

3 Finite-Elemente-Verfahren für elliptische Probleme

In diesem Kapitel werden wir die modernen Finite-Elemente-(Galerkin)-Methoden zur Lösung elliptischer Randwertaufgaben (RWA) diskutieren. Der Übersichtlichkeit halber werden wir uns dabei auf das Modellproblem der Poisson-Gleichung mit Dirichletschen Randbedingungen, d. h. auf die 1. RWA, beschränken:

$$Lu := -\Delta u = f \quad \text{in } \Omega, \quad u = g \quad \text{auf } \partial\Omega. \quad (3.0.1)$$

Das Definitionsgebiet $\Omega \in \mathbb{R}^2$ wird zunächst wieder als glatt berandet oder als konvexes Polygonebiet vorausgesetzt. Die Problemdata f, g sind ebenfalls glatt, so dass die im vorigen Kapitel beschriebenen Resultate anwendbar sind. Erweiterungen für Probleme mit variablen Koeffizienten oder anderen Randbedingungen sowie auf drei Raumdimensionen werden wieder in Bemerkungen berücksichtigt.

3.1 Allgemeine Projektionsverfahren

Ausgangspunkt ist die variationelle Formulierung der RWA. Wir erinnern an den oben diskutierten Ansatz zu einer allgemeinen Lösungstheorie. Eine „schwache“ bzw. „verallgemeinerte“ Lösung der 1. RWA des Laplace-Operators (zu den Randdaten $g \equiv 0$) ist definiert als das (eindeutige) Minimum auf dem Sobolew-Raum $H_0^1(\Omega)$ des Energiefunktional

$$E(v) := \frac{1}{2} \|\nabla v\|^2 - (f, v) \rightarrow \min.$$

Wir verwenden hier und im folgenden wieder die Bezeichnungen

$$(v, w) := \int_{\Omega} v(x)w(x) dx, \quad \|v\| := \left(\int_{\Omega} |v(x)|^2 dx \right)^{1/2}, \quad \|\nabla v\| := \left(\int_{\Omega} |\nabla v(x)|^2 dx \right)^{1/2}.$$

Über den Variationsansatz

$$\frac{d}{d\varepsilon} E(u + \varepsilon\varphi)|_{\varepsilon=0} = 0 \quad \forall \varphi \in H_0^1(\Omega),$$

erhalten wir die äquivalente Variationsgleichung (Stationaritätsbedingung)

$$u \in V : \quad (\nabla u, \nabla \varphi) = (f, \varphi) \quad \forall \varphi \in H_0^1(\Omega). \quad (3.1.2)$$

Wir erinnern daran, dass das natürliche Skalarprodukt des Raumes $H_0^1(\Omega)$ gerade durch das sogen. „Dirichlet-Produkt“ $(\nabla v, \nabla w)$ gegeben ist. Zum Nachweis der Definitheit dieses Ausdrucks haben wir die Poincarésche Ungleichung verwendet:

$$\|v\| \leq d_{\Omega} \|\nabla v\|, \quad v \in H_0^1(\Omega). \quad (3.1.3)$$

Bemerkung 3.1: Im Falle *inhomogener Randbedingungen* $u|_{\partial\Omega} = g$ geht man wie folgt vor. Wir nehmen an, dass die Randwerte als Spur einer Funktion $\bar{g} \in H^1(\Omega)$ gegeben

sind: $g = \bar{g}|_{\partial\Omega}$. Für die Funktion $v := u - \bar{g} \in H_0^1(\Omega)$ gilt dann im Falle $\Delta\bar{g} \in L^2(\Omega)$:

$$(\nabla v, \nabla\varphi) = (f, \varphi) - (\Delta\bar{g}, \varphi) =: (\tilde{f}, \varphi) \quad \forall \varphi \in H_0^1(\Omega),$$

d. h.: Die Funktion v genügt einer Variationsgleichung der Art (3.1.2). Im folgenden können wir also o.B.d.A. stets homogene Dirichlet-Randbedingungen annehmen.

Bemerkung 3.2: Im Fall von *Neumannschen Randbedingungen* $\partial_n u|_{\partial\Omega} = g$ wird der Sobolew-Raum $H^1(\Omega)$ (ohne Vorgabe von Randwerten) verwendet und die zugehörige variationelle Formulierung lautet

$$u \in H^1(\Omega) : \quad (\nabla u, \nabla\varphi) = (f, \varphi) + (g, \varphi)_{\partial\Omega} \quad \forall \varphi \in H^1(\Omega). \quad (3.1.4)$$

Um die eindeutige Lösbarkeit zu sichern, muss in diesem Fall noch eine Zusatzbedingung gestellt werden, um *konstante* Lösungen auszuschließen, z. B.: die Normierungsbedingung $(u, 1)_\Omega = 0$. Ferner muss die Verträglichkeitsbedingung $(f, 1) + (g, 1)_{\partial\Omega} = 0$ erfüllt sein. Jede hinreichend glatte Lösung von (3.1.4) erfüllt dann neben der Differentialgleichung $-\Delta u = f$ auch notwendig die in der variationellen Formulierung implizit enthaltene *natürliche* Randbedingung $\partial_n u|_{\partial\Omega} = g$ (Beweis ähnlich wie im Fall der 1. RWA durch partielle Integration und Variation der Testfunktion). Eine ähnliche Konstruktion liefert auch die variationelle Formulierung im Fall der 3. RWA, d. h. für Robinsche Randbedingungen.

Bemerkung 3.3: Nicht alle elliptischen RWAn lassen sich nicht über den Energieminimierungsansatz behandeln. Ein typisches Beispiel ist die Diffusions-Transport-Gleichung

$$-\Delta u + \partial_1 u = f \quad \text{in } \Omega, \quad u = g \quad \text{auf } \partial\Omega. \quad (3.1.5)$$

Ihre variationelle Formulierung lautet

$$(\nabla u, \nabla\varphi) + (\partial_1 u, \varphi) = (f, \varphi) \quad \forall \varphi \in H_0^1(\Omega). \quad (3.1.6)$$

Diese besitzt ebenfalls eine eindeutige Lösung $u \in H_0^1(\Omega)$, was sich weiter unten als Folgerung eines allgemeineren Resultats ergeben wird.

Die folgende Diskussion wird in einem etwas abstrakteren Rahmen durchgeführt, welche an den obigen Beispielen orientiert ist und diese als Sonderfälle beinhaltet. Sei V ein allgemeiner Hilbert-Raum mit Skalarprodukt $(\cdot, \cdot)_V$ und zugehöriger Norm $\|\cdot\|_V := (\cdot, \cdot)_V^{1/2}$ und $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$ eine *beschränkte* Bilinearform sowie $l(\cdot) : V \rightarrow \mathbb{R}$ eine *beschränkte* Linearform:

$$|a(v, w)| \leq \alpha \|v\|_V \|w\|_V, \quad |l(v)| \leq \gamma \|v\|_V, \quad v, w \in V. \quad (3.1.7)$$

Mit diesen Bezeichnungen betrachten wir die folgende allgemeine Variationsgleichung: Bestimme $u \in V$, so dass

$$a(u, \varphi) = l(\varphi) \quad \forall \varphi \in V. \quad (3.1.8)$$

Zum Nachweis, dass diese Aufgabe auch eine Lösung besitzt, postulieren wir, dass die Bilinearform $a(\cdot, \cdot)$ „(stark) V -elliptisch“ ist, d. h.:

$$a(v, v) \geq \kappa \|v\|_V^2, \quad v \in V, \quad (3.1.9)$$

mit einer Konstante $\kappa > 0$. Allgemeiner wird die Bilinearform $a(\cdot, \cdot)$ „koerzitiv“ (oder „regulär“) genannt, wenn gilt:

$$\sup_{\varphi \in V} \frac{a(v, \varphi)}{\|\varphi\|_V} \geq \gamma \|v\|_V, \quad \sup_{\varphi \in V} \frac{a(\varphi, v)}{\|\varphi\|_V} \geq \gamma \|v\|_V, \quad v \in V, \quad (3.1.10)$$

mit einer Konstante $\gamma > 0$.

Hilfssatz 3.1: (Lax¹-Milgram²-Lemma) *Unter den obigen Voraussetzungen besitzt die Gleichung (3.1.8) eine eindeutige Lösung $u \in V$, für welche die a priori Abschätzung gilt:*

$$\|u\|_V \leq \frac{1}{\kappa} \|l\|_{V^*}, \quad (3.1.11)$$

mit der „Dualnorm“ $\|l\|_{V^*} := \sup_{\{\varphi \in V, \|\varphi\|_V=1\}} |l(\varphi)|$.

Beweis: Für jedes feste $v \in V$ definiert $a(v, \cdot)$ ein lineares, stetiges Funktional auf V . Nach dem Rieszschens³ Darstellungssatz existieren Elemente $Av \in V$ und $f \in V$, so dass

$$a(v, \varphi) = (Av, \varphi)_V, \quad l(\varphi) = (f, \varphi)_V, \quad \varphi \in V.$$

Die Zuordnung $v \mapsto Av$ definiert eine lineare Abbildung mit der Eigenschaft

$$\|Av\|_{V^*} \leq \alpha \|v\|_V,$$

d. h.: A ist beschränkt. Die Aufgabe (3.1.8) ist offenbar äquivalent zu der Gleichung

$$Au = f. \quad (3.1.12)$$

Wir wollen zeigen, dass die Abbildung

$$v \in V \mapsto T_\delta v := v - \delta(Av - f) \in V$$

¹Peter David Lax (1926–): US-Amerikanischer Mathematiker ungarischer Abstammung; Prof. an der New York University und am Courant-Institut; wichtige Beiträge zur Analysis, insbesondere zu den partiellen Differentialgleichungen der Math. Physik, und zur Numerik.

²Arthur Norton Milgram (1912–1961): US-Amerikanischer Mathematiker; Prof. an der Univ. of Minnesota, Minneapolis, USA; Beiträge u. a. zur Funktionalanalysis und ihren Anwendungen in der Theorie partieller Differentialgleichungen; am besten bekannt durch das sog. „Lax-Milgram-Lemma“ zus. mit P. Lax (1954).

³Frigyes Riesz (1880–1956): Ungarischer Mathematiker; Prof. in Szeged und Budapest; fundamentale Beiträge zur Funktionalanalysis, insbesondere der Fourier-Analysis im Hilbert-Raum als theoretische Grundlage der frühen Quantenmechanik.

für einen geeigneten Wert $\delta > 0$ eine Kontraktion auf ganz V ist. Dann besitzt die Fixpunktgleichung

$$T_\delta v = v$$

eine eindeutige Lösung $u \in V$, welche wegen $0 = v - T_\delta v = \delta(Av - f)$ dann auch (eindeutige) Lösung von (3.1.12) bzw. (3.1.8) ist. Die Kontraktionseigenschaft ergibt sich aus der Beziehung

$$\begin{aligned} \|v - \delta Av\|_V^2 &= \|v\|_V^2 - 2\delta a(v, v) + \delta^2 \|Av\|_V^2 \\ &\leq (1 - 2\delta\kappa + \delta^2\alpha^2) \|v\|_V^2, \end{aligned}$$

für $0 < \delta < 2\kappa/\alpha^2$. Die a priori Abschätzung (3.1.11) ergibt sich dann direkt durch Testen mit $\varphi := u$ in der Variationsgleichung (3.1.8). Q.E.D.

Die beiden obigen Beispiele zur 1. RWA passen in diesen Rahmen mit den natürlichen Setzungen $V := H_0^1(\Omega)$, $l(v) := (f, \varphi)$ und

$$a(v, w) := (\nabla v, \nabla w), \quad a(v, w) := (\nabla v, \nabla w) + (\partial_1 v, w).$$

Die Beschränktheit dieser Formen ergibt sich direkt mit Hilfe der Hölderschen und der Poincaréschen Ungleichung. Ihre V -Elliptizität ergibt sich unmittelbar:

$$a(v, v) = \|\nabla v\|^2 = \|v\|_V^2,$$

bzw. unter Beachtung von $v|_{\partial\Omega} = 0$:

$$\begin{aligned} a(v, v) &= \|\nabla v\|^2 + (\partial_1 v, v) = \|\nabla v\|^2 + \frac{1}{2}(\partial_1 v^2, 1) \\ &= \|\nabla v\|^2 + \frac{1}{2}(n_1 v^2, 1)_{\partial\Omega} = \|\nabla v\|^2 = \|v\|_V^2. \end{aligned}$$

Durch geeignete Setzungen lassen sich auch die variationellen Formulierungen der 2. und 3. RWA in diesen abstrakten Rahmen einordnen.

Zur Approximation der Variationsgleichung (3.1.2) werden endlich dimensionale Teilräume

$$V_h \subset V \quad (0 < h \leq h_0)$$

ausgewählt, deren *Feinheit* durch einen Diskretisierungsparameter h (z. B.: Gitterweite) charakterisiert ist.

i) Im Fall einer symmetrischen Bilinearform $a(\cdot, \cdot)$ bestimmt das klassische „Ritzsche⁴ Projektions-Verfahren“ Näherungslösungen $u_h \in V_h$ durch die Vorschrift

$$E(u_h) = \min_{v_h \in V_h} E(v_h) \tag{3.1.13}$$

⁴Walter Ritz (1878–1909): Schweizer Physiker; Prof. in Zürich und Göttingen; Beiträge zu Spektraltheorie in der Kernphysik und Elektro-Magnetismus.

oder äquivalent durch die diskrete Variationsgleichung

$$a(u_h, \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h. \quad (3.1.14)$$

Die (eindeutige) Existenz der diskreten Lösung $u_h \in V_h$ folgt mit demselben Argument wie beim kontinuierlichen Problem. Diese Analogie der Schlußweisen von kontinuierlicher und diskreter (endlich dimensionaler) Situation ist die charakteristische Stärke der Projektionsmethoden im Gegensatz zu den Differenzenverfahren. Die Bezeichnung Projektionsverfahren ist motiviert durch die Beziehung

$$a(u - u_h, \varphi_h) = 0 \quad \varphi_h \in V_h, \quad (3.1.15)$$

welche man durch Subtraktion der Gleichungen (3.1.2) und (3.1.14) erhält. Sie kann geometrisch dahingehend interpretiert werden, dass der Fehler $e_h := u - u_h$ bzgl. des Skalarprodukts $a(\cdot, \cdot)$ senkrecht auf dem Ansatzraum V_h steht. Dies impliziert auch die sog. „Bestapproximationseigenschaft“ für den Approximationsfehler e_h bzgl. der natürlichen „Energie-Norm“ $\|\cdot\|_a := a(\cdot, \cdot)^{1/2}$, denn mit beliebigem $\varphi_h \in V_h$ gilt:

$$\|e_h\|_a^2 = a(e_h, e_h) = a(e_h, u - \varphi_h) + a(e_h, \varphi_h - u_h) \leq \|e_h\|_a \|u - \varphi_h\|_a$$

bzw.

$$\|e_h\|_a \leq \inf_{\varphi_h \in V_h} \|u - \varphi_h\|_a. \quad (3.1.16)$$

Da die Normen $\|\cdot\|_a$ und $\|\cdot\|_V$ auf V äquivalent sind, ist die Frage nach der Konvergenz des Projektionsverfahrens,

$$\|e_h\|_V \rightarrow 0 \quad (h \rightarrow 0), \quad (3.1.17)$$

damit zurückgeführt auf die Frage der Approximierbarkeit von Funktionen $u \in V$ durch Ansatzfunktionen $\varphi_h \in V_h$:

$$\inf_{\varphi_h \in V_h} \|u - \varphi_h\|_V \rightarrow 0 \quad (h \rightarrow 0). \quad (3.1.18)$$

ii) Wenn die Bilinearform $a(\cdot, \cdot)$ nicht symmetrisch ist, wie beim obigen Diffusions-Transport-Problem, kann die zugehörige RWA nicht mehr durch ein Minimierungsproblem charakterisiert werden. Das allgemeine „Galerkinsche⁵ (Projektions)-Verfahren“ geht direkt von der Variationsgleichung (3.1.2) aus und bestimmt Näherungen durch die Beziehung

$$a(u_h, \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h. \quad (3.1.19)$$

Wegen der V -Elliptizität der Bilinearform $a(\cdot, \cdot)$ auf dem endlich dimensionalen Teilraum V_h folgt unmittelbar die Existenz der (eindeutigen) Lösung $u_h \in V_h$. Die Ortho-

⁵Boris Grigorievich Galerkin (1871–1945): Russischer Bauingenieur und Mathematiker; Prof. in St. Petersburg; Beiträge zur Struktur-Mechanik, insbesondere zur Plattentheorie.

gonalitätsbeziehung (3.1.15) bleibt dabei gültig. Damit erschließen wir die Quasi-Best-Approximationseigenschaft (Übungsaufgabe)

$$\|e_h\|_V \leq \frac{\alpha}{\kappa} \min_{\varphi \in V_h} \|u - \varphi_h\|_V. \quad (3.1.20)$$

iii) Eine noch allgemeinere Variante, bei der Ansatzraum V_h^{ansatz} und Testraum V_h^{test} unterschiedlich gewählt werden,

$$u \in V_h^{\text{ansatz}} : \quad a(u_h, \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h^{\text{test}}, \quad (3.1.21)$$

ist das sog. „Petrow⁶ -Galerkin-Verfahren“. Wir werden später Beispiele für diesen unkonventionellen Ansatz kennenlernen.

