam}(\Omega) < 1$, die linke Seite in (4.20) ein Skalarprodukt (u.a. auch) auf $\mathcal{P}^0(\mathcal{E}_h)$ definiert. Insbesondere hat die Galerkin-Formulierung (4.20) bzw. das Gleichungssystem (4.22) eine eindeutige Lösung.

Literaturverzeichnis

[1] M. Hanke-Bourgeois: *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*, Teubner, Wiesbaden u.a. ^{2}2006.

[2] R. Plato: *Numerische Mathematik kompakt*, Vieweg, Braunschweig u.a. 2000.

[3] C. Überhuber: *Numerical Computation – Methods, Software and Analysis*, *Volume 2*, Springer, Heidelberg u.a. 1997.

[4] D. Werner: *Funktionalanalysis*, Springer, Heidelberg u.a. ^{3}2000.