Kapitel 3

Finite Elemente

In diesem Kapitel befassen wir uns genauer mit der Diskretisierung von partiellen Differentialgleichungen mit Finite Elemente (FE) Methoden. Der Fokus liegt dabei wieder auf elliptischen Gleichungen, da finite Elemente meist zur Ortsdiskretisierung verwendet werden. Die Erweiterung auf den ortsabhängigen Teil einer parabolischen ist dann z.B. im Rahmen einer horizontalen Linienmethode völlig klar, da ja dort in jedem Zeitschritt elliptische Probleme gelöst werden müssen. Zumindest teilweise lassen sich die hier vorgestellten Konzepte auch auf hyperbolische Probleme übertragen, eine genauere Diskussion dieser Probleme würde aber den Rahmen dieser Vorlesung sprengen.

3.1 Schwache Formulierung elliptischer Randwertprobleme

Wir beginnen mit der schwachen Formulierung elliptischer Randwertprobleme in Divergenzform und den dazugehörigen funktionalanalytischen Grundlagen. Der Einfachheit halber betrachten wir vor allem Gleichungen zweiter Ordnung, werden an einigen Stellen aber auch die Erweiterung auf höhere Ordnung diskutieren. Das Randwertproblem in einem Gebiet $\Omega \subset \mathbb{R}^d$ besteht aus der Gleichung

$$-\nabla \cdot (A\nabla u) + cu = f - \nabla \cdot h \qquad \text{in } \Omega \tag{3.1}$$

mit den Randbedingungen

$$u = g_D \qquad \text{auf } \Gamma_D \tag{3.2}$$

$$(A\nabla u) \cdot n = g_N \qquad \text{auf } \Gamma_N, \tag{3.3}$$

wobei $\partial\Omega = \Gamma_D \cup \Gamma_N$ gelten soll. In der obigen Formulierung nehmen wir wieder an, dass $A(x)$ für alle $x \in \Omega$ positiv definit mit minimalem Eigenwert grösser gleich $a_0 > 0$ ist, sowie dass die skalare Funktion c nichtnegativ ist.

Wir leiten zunächst die schwache Formulierung des Randwertproblems (3.1)-(3.3) her. Dazu multiplizieren wir (3.1) mit einer Testfunktion v und integrieren über Ω. Die beiden Divergenzterme auf der linken Seite können wir mit Hilfe des Gauss'schen Integralsatzes umformen und erhalten so

$$\int_\Omega ((A\nabla)u \cdot \nabla v + cuv)\ dx - \int_{\partial\Omega} (A\nabla u) \cdot nv\ d\sigma = \int_\Omega (fv + h \cdot \nabla v)\ dx - \int_{\partial\Omega} h \cdot nv\ d\sigma.$$

Nun schränken wir uns auf Testfunktionen ein, die auf Γ_D verschwinden, und setzen auf Γ_N die Randbedingung ein, woraus wir die Identität

$$\int_\Omega ((A\nabla)u \cdot \nabla v + cuv)\ dx = \int_\Omega (fv + h \cdot \nabla v)\ dx + \int_{\Gamma_N} (g_n - h \cdot n)v\ d\sigma. \tag{3.4}$$

erhalten. Wir sehen, dass wir es auf der linken Seite mit einer *Bilinearform*

$$B(u, v) := \int_\Omega ((A\nabla)u \cdot \nabla v + cuv)\ dx \tag{3.5}$$

und auf der rechten Seite mit einem *linearen Funktional* der Testfunktion,

$$\ell(v) := \int_\Omega (fv + h \cdot \nabla v)\ dx + \int_{\Gamma_N} (g_n - h \cdot n)v\ d\sigma \tag{3.6}$$

zu tun haben.

Funktionalanalytisch machen Bilinearformen und lineare Funktionale nur auf geeigneten Vektorräumen Sinn. Deshalb drängt sich sofort die Frage nach einer geeigneten Wahl von Räumen für die schwache Formulierung der Gleichung auf. Wie wir im folgenden sehen werden, führt dies in natürlicher Weise auf die sogenannten Sobolev-Räume.

3.1.1 Sobolev-Räume

Wir beginnen diesen Abschnitt mit einer kleinen funktionalanalytischen Erinnerung:

E1 Ein *normierter Raum* X ist ein Vektorraum (d.h. für $x, y \in X$ und $\alpha, \beta \in \mathbb{R}$ ist $\alpha x + \beta y \in X$) mit einer Norm, d.h. einer Abbildung $\|.\| : X \to \mathbb{R}_+$, sodass $\|x\| > 0$ für $x \neq 0$ und die Dreiecksungleichung $\|x + y\| \leq \|x\| + \|y\|$ gilt.

E2 Ein *Banachraum* X ist ein vollständiger normierter Raum, d.h. die Häufungspunkte jeder Folge $(x_n) \subset X$ liegen wieder in X.

E3 Ein *Hilbertraum* X ist ein Banachraum, dessen Norm von einem Skalarprodukt erzeugt wird, d.h. $\|x\| = \sqrt{\langle x, x\rangle}$. Das Skalarprodukt $\langle ., .\rangle : X \times X \to \mathbb{R}$ ist eine bilineare und symmetrische Abbildung, sodass $\langle x, x\rangle > 0$ für $x \neq 0$ gilt.

E4 Eine *Bilinearform* $B : X \times X \to \mathbb{R}$ ist eine in jeder Komponente lineare Abbildung. Ein Skalarprodukt ist ein Beispiel einer Bilinearform. Die Bilinearform B ist stetig, wenn $B(x, y) \leq C\|x\|\|y\|$ für alle $x, y \in X$ gilt, mit einer fixen Konstante C.

E5 Ein *lineares Funktional* $\ell : X \to \mathbb{R}$ ist eine lineare Abbildung nach $\mathbb{R}$. Das lineare Funktional heisst stetig, wenn $\ell(x) \leq C\|x\|$ für alle $x \in X$ gilt, mit einer fixen Konstante C.

E6 Der *Dualraum* X^* eines normierten Raumes X ist der Raum aller stetigen linearen Funktionale auf X. Mit

$$\|\ell\| = \sup_{x \in X \setminus \{0\}} \frac{|\ell(x)|}{\|x\|} = \sup_{x \in X, \|x\| \leq 1} |\ell(x)| = \inf\{\ C > 0 \mid \ell(x) \leq C\|x\|, \forall\ x \in X\ \}$$

ist X^* wieder ein normierter Raum. Falls X Banach-(Hilbert-)raum ist, dann ist auch X^* Banach-(Hilbert-)raum.

E7 In einem Hilbertraum gilt der *Riesz'sche Darstellungssatz:* Jedem linearen Funktional $\ell \in X^*$ kann ein eindeutiges Elements $x_\ell \in X$ zugeordnet werden, sodass

$$\ell(x) = \langle x_\ell, x\rangle \qquad \forall\ x \in X$$

gilt. Natürlich gilt auch die Umkehr, für jedes $y \in X$ ist $\ell_y(x) = \langle y, x\rangle$ ein lineares Funktional. Damit lässt sich (durch $\ell \leftrightarrow x_\ell$) der Raum X^* mit X identifizieren.

Die ersten unendlichdimensionalen Banach- und Hilberträume, die man üblicherweise betrachtet, sind die *Lebesgue-Räume* $L^p(\Omega)$. Für $1 \le p < \infty$ gilt

$$L^p(\Omega) = \{\ u : \Omega \to \overline{\mathbb{R}} \mid u \text{ messbar }, \int_\Omega |u|^p\ dx < \infty\ \},$$

wobei hier das Lebesgue-Integral verwendet wird. Die Norm in $L^p(\Omega)$ ist gegeben durch

$$\|u\|_p = \left(\int_\Omega |u|^p\ dx\right)^{1/p}.$$

Auf einem beschränkten Gebiet gilt klarerweise $L^p(\Omega) \subset L^q(\Omega)$ für $p \ge q$, auf unbeschränkten Gebieten gilt keine solche Inklusion.

Der Dualraum von $L^p(\Omega)$ $(1 < p < \infty)$ kann mit $L^{p_*}(\Omega)$ identifiziert werden, wobei $\frac{1}{p} + \frac{1}{p_*} = 1$ gilt. Die linearen Funktionale sind dann von der Form

$$\ell(u) = \int_\Omega uv\ dx$$

für ein $v \in L^{p_*}$. Mit der Hölder-Ungleichung überzeugt man sich leicht von der Stetigkeit solcher linearer Funktionale, es gilt ja

$$|\ell(u)| = |\int_\Omega uv\ dx| \le \left(\int_\Omega |u|^p\ dx\right)^{1/p} \left(\int_\Omega |v|^{p_*}\ dx\right)^{1/p_*} = \|u\|_p \|v\|_{p_*}.$$

Im Fall $p = 1$ erhält man aus der obigen Rechnung $p_* = \infty$. Der entsprechende Lebesgue-Raum ist

$$L^p(\Omega) = \{\ u : \Omega \to \overline{\mathbb{R}} \mid u \text{ messbar }, \operatorname{ess\,sup}_x |u(x)|\ \},$$

wobei dass *essentielle Supremum* definiert ist als

$$\operatorname{ess\,sup}_x |u(x)| = \inf\{\ N \mid N \text{ Nullmenge }\ \} \sup_x |u(x)|.$$

Die Norm in $L^\infty(\Omega)$ ist definiert durch

$$\|u\|_\infty = \operatorname{ess\,sup}_x |u(x)|.$$

Die Umkehrung des Dualraums gilt aber nicht, der Dualraum von $L^\infty(\Omega)$ ist echt grösser als $L^1(\Omega)$.