Zur praktischen Realisierung des Projektionsverfahrens muss die zunächst abstrakte Variationsgleichung (3.1.19) im Funktionenraum algebraisiert werden, d. h.: in ein äquivalentes algebraisches Gleichungssystem umgewandelt werden. Dazu wählen wir zunächst eine Basis $\{\varphi_h^{(i)}, i = 1, \dots, N\}$, $N := \dim V_h$, von V_h aus und machen für die zu bestimmende diskrete Lösung den Ansatz $u_h = \sum_{j=1}^N \xi_j \varphi_h^{(j)}$. Wird dies in (3.1.19) eingesetzt und lässt man die Testfunktionen $\varphi_h \in V_h$ alle Basisfunktionen durchlaufen, ergibt sich ein lineares (algebraisches) $N \times N$ -Gleichungssystem

$$\sum_{j=1}^N \xi_j a(\varphi_h^{(j)}, \varphi_h^{(i)}) = (f, \varphi_h^{(i)}), \quad i = 1, \dots, N,$$

für den Vektor $\xi = (\xi_j)_{j=1}^N$ der Entwicklungskoeffizienten, bzw. in kompakter Schreibweise

$$A_h \xi = b_h. \quad (3.1.22)$$

Dabei sind die Koeffizientenmatrix $A_h = (a_{ij})_{i,j=1}^N$ sowie die rechte Seite $b_h = (b_i)_{i=1}^N$ durch die spezielle Wahl der Basis bestimmt:

$$a_{ij} = a(\varphi_h^{(j)}, \varphi_h^{(i)}), \quad b_j = (f, \varphi_h^{(j)}).$$

Die Entwicklungskoeffizienten ξ_j können sehr unterschiedliche Bedeutung haben; z. B.: Monom-Koeffizienten einer Polynomdarstellung, Fourier-Koeffizienten einer trigonometrischen Entwicklung, Knotenwerte einer stückweise polynomialen Funktion, u.s.w.. Die Eigenschaften der Bilinearform $a(\cdot, \cdot)$ übertragen sich direkt auf die zugehörige Matrix A_h . Ist $a(\cdot, \cdot)$ symmetrisch, so auch A_h ,

$$a_{ij} = a(\varphi_h^{(j)}, \varphi_h^{(i)}) = a(\varphi_h^{(i)}, \varphi_h^{(j)}) = a_{ji},$$

⁶Georgi Iwanowitsch Petrow (1912–1987): Russischer Ingenieur; 1965–1973 Direktor des Instituts für Raumfahrtforschung; Publ.: „Application of the Galerkin method and the problem of flow stability of a viscous liquid“ (russ.), Prikl. Mat. Mekh. 4, 36–47 (1947)

und die V -Elliptizität von $a(\cdot, \cdot)$ impliziert die Definitheit von A_h , denn für $x \in \mathbb{R}^n \setminus \{0\}$ gilt:

$$\begin{aligned} (A_h x, x) &= \sum_{i,j=1}^N a_{ij} x_i x_j = \sum_{i,j=1}^N a(\varphi_h^{(j)}, \varphi_h^{(i)}) x_i x_j \\ &= a\left(\sum_{j=1}^N x_j \varphi_h^{(j)}, \sum_{i=1}^N x_i \varphi_h^{(i)}\right) \geq \kappa \left\| \sum_{i=1}^N x_i \varphi_h^{(i)} \right\|_V^2 > 0. \end{aligned}$$

3.1.1 Beispiele von Galerkin-Ansatzräumen

Wir wollen einige konkrete Realisierungen für den beschriebenen abstrakten Rahmen diskutieren. Bei der Wahl der Ansatzräume $V_h \subset V = H_0^1(\Omega)$ sowie der Basen zur Aufstellung der Gleichungssysteme (3.1.22) sind einige Bedingungen zu beachten:

- Die Berechnung der Matrixelemente $a_{ij} = a(\varphi_h^{(j)}, \varphi_h^{(i)})$ sowie die der rechten Seite $(f, \varphi_h^{(i)})$ sollte „billig“ sein.
- Aus Genauigkeitsgründen wird die Problemdimension in der Regel sehr groß sein: $N \gg 100$. Die Matrix A_h sollte daher möglichst dünn besetzt sein, d. h. möglichst viele Nullen enthalten.
- Die Matrix A_h sollte nicht zu schlecht konditioniert sein; akzeptabel sind z. B. beim vorliegenden Problem $\text{cond}_2(A_h) \approx O(N) - O(N^2)$, wogegen $\text{cond}_2(A_h) \approx O(N^4) - O(e^N)$ nicht praktikabel wären.

Beispiele von solchen Ansätzen sind:

1) *Globaler Polynomansatz*: Auf einem Quadrat $\Omega = (0, 1)^2$ wird der Tensor-Produkt-Ansatz gemacht

$$V_h := Q_m(\Omega) := \left\{ p(x, y) = \sum_{i,j=0}^m c_{ij} x^i y^j \right\}, \quad h := 1/m, \quad N = (m+1)^2.$$

Als Basen kommen dabei in Frage:

a) Monombasis $1, x, y, x^2, xy, y^2, \dots$; die zugehörige Matrix A mit den Elementen

$$a_{ij} = \int_0^1 \int_0^1 \nabla x^i \cdot \nabla y^j \, dx \, dy, \quad 0 \leq i, j \leq m,$$

verhält sich dann wie die bekannte Hilbert-Matrix mit exponentiell mit N wachsender Kondition $\text{cond}_2(A_h) = O(e^N)$. Dieser Ansatz ist also praktisch unbrauchbar.

b) Tensorprodukt-Basen L^2 -orthogonaler Polynome wie z. B. Legendre-Polynome oder Tschebyscheff-Polynome; die zugehörige Matrix A_h mit den Elementen

$$a_{ij} = \int_0^1 \int_0^1 \nabla L_i^{(m)} \cdot \nabla L_j^{(m)} \, dx \, dy, \quad 0 \leq i, j \leq m,$$

ist dann zwar voll besetzt, hat aber eine wesentlich günstigere Kondition, $\text{cond}_2(A_h) = O(N)$. Dieser Ansatz führt auf die sog. „Spektral-Galerkin-Verfahren“, welche bei Problemen auf geometrisch einfachen (rechteckigen) Gebieten sehr leistungsfähig sind. Die Bezeichnung „Spektralverfahren“ rührt daher, dass man die orthogonalen Polynome auch als Eigenfunktionen gewisser Differentialoperatoren 2. Ordnung charakterisieren kann. Wegen ihrer konzeptionellen Beschränkung auf einfache Geometrien wollen wir derartige Methoden hier nicht weiter diskutieren. Stichworte für Entwicklungen in Richtung auf eine Überwindung dieser Restriktion sind z. B. „Spektral-Elemente-Methoden“ und „ h/p -Finite-Elemente-Methode“.

2) *Globaler trigonometrischer Ansatz („echte Spektralverfahren“)*: Wieder auf einem Quadrat $\Omega = (0, 1)^2$ wird der Tensor-Produkt-Ansatz gemacht

$$V_h := T_m(\Omega) := \left\{ t(x, y) = \sum_{i,j=0}^m c_{ij} \sin(i\pi x) \sin(j\pi y) \right\}, \quad h := 1/m, \quad N = (m+1)^2.$$

Als Basen verwendet man dabei die trigonometrische Basis $\{1, \sin(n\pi x) \sin(m\pi y), \dots\}$. Die zugehörige Matrix A_h ist dann vergleichsweise gut konditioniert, $\text{cond}_2(A_h) = O(N)$. In diesem Fall gibt es mit der schnellen Fourier-Transformation („FFT“) einen fast optimalen Algorithmus mit der Komplexität $O(N \log(N))$ zur Lösung des Gleichungssystems (3.1.22). Der Nachteil dieses in Spezielfällen sehr leistungsfähigen Ansatzes ist wieder seine Beschränktheit auf einfache Rechteckgeometrien und sog. „separable“ Differentialoperatoren mit konstanten Koeffizienten.

3) *Stückweise polynomialer Ansatz („Finite Elemente“)*: Um das Problem der Approximation allgemeiner Gebiete zu lösen, werden Ansatzfunktionen (auch „Formfunktionen“ genannt) verwendet, welche bzgl. einer Zerlegung von $\bar{\Omega}$ in einfache Teilgebiete T , sog. „Zellen“, stückweise polynomial sind. Gängige Beispiele von Zellen sind Dreiecke oder (konvexe) Vierecke in zwei bzw. Tetraeder oder (konvexe) Hexaeder in drei Dimensionen. Der Parameter h ist in diesem Fall etwa der maximale Zelldurchmesser.

Wir illustrieren diesen Finite-Elemente-Ansatz anhand eines einfachen Beispiels. Die 1. RWA (3.1.2) sei auf einem (konvexen) polygonalen Gebiet $\Omega \subset \mathbb{R}^2$ mit homogenen Randwerten $u|_{\partial\Omega} = 0$ und rechter Seite $f \in L^2(\Omega)$ gestellt. Die zugehörige Lösung $u \in H_0^1(\Omega)$ ist dann auch im Sobolew-Raum $H^2(\Omega)$ und genügt der *a priori* Abschätzung

$$\|\nabla^2 u\| \leq c_S \|f\|, \quad (3.1.23)$$

wobei $c_S = 1$ im Falle eines konvexen Gebiets.

Weiter sei eine Folge von Zerlegungen $\mathbb{T}_h = \{T\}$ des Gebiets $\bar{\Omega}$ in abgeschlossene Dreiecke T („Triangulierung“) gegeben mit $h := \max_T \text{diam}(T) \rightarrow 0$. Wir stellen die folgenden Regularitätsbedingungen an diese Triangulierung:

i) *Strukturregularität*: Je zwei Dreiecke der Zerlegung $\bar{\Omega} = \bigcup \{T \in \mathbb{T}_h\}$ überlappen sich höchstens in gemeinsamen Eckpunkten oder in ganzen Seiten, d. h.: Sog. „hängende“ Knoten auf Dreiecksseiten sind hier nicht erlaubt.

ii) *Formregularität*: Alle Dreiecke der Triangulierungen $T \in \mathbb{T}_h$ sind von ähnlicher Gestalt, d.h.: Für den Inkreisradius ρ_T und Umkreisradius h_T eines jeden Dreiecks T gilt gleichmäßig für $h \rightarrow 0$:

$$\max_{T \in \mathbb{T}_h} \frac{h_T}{\rho_T} \leq c_1 \tag{3.1.24}$$

iii) *Größenregularität*: Alle Dreiecke einer Triangulierung \mathbb{T}_h sind von gleicher Größenordnung, d. h.: Es gilt gleichmäßig für $h \rightarrow 0$:

$$\max_{T \in \mathbb{T}_h} h_T \leq c_2 \min_{T \in \mathbb{T}_h} h_T. \tag{3.1.25}$$

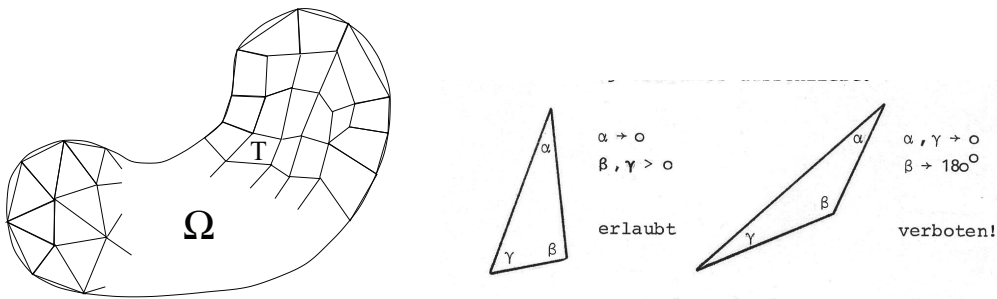


Abbildung 3.1: Finite-Elemente: Dreiecks- bzw. Vierecksgitter

Auf den Triangulierungen \mathbb{T}_h definieren wir Ansatzräume stückweise linearer Funktionen („lineare finite Elemente“: Verallgemeinerung des Konzepts eines Polygonzugs auf höhere Raumdimensionen):

$$V_h^{(1)} := \{v_h \in C(\overline{\Omega}) \mid v_{h|T} \in P_1(T), T \in \mathbb{T}_h, v|_{\partial\Omega} = 0\}.$$

Dabei bezeichnet allgemein $P_r(T)$ den Vektorraum der Polynome bis zum Grad $r \geq 0$ über T . Man überlegt sich leicht, dass dadurch tatsächlich Teilräume $V_h^{(1)} \subset H_0^1(\Omega)$ erklärt sind. Sei N die Anzahl der „inneren“ Knoten (Dreieckseckpunkte) der Triangulierung. Jedes $v_h \in V_h^{(1)}$ ist als stückweise lineare Funktion eindeutig durch Vorgabe ihrer Funktionswerte („Knotenwerte“) in den „inneren“ Dreieckseckpunkten („Knoten“) festgelegt. In den Eckpunkten auf dem Gebietsrand $\partial\Omega$ ist $v_h = 0$ wegen der Dirichlet-Randbedingung. In $V_h^{(1)}$ gibt es daher eine natürliche Basis, die sog. „Knotenbasis“ in Analogie zur „Lagrange⁷-Basis“ bei der eindimensionalen Lagrange-Interpolation. Jedem Knoten a_i wird durch die Bedingung

$$\varphi_h^i(a_j) = \delta_{ij}, \quad j = 1, \dots, N,$$

⁷Joseph Louis de Lagrange (1736–1813): Französischer Mathematiker; 1766-87 Direktor der mathem. Klasse der Berliner Akademie, dann Prof. in Paris; bahnbrechende Arbeiten zur Variationsrechnung, zur komplexen Funktionentheorie sowie zur theor. Mechanik und Himmelsmechanik.

eindeutig eine Funktion $\varphi_h^i \in V_h^{(1)}$ zugeordnet. Damit gilt dann für jedes $v_h \in V_h^{(1)}$ die Darstellung

$$v_h = \sum_{i=1}^N v_h(a_i) \varphi_h^{(i)}.$$

Daraus folgt, dass $\{\varphi_h^{(i)}, i = 1, \dots, N\}$ tatsächlich eine Basis von $V_h^{(1)}$ ist. Umgekehrt lässt sich jeder kontinuierlichen Funktion $v \in C(\overline{\Omega})$ durch die Vorschrift

$$I_h v := \sum_{i=1}^N v(a_i) \varphi_h^{(i)}$$

eindeutig eine (stückweise lineare) „Interpolierende“ $I_h v \in V_h^{(1)}$ zuordnen. Offenbar ist $I_h v_h \equiv v_h$ für $v_h \in V_h^{(1)}$.

Dieser Diskretisierungsansatz erfüllt offensichtlich die oben formulierten Anforderungen an ein brauchbares Galerkin-Verfahren: Die resultierende Systemmatrix A_h ist dünn besetzt (wegen der geringen Überlappung der Träger der Basisfunktionen), und ihre Elemente sind sehr leicht zu berechnen. Wir werden die praktische Berechnung von A_h im Zusammenhang mit allgemeineren Finite-Elemente-Ansätze dieser Art noch eingehender diskutieren.

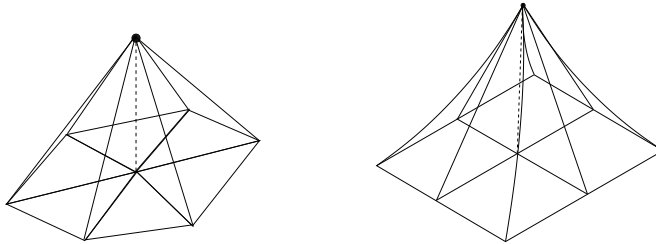


Abbildung 3.2: Knotenbasisfunktionen: Lineare und bilineare FE-Ansätze

Für die Kondition der Matrix A_h werden wir später in zwei Dimensionen $\text{cond}_2(A_h) = O(h^{-2}) = O(N)$ zeigen (ähnlich wie bei der 5-Punkte-Differenzdiskretisierung). Wir rekapitulieren für das Galerkin-Verfahren mit dem Finite-Elemente-Raum $V_h^{(1)}$ die asymptotische Abschätzung für den Fehler $e_h := u - u_h$:

$$\|\nabla e_h\| = \min_{\varphi_h \in V_h^{(1)}} \|\nabla(u - \varphi_h)\|. \quad (3.1.26)$$

Die Frage ist also, ob es $\varphi_h \in V_h^{(1)}$ gibt, so dass $\|\nabla(u - \varphi_h)\| \rightarrow 0$ ($h \rightarrow 0$). Wenn man nur weiß, dass $u \in H_0^1(\Omega)$ ist, dann kann man nur qualitative Konvergenz zeigen. Viel interessanter wäre es, wieder die Konvergenzgeschwindigkeit in Potenzen der Gitterweite h zu kennen. Hierzu ist aber natürlich mehr Regularität der Lösung u erforderlich. Für den stückweise linearen Ansatz werden wir später im Rahmen einer allgemeinen Theorie

die folgende Interpolationsabschätzung zeigen:

$$\|\nabla(v - I_h v)\| \leq c_I h \|\nabla^2 v\|, \quad v \in H_0^1(\Omega) \cap H^2(\Omega). \quad (3.1.27)$$

(Man vergleiche die analoge Abschätzung, welche in einer Dimension bei der Finite-Elemente-Approximation eindimensionaler Sturm-Liouville-Probleme verwendet wird.) Unter den bisher formulierten Voraussetzungen lässt sich eine erste quantitative Konvergenzaussage für das Finite-Elemente-Verfahren ableiten.

Satz 3.1 (Konvergenzsatz): *Für die Galerkin-Approximation von Problem (3.1.2) mit „linearen“ finiten Elementen gelten unter den obigen Voraussetzungen die Fehlerabschätzungen*

$$\|\nabla e_h\| \leq c_I c_S h \|f\|, \quad (3.1.28)$$

$$\|e_h\| \leq c_I^2 c_S^2 h^2 \|f\|, \quad (3.1.29)$$

mit den Konstanten c_I, c_S aus den Ungleichungen (3.1.27) und (3.1.23).

Beweis: (i) Die Abschätzung des sog. „Energienorm-Fehlers“ ergibt sich unmittelbar aus der Best-Approximationsbeziehung (3.1.26), der Interpolationsabschätzung (3.1.27) und der Regularitätsabschätzung (3.1.23):

$$\|\nabla e_h\| \leq \|\nabla(u - I_h u)\| \leq c_I h \|\nabla^2 u\| \leq c_I c_S h \|f\|.$$

(ii) Zum Beweis der Fehlerabschätzung in der L^2 -Norm verwenden wir ein sog. „Dualitätsargument“ („Aubin⁸ -Nitsche⁹ -Trick“). Sei $z \in H_0^1(\Omega)$ die (schwache) Lösung des Hilfsproblems

$$-\Delta z = \|e_h\|^{-1} e_h \text{ in } \Omega, \quad z|_{\partial\Omega} = 0.$$

Diese ist dann auch in $H^2(\Omega)$, und es gilt die *a priori* Abschätzung

$$\|\nabla^2 z\| \leq c_S \|\Delta z\| = c_S,$$

wobei wieder $c_S = 1$ auf konvexem Gebiet Ω . Nach Konstruktion folgt mit Hilfe der Galerkin-Orthogonalität:

$$\begin{aligned} \|e_h\| &= (\nabla e_h, \nabla z) = (\nabla e_h, \nabla(z - I_h z)) \\ &\leq \|\nabla e_h\| \|\nabla(z - I_h z)\| \leq c_I h \|\nabla e_h\| \|\nabla^2 z\| \leq c_I c_S h \|\nabla e_h\|. \end{aligned}$$

Mit dem Ergebnis (i) ergibt sich damit die gewünschte Abschätzung.

Q.E.D.

Die im Beweis von Satz 3.1 verwendete Schlussweise über ein Dualitätsargument ist „das“ zentrale Hilfsmittel bei der Konvergenzanalyse von Finite-Elemente-Verfahren. Die-

⁸Jean-Pierre Aubin (1939–): Französischer Mathematiker; Prof. an der Univ. Paris-Dauphine (2004) emeritiert); Beiträge zur Theorie partieller Differentialgleichungen und ihrer Numerik.

⁹Joachim A. Nitsche (1926–1996): Deutscher Mathematiker; Prof. in Freiburg; fundamentale Beiträge zur Theorie der Finite-Elemente-Methode (u. a. L^∞ -Fehlerabschätzungen).

ses abstrakte Argument entspricht der allgemeinen Regel, dass sich die Analyse der Projektionsverfahren eng an die abstrakten Hilbertraum-Methoden zur Behandlung des kontinuierlichen Problems anlehnt. Das zentrale Hilfsmittel bei der Untersuchung von Differenzenverfahren war dagegen das „(diskrete) Maximumprinzip“, welches sich mehr an den klassischen Techniken für partielle Differentialgleichungen orientiert. Wir wollen diesen Vergleich „Finite-Elemente (FEM) - Finite-Differenzen (FDM)“ anhand des Resultats von Satz 3.1 noch etwas weiterführen.