Für partielle Differentialgleichungen benötigt man Verallgemeinerungen der Lebesgue-Räume in denen auch Ableitungen vorkommen. Dazu definiert man zunächst die *distributionelle Ableitung*, wieder über lineare Funktionale. Die distributionelle Ableitung ist im allgemeinen ein Funktional auf $C_0^\infty(\Omega)$, dem Raum der unendlich oft differenzierbaren Funktionen mit kompaktem Träger in Ω. Man identifiziert w mit der Ableitung $\frac{u}{x_j}$, wenn

$$w(\varphi) = -\int_\Omega u \frac{\partial \varphi}{\partial x_j}\, dx, \qquad \forall\ \varphi \in C_0^\infty(\Omega)$$

gilt. Diese Definition ist konsistent mit der klassischen Definition einer Ableitung, denn in diesem Fall kann man durch partielle Integration (und Ausnutzen der Tatsache, dass φ kompakten Träger hat, also am Rand verschwindet) zeigen, dass $w(\varphi) = \int_\Omega \varphi \frac{\partial u}{\partial x_j}\, dx$ gilt. Nun kann man testen, ob w ein stetiges lineares Funktional auf einem Lebesgue-Raum $L^{p*}(\Omega)$ ist. Wenn ja, dann können wir das lineare Funktional als Element des Dualraums mit einem $v \in L^p(\Omega)$ identifizieren und man spricht von $v = \frac{\partial u}{\partial x_j}$ als schwache Ableitung. Diese Überlegung ist auch die Basis für die Definition von Sobolevräumen. Man definiert mittels der distributionellen Ableitung

$$W^{1,p}(\Omega) = \{\ u \in L^p(\Omega) \mid \frac{\partial u}{\partial x_j} \in L^p(\Omega), j = 1, \ldots, d\ \}. \tag{3.7}$$

$W^{1,p}$ ist ein normierter Raum mit Norm

$$\|u\|_{1,p} := \left(\|u\|_p + \sum_{j=1}^{d} \|\frac{\partial u}{\partial x_j}\|_p \right)^{1/p}. \tag{3.8}$$

Man verwendet üblicherweise die Notation

$$|u|_{1,p} := \left(\sum_{j=1}^{d} \|\frac{\partial u}{\partial x_j}\|_p \right)^{1/p}. \tag{3.9}$$

für die Halbnorm betreffend die ersten Ableitungen ($|u|_{1,p} = 0$ für u konstant). Die Konvergenz einer Folge $u_n \to u$ in der Norm von $W^{1,p}(\Omega)$ impliziert die Konvergenz $u_n \to u$ in $L^p(\Omega)$ und $\frac{\partial u}{\partial x_j} \to v_j$ in $L^p(\Omega)$. Weiters gilt für glatte Testfunktionen

$$-\int_\Omega u \frac{\partial \varphi}{\partial x_j}\, dx = \lim \left(-\int_\Omega u_n \frac{\partial \varphi}{\partial x_j}\, dx \right) = \lim \left(\int_\Omega \varphi \frac{\partial u_n}{\partial x_j}\, dx \right) = \int_\Omega \varphi v_j\, dx$$

und somit folgt $v_j = \frac{\partial u}{\partial x_j}$ im obigen Sinne. Damit gilt aber auch $u \in W^{1,p}(\Omega)$. Also ist $W^{1,p}(\Omega)$ vollständig und damit ein Banachraum.

Für $p = 2$ erhält man wie im Falle der Lebesgue-Räume sogar einen Hilbert-Raum, üblicherweise mit der Bezeichnung $H^1(\Omega) := W^{1,2}(\Omega)$. Das Skalarprodukt in $H^1(\Omega)$ ist gegeben durch

$$\langle u, v \rangle_1 := \int_\Omega \left(uv + \sum_{j=1}^{d} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_j} \right)\, dx,$$

also die Summe aus L^2-Skalarprodukten der Funktion und ihrer Ableitungen. Wie wir sehen werden, ist der Hilbert-Raum $H^1(\Omega)$ der wichtigste bei der Analysis linearer Gleichungen. Sobolevräume für $p \neq 2$ verwendet man meist bei nichtlinearen Gleichungen.

Durch Iteration der obigen Definitionen kann man analog höherere distributionelle Ableitungen als

$$w(\varphi) = (-1)^{|\alpha|} \int_\Omega u \frac{\partial^{|\alpha|}\varphi}{\partial x^\alpha}\, dx, \qquad \forall\ \varphi \in C_0^\infty(\Omega)$$

und Sobolevräume höherer Ordnung als

$$W^{k,p}(\Omega) = \{\ u \in L^p(\Omega \mid \frac{\partial^{|\alpha|}u}{\partial x^\alpha} \in L^p(\Omega),\ |\alpha| \leq k\ \}$$

definieren. Wiederum erhält man damit Banachräume mit der Norm

$$\|u\|_{k,p} := \left(\sum_{|\alpha|\leq k} \|\frac{\partial^{|\alpha|}u}{\partial x^\alpha}\|_p \right)^{1/p}$$

und definiert die Halbnorm

$$|u|_{k,p} := \left(\sum_{|\alpha|=k} \|\frac{\partial^{|\alpha|}u}{\partial x^\alpha}\|_p \right)^{1/p},$$

die auf Polynomen vom Grad kleiner oder gleich $k-1$ verschwindet. Im Fall $p = 2$ ist $H^k(\Omega) := W^{k,2}(\Omega)$ ein Hilbertraum mit dem Skalarprodukt

$$\langle u, v\rangle_k = \sum_{|\alpha|\leq k} \int_\Omega \frac{\partial^{|\alpha|}u}{\partial x^\alpha} \frac{\partial^{|\alpha|}v}{\partial x^\alpha}\, dx.$$

Die schwache Formulierung partieller Differentialgleichungen höherer als zweiter Ordnung führt in natürlicher Weise zu einem Sobolevraum mit $k > 1$.

Der Dualraum von $H^1(\Omega)$ wird üblicherweise mit $H^{-1}(\Omega)$ bezeichnet. Per Definition gilt ja die Einbettung $H^1(\Omega) \hookrightarrow L^2(\Omega)$. Damit definiert jedes Element $\psi \in L^2(\Omega)$ ein stetiges lineares Funktional $u \mapsto \int_\Omega u\psi\, dx$ auf $H^1(\Omega)$. Also folgt $H^{-1}(\Omega) \subset L^2(\Omega) \subset H^1(\Omega)$. Umgekehrt ist $H^1(\Omega)$ aber auch Hilbertraum und somit kann der Dualraum $H^{-1}(\Omega)$ wiederum mit $H^1(\Omega)$ identifiziert werden. Für ein lineares Funktional $w \in H^{-1}(\Omega)$ bedeutet diese Identifizierung ein Element $v \in H^1(\Omega)$ zu finden, sodass

$$\langle v, \varphi\rangle = w(\varphi) \qquad \forall\ \varphi \in H^1(\Omega)$$

gilt. Schreiben wir das Skalarprodukt aus, dann erhalten wir die Variationsgleichung

$$\int_\Omega (v\varphi + \nabla v \cdot \nabla\varphi)\, dx = w(\varphi) \qquad \forall\ \varphi \in H^1(\Omega).$$

Man erkennt, dass die Bilinearform auf der rechten Seite genau der schwachen Formulierung des Differentialoperators $-\Delta v + v$ entspricht. D.h., durch Anwendung des Riesz'schen Darstellungssatzes in $H^1(\Omega)$ lösen wir eigentlich eine elliptische Differentialgleichung zweiter

Ordnung mit rechter Seite in $H^{-1}(\Omega)$. Diese Sichtweise werden wir im nächsten Kapitel auf die Analysis schwacher Lösungen allgemeinerer Gleichungen übertragen.

Neben der Differentialgleichung in Ω haben wir immer auch Randbedingungen benötigt. Da Funktionen in Sobolev-Räumen über Lebesgue-Räume definiert werden, stellt sich sofort die Frage, ob bzw. in welchem Sinn die Auswertung von Funktionen aus $H^1(\Omega)$ definiert werden kann (der Rand ist ja eine Nullmenge). Dies geschieht durch sogenannte *Spursätze*, die jeder Funktion in einem Sobolevraum einen distributionellen Randwert zuordnen, obwohl die Punktauswertung am Rand keinen Sinn ergibt. Für glatte Funktionen u, ψ und φ mit $\frac{\partial\varphi}{\partial n} = \psi$ auf $\partial\Omega$ gilt ja nach dem Gauss'schen Integralsatz

$$\int_{\partial\Omega} u\psi \; d\sigma = \int_{\partial\Omega} u\frac{\partial\varphi}{\partial n} \; d\sigma = \int_\Omega \nabla\cdot(u\nabla\varphi) \; dx = \int_\Omega (u\Delta\varphi + \nabla u\cdot\nabla\varphi) \; dx.$$

Die rechte Seite kann auch sinnvoll auf Funktionen in $H^1(\Omega)$ erweitert werden, und damit erhält man sofort auch die distributionelle Definition eines Randwerts auf $\partial\Omega$, die sogenannte Spur (die man implizit auch immer mit dem Randwert gleichsetzt). Man kann zeigen, dass die Spur einer Funktion in $H^1(\Omega)$ immer einer Funktion in $L^2(\partial\Omega)$ entspricht. Es gilt sogar mehr, die Spur ist ein Element des (fraktionalen) Sobolevraums $H^{1/2}(\partial\Omega) \hookrightarrow L^2(\partial\Omega)$. Wie der Exponent $1/2$ schon nahelegt, ist $H^{1/2}(D)$ immer ein Zwischenraum zwischen $L^2(D)$ und $H^1(D)$, es gilt $L^2(D) \hookrightarrow H^{1/2}(D) \hookrightarrow H^1(D)$. Im Rahmen der Interpolationstheorie von Hilbert-Räumen (cf. [4]) kann man $H^{1/2}(D)$ tatsächlich als Interpolation von $L^2(D)$ und $H^1(D)$ betrachten. Wir gehen hier nicht weiter auf dieses fortgeschrittene Thema ein, erwähnen aber die Interpolationsgleichung

$$\|u\|_{H^{1/2}} = \sqrt{\|u\|_{L^2}\|u\|_{H^1}}.$$

Zum Abschluss dieser Diskussion liefern wir noch eine genaue Formulierung des Spursatzes:

Satz 3.1 (Spursatz). *Sei $\Gamma \subset \partial\Omega$ ein Teil des Randes mit positivem Maß und $\partial\Omega$ stückweise Lipschitz. Dann existiert ein stetiger linearer Operator $T : H^1(\Omega) \to H^{1/2}(\Gamma)$, sodass*

$$Tu = u|_\Gamma \qquad \forall u \in C(\overline{\Omega})$$

gilt. Der Operator T ist surjektiv, d.h. für jedes $v \in H^{1/2}(\Gamma)$ existiert ein $u \in H^1(\Omega)$ mit $Tu = v$.

Der Spursatz besagt, dass es eine Erweiterung der Randwerte von stetigen Funktionen auf Funktionen in $H^1(\Omega)$ gibt, und der Raum $H^{1/2}(\Gamma)$ genau aus diesen Randwerten besteht. Eine Erweiterung der Spur ist auch auf Sobolevräume $H^k(\Omega)$ möglich, es ist dann $T : H^k(\Omega) \to H^{k-1/2}(\Gamma)$ ein stetiger surjektiver Operator. Grob gesagt verliert man also beim Auswerten der Spur immer eine halbe Differentiationsordnung. Zur einfacheren Notation werden wir im Folgenden immer u für die Randwerte schreiben, damit aber eigentlich die Spur Tu assoziieren.

Neben den Randwerten von u (Dirichlet-Werte) haben wir auch die Normalableitungen $\frac{\partial u}{\partial n}$ (Neumann-Werte) in den Randbedingungen verwendet. Da nun eine Ableitung mehr auftritt, könnte man vermuten, dass $\frac{\partial u}{\partial n} \in H^{1/2-1}(\partial\Omega)$ definierbar ist. Dies ist auch tatsächlich der Fall, wenn man $H^{-1/2}(\Gamma)$ als Dualraum von $H^{1/2}(\Gamma)$ definiert (analog zur Definition von $H^{-1}(\Omega)$). Wir haben ja für glatte Funktionen nach dem Gauss'schen Integralsatz

$$\int_{\partial\Omega} \frac{\partial u}{\partial n} v \; dx = \int_\Omega (\nabla u\cdot\nabla v + \Delta uv) \; dx = \int_\Omega (\nabla u\cdot\nabla v + uv) \; dx - \int_\Omega (-\Delta u + u)v \; dx.$$

Die rechte Seite ist zunächst ein Skalarprodukt auf $H^1(\Omega)$, also auf Funktionen dieser Klasse auch sinnvoll definiert, und im zweiten Term kann man $-\Delta u + u \in H^{-1}(\Omega)$ wieder zu einem stetigen linearen Funktional auf $H^1(\Omega)$ erweitern. In dieser Sichtweise definiert also $\frac{\partial u}{\partial n}$ ein stetiges lineares Funktional auf den Spuren von Funktionen in $H^1(\Omega)$, nach dem Spursatz also genau in $H^{-1/2}(\Omega)$. Die Neumann-Randbedingung kann also in diesem Sobolev-Raum verstanden werden, und somit sollte man für Neumann-Daten auch $g_N \in H^{-1/2}(\Gamma_N)$ fordern.

Ein anderes interessantes Problem ist die Einbettung von Sobolev-Räumen in Räume stetiger Funktionen oder Hölder-Räume. Im eindimensionalen Fall haben wir ja gesehen, dass Funktionen in $H^1(\Omega)$ immer auch stetig sind, und sich die Supremumsnorm durch ein Vielfaches der H^1-Norm abschätzen lässt. In so einem Fall spricht man von einer Einbettung des Raumes $H^1(\Omega)$ in $C(\Omega)$. Solche Einbettungen sind aber immer dimensionsabhängig, was an einfachen Beispiel deutlich wird. Betrachten wir z.B. Funktionen der Form $u(x) = |x|^\alpha$ auf der Einheitskugel in $\mathbb{R}^d$ und mit $\alpha \in \mathbb{R}$. Dann gilt (in Polarkoordinaten)

$$\int_\Omega u(x)^2 \, dx = \int_\Omega |x|^{2\alpha} dx = S_d \int_0^1 r^{2\alpha} r^{d-1} dr,$$

wobei S_d die Oberfläche der Einheitssphäre in $\mathbb{R}^d$ ist. Da man Funktionen der Form r^β genau für $\beta > -1$ integrieren kann, folgt $u \in L^2(\Omega)$ für $\alpha > -\frac{d}{2}$. Der Gradient von u ist gegeben durch $\nabla u = \alpha x |x|^{\alpha-2}$ und es gilt

$$\int_\Omega |\nabla u|^2 \, dx = 2\alpha \int_\Omega |x|^{2\alpha-2} dx = \int_0^1 r^{2\alpha+d-3} dr.$$

Damit ist der Gradient integrierbar für $\alpha > \frac{1-d}{2}$. In Raumdimension 1 folgt damit $\alpha > 0$, was mit unserem vorigen Resultat über die Stetigkeit von H^1-Funktionen auch übereinstimmt. Für Raumdimensionen $d > 1$ ist auch $\alpha < 0$ möglich, und da $u(x)$ in diesem Fall in $x = 0$ unstetig ist, gibt es auch unstetige (sogar unbeschränkte) Elemente von $H^1(\Omega)$. Um eine Einbettung in $C(\Omega)$ zu erreichen, muss man dann zu einem $W^{1,p}(\Omega)$ mit höherem p oder zu einem $H^k(\Omega)$ mit höherem k übergehen. Allgemein gilt folgendes Resultat:

Satz 3.2 (Einbettungssatz für Sobolevräume). *Sei $\Omega \subset \mathbb{R}^d$ ein glattes, beschränktes Gebiet, und $kp > d$. Dann gilt die Einbettung $W^{k,p}(\Omega) \subset C(\overline{\Omega})$. Für $(k-j)p > n$ gilt weiters die Einbettung $W^{k,p}(\Omega) \subset C^j(\overline{\Omega})$.*

Der Einbettungssatz gibt nicht nur Auskunft über die Stetigkeit von Funktionen in Sobolevräumen, sondern auch über die Stetigkeit von Ableitungen. Man kann sich für eine Differentialgleichung zweiter (oder höherer) Ordnung anhand der jeweiligen Raumdimension ausrechnen, wie gross k bzw. p sein müssten, um aus einer Lösung in einem Sobolevraum durch Einbettung wieder eine stetige Lösung zu erhalten.

Aus den obigen Funktionen der Form $u(x) = |x|^\alpha$ sehen wir auch, dass Funktionen in Sobolevräume auch in Lebesgue-Räumen mit höherem Exponenten liegen. Es gilt z.B. $u \in H^1(\Omega)$ für $\alpha > \frac{1-d}{2}$. Umgekehrt ist $u \in L^p(\Omega)$ für $p\alpha + d > 0$. Setzen wir die untere Schranke für α ein, so folgt die Bedingung $p\alpha + d > 0$ sicher aus $p(1-d) + 2d > 0$, oder äquivalent $p < \frac{2d}{d-1}$. Man beachte, dass $\frac{2d}{d-1} = 2 + \frac{2}{d-1} > 2$ gilt, in Raumdimension 1 gelangt man sogar bis zu $p = \infty$. Allgemein gilt folgendes Resultat:

Satz 3.3 (Einbettung von Sobolev- in Lebesgueräume). *Sei $\Omega \subset \mathbb{R}^d$ ein beschränktes Gebiet, und $q \leq q_* = \frac{dp}{d-pk}$. Dann gilt die Einbettung $W^{k,p}(\Omega) \subset L^q(\Omega)$, diese ist sogar kompakt für $q < q_*$.*

Einbettung von Sobolev in Lebesgue-Räume lässt sich direkt (durch Einbettung der entsprechenden Ableitung) zur Einbettung von Sobolevräume $W^{k,p}(\Omega)$ in $W^{m,q}(\Omega)$ mit $m < k$ verallgemeinern. Die einfachste (und unten auch verwendete) Folgerung ist die kompakte Einbettung $W^{k,p}(\Omega) \hookrightarrow W^{m,q}(\Omega)$.

In manchen Fällen ist es wichtig äquivalente Normen zur obigen Sobolevraum-Norm zu verwenden, in dem man die Halbnorm mit anderen Halbnormen als der L^p-Norm der Funktion kombiniert. Solche Normen sind von der Form

$$|||u||| = \left(|u|_{k,p}^p + \sum_{j=1}^{m} f_i(u)^p \right)^{1/p}, \tag{3.10}$$

wobei f_i ein System von Halbnormen ist. Beispiele dafür sind

$$|||u||| = \left(|u|_{1,p}^p + |\int_\Omega u \, dx|^p \right)^{1/p},$$

oder

$$|||u||| = \left(|u|_{k,p}^p + \int_{\Gamma_d} |u(x)|^p \, dx \right)^{1/p}.$$

In diesen Fällen gilt der *Normierungssatz von Sobolev:*

Satz 3.4. *Sei $f_i : W^{k,p}(\Omega) \to \mathbb{R}$ ein System von Halbnormen, sodass*

$$0 \le f_i(u) \le C_i \|u\|_{k,p} \qquad \forall\ u \in W^{k,p}(\Omega)$$

gilt. Weiters soll für jedes Polynom $v \neq 0$ vom Grad kleiner oder gleich $k-1$ eine Halbnorm f_i existieren, sodass $f_i(v) \neq 0$. Dann sind die Normen $\|\|_{k,p}$ und $|||.|||$ wie in (3.10) äquivalent, d.h. es existieren Konstante $\alpha, \beta > 0$, sodass

$$\alpha|||u||| \le \|u\|_{k,p} \le \beta|||u||| \qquad \forall\ u \in W^{k,p}(\Omega).$$

Proof. Es gilt nach den obigen Voraussetzungen an f_i:

$$|||u|||^p = |u|_{k,p} + \sum f_i(u)^p \le \|u\|_{k,p} + \sum C_i^p \|u\|_{k,p}^p \le (1 + \sum C_i^p)\|u\|_{k,p}^p$$

und mit $\alpha := (1 + \sum C_i^p)^{-1/p}$ folgt die erste Ungleichung.

Die zweite Ungleichung beweisen wir indirekt durch Widerspruch. Wir nehmen an, es existiert keine solche Konstante β. Dann existiert eine Folge $u_n \in W^{k,p}(\Omega)$ sodass

$$\|u_n\|_{k,p} > n|||u_n|||.$$

Wir definieren nun $v_n := \frac{u_n}{\|u_n\|_{k,p}}$. Dann gilt $\|v_n\|_{k,p} = 1$ und $|||v_n||| < \frac{1}{n}$. Insbesonders folgt $|v_n|_{k,p} < \frac{1}{n}$. Da v_n uniform beschränkt in $W^{k,p}$ ist, existiert (wegen der Kompaktheit der Einbettung) eine konvergente Teilfolge $v_{n'}$ in $W^{k-1,p}(\Omega)$, deren Grenzwert wir mit v bezeichnen. Da aber die Ableitungen der Ordnung k gegen Null konvergieren ($|v_{n'}|_{k,p} < \frac{1}{n'}$) muss für den Grenzwert auch $|v|_{k,p} = 0$ gelten, d.h. alle Ableitungen der Ordnung k verschwinden. Damit ist v ein Polynom mit Grad kleiner gleich $k-1$. Wegen $f_i(v_{n'}) < \frac{1}{n'}$ folgt $f_i(v_{n'}) \to 0$. Es gilt aber auch wegen Dreiecksungleichung und Annahmen an f_i

$$|f_i(v_{n'}) - f_i(v)| \le f_i(v_{n'} - v) \le C_i \|v_{n'} - v\|_{k,p} \to 0,$$

und damit muss $f_i(v) = 0$ gelten für alle i. Aus den Annahmen über die f_i folgt dann aber sofort $v = 0$. Dies ist aber ein Widerspruch wegen

$$0 = \|v\|_{k,p} = \lim \|v_{n'}\|_{k,p} = 1.$$

□

Konsequenzen aus dem Normierungssatz von Sobolev sind unter anderem die Poincare-Ungleichung und die Friedrichs-Ungleichung. Die Poincare-Ungleichung ergibt sich für $k = 1$ mit der Seminorm $f_1(u) = |\int_\Omega u \, dx|$, d.h. es gilt

$$\|u\|_{1,p} \leq C \left(|\int_\Omega u \, dx|^p + |u|_{1,p}^p \right)^{1/p}.$$

Insbesonders folgt für Funktionen mit Mittelwert Null die Abschätzung

$$\|u\|_{1,p} \leq C|u|_{1,p}^p.$$

Friedrichs-Ungleichungen erhält man für $k = 1$ mit der Wahl $f_1(u)^p = \int_\Gamma |u|^p \, d\sigma$, und es folgt

$$\|u\|_{1,p} \leq C \left(\int_\Gamma |u|^p \, d\sigma + |u|_{1,p}^p \right)^{1/p}.$$

Für Funktionen mit homogenen Dirichlet-Randwerten auf Γ_D lässt sich wiederum die H^1-Norm durch die Halbnorm abschätzen. Diese Funktionen fasst man meist in einem eigenen Unterraum zusammen, mit der Notation

$$H_0^1(\Omega) := \{ \, u \in H^1(\Omega) \mid u = 0 \text{ auf } \Gamma_D \, \},$$

und verwendet darin wiederum das Skalarprodukt von $H^1(\Omega)$. Da die Spur stetig ist, kann eine Folge in $H_0^1(\Omega)$ nur Häufungspunkte mit verschwindender Spur auf Γ_D haben. Folglich ist $H_0^1(\Omega)$ abgeschlossen, also selbst ein Hilbertraum.