Die a priori Fehlerabschätzung (3.1.29) für das Finite-Elemente-Verfahren ist zu vergleichen mit der Abschätzung (2.2.37) für das Differenzenverfahren (5-Punkte-Diskretisierung mit Shortley-Weller-Randapproximation auf polygonalen Gebieten):

$$\max_{\bar{\Omega}} |e_h| \leq \frac{1}{24} d_{\Omega}^2 M_4(u) h^2 + \mathcal{O}(h^3), \quad (3.1.30)$$

mit der Schranke $M_4(u)$ für die vierten Ableitungen von u . Beide Abschätzungen zeigen dieselbe asymptotische Konvergenzordnung $\mathcal{O}(h^2)$, was aufgrund der verwendeten Diskretisierungsansätze auch zu erwarten ist. Die Unterschiede liegen zum einen in der Art der Norm, in der der Fehler gemessen wird, und zum anderen in der benötigten Regularität der approximierten Lösung. Beim Differenzenverfahren erhält man wegen der Verwendung des Maximumprinzips punktweise Abschätzungen, wie sie auch der Anwender gern hat. (Der Ingenieur ist z. B. an der maximalen Auslenkung einer belasteten Brückenkonstruktion interessiert.) Dagegen liefert die Hilbert-Raum-Theorie für das Finite-Elemente-Verfahren zunächst nur Abschätzungen im quadratischen Mittel, was etwa lokale „Ausreißer“ an kritischen Stellen nicht ausschließt. (Dem Brückenbauer genügt so etwas nicht, wenn Fehlerspitzen etwa in kritischen Lagerungspunkten der Brücke auftreten können.) Wir werden später die Frage diskutieren, ob und wie man auch für das Finite-Elemente-Verfahren Fehlerabschätzungen in der Maximumnorm herleiten kann. Die in der Abschätzung (3.1.30) geforderte hohe Regularität der Lösung ist ein sehr viel schwerwiegender Nachteil unserer Analyse des Differenzenverfahrens, da diese Regularitätsstufe i. Allg. auf Polygonebenen und unter realistischen Annahmen an die Problemdata nicht erwartet werden kann. Wir bemerken, dass man für das Finite-Elemente-Verfahren mit wesentlich mehr technischem Aufwand „optimale“ Maximumnorm-Fehlerabschätzungen der Form ($h \geq h_0 > 1$)

$$\max_{\bar{\Omega}} |e_h| \leq c M_2(u) h^2 |\log h| \quad (3.1.31)$$

beweisen kann. Allerdings ist auch die abgeschwächte Annahme $M_2(u) < \infty$ i. Allg. noch zu restriktiv. Für das Finite-Elemente-Verfahren ist auch noch unter der Minimalvoraussetzung $u \in H_0^1(\Omega)$ wenigstens qualitative Konvergenz gesichert. Seinen eigentlichen Vorteil, nämlich die große Flexibilität bei der Approximation von komplizierten Geometrien auf unstrukturierten Gittern werden wir später im Zusammenhang mit der Frage nach adaptiver Gittersteuerung und Fehlerkontrolle erkennen.

3.1.2 Diskretes Maximumprinzip für Finite-Elemente-Approximationen

Als nächstes wollen wir zeigen, dass Galerkin-Verfahren mit finiten Elementen als Ansatzfunktionen tatsächlich eng verwandt mit Differenzenverfahren sind. Dazu werden die Elemente der „Steifigkeitsmatrix“ A_h des Finite-Elemente-Verfahrens für den Fall „linearer“ Ansatzfunktionen auf einem Dreiecksgitter explizit bestimmt.

Dazu wird zunächst ein einzelnes Dreieck T mit den Eckpunkten P_i ($i = 1, 2, 3$) betrachtet (siehe Abb. 3.3). Die dem Eckpunkt P_i gegenüberliegende Seite sei mit S_i und die zugehörige Höhe mit H_i bezeichnet. Die Seiten werden dabei als im Gegenuhrzeigersinn und die Höhen gegen den Eckpunkt orientierte Vektoren aufgefasst. Weiter bezeichne ψ_i die (stückweise lineare) Knotenbasisfunktion zum Punkt P_i , welche auf T definiert ist durch $\psi_i(P_k) = \delta_{ik}$, $i, k = 1, 2, 3$.

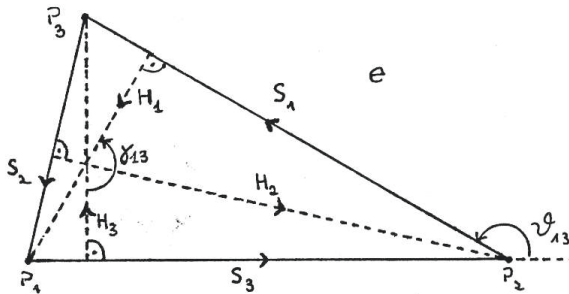


Abbildung 3.3: Dreiecksschema

Es gilt $\nabla\psi_h \equiv \text{konst.}$ und wegen $\psi_i(P_j) = \psi_i(P_k) = 0$, $i \neq j, k$, hat $\nabla\psi_i$ in Richtung S_i die Komponente Null. Folglich zeigt $\nabla\psi_i$ in Richtung H_i und hat wegen $\psi_i(P_i) = 1$ den Betrag $|H_i|^{-1}$:

$$\nabla\psi_i = \frac{H_i}{|H_i|^2}.$$

Wir erhalten also

$$(\nabla\psi_i, \nabla\psi_j)_T = |T| \frac{(H_i, H_j)}{|H_i|^2 |H_j|^2}.$$

Für den Winkel γ_{ij} zwischen den Höhen H_i und H_j gilt

$$\cos(\gamma_{ij}) = \frac{(H_i, H_j)}{|H_i| |H_j|}$$

und, da γ_{ij} gleich dem Winkel θ_{ij} zwischen den Seitenvektoren S_i und S_j ist, auch (Man beachte die Orientierung von S_i und S_j):

$$\cos(\gamma_{ij}) = \frac{(S_i, S_j)}{|S_i| |S_j|}$$

Damit folgt bei Beachtung von $2|T| = |H_i| |S_i|$ die Beziehung

$$(\nabla\psi_i, \nabla\psi_j)_T = |T| \frac{(S_i, S_j)}{|S_i| |S_j| |H_i| |H_j|} = \frac{(S_i, S_j)}{4|T|}.$$

Hieraus lesen wir ab, dass

$$(\nabla\psi_i, \nabla\psi_i)_T > 0, \quad (\nabla\psi_i, \nabla\psi_j)_T \leq 0, \quad i \neq j, \tag{3.1.32}$$

falls alle Winkel im Dreieck T kleiner oder gleich $\pi/2$ sind. Weiter ist nach Konstruktion

$$\sum_{j=1}^3 \psi_j \equiv 1, \quad (\nabla\psi_i, \sum_{j=1}^3 \nabla\psi_j)_T = 0,$$

bzw.

$$\sum_{j=1, j \neq i}^N |(\nabla\psi_i, \nabla\psi_j)_T| \leq (\nabla\psi_i, \nabla\psi_i)_T, \quad i = 1, 2, 3.$$

Für die Elemente $a_{ij} = \sum_{T \in \mathbb{T}_h} (\nabla\psi_j, \nabla\psi_i)_T$ der Matrix A_h erhalten wir somit, dass

$$a_{ii} > 0, \quad a_{ij} \leq 0 \quad (i \neq j),$$

wenn alle Dreiecksinnenwinkel in der Triangulierung kleiner oder gleich $\pi/2$ sind. Darüber hinaus gilt dann

$$\sum_{j \neq i} |a_{ij}| \leq a_{ii}, \quad \sum_{j \neq i_0} |a_{i_0 j}| < a_{i_0 i_0} \quad \text{für ein } i_0. \tag{3.1.33}$$

Die Steifigkeitsmatrix A_h ist in diesem Fall also „(irreduzibel) diagonal-dominant“, „von nicht-negativem Typ“ und eine „ M -Matrix“. Wir fassen die sich daraus ergebenden Konsequenzen in einem Satz zusammen.

Satz 3.2 (Maximumprinzip für finite Elemente): *Wenn alle Innenwinkel der Triangulierung \mathbb{T}_h kleiner oder gleich $\pi/2$ sind, genügt das Finite-Elemente-Schema mit stückweise linearen Ansatzfunktionen einem diskreten Maximumprinzip, d. h.:*

$$(\nabla v_h, \nabla \varphi_h^{(n)}) \leq 0 \quad (n = 1, \dots, N) \quad \Rightarrow \quad \max_{\overline{\Omega}} v_h \leq \max\{0, \max_{\partial\Omega} v_h\}. \tag{3.1.34}$$

Ferner ist die Steifigkeitsmatrix A_h eine M -Matrix, d. h.: Es gilt $A_h^{-1} \geq 0$ sowie

$$x \in \mathbb{R}^N, \quad A_h x \geq 0 \quad \Rightarrow \quad x \geq 0. \tag{3.1.35}$$

Dieses Resultat kann so interpretiert werden, dass unter den gegebenen Voraussetzungen auch die Finite-Elemente-Diskretisierung ein „diskretes Maximumprinzip“ erfüllt. Leider gilt die kritische Eigenschaft (3.1.32) praktisch nur in der oben beschriebenen Situation. Insbesondere Finite-Elemente-Ansätze höherer Ordnung erfüllen dies nicht (z. B. quadratische Ansätze nur auf „gleichseitigen“ Triangulierungen).

Bemerkung 3.4: Im Spezialfall einer gleichförmigen, kartesischen Triangulierung (mit Kantenlänge h) des Einheitsquadrats erhalten wir aus der obigen expliziten Darstellung für die Matrixelemente $a_{ij} = (\nabla\psi_i, \nabla\psi_j)$ die Beziehung:

$$a_{ii} = 4, \quad a_{i,i\pm 1} = -1, \quad a_{i,i\pm m} = -1;$$

alle anderen Elemente a_{ij} sind Null. In diesem Fall stimmt die Steifigkeitsmatrix A_h^{FEM} also bis auf den Faktor h^{-2} mit der Matrix A_h^{FDM} des „5-Punkte-Operators“ überein:

$$A_h^{FEM} = h^2 A_h^{FDM}. \quad (3.1.36)$$

Für die Elemente des zugehörigen „Lastvektors“ gilt entsprechend:

$$b_i^{FEM} = \int_{\Omega} f\psi_i dx \approx h^2 f(P_i) + \mathcal{O}(h^4) = h^2 b_i^{FDM} + \mathcal{O}(h^4). \quad (3.1.37)$$

Dies zeigt, dass aus algebraischer Sicht FEM und FDM eng verwandt sind. Für eine stückweise „bi-linearen“ Ansatz auf einer gleichförmigen Quadratzerlegung erhält man ein Analogon zu einem „kompakten 9-Punkte-Operator“ (siehe Abb. 3.4):

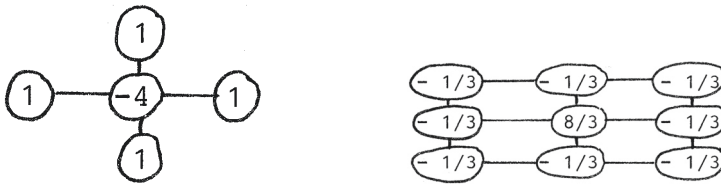


Abbildung 3.4: Differenzensterne: „5-Punkte-Stern“ (links) zur Approximation von Δ und „9-Punkte-Stern“ (rechts) zur Approximation von $-\Delta$.

3.1.3 Approximation krummer Ränder

Zum Abschluss dieser einführenden Diskussion wollen wir noch darstellen, wie in der FEM krumme Ränder approximiert werden. Dazu sei angenommen, dass der Rand $\partial\Omega$ regulär genug ist, dass die schwache Lösung $u \in V := H_0^1(\Omega)$ auch in $H^2(\Omega)$ ist und der a priori Abschätzung

$$\|u\|_{H^2} \leq c_S \|f\| \quad (3.1.38)$$

genügt.

i) Der „konvexe Fall“: Sei $\Omega \subset \mathbb{R}^2$ ein glatt berandetes, *konvexes* Gebiet. Dieses sei überdeckt durch eine reguläre Triangulierung $\mathbb{T}_h = \{T\}$, so dass alle Eckpunkte des Polygonebiets

$$\Omega_h := \bigcup \{T \in \mathbb{T}_h\} \subset \bar{\Omega},$$

auf dem Rand $\partial\Omega$ liegen. Die Länge der Polygonkanten von $\partial\Omega$ ist dann durch die Gitterweite h der Triangulierung \mathbb{T}_h beschränkt (siehe Abb. 3.5).

Auf $\bar{\Omega}$ wird nun zunächst der einfachste Finite-Elemente-Ansatz (mit linearen Formfunktionen) wie folgt definiert:

$$V_h^{(1)} := \{v_h \in C(\bar{\Omega}) \mid v_h|_T \in P_1(T), T \in \mathbb{T}_h, v_h|_{\bar{\Omega} \setminus \Omega_h} \equiv 0\} \subset V = H_0^1(\Omega).$$

Die zugehörigen Galerkin-Approximationen $u_h \in V_h^{(1)}$ sind durch die Variationsgleichung

$$(\nabla u_h, \nabla \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h \tag{3.1.39}$$

bestimmt. Wegen der Teilraumbeziehung $V_h \subset V$ gilt dann wieder für den Fehler $e_h := u - u_h$ die Bestapproximationsbeziehung (3.1.26).

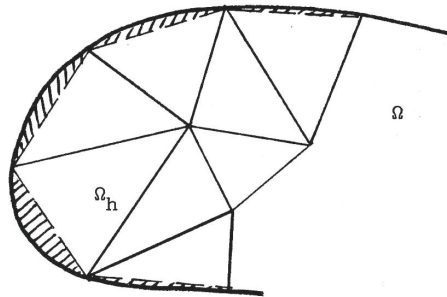


Abbildung 3.5: Polygonale Approximation eines krumm berandeten Gebiets; Randstreifen $S_h = \cup\{S_T\}$ schraffiert.

Satz 3.3 (FEM auf konvexem Gebiet): Für das FE-Schema (3.1.39) auf einem glatt berandeten, konvexen Gebiet Ω gelten die a priori Konvergenzabschätzungen

$$\|\nabla e_h\| \leq (c_I + c_\Omega) c_S h \|f\|, \tag{3.1.40}$$

$$\|e_h\| \leq (c_I + c_\Omega)^2 c_S^2 h^2 \|f\|, \tag{3.1.41}$$

mit Stabilitäts- und Interpolationskonstanten c_S, c_I und einer generischen Konstante c_Ω , die nur vom Gebiet Ω abhängt.

Beweis: i) Sei $I_h u \in V_h^{(1)}$ die natürliche Knoteninterpolierende von u , welche auf dem Streifen $S_h = \Omega \setminus \Omega_h$ zu Null gesetzt wird. Für diese gilt wieder auf jeder Zelle $T \in \mathbb{T}_h$ (wird später gezeigt werden)

$$\|\nabla(u - I_h u)\|_T \leq c_I h_T \|\nabla^2 u\|_T, \tag{3.1.42}$$

und folglich

$$\|\nabla(u - I_h u)\|_{\Omega_h} \leq c_I h \|\nabla^2 u\|_{\Omega_h}.$$

Mit Hilfe der Approximationsbeziehung (3.1.26) ergibt sich somit

$$\|\nabla e_h\|_{\Omega}^2 \leq c_I^2 h^2 \|\nabla^2 u\|_{\Omega_h}^2 + \|\nabla u\|_{S_h}^2. \tag{3.1.43}$$

Es bleibt, das Integral über den Randstreifen S_h zu behandeln.

ii) Für ein glatt berandetes Gebiet (d. h.: $\partial\Omega$ ist C^2 -parametrisiert.) ist nun $|S_h| = \mathcal{O}(h^2)$. Um dies zu sehen, nehmen wir an, dass der Randabschnitt $\partial\Omega_T$, welcher durch Γ von $\partial\Omega$ abgetrennt wird, als Graph einer Funktion $\psi(s)$ der Bogenlänge über Γ aufgefasst werden kann. Diese nehme ihr Maximum ψ_0 für $s = s_0$ an, so dass

$$(\psi - \psi_0)(s_0) = 0, \quad (\psi - \psi_0)'(s_0) = 0.$$

Durch Taylor-Entwicklung von $\psi - \psi_0$ um s_0 ergibt sich dann

$$\max_{\Gamma} |\psi(s)| = \max_{\Gamma} |\psi(s) - \psi_0| \leq \delta := \frac{1}{2} \max_{\Gamma} |\psi''| h_T^2.$$

Folglich ist $|S_h| \leq ch^2$. Damit ist bewiesen, dass

$$\|\nabla e_h\|_{\Omega} \leq ch \|u\|_{H^{2,p}}. \tag{3.1.44}$$

iii) Zur Abschätzung des Integrals über S_h gehen wir ähnlich vor wie beim Beweis der Poincaréschen Ungleichung. Sei $T \in \mathbb{T}_h$ ein Randdreieck und S_T der zugehörige Teilabschnitt des Randstreifens S_h , welcher von der Kante Γ von T begrenzt ist. O.B.d.A. sei angenommen, dass ein Rechteck Q_T mit Γ als kurzer Seite und Länge $L > 0$ (unabhängig von h) ganz in $\overline{\Omega}$ enthalten ist (s. Abb. 3.6).

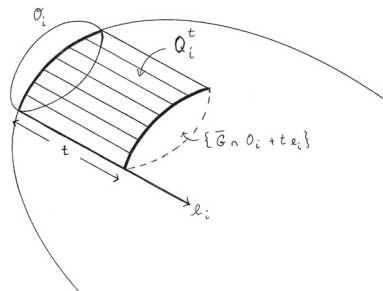


Abbildung 3.6: Schema der Randapproximation

Sei nun weiter $v \in C^1(\overline{\Omega})$ beliebig. Es bezeichne $n_{\Gamma}(x)$ den bzgl. S_h nach innen gerichteten Normaleneinheitsvektor zu Γ im Punkt $x \in \Gamma$ mit Parameterwert s . Damit gilt für $0 \leq t \leq \delta$:

$$v(x + t n_{\Gamma}(x)) = v(x) + \int_0^t \partial_r v(x + r n_{\Gamma}(x)) dr,$$

bzw.

$$|v(x + tn_\Gamma(x))|^2 \leq 2|v(x)|^2 + 2\delta \int_0^{\psi(s)} |\nabla v(x + rn_\Gamma(x))|^2 dr.$$

Wir integrieren dies zunächst über $0 \leq t \leq \psi(s) \leq \delta$ (Man beachte, dass für $x \in \Gamma$ gilt: $x + \psi(s)n_\Gamma(x) \in \partial\Omega_T$),

$$\int_0^{\psi(s)} |v(x + tn_\Gamma(x))|^2 dt \leq 2\delta|v(x)|^2 + 2\delta^2 \int_0^{\psi(s)} |\nabla v(x + rn_\Gamma(x))|^2 dr,$$

und dann über $x \in \Gamma$ und erhalten

$$\int_{S_T} |v(x)|^2 dx \leq 2\delta \int_\Gamma |v(x)|^2 ds + 2\delta^2 \int_{S_T} |\nabla v(x)|^2 dx.$$

Für das Randintegral rechts erhalten wir mit Hilfe der Spurabschätzung

$$\int_\Gamma |v(x)|^2 ds \leq c_\Omega \|v\|_{H^1(Q_T)}^2$$

mit einer von $\max_\Gamma |\psi'|$ abhängigen Konstante c_Ω . Damit gewinnen wir schließlich

$$\int_{S_h} |v(x)|^2 dx \leq c_\Omega h^2 \|v\|_{H^1(\Omega)}^2, \tag{3.1.45}$$

mit einer generischen Konstante c_Ω . Durch das übliche Stetigkeitsargument überträgt sich diese Abschätzung auf alle Funktionen $v \in H^1(\Omega)$. Wir wenden die Abschätzung (3.1.45) nun für die Funktion $|\nabla u| \in H^1(\Omega)$ an und erhalten

$$\|\nabla u\|_{S_h}^2 \leq c_\Omega h^2 \|u\|_{H^2}^2,$$

so dass sich mit (3.1.43) schließlich das erste gewünschte Resultat ergibt:

$$\|\nabla e\| \leq c_I h \|\nabla^2 u\| + c_\Omega h \|u\|_{H^2} \leq (c_I + c_\Omega) c_S h \|f\|. \tag{3.1.46}$$

iv) Zur Abschätzung des L^2 -Fehlers wird wieder ein Dualitätsargument verwendet. Sei $z \in H_0^1(\Omega) \cap H^2(\Omega)$ die Lösung des Hilfsproblems

$$-\Delta z = \|e\|^{-1} e \quad \text{in } \Omega, \quad z|_{\partial\Omega} = 0.$$

Wie im Fall eines Polygonebiets argumentieren wir nun wie folgt:

$$\|e\| = (\nabla e, \nabla z) = (\nabla e, \nabla(z - I_h z)) \leq \|\nabla e\| \|\nabla(z - I_h z)\|.$$

Mit Hilfe der Interpolationsabschätzung (3.1.42) sowie der Abschätzung (3.1.45) für $v := |\nabla z|$ folgt weiter

$$\|e\| \leq \|\nabla e\| \{c_I h \|\nabla^2 z\| + c_\Omega h \|z\|_{H^2}\} \leq (c_I + c_\Omega) h \|\nabla e\| \|z\|_{H^2}.$$

Hiermit folgt dann unter Verwendung des ersten Resultats (3.1.46) sowie der a priori Schranke $\|z\|_{H^2} \leq c_S$ auch die zweite gewünschte Ungleichung

$$\|e\| \leq (c_I + c_\Omega)^2 c_S^2 h^2 \|f\|. \quad (3.1.47)$$

Dies vervollständigt den Beweis.

Q.E.D.

ii) Der „nicht-konvexe Fall“: Sei $\Omega \subset \mathbb{R}^2$ ein glatt berandetes, aber nicht notwendig *konvexes* Gebiet. Dieses sei wieder überdeckt durch eine reguläre Triangulierung $\mathbb{T}_h = \{T\}$, so dass alle Eckpunkte des Polygongebiets $\Omega_h := \bigcup\{T \in \mathbb{T}_h\}$ auf dem Rand $\partial\Omega$ liegen (s. Abb. 3.7). Die Länge der Polygonkanten von $\partial\Omega_h$ ist dann durch die Gitterweite h der Triangulierung \mathbb{T}_h beschränkt. Ist Ω nicht konvex, so ist $\Omega_h \not\subset \bar{\Omega}$. Der auf $\Omega_h = \bigcup\{T \in \mathbb{T}_h\}$ definierte Finite-Elemente-Raum $V_h^{(1)}$ ist dann auch nicht in V enthalten, und die Approximation wird „nicht-konform“ (bzgl. $V = H_0^1(\Omega)$) genannt. Die Analyse dieser Approximation gestaltet sich technisch etwas schwieriger als im konformen Fall. Doch auch hierfür kann man Konvergenz mit der optimalen Ordnung beweisen.

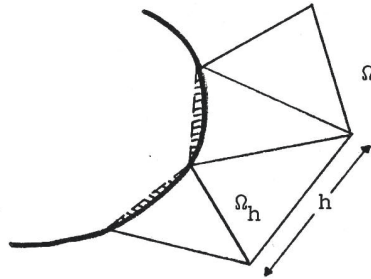


Abbildung 3.7: Approximation nicht-konvexer Randteile

Satz 3.4 (FEM auf nicht-konvexem Gebiet): Für das FE-Schema (3.1.39) auf einem glatt berandeten, nicht-notwendig konvexen Gebiet Ω gelten die a priori Konvergenzabschätzungen

$$\|\nabla e_h\|_\Omega \leq (c_I + c_\Omega) c_S h \|f\|_\Omega \quad (3.1.48)$$

$$\|e_h\|_\Omega \leq (c_I + c_\Omega)^2 c_S^2 h^2 \|f\|_\Omega, \quad (3.1.49)$$

mit Stabilitäts- und Interpolationskonstanten c_S , c_I und einer generischen Konstante c_Ω , die nur vom Gebiet Ω abhängt.

Beweis: i) Das Problem beim Beweis des Satzes besteht in der „Nicht-Konformität“ der Diskretisierung, d. h.: $V_h^{(1)} \not\subset V$. Zunächst gilt auch im Fall eines nichtkonvexen Gebiets

$$\|\nabla(u - I_h u)\|_\Omega \leq c_I h \|\nabla^2 u\|_\Omega \leq c_I c_S h \|f\|_\Omega. \quad (3.1.50)$$

Weiter haben wir

$$\begin{aligned}
 \|\nabla e_h\|_\Omega^2 &= (\nabla e_h, \nabla(u - I_h u))_\Omega + (\nabla e_h, \nabla(I_h u - u_h))_\Omega \\
 &= (\nabla e_h, \nabla(u - I_h u))_\Omega + (\nabla u, \nabla(I_h u - u_h))_\Omega - (\nabla u_h, \nabla(I_h u - u_h))_\Omega \\
 &= (\nabla e_h, \nabla(u - I_h u))_\Omega + (\nabla u, \nabla(I_h u - u_h))_\Omega - (f, I_h u - u_h)_\Omega \\
 &\leq (\nabla e_h, \nabla(u - I_h u))_\Omega + N_h(u) \|\nabla(I_h u - u_h)\|_\Omega,
 \end{aligned}$$

mit

$$N_h(u) := \sup_{\psi_h \in V_h^{(1)}} \frac{|(\nabla u, \nabla \psi_h)_\Omega - (f, \psi_h)_\Omega|}{\|\nabla \psi_h\|_\Omega}.$$

Hieraus ergibt sich

$$\|\nabla e_h\|_\Omega^2 \leq 3 \|\nabla(u - I_h u)\|_\Omega^2 + 2N_h(u)^2.$$

ii) Zur Abschätzung des Nichtkonformitätsterms schätzen wir wie folgt ab:

$$\begin{aligned}
 (\nabla u, \nabla \psi_h)_\Omega - (f, \psi_h)_\Omega &= (-\Delta u - f, \psi_h)_\Omega + (\partial_n u, \psi_h)_{\partial\Omega} \\
 &= (\partial_n u, \psi_h)_{\partial\Omega} \leq \|\partial_n u\|_{\partial\Omega} \|\psi_h\|_{\partial\Omega}.
 \end{aligned}$$

Zunächst gilt aufgrund einer bekannten Spurabschätzung und der üblichen Regularitätsabschätzung auf glatt-berandeten Gebieten

$$\|\partial_n u\|_{\partial\Omega} \leq c_\Omega \|u\|_{H^2} \leq c_\Omega c_S \|f\|_\Omega. \tag{3.1.51}$$

Zur Behandlung des Terms $\|\psi_h\|_{\partial\Omega}$ benötigen wir etwas zusätzliche Notation. Da wir nur lineare oder (isoparametrische) bilineare Ansatzfunktionen betrachten, gibt es auf $\partial\Omega$ ein stetiges (nach außen orientiertes) Richtungsvektorfeld $r(x)$, so dass jedes $\psi_h \in V_h^{(1)}$ entlang der von $r(x)$ aufgespannten Gerade $\{x + sr(x), s \in \mathbb{R}\}$ linear ist. Für den Schnittpunkt $x_h = x + d(x)r(x)$ dieser Gerade ausgehend von $x \in \partial\Omega$ mit dem Rand $\partial\Omega_h$ gilt wieder $|x_h - x| = d(x) \leq c_\Omega h^2$. Durch Taylor-Entwicklung folgt dann

$$0 = \psi_h(x_h) = \psi_h(x) + \partial_r \psi_h(x) d(x),$$

und bei Integration über $\partial\Omega$:

$$\|\psi_h\|_{\partial\Omega} \leq c_\Omega h^2 \|\partial_r \psi_h\|_{\partial\Omega}. \tag{3.1.52}$$

Dies impliziert insbesondere, dass

$$\|\psi_h\|_{\partial\Omega} \leq c_\Omega h \|\nabla \psi_h\|_\Omega,$$

und damit

$$N_h(u) \leq c_\Omega c_S h \|f\|_\Omega.$$

Dies in Verbindung mit der Interpolationsfehlerabschätzung (3.1.50) beweist die Energienormfehlerabschätzung.

iii) Zum Beweis der L^2 -Normfehlerabschätzung verwenden wir wieder ein Dualitätsragu-

ment. Sei $z \in H_0^1(\Omega) \cap H^2(\Omega)$ die (eindeutig bestimmte) Lösung des Hilfsproblems

$$-\Delta z = e_h \|e_h\|^{-1} \text{ in } \Omega, \quad z|_{\partial\Omega} = 0. \quad (3.1.53)$$

Auf dem Gebiet Ω gilt die a priori Abschätzung $\|z\|_{H^2} \leq c_\Omega \|\Delta z\|_\Omega \leq c_\Omega$. Damit gilt

$$\begin{aligned} \|e_h\|_\Omega &= (e_h, -\Delta z)_\Omega = (\nabla e_h, \nabla z)_\Omega - (e_h, \partial_n z)_{\partial\Omega} \\ &= (\nabla e_h, \nabla(z - I_h z))_\Omega + (\nabla e_h, \nabla I_h z)_\Omega - (e_h, \partial_n z)_{\partial\Omega} \\ &= (\nabla e_h, \nabla(z - I_h z))_\Omega + (\nabla u, \nabla I_h z)_\Omega - (f, I_h z)_\Omega - (e_h, \partial_n z)_{\partial\Omega} \\ &= (\nabla e_h, \nabla(z - I_h z))_\Omega + (\partial_n u, I_h z)_{\partial\Omega} + (u_h, \partial_n z)_{\partial\Omega}. \end{aligned}$$

Wie oben unter (i) und (ii) schätzen wir ab

$$|(\nabla e_h, \nabla(z - I_h z))_\Omega| \leq \|\nabla e_h\|_\Omega \|\nabla(z - I_h z)\|_\Omega \leq c_I h \|\nabla e_h\|_\Omega \|\nabla^2 z\|_\Omega \leq c_I c_S h \|\nabla e_h\|_\Omega.$$

Weiter folgt mit Hilfe der Abschätzung (3.1.52) und den schon oben verwendeten Interpolations-, Spur- und Regularitätsabschätzungen:

$$\begin{aligned} |(\partial_n u, I_h z)_{\partial\Omega}| &\leq \|\partial_n u\|_{\partial\Omega} \|I_h z\|_{\partial\Omega} \leq c_\Omega c_S \|f\|_\Omega h^2 \|\partial_r I_h z\|_{\partial\Omega} \\ &\leq c_\Omega c_S \|f\|_\Omega h^2 \{\|\partial_r(I_h z - z)\|_{\partial\Omega} + \|\partial_r z\|_{\partial\Omega}\} \\ &\leq c_\Omega c_S \|f\|_\Omega h^2 \{c_I h^{1/2} \|z\|_{H^2} + c_\Omega \|z\|_{H^2}\} \leq c_\Omega c_I c_S h^2 \|f\|_\Omega. \end{aligned}$$

Analog erschließen wir

$$\begin{aligned} |(u_h, \partial_n z)_{\partial\Omega}| &\leq \|u_h\|_{\partial\Omega} \|\partial_n z\|_{\partial\Omega} \leq c_\Omega c_S h^2 \|\partial_r u_h\|_{\partial\Omega} \|\Delta z\|_\Omega \\ &\leq c_\Omega c_S h^2 \{\|\partial_r(u_h - I_h u)\|_{\partial\Omega} + \|\partial_r(I_h u - u)\|_{\partial\Omega} + \|\partial_r u\|_{\partial\Omega}\} \\ &\leq c_\Omega c_S h^2 \{h^{-1/2} \|\nabla(u_h - I_h u)\|_\Omega + c_I h^{1/2} \|u\|_{H^2} + c_\Omega \|u\|_{H^2}\} \\ &\leq c_\Omega c_S h^2 \{h^{-1/2} \|\nabla e_h\|_\Omega + h^{-1/2} \|\nabla(u - I_h u)\|_\Omega + \|\partial_r(I_h u - u)\|_{\partial\Omega} + \|\partial_r u\|_{\partial\Omega}\} \\ &\leq c_\Omega c_S h^2 \{(c_I + c_\Omega) c_S h^{1/2} \|f\|_\Omega + c_I c_\Omega c_S h^{1/2} \|f\|_\Omega + c_\Omega c_S \|f\|_\Omega\} \\ &\leq (c_I + c_\Omega)^2 c_S^2 h^2 \|f\|_\Omega. \end{aligned}$$

Combination der obigen Abschätzungen liefert schließlich

$$\|e_h\|_\Omega \leq c_I c_S h \|\nabla e_h\|_\Omega + c_\Omega c_I c_S h^2 \|f\|_\Omega + (c_I + c_\Omega)^2 c_S^2 h^2 \|f\|_\Omega.$$

Dies zusammen mit der schon bewiesenen Energienormfehlerabschätzung (3.1.48) ergibt die behauptete L^2 -Fehlerabschätzung. Q.E.D.

3.2 Allgemeine Finite-Elemente-Ansätze

Wir wollen nun Finite-Elemente-Ansatzräume allgemeineren Typs konstruieren und Fragen der praktischen Realisierung der Methode diskutieren. Zunächst wird Ω als ein Polygonebiet (Polyeder in 3-D) angenommen. Seien \mathbb{T}_h Zerlegungen von $\overline{\Omega}$ in Dreiecke oder

Vierecke (Tetraeder oder Hexaeder in 3-D), welche den im vorigen Abschnitt formulierten Bedingungen genügen. Für die folgenden Konstruktionen von Finite-Elemente-Ansätzen verwenden wir die Bezeichnungen

$$P_r := \left\{ p(x) = \sum_{0 \leq i+j \leq r} c_{ij} x_1^i x_2^j \right\}, \quad Q_r := \left\{ q(x) = \sum_{0 \leq i,j \leq r} c_{ij} x_1^i x_2^j \right\},$$

für Polynom-Vektorräume im \mathbb{R}^2 (analog für solche im \mathbb{R}^3). Ein Finite-Elemente-Ansatz ist definiert durch Vorgabe eines Polynomraumes $P(T) \subset P_r(T)$ oder $P(T) \subset Q_r(T)$ auf $T \in \mathbb{T}_h$ sowie eines Satzes von „Knotenwerten“ (gegeben durch lineare „Knotenfunktionale“), z. B.:

$$\{v_h(a), v_h(m), \partial_n v_h(m), \nabla v_h(a), (v_h, 1)_\Gamma, (v_h, 1)_T, \dots\},$$

welcher Polynome aus $P(T)$ eindeutig bestimmt.

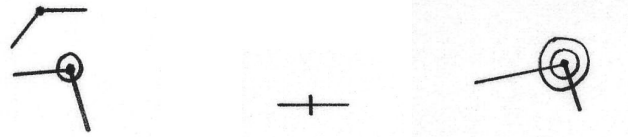


Abbildung 3.8: Funktionswerte $v_h(a)$ sowie $v_h(a), \nabla v_h(a)$ (links), Normalableitung $\partial_n v_h(m)$ in Seitenmitten (Mitte), Funktionswerte $v_h(a), \nabla v_h(a), \nabla^2 v_h(a)$ (rechts).

Wir verwenden die Bezeichnungen

$$\begin{aligned} \mathbb{T}_h &= \{T\} && \text{Zerlegung von } \bar{\Omega}, \\ \partial\mathbb{T}_h &= \{\Gamma\} && \text{Menge aller Kanten (bzw. Flächen),} \\ \partial^2\mathbb{T}_h &= \{a\} && \text{Menge aller Eckpunkte („Knoten“),} \end{aligned}$$

sowie die in Abb. 3.8 skizzierte Symbolik für einige typische Knotenwertevorgaben.

Definition 3.1 (Unisolvenz): Ein Polynomraum $P(T)$ und ein zugehöriger Satz von linearen „Knotenfunktionalen“ $K(T)$ heißen „unisolvent“, wenn jedes $p \in P(T)$ eindeutig durch die Vorgabe von $\chi(p)$ für alle $\chi \in K(T)$ bestimmt ist.

Definition 3.2 (Lagrange- und Hermite-Ansatz): Man spricht bei einem FE-Ansatz $P(T)$ mit zugehörigem Satz von Knotenfunktionalen $K(T) = \{\chi_r, r = 1, \dots, R\}$ von „Lagrange-Elementen“, wenn die Knotenfunktionale nur auf Funktionswerte zurückgreifen; werden auch Ableitungswerte verwendet, spricht man von „Hermite¹⁰-Elementen“.

¹⁰Charles Hermite (1822–1901): Französischer Mathematiker; Prof. an der École Polytechnique und der Sorbonne in Paris; Beiträge zur Zahlentheorie und zur Theorie elliptischer Funktionen; Beweis der Transzendenz von e .

Notwendig für Unisolvenz ist offenbar $\dim P(T) = \#K(T)$ und hinreichend, dass für ein $p \in P(T)$ aus $\chi(p) = 0$ für alle $\chi \in K(T)$ notwendig $p \equiv 0$ folgt. Dies wird in der Regel zum Nachweis von Unisolvenz verwendet.

Definition 3.3 (Interpolation): Für jede Zelle $T \in \mathbb{T}_h$ sei ein Polynomraum $P(T)$ mit Dimension R und ein Satz $K_T = \{\chi_r, r = 1, \dots, R\}$ von Knotenfunktionalen

$$\chi_r : H^m(T) \rightarrow \mathbb{R} \quad (r = 1, \dots, R),$$

spezifiziert, welche „unisolvant“ sind. Durch die Vorgabe

$$\chi_r(I_h v) = \chi_r(v), \quad r = 1, \dots, R,$$

ist dann eindeutig eine „Finite-Elemente-Interpolierende“ $I_h v \in P(T)$ definiert.

Durch Zusammensetzen der zunächst zellweise definierten Formfunktionen $v_T \in P(T)$ erhält man global definierte Funktionen

$$v_h : \bar{\Omega} \rightarrow \mathbb{R}, \quad v_{h|T} := v_T, \quad T \in \mathbb{T}_h,$$

mit denen der Finite-Elemente-Ansatzraum V_h gebildet wird. Durch Gleichsetzen geeigneter Knotenwerte auf dem gemeinsamen Rand $\Gamma = T \cap T'$ jeweils benachbarter Zellen wird Stetigkeit, Differenzierbarkeit und auf analogem Wege auch die Randbedingung $v_{h|_{\partial\Omega}} = 0$ implementiert. Die Dimension des endlich dimensionalen Teilraumes $V_h \subset V$ ist dann gleich der Anzahl der Knotenfunktionalwerte zur eindeutigen Festlegung einer Funktion $v_h \in V_h$.

Definition 3.4 (Konformität): Ein Finite-Elemente-Ansatzraum V_h dieser Art heißt „ H_0^1 -konform“, wenn $V_h \subset H_0^1(\Omega)$, und andernfalls „nicht-konform“.

A) Dreieckselemente im \mathbb{R}^2 : Als erstes betrachten wir Finite-Elemente-Ansätze auf Triangulierungen von Polygonegebieten.

1) *Konstanter Ansatz:* $P(T) = P_0$, $\dim P(T) = 1$.