3.1.2 Schwache Lösungen

Aus der obigen Theorie der Sobolevräume können wir nun die schwache Formulierung in Funktionenräumen angeben. Dazu suchen wir uns zunächst eine beliebige Funktion in $H^1(\Omega)$ mit der Spur g_D auf Γ_D (so eine Funktion existiert nach dem Spursatz für $g_D \in H^{1/2}(\Gamma_D)$), die wir wieder mit g_D bezeichnen. Dann können wir eine Lösung $u \in g_D + H_0^1(\Omega)$ suchen, sodass

$$B(u, v) = \ell(v) \qquad \forall \, v \in H_0^1(\Omega) \tag{3.11}$$

gilt. Wir sehen sofort, dass es genügt $A \in L^\infty(\Omega; \mathbb{R}^{d \times d})$ und $c \in L^\infty(\Omega)$ zu fordern, um Stetigkeit der Bilinearform B zu erhalten. Es gilt ja

$$B(u, v) = |\int_\Omega (A \nabla u \nabla v + cuv) \, dx| \leq \max\{\|A\|_\infty, \|c\|_\infty\} \int_\Omega (|\nabla u \cdot \nabla v| + |uv|) \, dx$$

und das Integral können wir mit der Cauchy-Schwarz Ungleichung durch $\|u\|_{1,2}\|v\|_{1,2}$ abschätzen. Die Anforderung an c könnte sogar noch abgeschwächt werden auf $c \in L^q(\Omega)$ für dimensionsabhängiges $q < \infty$, da man ja durch Einbettungssätze $u, v \in L^p(\Omega)$ mit $p > 2$ erhält (und die entsprechenden Normen können durch die H^1-Norm abgeschätzt werden).

Unter den üblichen Elliptizitätsannahmen $A(x) \geq \alpha I$ und $c \geq 0$ kann man die *Koerzivität der Bilinearform* zeigen. Dazu zuerst die passende Definition:

Definition 3.5. Eine Bilinearform $B : X \times X \to \mathbb{R}$ auf einem Hilbertraum X heisst koerziv, wenn eine Konstante $\gamma > 0$ existiert, sodass

$$B(u,u) \geq \gamma \|u\|^2 \qquad \forall\ u \in X$$

gilt.

Wir sehen, dass

$$B(u,u) = \int_\Omega (A\nabla u \cdot \nabla u + cu^2)\ dx \geq \alpha |u|_{1,2} + \int_\Omega cu^2\ dx.$$

Gilt $c(x) \geq c_0$ auf einer Menge von positivem Maß, dann sehen wir sofort, dass $f_1(u) = \sqrt{\int_\Omega cu^2\ dx}$ die Bedingungen des Sobolev'schen Normierungssatzes erfüllt, und damit können wir Koerzivität mit einer Konstante γ abhängig von c und α zeigen.

Im Fall $c \equiv 0$ kann man immer noch Koerzivität erhalten, nämlich durch die Randbedingung. Ist Γ_D eine Menge von positivem Maß, so können wir mit $\tilde{u} = u - g_D$ das Problem als Variationsgleichung für $\tilde{u}$ in $H^1_0(\Omega)$ schreiben (mit gleicher Bilinearform und geänderter rechter Seite). Dann verwenden wir einfach die Friedrichs-Ungleichung, die auf $H^1_0(\Omega)$ impliziert, dass

$$|u|_{1,2} \geq C\|u\|_{1,2}$$

gilt, da ja das Randintegral verschwindet.

Um auf der rechten Seite ein stetiges lineares Funktional zu erhalten, würde es genügen, dass $h \in L^2(\Omega;\mathbb{R}^d)$, $g_N \in H^{-1/2}(\Gamma_N)$, und $f \in L^r(\Omega)$ gilt, mit dimensionsabhängigem $r < 2$ (wieder erhalten aus passenden Einbettungssätzen).

Wir finden nun folgende Situation: wegen der Stetigkeit und Koerzivität gilt

$$c\|u\|^2 \leq B(u,u) \leq C\|u\|^2,$$

d.h. B definiert ein äquivalentes Skalarprodukt auf $H^1(\Omega)$. In $H^1(\Omega)$ mit dem neuen Skalarprodukt B gibt es nach dem Riesz'schen Darstellungssatz wieder ein eindeutiges Element $u \in H^1(\Omega)$, dass das lineare Funktional darstellt, d.h. eine schwache Lösung $u \in H^1(\Omega)$ von (3.11).

Damit haben wir direkt die Existenz und Eindeutigkeit einer schwachen Lösung gezeigt, allerdings ist dieser Beweis über den Riesz'schen Darstellungssatz nicht konstruktiv. Deshalb geben wir noch einen zweiten Beweis an, der auch gleichzeitig die Konvergenz einer einfachen iterativen Methode zur Lösung der Variationsgleichung liefert.

Satz 3.6 (Lax-Milgram Lemma). *Sei $B : X \times X \to \mathbb{R}$ eine symmetrische, stetige und koerzive Bilinearform, d.h. es existieren Konstante $c, C \in \mathbb{R}^+$, sodass*

$$B(u,v) \leq C\|u\|\|v\| \qquad \forall\ u,v \in X \tag{3.12}$$

und

$$B(u,u) \geq c\|u\|^2 \qquad \forall\ u \in X. \tag{3.13}$$

Dann existiert für jedes $\ell \in X^*$ eine eindeutige Lösung $\hat{u} \in X$ der Variationsgleichung

$$B(u,v) = \ell(v) \qquad \forall\ v \in X. \tag{3.14}$$

Die Fixpunktiteration (Richardson-Verfahren)

$$\langle u^{k+1}, v\rangle = \langle u^k, v\rangle - \tau(B(u^k,v) - \ell(v)) \qquad \forall\ v \in X \tag{3.15}$$

liefert eine konvergente Folge $u^{k+1} \to \hat{u}$ für $\tau < \frac{1}{C}$.

Proof. Wir betrachten den Fixpunktoperator $A : X \to X$, in schwacher Form definiert durch

$$\langle Au, v\rangle := \langle u, v\rangle - \tau B(u, v) \qquad \forall\ v \in X.$$

Weiters erhalten wir eine rechte Seite $f \in X$, definiert durch

$$\langle f, v\rangle = \tau \ell(v) \qquad \forall\ v \in X.$$

Dann können wir die Fixpunktiteration auch als

$$u^{k+1} = Au^k - f$$

schreiben, und erhalten sofort die Lösung

$$u^k = A^k u^0 - \sum_{j=0}^{k-1} A^j f.$$

Für $\|A\| < 1$ konvergiert diese Neumann'sche Reihe gegen

$$\hat{u} = -\sum_{j=0}^{\infty} A^j f = (A - I)^{-1} f,$$

also folgt $(A - I)\hat{u} = f$, bzw. in schwacher Form

$$\langle \hat{u}, v\rangle - \tau B(\hat{u}, v) - \langle \hat{u}, v\rangle = -\tau \ell(v) \qquad \forall\ v \in X,$$

was äquivalent zu (3.14) ist.

Wir müssen also nur mehr die Operatornorm von A abschätzen. Dazu verwenden wir zunächst, dass A ein selbst-adjungierter Operator ist, d.h. es gilt

$$\langle Au, v\rangle = \langle Av, u\rangle \qquad \forall\ u, v \in X,$$

was man wegen der Symmetrie von B sofort sieht. Wegen der Stetigkeit der Bilinearform ist A beschränkt, und aus $\tau C < 1$ folgt

$$\langle Au, u\rangle = \langle u, u\rangle - \tau B(u, u) \geq \|u\|^2 (1 - \tau C) > 0.$$

Es gilt dann auch für $u, v \in X$, dass

$$0 \leq \langle A(u - v), A(u - v)\rangle = \langle Au, u\rangle + \langle Av, v\rangle - 2\langle Au, v\rangle.$$

Also folgt

$$\|A\| = \sup_{\|u\|\leq 1, \|v\|\leq 1} \langle Au, v\rangle \leq \frac{1}{2} \left(\sup_{\|u\|\leq 1} \langle Au, u\rangle + \sup_{\|v\|\leq 1} \langle Av, v\rangle \right) = \sup_{\|u\|\leq 1} \langle Au, u\rangle.$$

Wegen der Koerzivität folgt aber

$$\langle Au, u\rangle = \langle u, u\rangle - \tau B(u, u) \leq \|u\|^2 - \tau c \|u\|^2,$$

und damit $\|A\| \leq 1 - \tau c < 1$. □

Wir haben nun Existenz, Eindeutigkeit einer Lösung sowie bereits die Konvergenz einer iterativen Methode nachgewiesen, es fehlt nur noch die Stabilität der Lösung. Diese erhalten wir sehr direkt aus der Koerzivität:

Satz 3.7. *Es gelten die Voraussetzungen von Satz 3.6. Dann gilt die Stabilitätsabschätzung*

$$\|\hat{u}\| \leq \frac{1}{c}\|\ell\| \tag{3.16}$$

Proof. Sei $\hat{u} \in X$ die Lösung von (3.14). Dann erhalten wir mit der Testfunktion $v = \hat{u}$:

$$c\|\hat{u}\| \leq B(\hat{u}, \hat{u}) = \ell(\hat{u}) \leq \|\ell\|\|\hat{u}\|,$$

was direkt (3.16) impliziert. □

Wir sehen, dass die Stabilitätskonstante indirekt proportional zur Koerzivitätskonstante c und proportional zur Norm der rechten Seite ist. Die Konstante C aus der Stetigkeitsbedingung kommt hier nicht vor, sie ist aber in der Norm der rechten Seite versteckt, da man $\|\ell\|$ immer als relative Grösse zu B sehen sollte. Skaliert man die Bilinearform (durch Division durch C) so, dass die Norm von B eins ergibt, dann ist die effektive Koerzivitätskonstante $\frac{C}{c}$, diese Grösse ist analog zur Konditionszahl einer Matrix.

Da wir die Bedingungen des Lax-Milgram Lemmas für die schwache Formulierung von (3.1)-(3.3) oben unter vernünftigen Bedingungen an die Daten nachgeprüft haben, erhalten wir nun sofort die Existenz, Eindeutigkeit, und Stabilität für die elliptische Differentialgleichung zweiter Ordnung.