Mit konstanten Polynomansätzen kann offensichtlich für die Variationsgleichung (3.1.2) keine konforme Diskretisierung gewonnen werden. Diese spielen aber bei anderen Typen von variationellen Formulierungen, den sog. „dual-gemischten“, eine Rolle

2a) *Linearer Ansatz:* $P(T) = P_1$, $\dim P(T) = 3$.

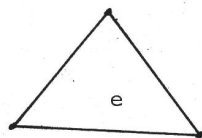


Abbildung 3.9: Knotenpunkte des (konformen) linearen Ansatzes ($e = T$)

Der Polynomraum $P_1(T)$ und der Satz von Knotenfunktionalen $\{\chi_i(p) = p(a_i), i = 1, 2, 3\}$ sind unisolvent, denn für $p \in P_1(T)$ mit $p(a_i) = 0$ gilt notwendig $p|_{\partial T} \equiv 0$ und damit auch $p \equiv 0$. Mit dem Ansatz

$$V_h^{(1)} = \{v_h : \bar{\Omega} \rightarrow \mathbb{R} \mid v_h|_T \in P_1, v_h \text{ stetig in Eckpunkten, } v_h = 0 \text{ in Eckpunkten auf } \partial\Omega\}.$$

erhält man einen H_0^1 -konformen Finite-Elemente-Raum. Denn der Sprung von $[v_h]$ über eine gemeinsame Kante Γ zweier Dreiecke T_1 und T_2 ist in $P_1(\Gamma)$. Folglich ist v_h stetig, da seine Restriktionen auf T_1 und T_2 gemeinsame Werte in den beiden Endpunkten von Γ haben.

2b) Einen *nicht-konformen* linearen Ansatz erhält man durch

$$\tilde{V}_h^{(1)} = \{v_h : \bar{\Omega} \rightarrow \mathbb{R} \mid v_h|_T \in P_1, v_h \text{ stetig in Kantenmitten, } v_h = 0 \text{ in Kantenmitten auf } \partial\Omega\}.$$

Die Unisolvenz folgt analog wie die des entsprechenden konformen Ansatzes. Dieses nicht-konforme „lineare“ Element spielt z. B. eine Rolle bei der Diskretisierung der Navier-Stokes-Gleichungen in der Strömungsmechanik.

3) *Quadratischer Ansatz*: $P(T) = P_2$, $\dim P(T) = 6$.

Der Polynomraum $P_2(T)$ und der Satz von Knotenfunktionalen $\{\chi_i(p) = p(a_i), \psi_i(p) = p(m_i),$

$i = 1, 2, 3\}$ sind unisolvent. Für $p \in P_2(T)$ mit $p(a_i) = p(m_i) = 0$ gilt notwendig $p|_{\partial T} \equiv 0$. Dies impliziert, dass $\nabla p(a_i) = 0$, woraus wegen $\partial_i p \in P_1(T)$ wiederum $\nabla p \equiv 0$ und schließlich $p \equiv 0$ folgt. Mit dem Ansatz

$$V_h^{(2)} = \{v_h : \bar{\Omega} \rightarrow \mathbb{R} \mid v_h|_T \in P_2, v_h \text{ stetig in Eckpunkten und Kantenmitten, } v_h = 0 \text{ in solchen Punkten auf } \partial\Omega\}.$$

erhält man einen H_0^1 -konformen Finite-Elemente-Raum. Denn der Sprung von $[v_h]$ über eine gemeinsame Kante Γ zweier Dreiecke T_1 und T_2 ist in $P_2(\Gamma)$. Folglich ist v_h stetig, da seine Restriktionen auf T_1 und T_2 gemeinsame Werte in den beiden Endpunkten und dem Mittelpunkt von Γ haben. Alternativ zu den Werten in den Kantenmitten m kann man auch die Mittelwerte $|\Gamma|^{-1} \int_{\Gamma} v_h ds$ über Kanten $\Gamma \in \partial T_h$ als Knotenwerte verwenden.

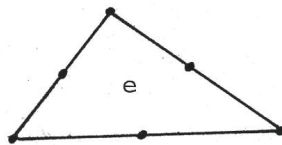


Abbildung 3.10: Knotenpunkte des (konformen) quadratischen Ansatzes ($e = T$)

Der naheliegende *nicht-konforme* Ansatz

$$\tilde{V}_h^{(2)} = \{v_h : \bar{\Omega} \rightarrow \mathbb{R} \mid v_h|_T \in P_2, v_h \text{ stetig in jeweils zwei Gauß-Punkten auf Kanten}, \\ v_h = 0 \text{ in solchen Punkten auf } \partial\Omega\}$$

ist *nicht* unisolvant, denn es existieren nicht-triviale, stückweise quadratische Funktionen, welche in allen Knotenpunkten verschwinden. Man betrachte dazu auf jeder Kante die Legendre-Polynome L_2 zweiten Grades, deren Nullstellen gerade die beiden Gauß-Punkte m_1, m_2 sind. Sie können wegen ihrer Symmetrie zum Kantenmittelpunkt so normiert werden, dass sie in den Eckpunkten a gleiche Werte $L_2(a) = 1$ haben. Dann lässt sich eine Funktion $L \in P_2(T)$ finden, durch konforme Interpolation in Eckpunkten und Seitenmitten: $L(a) = 1, L(m) = L_2(m)$. Auf jeder Kante ist dann $L \equiv L_2$, so dass sich ein Widerspruch zur Unisolvenz des Ansatzes ergibt.

Eine weitere *nicht-konforme* Variante des quadratischen Elements (das sog. „Morley¹¹-Plattenelement“) erhält man bei Wahl der Knotenwerte $\{v_h(a), \partial_n v_h(m)\}$. Dieser Ansatz ist wieder unisolvant. Für $p \in P_2(T)$ folgt aus $p(a_i) = \partial_n p(m_i) = 0, i = 1, 2, 3$, notwendig $\partial_\tau p(m_i) = 0$ und somit $\nabla p(m_i) = 0$. Wegen $\partial_i p \in P_1(T)$ folgt $\nabla p \equiv 0$ und damit auch $p \equiv 0$.

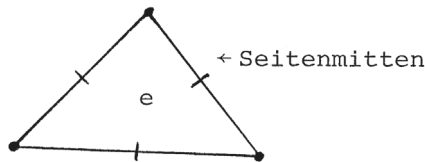


Abbildung 3.11: Knotenpunkte des (nicht-konformen) quadratischen „Morley-Elements“ ($e = T$)

Das Morley-Element ist zwar nicht-konform bzgl. der H^1 - wie auch der H^2 -Norm, trotzdem kann es bei geeigneter Modifikation der variationellen Formulierungen,

$$u_h \in V_h^M : \quad \sum_{T \in \mathbb{T}_h} (\nabla^2 u_h, \nabla^2 \varphi_h)_T = (f, \varphi_h) \quad \forall \varphi_h \in V_h^M,$$

sogar zur Approximation des biharmonischen Operators $\Delta^2 u = f$ verwendet werden.

5) *Kubischer Ansatz* (sog. „kubisches Membran-Element“: $P(T) = P_3$, $\dim P(T) = 10$). Mit dem Ansatz

$$V_h^{(3)} = \{v_h : \bar{\Omega} \rightarrow \mathbb{R} \mid v_h|_T \in P_3, v_h \text{ stetig in Eckpunkten und in je} \\ \text{zwei Gauß-Punkten auf Kanten}, v_h = 0 \text{ in solchen Punkten auf } \partial\mathbb{T}_h\}.$$

¹¹Leslie Sydney Dennis Morley (1924–2011): Englischer Ingenieur; wirkte an der Brunel University in Uxbridge, England; Beiträge zur FEM für nichtlineare Schalenmodelle.

ist H_0^1 -konform. Denn der Sprung von $[v_h]$ über eine gemeinsame Kante Γ zweier Dreiecke T_1 und T_2 ist in $P_3(\Gamma)$. Folglich ist v_h stetig, da seine Restriktionen auf T_1 und T_2 gemeinsame Werte in den beiden Endpunkten und zwei Gauß-Punkten auf Γ haben. Alternativ zu den Gauß-Punkten auf den Kanten kann man auch in jedem Knoten die beiden partiellen Ableitungen, d.h. den Gradienten $\nabla v_h(a)$, als Knotenwerte verwenden. Auch damit erhält man einen H_0^1 -konformen Ansatzraum $\hat{V}_h^{(3)}$; dieser ist offenbar echt kleiner als $V_h^{(3)}$ (Übungsaufgabe).



Abbildung 3.12: Knotenpunkte der (konformen) kubischen Ansätze ($e = T$)

Zur Diskretisierung der biharmonischen Gleichung kann man wieder eine *nicht-konforme* Variante mit den Knotenwerten $\{v_h(a), \partial_n v_h(m_1), \partial_n v_h(m_2), v_h(z)\}$ (m_1, m_2 zwei Gauß-Punkte auf Γ und z der Mittelpunkt von T) verwenden. Ein H^2 -konformes Platten-Element erhält man durch den sog. „Clough-Tocher-Ansatz“ als „zusammengesetztes“ Element ($v_h \in C^1(T)$ stückweise kubisch).

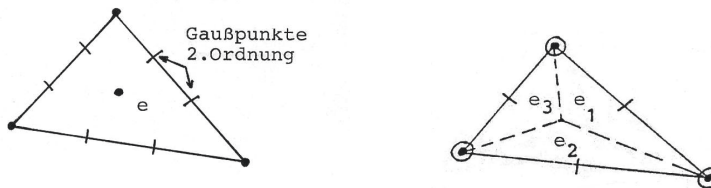


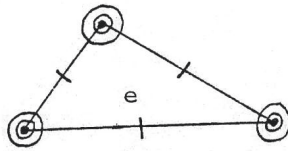
Abbildung 3.13: Knotenpunkte des (nicht-konformen) kubischen „Platten-Elemente“ ($e_i = T_i$)

6) *Quartischer Ansatz*: $P(T) = P_4(T)$, $\dim P(T) = 15$.

Mit dem Satz von Knotenwerten $\{v_h(a), \nabla v_h(a), v_h(m_1), v_h(m_2)\}$ (m_1, m_2 die beiden Gauß-Punkte auf jeder Kante $\Gamma \in \partial\mathbb{T}_h$) erhält man hier einen H^1 -konformen Ansatz (Übungsaufgabe).

7) *Quintischer Ansatz* (sog. „Argyris¹²-Plattenelement“): $P(T) = P_5(T)$, $\dim P(T) = 21$. Mit dem Satz von Knotenwerten $\{v_h(a), \nabla v_h(a), \nabla^2 v_h(a), \partial_n v_h(m)\}$ erhält man hier sogar einen H^2 -konformen Ansatz (Übungsaufgabe). Dieses finite Element ist ein Beispiel für einen konformen Ansatz zur Lösung der biharmonischen Gleichung $\Delta^2 u = f$, für welche stetig differenzierbare Übergänge von Zelle zu Zelle erforderlich sind.

¹²John Hadji Argyris (1913–2004): Griechischer Bauingenieur; Prof. in Stuttgart; einer der „Erfinder“ der Finite-Elemente-Methode.

Abbildung 3.14: Knotenpunkte des (konformen) quintischen „Argyris-Elements“ ($e = T$)

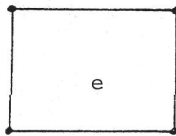
B) Viereckselemente: Als nächstes betrachten wir Finite-Elemente-Ansätze auf (kartesischen) Rechteckszerlegungen.

1) *Bi-linearer Ansatz:* $P(T) = Q_1 = \text{span}\{1, x_1, x_2, x_1x_2\}$, $\dim P(T) = 4$.

Der Polynomraum $Q_1(T)$ und der Satz von Knotenfunktionen $\{\chi_i(p) = p(a_i), i = 1, \dots, 4\}$ sind unisolvent. Ein $p \in Q_1(T)$ ist entlang der Kanten von T linear. Aus $p(a_i) = 0$ folgt also $p|_{\partial T} \equiv 0$, und weiter $\nabla p(a_i) = 0$. Wegen $\partial_i p \in P_1(T)$ impliziert dies $\nabla p \equiv 0$ und schließlich $p \equiv 0$. Der Ansatz

$$V_h^{(1)} = \{v_h : \bar{\Omega} \rightarrow \mathbb{R} \mid v_h|_T \in Q_1, v_h \text{ stetig in Eckpunkten}, v_h = 0 \text{ in Eckpunkten auf } \partial\Omega\}$$

ist H_0^1 -konform, da die Sprünge von v_h entlang von Kanten linear sind.

Abbildung 3.15: Knotenpunkte des (konformen) bi-linearen Ansatzes ($e = T$)

Der naheliegende *nicht-konforme* Ansatz

$$\tilde{V}_h^{(1)} = \{v_h : \bar{\Omega} \rightarrow \mathbb{R} \mid v_h|_T \in Q_1, v_h \text{ stetig in Kantenmitten}, v_h = 0 \text{ in diesen auf } \partial\Omega\}$$

ist aber i. Allg. *nicht* unisolvent, da z. B. die Funktion $v_h(x_1, x_2) = x_1x_2$ in den Kantenmitten des Quadrats $T_1 = [-1, 1] \times [-1, 1]$ verschwindet. Auf T_1 erhält man aber durch $P(T) = \text{span}\{1, x_1, x_2, x_1^2 - x_2^2\}$ einen mit den Kantenmitten als Knotenfunktionen unisolventen Ansatz. Alternativ zu den Funktionswerten in den Seitenmitten kann man auch die Mittelwerte $|\Gamma|^{-1} \int_{\Gamma} v_h ds$ über die Kanten als Knotenwerte verwenden. Dies ergibt aber einen von $\tilde{V}_h^{(1)}$ verschiedenen Ansatzraum. Beide Ansätze sind offensichtlich *nicht-konform*.

2) *Bi-quadratischer Ansatz:* $P(T) = Q_2 = \text{span}\{1, x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^2x_2, x_1x_2^2, x_1^2x_2^2\}$, $\dim P(T) = 9$. Die Konstruktion eines „zulässigen“ Satzes von Knotenwerten ist Übung.

3) *Reduzierter bi-quadratischer Ansatz* (sog. „Wilson¹³-Membran-Element“):

$P(T) = P_2(T) \oplus \text{span}\{x_1^2x_2, x_1x_2^2\}$, $\dim P(T) = 8$. Dieser Ansatz wird mit den Knotenwerten $\{v_h(a), v_h(m)\}$ H^1 -konform.

4) *Bi-kubischer Ansatz*: $P(T) = Q_3 = \text{span}\{1, x_1, x_2, x_1^2, \dots, x_1^3x_2^3\}$, $\dim P(T) = 16$.

Die Konstruktion eines „zulässigen“ Satzes von Knotenwerten wird als Übung gestellt.

5) *Reduzierter bi-kubischer Ansatz* (sog. „Adini¹⁴-Platten-Element“):

$P(T) = P_3(T) \oplus \text{span}\{x_1^3x_2, x_1x_2^3\}$, $\dim P(T) = 12$. Dieser Ansatz wird mit den Knotenwerten $\{v_h(a), \nabla v_h(a)\}$ unisolvent und H^1 -konform, ist aber nicht H^2 -konform.

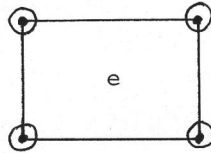


Abbildung 3.16: Knotenpunkte des (nicht-konformen) „Adini-Plattenelements“ ($e = T$)

Viele der aufgeführten zwei-dimensionalen Finite-Elemente-Ansätze haben natürliche Erweiterungen auf drei Dimensionen. Die gebräuchlichsten Beispiele sind:

1) *Lineares Tetraederelement*: $P(T) = \text{span}\{1, x_1, x_2, x_3\}$, $\dim P(T) = 4$. Mit den Funktionswerten in den Eckpunkten als Knotenwerte ist dieser Ansatz unisolvent und H^1 -konform.

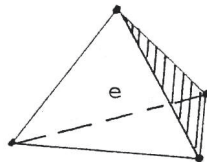


Abbildung 3.17: Knotenpunkte des (konformen) linearen Ansatzes ($e = T$)

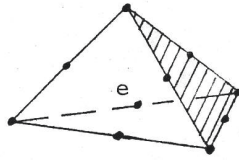
2) *Nicht-konf., lineares Tetraederelement*: $P(T) = \text{span}\{1, x_1, x_2, x_3\}$, $\dim P(T) = 4$.

Mit den Funktionswerten in den Flächenmitten als Knotenwerte ist dieser Ansatz unisolvent, aber *nicht-konform*.

3) *Quadratisches Tetraederelement*: $P(T) = \text{span}\{1, x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1^2, x_2^2, x_3^2\}$, $\dim P(T) = 10$. Mit den Funktionswerten in den Ecken und den Kantenmitten als Knotenwerte ist dieser Ansatz unisolvent und H^1 -konform.

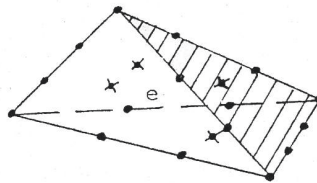
¹³Edward L. Wilson (1931–): US-Amerikanischer Ingenieur; Prof. für Ingenieurwissenschaften an der Univ. of California (Berkeley, USA); ein früherer Pionier der (praktischen) Finite-Elemente-Methode; Koautor des Buches „Numerical Methods in Finite Element Analysis“ (zus. mit K. J. Bathe), 1976.

¹⁴Avner Adini (1911–): Promotion 1961 an der Univ. of California (Berkeley, USA) als Bauingenieur; Beiträge u. a. zur Finite-Elemente-Methode in der Plattenstatik.

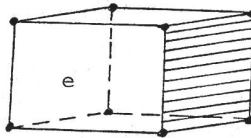
Abbildung 3.18: Knotenpunkte des (konformen) quadratischen Ansatzes ($e = T$)

4) *Kubisches Tetraederelement*: $P(T) = P_3$, $\dim P(T) = 20$.

Mit den Funktionswerten in den Ecken, in zwei Gauß-Punkten auf den Kanten und in den Seitenmitten als Knotenwerte ist dieser Ansatz unisolvent und H^1 -konform.

Abbildung 3.19: Knotenpunkte des (konformen) Kubisches Tetraederelement ($e = T$)

5) *Tri-lineares Quaderelement*: $P(T) = \text{span}\{1, x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1x_2x_3\}$, $\dim P(T) = 8$. Mit den Funktionswerten in den Eckpunkten als Knotenwerte ist dieser Ansatz unisolvent und H^1 -konform.

Abbildung 3.20: Knotenpunkte des (konformen) tri-linearen Quaderelements ($e = T$)

Ein zugehöriges *nicht-konformes* Quaderelement erhält man durch den Ansatz $P(T) = \text{span}\{1, x_1, x_2, x_3, x_1^2 - x_2^2, x_1^2 - x_3^2\}$, $\dim P(T) = 6$. Mit den Funktionswerten in den Flächenmitten als Knotenwerte ist dieser Ansatz unisolvent.

6) *Tri-quadratisches Quaderelement*: $P(T) = \text{span}\{1, x_1, x_2, x_3, \dots, x_1^2x_2^2x_3^2\}$, $\dim P(T) = 27$. Mit den Funktionswerten in den Eckpunkten, den Kantenmitten, den Seitenmitten und dem Mittelpunkt als Knotenwerte ist dieser Ansatz unisolvent und H^1 -konform.

In allgemeinen Situationen können die Zellen bzgl. des Koordinatensystems gedreht oder gezerzt werden. Daraus folgt, dass es Fälle gibt, in denen man für zwei Zellen der

Zerlegung \mathbb{T}_h nicht denselben Ansatz nehmen kann (z. B. das bi-lineare Viereckselement). Deshalb müssen wir uns bei der Definition von Finite-Elemente-Ansätzen von dem festen Koordinatensystem befreien. Dies induziert die Idee des „Referenzelements“. Wir verwenden als Referenzelement ein natürliches Einheitsselement \hat{T} (Einheitsdreieck, Einheitsviereck, ...) und definieren zunächst einen Polynomansatz $\hat{P}(\hat{T})$ auf diesem Referenzelement.

Sei σ_T eine (polynomiale) Transformation des Referenzelements auf das („physikalische“) Element mit der Inversen $\sigma_T^{-1} : T \rightarrow \hat{T}$. Der Ansatz auf der Zelle T ist dann gegeben durch

$$P(T) := \{v_h : T \rightarrow \mathbb{R} \mid v_h(\sigma_T(\cdot)) \in \hat{P}(\hat{T})\}. \quad (3.2.54)$$

Der Funktionenraum $P(T)$ ist nicht notwendig ein Raum von Polynomen, auch wenn $\hat{P}(\hat{T})$ ein solcher ist. Dies liegt daran, dass i. Allg. die inverse Abbildung σ_T^{-1} und damit die Funktion $v_h(x) = v_h(\sigma(\sigma^{-1}(\cdot)))$ nicht polynomial ist, z. B. im Fall (echt) bilinearer Abbildungen σ_T . Wenn man T aus \hat{T} durch eine Verschiebung, eine Rotation, eine Scherung und eine Skalierung gewinnen kann, so ist σ_T eine affin-lineare Transformation:

$$\sigma_T(\hat{x}) = B_T \hat{x} + b_T$$

mit einer Matrix $B_T \in \mathbb{R}^{d \times d}$ und einem Verschiebungsvektor $b_T \in \mathbb{R}^d$. Dies ist möglich bei Dreiecken (in 2-D) bzw. Tetraedern (in 3-D) sowie bei Parallelogrammen (in 2-D) bzw. Parallelepipeden (in 3-D). Für allgemeine (konvexe) Vierecke (in 2-D) oder Hexaeder (in 3-D) benötigt man für die Transformation echt bi- bzw. tri-lineare Abbildungen.

Wir gehen nun zum allgemeinen d -dimensionalen Fall über und betrachten Zerlegungen $\mathbb{T}_h = \{T\}$ des Abschlusses eines Gebiets $\Omega \subset \mathbb{R}^d$ in d -Simplizes. Dabei ist ein (nicht degeneriertes) Simplex $T \subset \mathbb{R}^d$ die konvexe lineare Hülle von $d + 1$ *linear unabhängigen* Punkten $a^i \in \mathbb{R}^d$, $i = 0, \dots, d$:

$$T = \left\{ x \in \mathbb{R}^d \mid x = \sum_{i=0}^d \lambda_i a^i, \sum_{i=0}^d \lambda_i = 1, 0 \leq \lambda_i \leq 1 \right\}. \quad (3.2.55)$$

Das System $\{a^i, i = 0, \dots, d\}$ heißt linear unabhängig, wenn die erzeugenden Vektoren $\{w^i = a^i - a^0, i = 1, \dots, n\}$ eine Basis des \mathbb{R}^d bilden. Die einfachsten Beispiele sind wieder Dreiecke für $d = 2$ und Tetraeder für $d = 3$. Sei \hat{T} das Einheitssimplex im \mathbb{R}^d , welches von den Punkten $e^0 := 0$, $e^i = (\delta_{1i}, \dots, \delta_{di})^T$, $i = 1, \dots, d$, aufgespannt wird.

Hilfssatz 3.2 (Referenztransformation): *Jedes (nicht degenerierte) Simplex $T \subset \mathbb{R}^d$ lässt sich mittels einer umkehrbaren affinen Abbildung*

$$x = \sigma_T(\hat{x}) := B_T \hat{x} + b_T, \quad B_T \in \mathbb{R}^{d \times d}, \quad b_T \in \mathbb{R}^d, \quad (3.2.56)$$

aus dem Einheitssimplex \hat{T} gewinnen: $T = B_T \hat{T} + b_T$. Dabei ist die Umkehrabbildung gegeben durch $\hat{x} = \sigma_T^{-1}(x) = B_T^{-1}x - B_T^{-1}b_T$.

Beweis: Wir lassen im Folgenden den Zusatz T weg. Sei $A \in \mathbb{R}^{d \times d}$ die reguläre Matrix, welche die Basis $\{w^i = a^i - a^0, i = 1, \dots, d\}$ auf die kartesische Einheitsbasis $\{e^i, i = 1, \dots, d\}$ abbildet: $e^i = Aw^i, i = 1, \dots, d$. Man gewinnt ihre Elemente $a_{\nu\mu}$ als Lösungen der Gleichungssysteme

$$\sum_{\mu=1}^n a_{\nu\mu} w_{\mu}^i = e_{\nu}^i, \quad i = 1, \dots, d,$$

für $\nu = 1, \dots, d$. Die Koeffizientenmatrix $(w_{\mu}^i)_{i,\mu=1}^d$ enthält die linear unabhängigen Vektoren $w^i, i = 1, \dots, d$, als Zeilenvektoren und ist folglich regulär. Die affine Abbildung $\hat{x} = Ax - Aa^0$ ist dann umkehrbar und bildet das Simplex T auf das Einheits-simplex \hat{T} ab, denn für $x = \sum_{i=0}^d \lambda_i a^i \in T$ ist (Man beachte $\sum_{i=1}^d \lambda_i = 1$.)

$$Ax - Aa^0 = \sum_{i=0}^d \lambda_i A(a^i - a^0) = \sum_{i=0}^d \lambda_i e^i \in \hat{T},$$

und umgekehrt für $\hat{x} = \sum_{i=0}^d \lambda_i e^i \in \hat{T}$

$$A^{-1}\hat{x} + a^0 = \sum_{i=0}^d \lambda_i A^{-1}e^i + a^0 = \sum_{i=0}^d \lambda_i w^i + a^0 = \sum_{i=0}^d \lambda_i a^i + (1 - \sum_{i=1}^d \lambda_i)a^0 \in T.$$

Dies komplettiert den Beweis mit $B_T := A^{-1}$ und $b_T := a_0$.

Q.E.D.

Bemerkung 3.5: Zu Hilfssatz 3.2 gibt es ein Analogon für Vierecks-Zerlegungen im \mathbb{R}^2 sowie Hexaeder-Zerlegungen im \mathbb{R}^3 . Zu jedem konvexen Viereck $T \in \mathbb{R}^2$ oder Hexaeder $T \in \mathbb{R}^3$ („6-Flächner“) existieren bi- bzw. tri-lineare Abbildungen $\sigma_T : \hat{T} \rightarrow T$ des Einheitsquadrats bzw. Einheitswürfels \hat{T} auf T . Dabei werden die Eckpunkte von \hat{T} auf die Eckpunkte von T sowie die Kanten bzw. Seitenflächen von \hat{T} auf die Kanten bzw. Seitenflächen von T abgebildet (jeweils in derselben Orientierung).

Die Erzeugung von Finite-Elemente-Ansätzen über Transformation von einem Referenzelement hat auch den Zweck, auf allgemeinen Vierecks- oder Hexaeder-Zerlegungen *konforme* Ansätze zu gewinnen. Wir wollen das anhand des bi-linearen Ansatzes diskutieren:

$$V_h^{(1)} = \{v_h : \bar{\Omega} \rightarrow \mathbb{R} \mid v_h|_T \in Q_1, v_h \text{ stetig in Knoten, } v_h = 0 \text{ in Knoten auf } \partial\Omega\}.$$

Wir betrachten allgemeine Vierecke T und T' , die eine gemeinsame Kante Γ haben, und setzen $P(T) = P_1 \oplus \text{span}\{x_1 x_2\}$. Der Sprung von v_h ist zwar gleich Null in den Endpunkten von Γ , doch ist er i. Allg. nicht linear auf Γ , so dass die Stetigkeit über Γ nicht gesichert ist. Hierfür muss erreicht werden, dass die Restriktion von $v_h \in V_h^{(1)}$ auf alle Kanten $\Gamma \in \partial\mathbb{T}_h$ *linear* ist. Dies kann durch Konstruktion von $V_h^{(1)}$ mit Hilfe der bilinearen Transformationen $\sigma_T : \hat{T} \rightarrow T$ erreicht werden.

Für eine bi-lineare Transformation $\sigma_T : \hat{T} \rightarrow T$ ist die Inverse $\sigma_T^{-1} : T \rightarrow \hat{T}$ i. Allg. nicht bi-linear. In diesem Fall ist der gemäß (3.2.54) erzeugte lokale Ansatzraum $P(T)$

auch kein Polynomraum. Dennoch ist $v|_\Gamma$ auf jeder Kante $\Gamma \in \partial\mathbb{T}_h$ linear, so dass Stetigkeit in allen Eckpunkten auch automatisch globale Stetigkeit auf $\overline{\Omega}$ sowie $v|_{\partial\Omega} = 0$ garantiert. Dieser Transformationsansatz löst also das Problem der globalen Stetigkeit.

Definition 3.5 (Parametrischer Ansatz): Der Ansatz $P(T)$ wird „parametrisch“ genannt, wenn er durch Transformationen $\sigma_T : \hat{T} \rightarrow T$ von einem Referenzelement \hat{T} erzeugt wird. Er heißt „isoparametrisch“, wenn die Transformation σ_T vom selben Polynomtyp wie die Ansatzfunktionen in $\hat{P}(\hat{T})$ ist.

Der Begriff des „isoparametrischen“ Finite-Element-Ansatzes lässt sich auf höheren Polynomgrad $r \geq 2$ übertragen. Dies wird wichtig bei der Approximation eines krummen Randes bei Verwendung von Ansätzen höherer Ordnung. (s. Abb. 3.21). Die einfache Polygonzugapproximation würde hier zu einer Ordnungsreduktion führen:

$$\|\nabla(u - u_h)\| \leq c\{h^r \|u\|_r + h \|u\|_2\},$$

im Gegensatz zur optimalen Abschätzung

$$\|\nabla(u - u_h)\| \leq ch^r \|u\|_r.$$

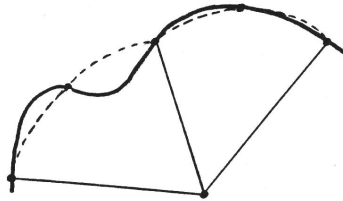


Abbildung 3.21: Parametrische Randapproximation

3.3 Interpolation mit finiten Elementen

Dieser Abschnitt ist dem grundlegenden Aspekt bei der mathematischen Analyse der Methode der finiten Elemente gewidmet: Wie gut lassen sich hinreichend glatte Funktionen durch stückweise polynomiale approximieren? Wir gehen von der im vorigen Abschnitt anhand von Beispielen beschriebenen abstrakten Situation aus. Sei $T \subset \mathbb{R}^d$ eine Zelle eines Finite-Elemente-Gitters \mathbb{T}_h ; der Durchmesser von T wird wieder mit $\text{diam}(T) = h_T$ und der Radius einer (maximalen) einbeschriebenen Kugel mit ρ_T bezeichnet.

Wir betrachten im Folgenden ausschließlich „parametrische“ Finite-Elemente-Ansätze. Jedes $T \in \mathbb{T}_h$ sei Bild eines „Referenz-Einheitslements“ $\hat{T} \subset \mathbb{R}^d$ mit Durchmesser $\text{diam}(\hat{T}) = \hat{h} \approx 1$ und Inkugelradius $\hat{\rho} > 0$. Die zugehörigen Abbildungen $\sigma_T : \hat{T} \rightarrow T$

seien der Einfachheit halber als affin-linear angenommen:

$$x = \sigma_T(\hat{x}) = B_T \hat{x} + b_T, \quad B_T \in R^{d \times d}, \quad b_T \in R^d. \quad (3.3.57)$$

Der Fall allgemeiner Vierecke mit erzeugenden *bi-linearen* Transformationen wird gegebenenfalls in Bemerkungen berücksichtigt werden.

Allgemeine Interpolationsaufgabe: Auf einem beliebigen, aber festen Element T (z. B. dem Einheitsselement \hat{T}) seien ein Vektorraum $P(T)$ von Polynomen über T mit $\dim P(T) = R$ sowie ein System von linearen „Knotenfunktionalen“ $K_T = \{\chi_r, r = 1, \dots, R\}$ gegeben, so dass die folgenden Bedingungen erfüllt sind:

i) Der Ansatz ist unisolvent:

$$q \in P(T) : \chi_r(q) = 0 \quad (r = 1, \dots, R) \quad \Rightarrow \quad q = 0. \quad (3.3.58)$$

ii) Für ein $m \geq 1$ gilt $P_{m-1} \subset P(T)$.

iii) Die Knotenfunktionale aus K_T sind stetig auf $H^m(T)$:

$$|\chi_r(v)| \leq c_b \|v\|_{m;T}, \quad v \in H^m(T), \quad r = 1, \dots, R. \quad (3.3.59)$$

Unter der Bedingung (i) ist die zugehörige Lagrangesche bzw. Hermitesche Interpolationsaufgabe eindeutig lösbar, d. h.: Zu jeder Funktion $v \in H^m(T)$ existiert ein eindeutig bestimmtes „Interpolationspolynom“ $I_T v \in P(T)$ mit den Eigenschaften

$$\chi_r(I_T v) = \chi_r(v), \quad r = 1, \dots, R.$$

Wenn die Knotenfunktionale zu „singular“ sind, um für Funktionen aus $H^m(\Omega)$ definiert zu sein (z. B. die Ableitung $\partial_n^{m-1} v(m)$ in einer Seitenmitte $m \in \partial T$), kann auch ein stärkerer Sobolew-Raum $H^{m,p}(\Omega)$ mit $p > d$ verwendet werden. Wir werden diesen Fall hier aber nicht weiter verfolgen.

Notation: Im folgenden verwenden wir eine gebräuchliche „Multiindex“-Schreibweise für mehrfach indizierte Größen. Für einen Indexvektor $\alpha = (\alpha_1, \dots, \alpha_d)^T \in \mathbb{N}_0^d$ mit ganzzahligen, nichtnegativen Komponenten setzen wir

$$|\alpha| := \sum_{i=1}^d \alpha_i, \quad x^\alpha := \prod_{i=1}^d x_i^{\alpha_i}, \quad D^\alpha := \prod_{i=1}^d \partial_i^{\alpha_i}, \quad P_k = \left\{ q(x) = \sum_{|\alpha| \leq k} a_\alpha x^\alpha \right\}.$$

Mit dieser Notation schreiben sich z. B. die Sobolew-Normen bzw. -Halbnormen über T in der Form

$$\|v\|_{m;T} = \left(\sum_{0 \leq |\alpha| \leq m} \|D^\alpha v\|_T^2 \right)^{1/2}, \quad |v|_{m;T} = \left(\sum_{|\alpha|=m} \|D^\alpha v\|_T^2 \right)^{1/2}.$$

Wir leiten nun eine Reihe von technischen Hilfssätzen ab, die am Schluss zu den

gewünschten allgemeinen Abschätzungen für den Interpolationsfehler bei Finite-Elemente-Ansätze führen werden. Die dabei verwendete Schlussweise geht in diesem Zusammenhang auf Bramble¹⁵ und Hilbert¹⁶ (1971) zurück, weswegen die ganze Theorie auch „Bramble-Hilbert-Theorie“ und das Hauptresultat „Bramble-Hilbert-Lemma“ genannt werden.

Hilfssatz 3.3 (Nullraum von Ableitungsoperatoren): *Jede Funktion $v \in H^m(T)$ mit der Eigenschaft*

$$D^\alpha v = 0, \quad |\alpha| = m, \quad (3.3.60)$$

ist fast überall gleich einem Polynom aus $P_{m-1}(T)$.

Beweis: Aus den Voraussetzungen folgt $D^\beta D^\alpha v \equiv 0$ für beliebiges β und somit $v \in \bigcap_{k=1}^{\infty} H^k(T)$. Nach dem Sobolewschen Einbettungssatz ist damit $v \in C^m(T)$, so dass sich die Behauptung mit Hilfe „klassischer“ Argumente ergibt. Q.E.D.

Hilfssatz 3.4 (Polynomprojektion): *Zu jeder Funktion $v \in H^m(T)$ existiert ein eindeutig bestimmtes Polynom $q \in P_{m-1}(T)$ mit der Eigenschaft*

$$\int_T D^\alpha (v - q) dx = 0, \quad 0 \leq |\alpha| \leq m - 1. \quad (3.3.61)$$

Beweis: Zur Lösung der Aufgabe machen wir den Ansatz

$$q(x) := \sum_{|\beta| \leq m-1} \xi^\beta x^\beta \in P_{m-1}(T)$$

mit unbekanntem Koeffizienten $\xi = (\xi^\beta)_{|\beta| \leq m-1}$ (bei lexikographischer Anordnung der Indexkomponenten). Dies führt auf das quadratische, lineare Gleichungssystem

$$\sum_{0 \leq |\beta| \leq m-1} \xi^\beta \int_T D^\alpha x^\beta dx = \int_T D^\alpha v dx, \quad 0 \leq |\alpha| \leq m - 1.$$

Dessen Koeffizientenmatrix

$$M = \left(\int_T D^\alpha x^\beta dx \right)_{0 \leq |\alpha|, |\beta| \leq m-1}$$

ist regulär. Andernfalls gäbe es ein $\xi = (\xi^\beta)_{|\beta| \leq m-1} \neq 0$ mit $M\xi = 0$. Das damit gebildete Polynom $q(x) = \sum_{0 \leq |\beta| \leq m-1} \xi^\beta x^\beta \in P_{m-1}(T)$ hätte dann die Eigenschaft

$$\int_T D^\alpha q dx = 0, \quad 0 \leq |\alpha| \leq m - 1, \quad (3.3.62)$$

¹⁵James H. Bramble (1932–): US-Amerikanischer Mathematiker: Prof. an der Cornell University und der Texas A&M University; fundamentale Beiträge zur Theorie der Finite-Elemente-Methode und von Iterationsverfahren, insbesondere Mehrgitterverfahrens.

¹⁶Stephen R. Hilbert (1907–): US-amerikanischer Mathematiker: Prof. am Ithaca College, New York; Student von J. Bramble; bekannt durch das sog. „Bramble-Hilbert-Lemma“ (1970).

woraus offensichtlich $q \equiv 0$ und damit der Widerspruch $\xi \equiv 0$ folgte. Also existiert ein eindeutig bestimmtes Polynom mit den verlangten Eigenschaften. Q.E.D.