3.1.3 Variationsprinzip

Wir werden nun noch kurz ein Variationsprinzip diskutieren, das die Variationsgleichung mit einer Energieminimierung in Beziehung setzt. Dazu definieren wir die (quadratische) Energie

$$E(u) := \frac{1}{2}B(u,u) - \ell(u). \tag{3.17}$$

Nun berechnen wir die Richtungsableitungen des Energiefunktionals in eine beliebige Richtung $v \in X$, d.h.

$$E'(u)v = \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon}\left(E(u+\epsilon v) - E(u)\right)$$

Wegen der Bilinearität von B und der Linearität von ℓ folgt

$$\begin{aligned} E(u+\epsilon v) &= \frac{1}{2}B(u+\epsilon v, u+\epsilon v) + \ell(u+\epsilon v) \\ &= \frac{1}{2}B(u,u) + \epsilon B(u,v) + \frac{\epsilon^2}{2}B(v,v) + \ell(u) + \epsilon\ell(v) \\ &= E(u) + \epsilon(B(u,v) - \ell(v)) + \frac{\epsilon^2}{2}B(v,v) \end{aligned}$$

und im Grenzwert erhalten wir

$$E'(u)v = B(u,v) - \ell(v).$$

Betrachten wir die Optimalitätsbedingung erster Ordnung für ein Minimum $E'(u)v = 0$ für alle $v \in X$, so erkennen wir wieder die Variationsgleichung, durch die wir also einen stationären Punkt des Energiefunktionals berechnen. Nun betrachten wir noch die zweite Ableitung, also

$$E''(u)(v,w) = \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \left(E'(u+\epsilon w)v - E'(u)v \right).$$

Es gilt

$$E'(u+\epsilon w)v = B(u+\epsilon w, v) - \ell(v) = B(u,v) + \ell(v) + \epsilon B(w,v) = E'(u)v + \epsilon B(v,w).$$

Also erhalten wir $E''(u)(v,w) = B(v,w)$, und damit gilt wegen der Koerzivität

$$E''(u)(v,v) = B(v,v) \geq c\|v\|^2 > 0.$$

Damit erwarten wir natürlich, dass E konvex und die Lösung der Variationsgleichung damit ein Minimum des Energiefunktionals ist. Dies beweisen wir im nächsten Satz:

Satz 3.8. *Die Lösung $\hat{u} \in X$ der Variationsgleichung* (3.14) *ist das eindeutige Minimum des Energiefunktionals E in X und umgekehrt.*

Proof. Sei $\hat{u}$ die eindeutige Lösung von (3.14). Dann gilt

$$\begin{aligned} E(v) &= \frac{1}{2}B(\hat{u}+(v-\hat{u}), \hat{u}+(v-\hat{u})) - \ell(\hat{u}+(v-\hat{u})) \\ &= \frac{1}{2}B(\hat{u},\hat{u}) - \ell(\hat{u}) + \underbrace{B(\hat{u}, v-\hat{u}) - \ell(v-\hat{u})}_{=0} + \frac{1}{2}B(v-\hat{u}, v-\hat{u}) \\ &= E(\hat{u}) + \frac{1}{2}B(v-\hat{u}, v-\hat{u}) \\ &\geq E(\hat{u}) + \frac{c}{2}\|v-u\|^2. \end{aligned}$$

Daraus folgt für $v \neq \hat{u}$ die Ungleichung $E(v) > E(\hat{u})$, d.h. $\hat{u}$ ist das eindeutige Minimum.

Ist umgekehrt $\hat{u}$ Minimum von E, so ist $\hat{u}$ auch stationärer Punkt, d.h. $E'(u)v = 0$ für alle $v \in X$ und wie wir oben gesehen haben ist diese Bedingung genau (3.14). □

Wollen wir das ganze auch direkt auf konvexen Teilmengen $C \subset X$ und $g \in X$ anwenden, d.h. die Energie E über C minimieren, dann folgt aus der selben Rechnung wie oben, dass $\hat{u}$ ein Minimum ist, falls

$$B(\hat{u}, v-\hat{u}) - \ell(v-\hat{u}) \geq 0 \qquad \forall\, v \in C$$

gilt. Damit erhält man eine sogenannte *Variationsungleichung*. Dies können wir auch für den Fall inhomogener Dirichlet-Randbedingungen anwenden, d.h. für $C = g_D + H_0^1(\Omega)$. In diesem Fall ist das Minimum $\hat{u} \in g_D + H_0^1(\Omega)$ durch die obige Variationsgleichung für $v \in g_D + H_0^1(\Omega)$ charakterisiert. Es gilt dann $v - \hat{u} \in H_0^1(\Omega)$ und wir können offensichtlich jedes Element in $H_0^1(\Omega)$ durch geeignete Wahl von v erhalten. Damit erhalten wir

$$B(\hat{u}, w) - \ell(w) \geq 0 \qquad \forall\, w \in H_0^1(\Omega).$$

Da wir im Unterraum $H_0^1(\Omega)$ auch $-w$ als Testfunktion wählen können, wird aus der Ungleichung eine Gleichung, d.h. wir erhalten genau die Variationsgleichung mit Testfunktionen in $H_0^1(\Omega)$ wie auch oben gewählt.

Zum Abschluss geben wir noch den quadratischen Teil der Energie im Fall der partiellen Differentialgleichung (3.1) an, nämlich

$$B(u,u) = \int_\Omega (A\nabla u \cdot \nabla u + cu^2)\, dx.$$

Die Minimierung von E impliziert auch, dass eine gewichtete L^2-Norm des Gradienten und der Funktion klein werden, deshalb ist es auch nicht überraschend, dass genau diese Norm in der Stabilitätsabschätzung kontrolliert werden kann. Aus dem Variationsprinzip könnte man die Stabilität aus der Eigenschaft

$$\frac{1}{2}B(u,u) \le E(u) + \ell(u) \le E(0) + \ell(u) = \ell(u)$$

mit anschliessender Verwendung der Koerzivität und Stetigkeit von ℓ herleiten.

3.2 Galerkin-Approximation

Nachdem wir die Grundlage für die Theorie schwacher Lösungen nun detailliert diskutiert haben, wenden wir uns nun dem eigentlich Zweck dieses Kapitels zu, nämlich der Diskretisierung der Variationsgleichung. Dazu betrachten wir die so genannte *Galerkin-Methode*, d.h. wir wählen einen endlichdimensionalen Teilraum $X^h \subset X$ und suchen eine Lösung $u^h \in X^h$ der Variationsgleichung

$$B(u^h, v) = \ell(v) \qquad \forall\ v \in X^h. \tag{3.18}$$

Die Diskretisierung mit der Galerkin-Methode schränkt die gesamte Varationsgleichung (Lösung und Testfunktion) auf einen endlich-dimensionalen Teil ein. Es ist einfach zu zeigen, dass (3.18) äquivalent zur Minimierung der Energie E über dem Teilraum X^h ist.

Wir nehmen an die Dimension von X^h ist gleich N, und $\{\varphi_j\}_{j=1,\dots,N}$ ist eine Basis von X^h. Bezüglich der Basis können wir u^h in der Form

$$u^h = \sum_{j=1}^N U_j^h \varphi_j$$

entwickeln, mit einem Koeffizientenvektor $U^h \in \mathbb{R}^N$. Durch Verwendung der speziellen Testfunktionen $v = \varphi_k$ in (3.18) erhalten wir dann das $N \times N$ Gleichungssystem

$$\sum_{j=1}^N U_j^h B(\varphi_j, \varphi_k) = \ell(\varphi_k) \qquad k = 1, \dots, N,$$

für den Koeffizientenvektor U^h. Mit der Setzung

$$(K_h)_{jk} := B(\varphi_j, \varphi_k), \qquad (F_h)_j := \ell(\varphi_j) \tag{3.19}$$

für die Matrix $K_h \in \mathbb{R}^{N\times N}$ und die rechte Seite $F_h \in \mathbb{R}^N$ können wir die diskrete Variationsgleichung äquivalent als $K_h U^h = F_h$ schreiben. Für die spezielle Form von B und ℓ im Fall der schwachen Formulierung von (3.1)-(3.3) sehen wir, dass N^2 Gebietsintegrale berechnet werden müssen, um die Matrix K_h aufzustellen, sowie zumindest N Integrale um F_h zu berechnen. Dies kann einen enormen Rechenaufwand bedeuten, der das Verfahren ineffizient

machen würde. Deshalb entstand die Idee der finiten Elemente, d.h. von Ansatzfunktionen φ_j mit lokalem Träger. Durch die lokalen Träger verschwindet in den meisten Fällen eine der beiden Funktionen in der Integration von $B(\varphi_j, \varphi_k)$ und die Matrix K_h wird folglich dünnbesetzt. Bei den nicht verschwindenden Integralen muss nur über einen (kleinen) lokalen Träger integriert werden, was den Rechenaufwand weiter senkt. Wir werden die exakte Konstruktion (Knoten-basierter) finiter Elemente im nächsten Abschnitt genauer diskutieren. Zuvor sammeln wir noch einige allgemeine Resultate zur Analysis der Galerkin-Diskretisierung:

Satz 3.9 (Existenz, Eindeutigkeit, Stabilität). *Es gelten die Voraussetzungen von Satz 3.6 und $X^h \subset X$ sei ein endlichdimensionaler Teilraum. Dann besitzt die diskrete Variationsgleichung* (3.18) *eine eindeutige Lösung $\hat{u}^h \in X^h$ und es gilt die Stabilitätsabschätzung*

$$\|\hat{u}^h\| \leq \frac{1}{c}\|\ell\|. \tag{3.20}$$

Proof. Da X^h mit derselben Norm wie X wieder ein Hilbert-Raum ist (endlichdimensionale Teilräume sind immer abgeschlossen), können wir das Lax-Milgram Lemma auch in X^h anwenden um Existenz und Eindeutigkeit zu zeigen. Für die Stabilitätsabschätzung verwenden wir die Testfunktion $v = \hat{u}^h$ und die Koerzivität. □

Wir sehen, dass sich Existenz, Eindeutigkeit, und Stabilität direkt von der schwachen Formulierung der Differentialgleichung auf die Diskretisierung vererbt. Eine weitere interessante Eigenschaft ist die sogenannte *Galerkin-Orthogonalität*, die wir durch Wahl einer Testfunktion $v \in X^h \subset X$ sowohl in (3.14) und (3.18) erhalten. Subtrahieren wir die resultierenden Gleichungen, so folgt

$$B(\hat{u} - \hat{u}^h, v) = 0 \qquad \forall\ v \in X_h. \tag{3.21}$$

Dies bedeutet, dass der Fehler $\hat{u} - \hat{u}^h$ im durch B definierten Skalarprodukt auf den Teilraum X_h orthogonal steht. Dies ist äquivalent zur Aussage, dass $\hat{u}^h$ die Projektion von $\hat{u}$ auf X^h im durch B definierten Skalarprodukt ist. Um eine Abschätzung im ursprünglichen Skalarprodukt zu erhalten, verwenden wir in der Galerkin-Orthogonalität die Testfunktion $v - \hat{u}^h \in X^h$ und addieren links und rechts $B(\hat{u} - \hat{u}^h, \hat{u})$. Damit folgt

$$B(\hat{u} - \hat{u}^h, \hat{u} - \hat{u}^h) = B(\hat{u} - \hat{u}^h, \hat{u} - v).$$