Hilfssatz 3.5 (Verallg. Poincarésche Ungleichung): Für jede Funktion $v \in H^m(T)$ mit der Eigenschaft

$$\int_T D^\alpha v \, dx = 0, \quad 0 \leq |\alpha| \leq m-1, \quad (3.3.63)$$

gilt mit einer Konstante $c_0 = c(d, m, T)$

$$\|v\|_{m;T} \leq c_0 |v|_{m;T}. \quad (3.3.64)$$

Beweis: Angenommen, die Behauptung ist falsch. Dann existiert eine Folge von Funktionen $v_k \in H^m(T)$, $k \in \mathbb{N}$ mit den Eigenschaften

$$1 = \|v_k\|_{m;T} \geq k |v_k|_{m;T}, \quad k \in \mathbb{N}. \quad (3.3.65)$$

Aufgrund der Kompaktheit der Einbettung von $H^m(T)$ in $H^{m-1}(T)$ konvergiert eine Teilfolge, welche wir wieder mit $(v_k)_{k \in \mathbb{N}}$ bezeichnen, in $H^{m-1}(T)$ gegen ein $v \in H^{m-1}(T)$:

$$\|v_k - v\|_{m-1;T} \rightarrow 0 \quad (k \rightarrow \infty). \quad (3.3.66)$$

Mit der Annahme folgt $|v_k|_{m;T} \rightarrow 0$ ($k \rightarrow \infty$). Also ist $(v_k)_{k \in \mathbb{N}}$ Cauchy-Folge in $H^m(\Omega)$ mit Limes $\tilde{v} \in H^m(T)$. Wegen $v_k \rightarrow_{H^{m-1}} v$ muss $\tilde{v} = v$ sein. Damit folgt $|v|_{m;T} = 0$. Nach Hilfssatz 3.3 ist also $v \in P_{m-1}(T)$ und besitzt die Eigenschaft

$$\int_T D^\alpha v \, dx = \lim_{k \rightarrow \infty} \int_T D^\alpha v_k \, dx = 0, \quad 0 \leq |\alpha| \leq m-1. \quad (3.3.67)$$

Dies bedeutet aber wegen Hilfssatz 3.4 notwendig $v \equiv 0$, was im Widerspruch zur Annahme $\|v\|_{m;T} = \lim_{k \rightarrow \infty} \|v_k\|_{m;T} = 1$ steht. Q.E.D.

Nach diesen Vorbereitungen können wir das zentrale Resultat dieses Abschnitts, das sog. „Bramble-Lemma“, beweisen.

Satz 3.5 (Bramble-Hilbert-Lemma): Sei $F(\cdot) : H^m(T) \rightarrow \mathbb{R}$ ein beschränktes, sublineares Funktional, welches auf $P_{m-1}(T)$ verschwindet, d. h.:

- i) $|F(v)| \leq c_1 \|v\|_{m;T}$ (Beschränktheit),
- ii) $|F(u+v)| \leq c_2 \{|F(u)| + |F(v)|\}$ (Sublinearität),
- iii) $F(q) = 0, \quad q \in P_{m-1}(T)$ (Annulierungseigenschaft).

Dann gilt mit der Konstante c_0 aus Hilfssatz 3.5:

$$|F(v)| \leq c_0 c_1 c_2 |v|_{m;T}, \quad v \in H^m(T). \quad (3.3.68)$$

Beweis: Für ein $v \in H^m(T)$ gilt mit beliebigem $q \in P_{m-1}(T)$:

$$|F(v)| = |F(v - q + q)| \leq c_2\{|F(v - q)| + |F(q)|\} \leq c_1c_2\|v - q\|_{m;T}.$$

Wir wählen nun $q \in P_{m-1}(T)$ als das gemäß Hilfssatz 3.4 zu v gehörende Polynom, so dass gemäß Hilfssatz 3.5 folgt:

$$\|v - q\|_{m;T} \leq c_0 |v - q|_{m;T} = c_0 |v|_{m;T}.$$

Dies impliziert dann

$$|F(v)| \leq c_3 |v|_{m;T},$$

mit $c_3 := c_0c_1c_2$, was zu beweisen war.

Q.E.D.

Korollar 3.1 (Allgemeiner Interpolationssatz): Seien die obigen Voraussetzungen erfüllt. Für jede Funktion $v \in H^m(T)$ und das zugehörige interpolierende Polynom $I_T v \in P(T)$ gilt mit einer beliebigen beschränkten Halbnorm $|\cdot|$ auf $H^m(T)$:

$$|v - I_T v| \leq c |v|_{m;T} \tag{3.3.69}$$

mit einer Konstante $c = c(d, m, R, T, |\cdot|)$.

Beweis: O.B.d.A. sei $|v| \leq \|v\|_{m;T}$, $v \in H^m(T)$. Durch $F(v) := |v - I_T v|$ wird auf $H^m(T)$ ein sublineares Funktional definiert. Die Interpolierende $I_T v$ besitzt die Darstellung

$$I_T v = \sum_{r=1}^R \chi_r(v) \varphi^{(r)}$$

mit der durch die Bedingung $\chi_r(\varphi^{(s)}) = \delta_{rs}$, ($r, s = 1, \dots, R$) eindeutig bestimmten verallgemeinerten Lagrange-Basis $\{\varphi^{(r)}, r = 1, \dots, R\}$ des Polynomraums $P(T)$. Wegen der Beschränktheit der Knotenfunktionale χ_r folgt

$$|F(v)| \leq |v| + |I_T v| \leq |v| + \sum_{r=1}^R |\chi_r(v)| |\varphi^{(r)}| \leq (1 + Rc_b \max_{r=1, \dots, R} |\varphi^{(r)}|) \|v\|_{m;T},$$

und damit die Beschränktheit von $F(\cdot)$. Wegen $I_T q = q$ für $q \in P(T)$ gilt weiter

$$F(q) = 0, \quad q \in P_{m-1}(T).$$

Aus Satz 3.5 folgt damit die behauptete Abschätzung.

Q.E.D.

Beispiele von Halbnormen, für die das obige Resultat angewendet wird, sind etwa:

1. L^2 -Norm über T :

$$|v - I_T v| := \left(\int_T |v - I_T v|^2 dx \right)^{1/2}.$$

2. L^2 -Norm über den Rand ∂T :

$$|v - I_T v| := \left(\int_{\partial T} |v - I_T v|^2 dx \right)^{1/2}.$$

3. Mittelwert über eine Kante $\Gamma \subset \partial T$:

$$|v - I_T v| := \left| \int_{\Gamma} (v - I_T v) dx \right|.$$

4. Maximum-Norm über T :

$$|v - I_T v| := \max_T |v - I_T v|.$$

5. Wert in einem Punkt $P \in \Omega$:

$$|v - I_T v| := |(v - I_T v)(P)|.$$

Als nächstes greifen wir nun unser eigentliches Problem an, nämlich den Interpolationsfehler auf den einzelnen Zellen T der Zerlegung \mathbb{T}_h abzuschätzen. Wir tun dies für den repräsentativen Spezialfall der klassischen Lagrange/Hermite-Interpolation, bei der die Knotenfunktionale χ_r als Punktfunktionale für Funktionswerte sowie Ableitungswerte in gewissen Knotenpunkten $\hat{a}_r \in \hat{T}$ ($r=1, \dots, R$) gegeben sind.

Sei \hat{T} das Referenzelement der Größe $\hat{h} := \text{diam}(\hat{T}) = 1$ und Inkreisradius $\hat{\rho} > 0$. Für eine einzelne Zelle $T \in \mathbb{T}_h$ bezeichnen wir mit $a_r = B\hat{a}_r + b$ die aus den Stützstellen $\hat{a}_r \in \hat{T}$ durch Anwendung der Transformation $\sigma(\hat{x}) = B\hat{x} + b$ erzeugten Punkte $a_r \in T$. Entsprechend seien $h_T := \text{diam}(T)$ sowie $\rho_T > 0$ der Inkreisradius von T . Die Umkehrabbildung $\sigma^{-1} : T \rightarrow \hat{T}$ hat die Darstellung $\sigma^{-1}(x) = B^{-1}x - B^{-1}b$ mit der inversen Matrix $B^{-1} = (b_{ij}^{(-1)})_{i,j=1}^d$. Unter Verwendung dieser Abbildung $x = \sigma(\hat{x})$ werden für Funktionen $v : T \rightarrow \mathbb{R}$ und $\hat{w} : \hat{T} \rightarrow \mathbb{R}$ zugehörige Funktionen $\hat{v} : \hat{T} \rightarrow \mathbb{R}$ und $w : T \rightarrow \mathbb{R}$ definiert durch

$$\hat{v}(\hat{x}) := v(x), \quad w(x) := \hat{w}(\hat{x}). \quad (3.3.70)$$

Entsprechend lassen sich die partiellen Ableitungen nach \hat{x} und x durch die jeweils anderen ausdrücken:

$$\hat{\partial}_i := \sum_{j=1}^d b_{ij} \partial_j, \quad \partial_i := \sum_{j=1}^d b_{ij}^{(-1)} \hat{\partial}_j.$$

Wir nehmen an, dass der Polynomansatz $\hat{P}(\hat{T})$ unisolvent ist mit einem Satz von Knotenfunktionalen der Form $\hat{D}_r \hat{v}(\hat{a}_r)$, $r = 1, \dots, R$, wobei die Punkte \hat{a}_r auch mehrfach auftreten können und die Ableitungsoperatoren die Gestalt $\hat{D}_r = \hat{\partial}^{\alpha^r}$ mit geeigneten Multiindizes $\alpha^r = (\alpha_i^r, \dots, \alpha_d^r)$ haben. Der Ansatzraum $\hat{P}(\hat{T})$ habe die Lagrange-Basis $\{\hat{\varphi}_r, r = 1, \dots, R\}$, d.h.:

$$\hat{D}_r \hat{\varphi}_s(\hat{a}_r) = \delta_{rs}.$$

Wir nehmen weiter an, dass alle auftretenden Ableitungen \hat{D}_r auf $H^m(\hat{T})$ wohl definiert und stetig sind:

$$|\hat{D}_r \hat{v}(\hat{a}_r)| \leq c \|\hat{v}\|_{m;\hat{T}}, \quad \hat{v} \in H^m(\hat{T}).$$

Die lokalen Polynomräume $P(T)$ werden wieder erzeugt via Koordinatentransformation aus dem Ansatzraum $\hat{P}(\hat{T})$ auf dem Referenzelement:

$$P(T) := \{q : T \rightarrow \mathbb{R} \mid q(\sigma(\cdot)) \in \hat{P}(\hat{T})\}.$$

Dasselbe gilt für die zugehörigen Basen $\{\varphi^{(r)}, r = 1, \dots, R\}$ von $P(T)$

$$\varphi_r(x) := \hat{\varphi}_r(\sigma^{-1}(x)), \quad x \in T.$$

Für Funktionen $v \in H^m(\Omega)$ erhält man dann durch Setzung

$$I_T v := \sum_{s=1}^R \hat{D}_s \hat{v}(\hat{a}_s) \varphi_s \in P(T)$$

eine zellweise Lagrange/Hermite-Interpolierende $I_T v \in P(T)$ mit den Eigenschaften (Man beachte $\hat{D}_r \hat{\varphi}_s(\hat{a}_r) = \delta_{rs}$):

$$\begin{aligned} \hat{D}_r I_T v(a_r) &= \sum_{s=1}^R \hat{D}_s \hat{v}(\hat{a}_s) \hat{D}_r \varphi_s(a_r) = \sum_{s=1}^R \hat{D}_s \hat{v}(\hat{a}_s) \hat{D}_r \hat{\varphi}_s(\hat{a}_r) \\ &= \hat{D}_r \hat{v}(\hat{a}_r) = \hat{D}_r v(a_r), \quad r = 1, \dots, R. \end{aligned}$$

Je nach Art der Ableitungsoperatoren \hat{D}_r lässt sich dies gegebenenfalls auch als eine Interpolation mit lokal auf den Einzelzellen T definierten Ableitungsoperatoren D_r bzgl. der (physikalischen) Variablen x ausdrücken; z. B. in den einfachsten Fällen $D_r v(a_r) := v(a_r)$ bzw. $D_r v(a_r) \approx \nabla v(a_r)$.

Wir beweisen nun den Hauptsatz dieses Abschnittes zur Polynominterpolation.

Satz 3.6 (Spezieller Interpolationssatz): Für jedes $v \in H^m(T)$ und die zugehörige Interpolierende $I_T v \in P(T)$ auf der Zelle $T \in \mathbb{T}_h$ gilt:

$$|v - I_T v|_{k;T} \leq c_I \frac{h_T^m}{\rho_T^k} |v|_{m;T}, \quad 0 \leq k \leq m, \quad (3.3.71)$$

mit Durchmesser und Inkreisradius h_T bzw. ρ_T von T und einer Konstante $c_I = c_I(d, m, \hat{T})$.

Beweis: i) Für eine Funktion $f \in L^1(T)$ bezeichnet $\hat{f} \in L^1(\hat{T})$ die zugehörige Transformierte

$$\hat{f}(\hat{x}) := f(x), \quad \hat{x} = \sigma^{-1}(x) = B^{-1}x - B^{-1}b \in \hat{T}. \quad (3.3.72)$$

Die Transformation σ^{-1} hat die Funktionalmatrix $\nabla\sigma^{-1} = B^{-1}$, so dass gilt:

$$\int_{\hat{T}} \hat{f}(\hat{x}) d\hat{x} = |\det B^{-1}| \int_T \hat{f}(\sigma^{-1}(x)) dx = |\det B|^{-1} \int_T f(x) dx. \quad (3.3.73)$$

Wir schätzen nun die Elemente b_{ij} sowie $b_{ij}^{(-1)}$ der Matrizen B und B^{-1} ab. Jedes $x \in \mathbb{R}^d$ mit $|x| = \rho$ gestattet mit zwei Punkten $\xi, \eta \in T$ die Darstellung $x = \xi - \eta$, wobei ξ etwa als Mittelpunkt der Inkugel von T gewählt werden kann. Dann sind $\hat{\xi} = B^{-1}\xi - B^{-1}b$, $\hat{\eta} = B^{-1}\eta - B^{-1}b \in \hat{T}$, und wir erhalten

$$|B^{-1}x| = |B^{-1}\xi - B^{-1}b - B^{-1}\eta + B^{-1}b| = |\hat{\xi} - \hat{\eta}| \leq \hat{h}. \quad (3.3.74)$$

Da alle Matrixnormen auf $\mathbb{R}^{d \times d}$ äquivalent sind, gilt mit einer Konstante $c = c(d)$

$$\max_{i,j=1,\dots,d} |b_{ij}^{(-1)}| \leq c \sup_{x \in \mathbb{R}^d} \frac{|B^{-1}x|}{|x|} = c \sup_{|x|=\rho} \frac{|B^{-1}x|}{|x|} \leq c \frac{\hat{h}}{\rho}. \quad (3.3.75)$$

Analog wird bewiesen:

$$\max_{i,j=1,\dots,d} |b_{ij}| \leq c \frac{h}{\rho}. \quad (3.3.76)$$

ii) Es seien nun $T \in \mathbb{T}_h$ sowie $v \in H^m(T)$ beliebig gegeben. Es ist $I_T v \in P(T)$ die Interpolierende von v auf T und

$$I_{\hat{T}} \hat{v}(\cdot) = I_T v(\sigma(\cdot)) \in P(\hat{T}) \quad (3.3.77)$$

die Interpolierende der Transformaten $\hat{v} \in H^m(\hat{T})$ auf \hat{T} . Nach Korollar 3.1 gilt auf \hat{T} :

$$|\hat{v} - I_{\hat{T}} \hat{v}|_{k;\hat{T}} \leq \hat{c} |\hat{v}|_{m;\hat{T}}, \quad 0 \leq k \leq m, \quad (3.3.78)$$

mit einer festen Konstante $\hat{c} = \hat{c}(d, m, \hat{T})$. Durch Koordinatentransformation zwischen \hat{T} und T werden wir nun zeigen, dass

$$\begin{aligned} |v - I_T v|_{k;T} &\leq \frac{c}{\rho^k} |\det B|^{1/2} |\hat{v} - I_{\hat{T}} \hat{v}|_{k;\hat{T}}, \\ |\hat{v}|_{m;\hat{T}} &\leq c h^m |\det B|^{-1/2} |v|_{m;T}, \end{aligned}$$

mit Konstanten $c = c(d, m, \hat{T})$. Diese beiden Beziehungen ergeben dann zusammen mit (3.3.78) die Behauptung.

iii) Für die Ableitungen von v bzw. \hat{v} gilt mit $\hat{x} = \sigma^{-1}(x) = B^{-1}x - B^{-1}b$:

$$\begin{aligned} \widehat{\partial}_i v(\hat{x}) &= \partial_i v(x) = \partial_i \hat{v}(\sigma^{-1}(x)) = \sum_{j=1}^d \hat{\partial}_j \hat{v}(\hat{x}) \partial_i \sigma_j^{-1}(x) = \sum_{j=1}^d \hat{\partial}_j \hat{v}(\hat{x}) b_{ji}^{(-1)}, \\ \hat{\partial}_i \hat{v}(\hat{x}) &= \hat{\partial}_i v(\sigma(\hat{x})) = \sum_{j=1}^d \partial_j v(x) \hat{\partial}_i \sigma_j(\hat{x}) = \sum_{j=1}^d \partial_j v(x) b_{ji} = \sum_{j=1}^d \widehat{\partial}_j v(\hat{x}) b_{ji}, \end{aligned}$$

und folglich

$$|\widehat{\partial}_i v(\hat{x})| \leq c \frac{\hat{h}}{\rho} \max_{j=1,\dots,d} |\hat{\partial}_j \hat{v}(\hat{x})| \quad |\hat{\partial}_i \hat{v}(\hat{x})| \leq c \frac{h}{\hat{\rho}} \max_{j=1,\dots,d} |\widehat{\partial}_j v(\hat{x})|.$$

Für allgemeine Ableitungen D^α bzw. \hat{D}^α der Ordnung $k = |\alpha|$ gewinnen wir durch k -malige Anwendung dieser Beziehungen:

$$\begin{aligned} |\widehat{D}^\alpha v(\hat{x})| &\leq c \max_{i,j=1,\dots,d} |b_{ij}^{(-1)}|^{\alpha} \max_{|\beta|=|\alpha|} |\hat{D}^\beta \hat{v}(\hat{x})| \leq c \left(\frac{\hat{h}}{\rho}\right)^{|\alpha|} \max_{|\beta|=|\alpha|} |\hat{D}^\beta \hat{v}(\hat{x})|, \\ |\hat{D}^\alpha \hat{v}(\hat{x})| &\leq c \max_{i,j=1,\dots,d} |b_{ij}|^{|\alpha|} \max_{|\beta|=|\alpha|} |\widehat{D}^\beta v(\hat{x})| \leq c \left(\frac{h}{\hat{\rho}}\right)^{|\alpha|} \max_{|\beta|=|\alpha|} |\widehat{D}^\beta v(\hat{x})|. \end{aligned}$$

Mit Hilfe dieser Abschätzungen und der Substitutionsformel (3.3.73) folgt nun

$$\begin{aligned} \int_T |D^\alpha v(x)|^2 dx &= |\det B| \int_{\hat{T}} |\widehat{D}^\alpha v(\hat{x})|^2 d\hat{x} \\ &\leq c |\det B| \left(\frac{\hat{h}}{\rho}\right)^{2|\alpha|} \max_{|\beta|=|\alpha|} \int_{\hat{T}} |\hat{D}^\beta \hat{v}(\hat{x})|^2 d\hat{x} \\ \int_{\hat{T}} |\hat{D}^\alpha \hat{v}(\hat{x})|^2 d\hat{x} &\leq c \left(\frac{h}{\hat{\rho}}\right)^{2|\alpha|} \max_{|\beta|=|\alpha|} \int_{\hat{T}} |\widehat{D}^\beta v(\hat{x})|^2 d\hat{x} \\ &\leq c |\det B|^{-1} \left(\frac{h}{\hat{\rho}}\right)^{2|\alpha|} \max_{|\beta|=|\alpha|} \int_T |D^\beta v(x)|^2 dx. \end{aligned}$$

Damit ist für $0 \leq k \leq m$ bewiesen:

$$|v|_{k;T} \leq c |\det B|^{1/2} \left(\frac{\hat{h}}{\rho}\right)^k |\hat{v}|_{k;\hat{T}}, \quad |\hat{v}|_{k;T} \leq c |\det B|^{-1/2} \left(\frac{h}{\hat{\rho}}\right)^k |v|_{k;T}.$$

Anwendung dieser Beziehungen für v und $v - I_T v$ ergibt schließlich wegen $0 < \rho \leq h \leq 1$ die behauptete Abschätzungen (3.3.71). Q.E.D.