Die linke Seite dieser Identität können wir unter Verwendung der Koerzivität durch $c\|\hat{u} - \hat{u}^h\|$ nach unten abschätzen, die rechte Seite unter Verwendung der Stetigkeit durch $C\|\hat{u} - \hat{u}^h\|\|\hat{u} - v\|$ nach oben. Insgesamt folgt also die Aussage

Lemma 3.10. *Es gelten die Voraussetzungen von Satz 3.9. Dann gilt*

$$\|\hat{u} - \hat{u}^h\| \leq \frac{C}{c}\|\hat{u} - v\| \qquad \forall\ v \in X^h \tag{3.22}$$

und damit auc

$$\|\hat{u} - \hat{u}^h\| \leq \frac{C}{c} \inf_{v \in X^h} \|\hat{u} - v\|. \tag{3.23}$$

Wir sehen also, dass bis auf eine multiplikative Konstante der Fehler bei der Galerkin-Diskretisierung der kleinstmögliche im gewählten endlichdimensionalen Teilraum ist. Man spricht in diesem Fall von einer Approximation *optimaler Ordnung*, der Fehler des Verfahrens lässt sich durch ein Vielfaches des Projektionsfehlers abzuschätzen. Um den Fehler nun weiter

und auch konkreter abzuschätzen, verwendet man geeignete Elemente v^h in Abhängigkeit von der jeweiligen Struktur der Basisfunktionen φ_j und der exakten Lösung $\hat{u}$. Wir werden später sehen, dass im Fall finiter Elemente zumeist eine geeignete Interpolierende gewählt wird, da der Interpolationsfehler am leichtesten abgeschätzt werden kann. Vom Standpunkt in Kapitel 2 gesehen liefert die Koerzivität genau die Stabilität des Verfahrens, während der noch abzuschätzende Teil $\inf_{v\in X^h} \|\hat{u}-v\|$ als Konsistenz bzw. Konsistenzordnung interpretiert werden kann.

3.3 Finite Elemente

Abstrakt definiert man ein finites Element als Tripel $(K, \mathcal{P}, \mathcal{X})$ mit den Eigenschaften:

(i) $K \subset \mathbb{R}^d$ ist ein kompakte Menge mit nichtleerem Inneren und stückweise stetigem Rand (das Elementgebiet bzw. Referenzelement).

(ii) $\mathcal{P}$ ist ein endlichdimensionaler Raum von Funktionen auf K (der Raum der Formfunktionen)

(iii) $\mathcal{X} = \{X_1, \ldots, X_N\}$ sei eine Basis von $\mathcal{P}'$ (die Knoten).

Die Knotenbasis $\{\varphi_1, \ldots, \varphi_N\}$ erhält man dann als Basis von $\mathcal{P}$ dual zu $\mathcal{P}'$, d.h. $X_j(\varphi_i) = \delta_{ij}$. Im oben diskutierten eindimensionalen Fall finiter Elemente kann man $K = (0, 1)$ als Referenzelement wählen, $\mathcal{P}$ als den Raum der affin-linearen Funktionen auf K und $\mathcal{X}$ entspricht den Randknoten $\{0, 1\}$ oder eigentlich genauer den Distributionen $X_1 = \delta$ und $X_2 = \delta(. - 1)$. Dann erhält man als Knotenbasis $\varphi_1(x) = 1-x$ und $\varphi_2(x) = x$. Durch (lineare) Transformation des Referenzelements $(0, 1)$ auf beliebige Teilintervalle lassen sich dann die Basisfunktionen auf einem Gitter konstruieren, so wie im ersten Kapitel verwendet.

Für das Verständnis finiter Elemente ist die obige Definition jedoch eher unhandlich. Wir werden uns deshalb im folgenden stark einschränken und eine einfachere und weniger allgemeine Sichtweise verwendet. Unter finiten Elementen versteht man normalerweise stückweise polynomiale Ansatzfunktionen auf einem Dreiecksgitter (Tetraedergitter in $3D$) mit lokalem Support. Wir werden uns im folgenden der Einfachheit halber auch auf den zweidimensionalen Fall einschränken, analoges Vorgehen ist aber auch in höheren Dimensionen möglich.

Der erste Schritt zur Konstruktion einer finite Elemente Methode ist dann eine Triangularisierung des Gebiets Ω (bzw. einer polygonalen Approximation $\hat{\Omega}$). Der Vorteil einer Triangularisierung gegenüber rechteckigen Gittern ist die grössere Flexibilität bei der Approximation krummliniger Ränder. Wir nehmen also an, es gilt

$$\overline{\Omega} = \bigcup_{j=1}^{M} T_j \tag{3.24}$$

wobei alle T_j Dreiecke sind, und $T_j \cap T_i$ für $i \neq j$ Mengen vom MaßNull sind. Die Schnittmenge $T_i \cap T_j$ soll entweder leer sein, genau einen Punkt (Knoten), oder genau eine Kante enthalten. Damit nehmen wir implizit an, dass es N Knoten P_i gibt, die durch Kanten zu Dreiecken verbunden sind. Jedem Knoten P_i lässt sich dann seine *Nachbarschaft* $N(P_i)$ zuordnen als

$$N(P_i) = \bigcup_{P_i \in \partial T_j} T_j.$$

Es ist in dieser Sichtweise (knotenbasierte Elemente) natürlich, die Freiheitsgrade in die Knoten P_i zu legen. Dies passiert in dem man für jeden Knoten eine Ansatzfunktion (finites Element) mit den folgenden Eigenschaften konstruiert:

$$\begin{array}{ll} \varphi \in C(\overline{\Omega}) & \\ \varphi_i|_{P_k} = \delta_{ik} & \\ \varphi_i(x) = 0 & x \notin N(P_i) \\ \varphi_i|T_j \in \mathcal{P}_K(T_j) & T_j \subset N(P_i) \end{array} \tag{3.25}$$

Hier bezeichnet $\mathcal{P}_\|(T_j)$ die Menge der Polynome vom Grad kleiner gleich k auf T_j. Die einfachste Wahl für knotenbasierte Elemente ist $k = 1$, was stückweise lineare Elemente liefert. Diese sind durch die Bedingungen in (3.25) eindeutig festgelegt, da es für eine lineare Funktion in zwei Dimensionen genügt die Funktionswerte in der Eckpunkten der Dreiecke zu spezifizieren.

In manchen Anwendungen werden auch Elemente vom Grad $k = 0$ verwendet, allerdings liefern diese keine stetigen Ansatzfunktionen mehr und sind deshalb als knotenbasierte Elemente ungeeignet. Man interpretiert solche unstetigen Ansatzfunktionen als volumsbasierte Elemente, der Freiheitsgrad wird dem Elementmittelpunkt zugeordnet. Eine dritte Klasse von finiten Elementen sind kantenbasierte Elemente, in denen üblicherweise der Freiheitsgrad den Kantenmittelpunkten des Dreiecks zugeordnet wird.

Wir werden jedes Dreieck T_j als Transformation des Einheitdreiecks

$$\hat{T} := \{\ (x_1, x_2) \in [0,1]^2 \mid x_1 + x_2 \leq 1\ \}$$

betrachten, d.h. $T_j = S_j^h(\hat{T})$, wobei die Transformation S_j^h natürlich von der Gittergrösse

$$h := \max_j \ \text{diam}\ T_j \tag{3.26}$$

abhängt. Diese Transformation und ihrer Skalierung kann später vor allem zur Abschätzung des Interpolationsfehlers effektiv verwendet werden.

Der diskrete Teilraum $X^h \subset H^1(\Omega)$, den wir verwenden werden ist gegeben durch

$$X^h = \{\ \sum U_i^h \varphi_i \mid U^h \in \mathbb{R}^N\ \}. \tag{3.27}$$

Wir beginnen mit der Verifikation, dass X^h tatsächlich ein Teilraum von $H^1(\Omega)$ ist.

Lemma 3.11. *Sei $\{\varphi_i\}_{i=1,\ldots,N} \subset C(\Omega)$ eine Familie von Ansatzfunktionen, die* (3.25) *erfüllen. Dann ist X^h definiert durch* (3.27) *ein Teilraum von $H^1(\Omega)$ und $\{\varphi_i\}_{i=1,\ldots,N}$ ist eine Basis von X^h.*

Proof. Da Polynome beschränkt sind, gilt offensichtlich $\varphi_i \in L^\infty(\Omega) \subset L^2(\Omega)$. Um die distributionellen Ableitungen zu berechnen betrachten wir für $\psi \in C_0^\infty(\Omega)$ die Funktionale

$$w(\psi) = -\int_\Omega \varphi_i \frac{\partial \psi}{\partial x_k}\ dx.$$

Setzen wir die spezielle Form von ψ_j ein so folgt

$$w(\psi) = -\int_{N(P_i)} \varphi_i \frac{\partial \psi}{\partial x_k}\ dx = -\sum_{T_j \in N(P_i)} \int_{T_j} \varphi_i \nabla \cdot (\psi e_k)\ dx$$

mit dem Einheitsvektor e_k. Mit dem Gauss'schen Integralsatz erhalten wir daraus

$$w(\psi) = \sum_{T_j \in N(P_i)} \left(\int_{T_j} \frac{\partial \varphi_i|_{T_i}}{\partial x_k} \psi \, dx - \int_{\partial T_j} \varphi_i \psi e_k \cdot n \, d\sigma \right).$$

Für jede Kante $E \subset \partial T_j$ gilt entweder $E \subset \partial N(P_i)$ und damit $\varphi_i|_E = 0$ oder die Kante trennt zwei Dreiecke T_j und T_m. Da die Normalvektoren an ∂T_i und ∂T_j entgegengesetzt orientiert und φ_i, ψ stetig sind, erhalten wir zweimal das selbe Integral über E mit verschiedenen Vorzeichen. Aus diesen Argumenten sehen wir, dass

$$\sum_{T_j \in N(P_i)} \int_{\partial T_j} \varphi_i \psi e_k \cdot n \, d\sigma = 0$$

gilt. Damit folgt

$$w(\psi) = \sum_{T_j \in N(P_i)} \int_{T_j} \frac{\partial \varphi_i|_{T_i}}{\partial x_k} \psi \, dx = \int_\Omega \frac{\partial \varphi_i}{\partial x_k} \psi \, dx,$$

und damit stimmt die distributionelle Ableitung mit den stückweise Ableitungen in den Elementen überein. Da die stückweisen Ableitungen wieder Polynome sind, folgt auch deren Beschränktheit und damit $\varphi_i \in W^{1,\infty}(\Omega) \subset H^1(\Omega)$. Die lineare Hülle der φ_i ist folglich ein Teilraum von $H^1(\Omega)$.

Um die lineare Unabhängigkeit zu zeigen nehmen wir an es gilt

$$\sum \alpha_i \varphi_i = 0$$

für $\alpha_i \in \mathbb{R}$. Da alle Ansatzfunktionen stetig sind, folgt damit auch

$$\alpha_k = \sum \alpha_i \varphi_i(P_k) = 0, \qquad k = 1, \ldots, N.$$

Also sind die φ_j linear unabhängig. □

3.3.1 Assemblierung von Matrizen und Vektoren

Das erste Problem bei der praktischen Durchführung von finite Elemente Methoden ist die Triangulierung des Gebiets Ω (oder einer sinnvollen Approximation im Fall nichtpolygonaler Gebiete). Die Erzeugung eines Gitters ist ein nichttriviales informatisches Problem, vor allem da das Gitter gewisse Eigenschaften erfüllen sollte, wie wir bei der Analysis im nächsten Abschnitt sehen werden. In jedem Fall liefert ein typischer Gittergenerator Information über die Knoten, Kanten, und die Dreiecksflächen (oft als Elemente bezeichnet), in 3D über Knoten, Kanten, Flächen und Volumen. Diese werden normalerweise global indiziert, und durch entsprechende Zuordnungsvorschriften werden diese Indizes miteinander in Beziehung gesetzt, z.B. um zu klären bei welchen Dreieck ein Knoten als Eckpunkt auftritt.

Wir sehen, dass wir zum Aufstellen des diskreten Problems $K_h U^h = F_h$ Produkte von Testfunktionen und gegebenen Funktionen integrieren müssen. In den wenigsten Fällen sind diese Berechnungen analytisch möglich, weshalb man eine geeignete numerische Integrationsformel auf den Dreiecken verwenden sollte. Diese sollte zumindest eine so hohe Ordnung aufweisen wie die Approximationsordnung des benutzten Ansatzraumes um nicht den eigentlichen Fehler des finite Elemente Verfahrens durch den Integrationsfehler zu dominieren (den Approximationsfehler werden wir im nächsten Abschnitt abschätzen).

Neben der Frage der geeigneten numerischen Integration, stellt sich auch die Frage, wie die Matrix K_h und der Vektor F_h aufgebaut (assembliert werden sollen). Nach der obigen Definition scheint ein knotenbasiertes Vorgehen auf den ersten Blick naheliegend. Dabei würde man alle Knoten des Netzes einmal durchlaufen, und die entsprechende Basisfunktion φ_j (also jene die jeweils im aktuellen Knoten den Wert 1 annimmt) in Produkten mit allen anderen umgebenden bzw. mit der rechten Seite zu integrieren. Dies führt zu einem zeilenweisen Aufbau der Matrix K_h und des Vektors F_h. Dieser Zugang stellt sich jedoch nicht als der effizienteste heraus. Ein erstes Problem besteht darin, dass man sich zu jedem Knoten alle benachbarten suchen muss, um herauszufinden, welche der Terme $B(\varphi_i, \varphi_j)$ wirklich integriert werden müssen. Zweitens muss sich dann zu den Knoten auch noch die entsprechenden Elemente (und die darin liegenden Stützstellen für die numerische Integration suchen) um die Assemblierung wirklich durchzuführen.

Für eine effizientere Assemblierung verwendet man deshalb ein elementweises Vorgehen. Grundlage dafür ist die Aufspaltung

$$(K_h)_{ij} = B(\varphi_i, \varphi_j) = \sum_{T_k} \int_{T_k} (A\nabla\varphi_i \cdot \nabla\varphi_j + c\varphi_i\varphi_j).$$

Man kann also auch eine Schleife über die Dreiecke (in 3D Tetraeder) durchführen und die einzelnen Integrale über T berechnen. Diese können dann jeweils zum aktuellen Matrixeintrag ($(K_h)_{ij}$, initialisiert mit 0) addiert werden. In dieser Reihenfolge ist auch klar, welche Integrale jeweils für dieses Element berechnet werden, nämlich genau jene mit Basisfunktion in einem der Randknoten (9 pro Element für Dreiecke in $2D$). Ein völlig analoges Vorgehen ist natürlich auch für die rechte Seite F_h möglich.

Um diese Assemblierung durchzuführen ist es wichtig eine lokale Abbildung zwischen den Elementen und Knoten zu speichern, die jedem Element die Indizes seiner Randknoten und Kanten zuordnet.

3.3.2 Fehlerabschätzungen

Die Grundlage für jede finite Elemente Fehlerabschätzung ist die allgemeine Aussage für Galerkin-Approximationen in Lemma 3.10. Nach dieser Aussage genügt es den Projektionsfehler für die exakte Lösung abzuschätzen. Wir werden diese Abschätzung in diesem Abschnitt exemplarisch in der Dimension $d = 2$ für stückweise lineare Basisfunktionen durchführen, aber einige allgemeine Grundideen dabei betonen.

Wir bemerken auch, dass in der Praxis nicht wirklich die exakte Bilinearform und das wirkliche lineare Funktional ℓ verwendet werden, falls z.B. eine numerische Integration durchgeführt oder der Rand approximiert wird. In Wirklichkeit erhält man dann eine Bilinearform $B_h : X_h \times X_h$, sodass

$$B_h(u^h, v) = \langle \ell^h(v) \qquad \forall\ v \in X_h$$

gilt. Man kann aber dieses Variationsproblem als

$$B(u^h, v) = \ell(v) + r_h(v) \qquad \forall\ v \in X_h$$

schreiben, mit dem Restfunktional

$$r^h(v) = B(u^h, v) - B_h(u^h, v) + \ell_h(v) - \ell(v).$$

Nun kann man die Galerkin-Orthogonalität zu

$$B(\hat{u} - \hat{u}^h, \hat{u} - \hat{u}^h) = B(\hat{u} - \hat{u}^h, \hat{u} - v) + r_h(v)$$

modifizieren. Es genügt dann r^h quantitativ abzuschätzen (z.B. mit den bekannten Methoden für numerische Integration) um wieder eine Fehlerabschätzung für $\hat{u} - \hat{u}^h$ zu erhalten. Als obere Schranke erhält man dann die Summe aus Projektionsfehler und einer von r_h abhängigen Schranke.

Wir wenden uns nun also der Abschätzung des Interpolationsfehlers zu. Wir sehen sofort, dass der Interpolationsoperator bei stückweise linearen Basisfunktionen gegeben ist durch

$$I_h u = \sum_j u(x_j)\varphi_j(x) \qquad \forall\ u \in C(\Omega).$$

Der Fehler bei der Interpolation ist also $v = u - I_h u$. Wir nehmen in der Folge an, dass für die schwache Lösung der Variationsgleichung die Regularität $\hat{u} \in H^2(\Omega)$ gilt, und nach den oben diskutierten Einbettungssätzen folgt (in Dimension zwei) $\hat{u} \in C(\Omega)$. Der Interpolationsfehler ist dann natürlich auch sinnvoll definiert.

Im folgenden werden wir häufig die Transformation $\overset{h}{j}$ zwischen Dreiecken T_j und dem Referenzdreieck $\hat{T}$ verwenden. Nach der Transformationsregel für Integrale gilt

$$\int_{T_j} \varphi(x)\ dx = \frac{1}{\delta_j^h} \int_{\hat{T}} \varphi(S_j^h(y))\ dy,$$

wobei $\delta_j^h = |\det(\nabla S_j^h)|$. Wir berechnen leicht $\nabla S_j^h = \mathcal{O}(h)$ und damit $\delta_j^h = \mathcal{O}(h^2)$. Wir nehmen nun an, dass

$$c_1 h \leq \lambda_{\min}(\nabla S_j^h) \leq \lambda_{\max}(\nabla S_j^h) \leq c_2 h. \tag{3.28}$$

Damit folgt $c_1^2 h^2 \leq \delta_j^h \leq c_2^2 h^2$. Für Normen auf $\hat{T}$ erhalten wir folgende Transformation:

$$\begin{aligned} \int_{T_j} \varphi(x)^2\ dx &= \frac{1}{\delta_j^h} \int_{\hat{T}} \varphi(S_j^h(y))^2\ dy \\ \int_{T_j} |\nabla\varphi(x)|^2\ dx &= \frac{1}{\delta_j^h} \int_{\hat{T}} |(\nabla S_j^h)\nabla\varphi(S_j^h(y))|^2\ dy. \end{aligned}$$

Wir betrachten den Interpolationsfehler in T_j zurücktransformiert auf $\hat{T}$, d.h. $\varphi = (u - I_h u) \circ S_j^h$, für $u \in H^2(\Omega)$. Wegen $u|_{T_j} \in H^2(T_j)$ und da $(I_h u)|_{T_j}$ linear ist, folgt $\varphi \in H^2(\hat{)}$. Also können wir die H^2-Norm von φ betrachten, bzw. eine weitere Norm

$$|||\varphi||| = \sqrt{|\varphi|_{2,2}^2 + \varphi(P_1)^2 + \varphi(P_2)^2 + \varphi(P_3)^2},$$

wobei $P_1 = (0,0)$, $P_2 = (1,0)$ und $P_3 = (0,1)$ gilt. Sei $f_i(\varphi) = |\varphi(P_i)|$, dann ist f_i eine Halbnorm auf $H^2(\hat{T})$ und das System der Halbnormen (f_1, f_2, f_3) erfüllt die Bedingungen des Sobolevschen Normierungssatz, d.h. für eine lineare Funktion (Polynom vom Grad kleiner gleich eins) auf $\hat{T}$ gibt es immer ein P_i mit $\varphi(P_i) \neq 0$. Folglich ist die Norm $|||.|||$ äquivalent zur H^2-Norm auf $\hat{T}$, d.h. es existiert eine Konstante γ, sodass

$$\|\varphi\|_{1,2} \leq \|\varphi\|_{2,2} \leq \gamma |||\varphi|||$$

gilt.

Nun betrachten wir die Abschätzung näher für $\varphi = (u - I_h u) \circ S_j^h$. Da $(I_h u) \circ S_j^h$ immer noch linear auf $\hat{T}$ ist, verschwinden alle zweiten Ableitung. Also gilt $|\varphi|_{2,2} = |u \circ S_j^h|_{2,2}$. Weiter ist wegen der Interpolationseigenschaft $\varphi(P_k) = (u - I_h u)(S_j^h(P_i)) = 0$ (beachte $S_j^h(P_i)$ ist Eckpunkt des Dreiecks T_j und $I_h u$ interpoliert u in den Eckpunkten). Also folgt $|||\varphi||| = |u \circ S_j^h|_{2,2}$ und damit

$$\|\varphi\|_{1,2} \leq \gamma |u \circ S_j^h|_{2,2}.$$

Nun können wir die Rücktransformation der Normen auf das Dreieck T_j benutzen, es gilt ja

$$\|\varphi\|_{1,2}^2 = \int_{\hat{T}} (\varphi(x)^2 + |\nabla \varphi(x)|^2)\ dx = \frac{1}{\delta_j^h} \int_{T_j} ((u - I_h u)^2 + |(\nabla S_j^h)(\nabla u - \nabla(I_h u))|^2)\ dx.$$

Nun können wir die obigen Abschätzungen für die Eigenwerte von S_j^h benutzen, um weiter zu zeigen

$$\|\varphi\|_{1,2}^2 \geq \frac{1}{\delta_j^h} \int_{T_j} ((u - I_h u)^2 + c_1^2 h^2 |\nabla u - \nabla(I_h u)|^2)\ dx.$$

Analog können wir die Transformation für die zweiten Ableitungen ausrechnen. Es gilt

$$|u \circ S_j^h|_{2,2}^2 = \int_{\hat{T}} \sum_k |\nabla \partial_{x_k} (u \circ S_j^h)|^2\ dx = \frac{1}{\delta_j^h} \int_{T_j} \sum_k |(\nabla S_j^h) \nabla ((\nabla S_j^h) \nabla u)_k|^2\ dx,$$

und mit den obigen Abschätzungen folgt

$$\begin{aligned} |u \circ S_j^h|_{2,2}^2 &\leq c_2 h^2 \frac{1}{\delta_j^h} \int_{T_j} \sum_k |\nabla ((\nabla S_j^h) \nabla u)_k|^2\ dx \leq d c_2^2 h^4 \frac{1}{\delta_j^h} \int_{T_j} \sum_k |\nabla \partial_{x_k} u|^2\ dx \\ &= d c_2^2 h^4 \frac{1}{\delta_j^h} |u|_{H^2(T_j)}^2. \end{aligned}$$

Durch Kombination der obigen Abschätzungen erhalten wir also

$$\int_{T_j} ((u - I_h u)^2 + c_1^2 h^2 |\nabla u - \nabla(I_h u)|^2)\ dx \leq \gamma d c_2^2 h^4 |u|_{H^2(T_j)}^2.$$

Nun summieren wir noch über die Dreiecke T_j und folgern (für h hinreichend klein, sodass $c_1 h \geq 1$)

$$\begin{aligned} \|u - I_h u\|_{1,2}^2 &= \sum_j \int_{T_j} ((u - I_h u)^2 + c_1^2 h^2 |\nabla u - \nabla(I_h u)|^2)\ dx \\ &\leq \frac{1}{c_1^2 h^2} \sum_j \int_{T_j} ((u - I_h u)^2 + c_1^2 h^2 |\nabla u - \nabla(I_h u)|^2)\ dx \\ &\leq \gamma d \frac{c_2^2}{c_1^2} h^2 \sum_j |u|_{H^2(T_j)}^2 \\ &= \gamma d \frac{c_2^2}{c_1^2} h^2 |u|_{2,2}^2. \end{aligned}$$

Ziehen wir nun abschliessend noch die Wurzel und nennen $\tilde{C} = \sqrt{\gamma d} \frac{c_2}{c_1}$, dann haben wir folgendes Resultat bewiesen.

Proposition 3.12. *Seien B und ℓ wie oben und $u \in H^1(\Omega)$ die Lösung des Variationsproblems*

$$B(u, v) = \ell(v) \qquad \forall\ v \in H^1(\Omega),$$

mit der zusätzlichen Regularität $u \in H^2(\Omega)$. Weiter sei (3.28) mit Konstanten c_1, c_2 unabhängig von h erfüllt. Dann existiert eine von h unabhängige Konstante $C > 0$, sodass

$$\|u - I_h u\|_{1,2} \leq \tilde{C}|u|_{2,2}h \tag{3.29}$$

gilt.

Aus der Abschätzung des Interpolationsfehlers und der obigen Galerkin-Fehlerabschätzung können wir nun eine quantitative Fehlerabschätzung für die finite Elemente Diskretisierung angeben.

Satz 3.13. *Seien die Bedingungen von Proposition erfüllt und sei $u^h \in X^h$ die Lösung des diskretisierten Problems*

$$B(u^h, v) = \ell(v) \qquad \forall\ v \in X^h,$$

wobei X^h der Ansatzraum der stückweise linearen finiten Elemente ist. Dann gilt die Fehlerabschätzung

$$\|u - u^h\|_{1,2} \leq \frac{C}{c}\|u - I_h u\|_{1,2} \leq \frac{C\tilde{C}}{c}|u|_{2,2}h = \mathcal{O}(h). \tag{3.30}$$

Die obige Technik zur Herleitung von Fehlerabschätzungen kann sowohl bezüglich des Polynomgrads der Elemente, der Ordnung der Appproximation, als auch der Ordnung der Ableitungen verallgemeinert werden. Wichtig ist immer auf dem Referenzelement eine entsprechende Abschätzung einer Sobolev-Norm durch eine Halbnorm in einem Sobolev-Raum höherer Ableitungsordnung zu erhalten. Die unterschiedlichen Skalierungseigenschaften der Ableitungen bei der Rücktransformation auf die Dreiecke im Gitter liefern dann Abschätzungen bezüglich h.

Man sieht auch, dass die Eigenschaft (3.28) mit Konstanten c_1, c_2 unabhängig von h essentiell für die Güte der Fehlerabschätzung ist. Diese Bedingung lässt sich als eine Bedingung an die Dreiecke im Gitter auffassen, diese sollte (insbesonders mit kleiner werdendem h) nicht zu weit entarten. Wie wir in der Übung sehen werden, kann (3.28) auch als Bedingung an In- und Umkreisradius, sowie auch an den kleinsten und grössten Winkel im Dreieck interpretiert werden.

Die Abschätzung (3.30) ist ein Beispiel einer *a-priori* Abschätzung, d.h. es werden Eigenschaften über die Lösung u a-priori vorausgesetzt, und damit sind auch die Konstanten in der Abschätzung nicht explizit berechenbar (z.B. kennt man $|u|_{2,2}$ nicht, da man ja die Lösung u nicht kennt). Alternativ lassen sich sogenannte *a-posteriori* Abschätzungen herleiten, bei denen die Schranken nur von der diskreten Lösung u^h (die man wirklich berechnet) abhängen. Diese Abschätzungen werden dann vor allem zur adaptiven Verfeinerung des Gitters verwendet. Wir werden diese Aspekte in dieser Vorlesung nicht weiter diskutieren und verweisen auf weiterführende Lehrveranstaltungen zur Numerik partieller Differentialgleichungen.

3.3.3 Eigenwerte und Kondition von K_h

Zum Abschluss dieses Kapitels betrachten wir noch einmal die Eigenwerte bzw. Konditionszahl der Systemmatrix K_h. Da K_h symmetrisch und positiv definit ist, gilt nach dem

Rayleigh-Prinzip den grössten und kleinsten Eigenwert aus

$$\lambda_{\min} = \min_{V\in\mathbb{R}^N\setminus\{0\}} \frac{V^T K_h V}{V^T V} \tag{3.31}$$

$$\lambda_{\max} = \max_{V\in\mathbb{R}^N\setminus\{0\}} \frac{V^T K_h V}{V^T V} \tag{3.32}$$

berechnen, sowie die Konditionszahl als $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$.

Wir definieren nun die Funktion $v = \sum V_j \varphi_j$, dann gilt

$$V^T K_h V = \sum_{j,k} V_j V_k B(\varphi_j, \varphi_k) = B(v, v).$$

Wie wir in der Übung sehen werden, folgt aus (3.28) die Existenz von Konstanten β_1 und β_2, sodass

$$\beta_1 \int_\Omega v^2 \, dx = \beta_1 \sum_j \int_{T_j}^2 v^2 \, dx \le h^2 \sum_i V_i^2 \le \beta_2 \sum_j \int_{T_j}^2 v^2 \, dx = \beta_2 \int_\Omega v^2 \, dx.$$

Verwenden wir nun noch die Koerzivität und Stetigkeit von B, so folgt

$$\frac{c}{\beta_2} h^2 \frac{\|v\|_{1,2}^2}{\|v\|_2^2} \le \frac{V^T K_h V}{V^T V} \le \frac{C}{\beta_1} h^2 \frac{\|v\|_{1,2}^2}{\|v\|_2^2}. \tag{3.33}$$

Wegen der Normabschätzung $\|v\|_{1,2} \ge \|v\|_2$ folgt

$$\lambda_{\min} = \min_{V\in\mathbb{R}^N\setminus\{0\}} \frac{V^T K_h V}{V^T V} \ge \min_{v\in X^h\setminus\{0\}} \frac{c}{\beta_2} h^2 \frac{\|v\|_{1,2}^2}{\|v\|_2^2} \ge \frac{c}{\beta_2} h^2.$$

Die Abschätzung des grössten Eigenwerts ist schwieriger, da wir im Allgemeinen die H^1-Norm nicht durch die L^2-Norm nach oben abschätzen können. Deswegen verwenden wir wieder die Transformation auf das Referenzelement und bezeichnen mit $\tilde{v}$ die transformierte Funktion, die wiederum linear auf dem Einheitsdreieck ist. Da der Raum der linearen Funktionen auf dem Einheitsdreieck endlichdimensional ist, und alle Normen auf einem endlichdimensionalen Raum äquivalent sind, existiert eine Konstante $\alpha > 0$, sodass

$$\|\varphi\|_{H^1(T)} \le \alpha \|\varphi\|_{L^2(T)}$$

für alle linearen Funktionen φ auf T gilt. Wie oben erhalten wir

$$\|\tilde{v}\|_{L^2(T)} = \frac{1}{\sqrt{\delta_j^h}} \|v\|_{L^2(T_j)}$$

und

$$\|\tilde{v}\|_{H^1(T)} \ge \frac{c_1 h}{\sqrt{\delta_j^h}} \|v\|_{H^1(T_j)},$$

und folglich

$$\|v\|_{H^1(T_j)} \le \frac{\alpha}{c_1 h} \|v\|_{L^2(T_j)}.$$

Nach Summation über die Dreiecke folgt die sogenannte *inverse Ungleichung*

$$\|v\|_{1,2} \leq \frac{\alpha}{c_1 h}\|v\|_2. \tag{3.34}$$

Nun können wir auch den grössten Eigenwert abschätzen, und zwar als

$$\lambda_{\max} = \max_{V\in\mathbb{R}^N\setminus\{0\}} \frac{V^T K_h V}{V^T V} \leq \max_{v\in X^h\setminus\{0\}} \frac{C}{\beta_1} h^2 \frac{\|v\|_{1,2}^2}{\|v\|_2^2} \leq \frac{C\alpha^2}{\beta_1 c_1^2}.$$

Abschliessend können wir nun auch die Konditionszahl abschätzen durch

$$\kappa \leq \frac{C\beta_2\alpha^2}{c\beta_1 c_1^2}\frac{1}{h^2}.$$

Neben den vom FE-Raum abhängigen Konstanten α, β_1 und c_1 sehen wir wieder den Quotienten $\frac{C}{c}$, der ein Maß für die Stabilität des ursprünglichen Problems ist (siehe Stabilitätsabschätzung nach Lax-Milgram). Zusätzlich tritt noch der Faktor $\frac{1}{h^2}$ auf, der bei kleinem h für eine sehr schlechte Kondition sorgt, und damit auch besondere Sorgfalt bei der Wahl iterativer Lösungsverfahren erfordert wie wir in Kapitel 5 noch sehen werden.