Bemerkung 3.6: Die Fehlerabschätzung (3.3.71) für die Polynominterpolation hat natürliche Verallgemeinerungen auf andere Halbnormen $|\cdot|_T$ sowie auf andere Regularitätsstufen $v \in H^{m,p}(T)$, $1 \leq p \leq \infty$. Wir geben ohne Beweis das folgende allgemeine Resultat an:

$$|v - I_T v|_{k,q;T} \leq c_I \frac{h_T^{m-d/p}}{\rho_T^{k-d/q}} |v|_{m,p;T}, \tag{3.3.79}$$

für $0 \leq k \leq m$, $1 \leq p \leq q \leq \infty$, mit einer Konstante $c_I = c_I(d, k, m, p, q, \hat{T})$. Für spätere Zwecke sind hiervon insbesondere die Fälle $p = q = 1$ (für $m \geq 2$), $p = q = \infty$ sowie $p = 2, q = \infty$ (für $k \leq m + 2$) von Interesse; z. B. gilt für $d = 2, m = 2, q = \infty, p = 2$:

$$\max_T |v - I_T v| \leq c_I h_T |v|_{2,2;T}. \tag{3.3.80}$$

Auf ähnliche Art wie im Beweis von Satz 3.6 gewinnt man durch Transformation auf das Einheitselement \hat{T} die folgende sog. „inverse Beziehung“ für finite Elemente:

Satz 3.7 (Inverse Beziehung): *Unter den obigen Voraussetzungen gilt auf jeder Zelle $T \in \mathbb{T}_h$ für Finite-Elemente-Funktionen $v \in P(T)$ mit einer Konstante $c = c(d, \rho_T, k, s)$:*

$$|v|_{k;T} \leq c \frac{h_T^s}{\rho_T^k} |v|_{s;T}, \quad 0 \leq s \leq k \leq m. \quad (3.3.81)$$

Beweis: Für Polynome $\hat{q} \in P(\hat{T})$ gilt wegen der Äquivalenz von Normen auf dem (endlich dimensionalen) Quotientenraum $P(T)/P_{k-1}(T)$ mit einer Konstante $\hat{c} = \hat{c}(d, m, \hat{T})$:

$$|\hat{q}|_{k;\hat{T}} \leq \hat{c} |\hat{q}|_{s;\hat{T}}, \quad 0 \leq s \leq k \leq m. \quad (3.3.82)$$

Die Behauptung ergibt sich nun wieder durch Transformation auf die Zelle T . Q.E.D.

Bemerkung 3.7: Analoge Abschätzungen wie die in Satz 3.6 und Satz 3.7 gelten auch für Tensorprodukt-Polynomansätze auf Vierecken bzw. Hexaedern.

Wir wollen diese Resultate auf die obigen konkreten Beispiele anwenden. In diesen sind alle formulierten Bedingungen erfüllt. Insbesondere gilt die „uniform shape“ Bedingung

$$\sup_{h>0} \max_{T \in \mathbb{T}_h} \frac{h_T}{\rho_T} \leq c.$$

Dann folgt aus Satz 3.6 für die FE-Ansatzräume $V_h^{(m-1)} \subset H_0^1(\Omega)$ vom Polynomgrad $m-1 \in \mathbb{N}_0$ die Interpolationsfehlerabschätzungen

$$|v - I_T v|_{k;T} \leq c h_T^{m-k} |v|_{m;T}, \quad k = 0, \dots, m, \quad T \in \mathbb{T}_h, \quad (3.3.83)$$

mit Konstanten $c = c(d, m, \hat{T})$. Manchmal möchte man den Interpolationsfehler auch über den Rand der Zelle oder punktweise abschätzen. Hierfür gilt

$$\|v - I_T v\|_{\partial T} \leq c h_T^{m-1/2} \|\nabla^2 v\|_{m;T}, \quad (3.3.84)$$

und z. B. in zwei Dimensionen

$$\max_{x \in T} |v - I_T v| \leq c h_T^{m-1} \|\nabla^2 v\|_{m;T}. \quad (3.3.85)$$

Als Folgerung aus diesen lokalen Abschätzungen erhalten wir die folgenden globale Approximationsabschätzungen

$$|v - I_h v|_k \leq c h^{m-k} |v|_m, \quad k = 0, \dots, m. \quad (3.3.86)$$

Für „lineare“ finite Elemente gilt also speziell

$$\|u - I_h u\| + h \|\nabla(u - I_h u)\| \leq c h^2 \|\nabla^2 u\|. \quad (3.3.87)$$

In diesem Fall ergibt die „inverse“ Beziehung (3.3.81):

$$\|\nabla v_h\| \leq h^{-1}\|v_h\|, \quad v_h \in V_h^{(1)}. \quad (3.3.88)$$

Entsprechend gilt für „quadratische“ finite Elemente

$$\|u - I_h u\| + h\|\nabla(u - I_h u)\| \leq ch^3\|\nabla^3 u\|, \quad (3.3.89)$$

und

$$\|\nabla^2 v_h\| \leq h^{-2}\|v_h\|, \quad v_h \in V_h^{(2)}. \quad (3.3.90)$$

Damit ist jetzt die theoretische Grundlage für die *a priori* Fehlerabschätzungen für das Galerkin-Finite-Elemente-Verfahren aus Satz 3.1 geschaffen.

3.4 A priori Fehleranalyse

Die in Abschnitt 3.3 hergeleiteten Abschätzungen für den Fehler bei der Interpolation mit stückweise polynomialen Funktionen sind die Basis für die *a priori* Fehleranalyse des Finite-Elemente-Verfahrens. Wir formulieren die folgende Verallgemeinerung von Satz 3.1 für FE-Approximationen allgemeiner Ordnung $m \geq 2$,

$$(\nabla u_h, \nabla \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h, \quad (3.4.91)$$

der Poisson-Gleichung. Dabei ist die sog. „Ordnung“ eines FE-Verfahrens im wesentlichen durch den Polynomgrad der verwendeten Ansatzfunktionen bestimmt, d. h. durch die Beziehung $P_{m-1} \subset P(T)$ für alle $T \in \mathcal{T}_h$.

Korollar 3.2 (Allgemeine FE-Konvergenz): *Für den Fehler $e_h := u - u_h$ einer FE-Methode mit Ansatzräumen $V_h \subset H_0^1(\Omega)$ der Ordnung $m \geq 2$ zur Approximation der Poisson-Gleichung gilt die a priori Fehlerabschätzung:*

$$\|e_h\| + h\|\nabla e_h\| \leq ch^m \|\nabla^m u\|. \quad (3.4.92)$$

Wir wollen eine Schwäche dieses Resultats nicht unerwähnt lassen. Im allgemeinen sind Lösungen $u \in H_0^1(\Omega)$ elliptischer Gleichungen zweiter Ordnung über Gebieten mit Polygonrand nicht aus $H^m(\Omega)$ für $m \geq 3$. An den Ecken von $\partial\Omega$ treten starke Singularitäten der Ableitungen von u auf. Die obige Voraussetzung $u \in H^m(\Omega)$ ist also für $m > 2$ unrealistisch. Finite Elemente höherer Ordnung $m > 2$ können jedoch bei Gebieten mit hinreichend glattem Rand $\partial\Omega$ erfolgreich verwendet werden, denn in diesem Fall kann die Regularitätstheorie meist $u \in H^m(\Omega)$ sichern. Dabei sind aber besondere Maßnahmen, wie z. B. die Verwendung isoparametrischer Ansätze, zur Approximation entlang des krummen Randes erforderlich.

Wir haben gesehen, dass der Fehler z. B. bei „linearen“ finiten Elementen gemessen in der L^2 -Norm eine verbesserte Konvergenzordnung $O(h^2)$ gegenüber der von $O(h)$ in der

Energie-Norm zulässt. Es stellt sich die Frage, ob man durch weitere Abschwächung der Norm vielleicht noch höhere Konvergenzordnungen erzielen kann. Diese Hoffnung wird zunächst bestärkt durch die Beobachtung, dass dies für die L^2 -Projektion durchaus der Fall ist. Die in Frage kommenden Normen werden illustrativ als „negative“ Sobolew-Normen (genauer als Sobolew-Normen mit „negativer Ordnung“) bezeichnet und sind in den einfachsten Fällen definiert durch

$$\|v\|_{-1} := \sup_{\varphi \in V} \frac{(v, \varphi)}{\|\varphi\|_1}, \quad \|v\|_{-2} := \sup_{\varphi \in V \cap H^2(\Omega)} \frac{(v, \varphi)}{\|\varphi\|_2}.$$

wobei wieder $V := H_0^1(\Omega)$.

Hilfssatz 3.6 (L^2 -Projektion): Für die L^2 -Projektion $P_h : V \rightarrow V_h$ auf den Raum $V_h^{(1)}$ der „linearen“ finiten Elemente gilt die Fehlerabschätzungen

$$\|u - P_h u\|_{-2} + h\|u - P_h u\|_{-1} + h^2\|u - P_h u\| \leq ch^4 \|\nabla^2 u\|. \quad (3.4.93)$$

Beweis: Zunächst rekapitulieren wir die „Bestapproximationseigenschaft“ der L^2 -Projektion:

$$\|u - P_h u\| = \min_{\varphi_h \in V_h^{(1)}} \|u - \varphi_h\|. \quad (3.4.94)$$

Daraus folgt mit der lokalen Interpolationsabschätzung (3.3.83) die Beziehung

$$\|u - P_h u\| \leq ch^2 \|\nabla^2 u\|. \quad (3.4.95)$$

Mit einem beliebigen $\varphi \in H_0^1(\Omega)$ gilt entsprechend für $k \in \{1, 2\}$:

$$\begin{aligned} (u - P_h u, \varphi) &= (u - P_h u, \varphi - P_h \varphi) \\ &\leq \|u - P_h u\| \|\varphi - P_h \varphi\| \\ &\leq ch^{2+k} \|\nabla^2 u\| \|\nabla^k \varphi\|. \end{aligned}$$

Dies impliziert

$$\sup_{\varphi \in H_0^1(\Omega) \cap H^k(\Omega)} \frac{(u - P_h, \varphi)}{\|\varphi\|_k} \leq ch^{2+k} \|\nabla^2 u\|,$$

was zu beweisen war.

Q.E.D.

Der Beweis von Hilfssatz 3.6 zeigt, dass eine weitere Erhöhung jenseits $\mathcal{O}(h^4)$ der Approximationsordnung der L^2 -Projektion auf den Ansatzraum $V_h^{(1)}$ auch in einer noch schwächeren Norm nicht mehr möglich ist. Für die „Ritz-Projektion“ $R_h : V \rightarrow V_h^{(1)}$ ist in diesem Fall sogar $\mathcal{O}(h^2)$ die Obergrenze für die erreichbare Ordnung. Dies zeigt der folgende Satz.

Satz 3.8 (Ritz-Projektion): Für die Ritz-Projektion $R_h : V \rightarrow V_h^{(1)}$ auf den Raum der „linearen“ finiten Elemente gilt die Abschätzung

$$\|u - R_h u\|_{-1} \geq c(f) \|\nabla(u - R_h u)\|^2, \quad (3.4.96)$$

mit einer positiven Konstante $c(f) > 0$.

Beweis: Sei $f \in H_0^1(\Omega)$. Für die Lösung $u \in V$ der Gleichung $-\Delta u = f$ gilt unter Ausnutzung der „Galerkin-Orthogonalität“:

$$(u - R_h u, f) = (\nabla(u - R_h u), \nabla u) = \|\nabla(u - R_h u)\|^2.$$

Wegen

$$\sup_{\varphi \in V} \frac{(u - R_h u, \varphi)}{\|\varphi\|_1} \geq \frac{(u - R_h u, f)}{\|f\|_1}$$

impliziert dies die Behauptung. Q.E.D.

Da die Energie-Fehlerabschätzung

$$\|\nabla(u - R_h u)\| \leq ch \|\nabla^2 u\|$$

bzgl. der Ordnung bestmöglich ist, folgt aus (3.4.96) die Unmöglichkeit einer Fehlerabschätzung mit einer Ordnung größer als zwei für „lineare“ finite Elemente.

3.4.1 Punktweise Fehlerabschätzung

In den vorangegangenen Abschnitten haben wir gesehen, dass die Methode der finiten Elemente als Projektionsmethode zunächst ganz natürlich zu a priori Abschätzungen in der „Energienorm“ $\|\nabla e\|$ und dann über ein Dualitätsargument auch zu verbesserten Abschätzungen in der L^2 -Norm $\|e\|$ führt. Diese Konvergenzaussagen im „quadratischen Mittel“ gestatten aber noch keinen unmittelbaren Schluss auf die punktweise Konvergenz des Verfahrens. Dieser Frage soll jetzt wieder anhand des einfachen Modellproblems „Poisson-Gleichung“,

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ auf } \partial\Omega, \quad (3.4.97)$$

auf einem (konvexen) Polygonebiet $\Omega \subset \mathbb{R}^2$ nachgegangen werden. Wir beschränken uns dabei auf die Approximation (3.4.91) mit stückweise linearen Ansätzen $V_h^{(1)} \subset V$ auf regulären Triangulierungen. Wir rekapitulieren hierfür die a priori Fehlerabschätzung

$$\|e\| + h\|\nabla e\| \leq ch^2 \|\nabla^2 u\|, \quad (3.4.98)$$

wobei die Konstante c wesentlich durch die Konstante in der Interpolationsfehlerabschätzung

$$\|\nabla(u - I_h u)\|_T \leq c_I h \|\nabla^2 u\|_T, \quad T \in \mathbb{T}_h, \quad (3.4.99)$$

bestimmt ist. Wir erinnern daran, dass im vorliegenden Fall eines konvexen Grundgebiets jedes $v \in V$ mit $\Delta v \in L^2(\Omega)$ automatisch in $H^2(\Omega)$ ist und der a priori Abschätzung genügt:

$$\|\nabla^2 v\| \leq \|\Delta v\|. \quad (3.4.100)$$

Satz 3.9 (Sub-optimale L^∞ -Norm-Fehlerabschätzung): *Unter den obigen Voraussetzungen konvergiert die Methode der finiten Elemente punktweise mit der Ordnung $\mathcal{O}(h)$:*

$$\max_{\bar{\Omega}} |e| \leq ch \|\nabla^2 u\|. \quad (3.4.101)$$

Beweis: Sei T ein beliebiges Dreieck aus \mathbb{T}_h . Für eine Funktion $v_h \in V_h^{(1)}$ gilt dann

$$\max_T |v_h| \leq c|T|^{-1} \int_T |v_h| dx. \quad (3.4.102)$$

Dies folgt leicht mit Hilfe der Transformation auf die Referenzzelle \hat{T} ($dx \approx |T|d\hat{x}$) und der dort geltenden Beziehung (Äquivalenz von Normen)

$$\max_{\hat{T}} |\hat{v}_h| \leq \hat{c} \int_{\hat{T}} |\hat{v}_h| d\hat{x}.$$

Mit der Knoteninterpolierenden $I_h u \in V_h^{(1)}$ zu $u \in V \cap H^2(\Omega)$ gilt

$$\max_T |u - I_h u| \leq ch_T \|\nabla^2 u\|_T. \quad (3.4.103)$$

Damit erschließen wir dann:

$$\begin{aligned} \max_T |e| &\leq \max_T |u - I_h u| + \max_T |I_h e| \\ &\leq \max_T |u - I_h u| + c|T|^{-1} \int_T |I_h e| dx \\ &\leq \max_T |u - I_h u| + c|T|^{-1} \int_T |e| dx \leq ch_T \|\nabla^2 u\|_T + ch_T^{-1} \|e\|_T. \end{aligned}$$

Mit Hilfe der L^2 -Fehlerabschätzung (3.4.98) folgt also die Behauptung. Q.E.D.

Bemerkung 3.8: Unter der bloßen Annahme $u \in V \cap H^2(\Omega)$ ist die Fehlerabschätzung (3.4.101) bzgl. der h -Potenz optimal. Um dies einzusehen, betrachte man z. B. auf dem Einheitskreis $B_1 = \{x \in \mathbb{R}^2 \mid |x| < 1\}$ die Funktionen

$$u^h(x) := (|x|^2 + h^4)^{1/2}$$

Offenbar ist $u^h \in H^2(B_1)$, und es gilt

$$\|\nabla^2 u^h\|_{B_1} \leq c \left(\int_{B_1} (|x|^2 + h^4)^{-1} dx \right)^{1/2} \leq c |\ln(h)|^{1/2}.$$

Wir nehmen nun an, dass unser Lösungsgebiet Ω den Kreis B_1 enthält und die Funktionen u^h geeignet zu Funktionen $\bar{u}^h \in H_0^1(\Omega) \cap H^2(\Omega)$ fortgesetzt sind. Ferner gehöre zu jeder der Triangulierungen \mathbb{T}_h ein Dreieck T_0 mit Durchmesser h und dem Inkreismittelpunkt $a_0 = 0$. Dann ist in den Eckpunkten a_i und dem Mittelpunkt a_0 von T_0 stets

$$\bar{u}^h(a_i) \geq h, \quad \bar{u}^h(a_0) = h^2.$$

Würde nun für die Ritz-Projektion $\bar{u}_h^h \in V_h$ zu \bar{u}^h mit einem $\varepsilon > 0$ gelten

$$\max_{T_0} |\bar{u}^h - \bar{u}_h^h| \leq ch^{1+\varepsilon} \|\nabla^2 \bar{u}^h\|_{\Omega},$$

so ergäbe sich für hinreichend kleines h im Widerspruch zur Linearität der \bar{u}_h^h auf T_0 :

$$\begin{aligned} |\bar{u}_h^h(a_0)| &\leq |\bar{u}_h^h(a_0) - \bar{u}^h(a_0)| + |\bar{u}^h(a_0)| \leq ch^{1+\varepsilon_1}, \quad 0 < \varepsilon_1 < \varepsilon, \\ |\bar{u}_h^h(a_i)| &\geq |\bar{u}^h(a_i)| - |\bar{u}^h(a_i) - \bar{u}_h^h(a_i)| \geq ch. \end{aligned}$$

Numerische Experimente zeigen, dass im Falle höherer Regularität von u (etwa $u \in C^2(\bar{\Omega})$) die optimale Ordnung $O(h^2)$ des L^2 -Fehlers auch für die punktweise Konvergenz vorliegt. Ein ähnliches Phänomen haben wir bereits bei der Differenzenapproximation mit dem 5-Punkte-Operator gesehen, bei dem sich auf gleichförmigen Gittern die Konvergenzordnung $O(h)$ im Falle $u \in C^3(\bar{\Omega})$ auf $O(h^2)$ im Falle $u \in C^4(\bar{\Omega})$ erhöht. Für die Methode der finiten Elemente beweisen wir nun das folgende optimale Resultat auf allgemeinen regulären Gittern:

Satz 3.10 (Optimale L^∞ -Abschätzung): *Im Falle $u \in V \cap C^2(\bar{\Omega})$ gilt die Konvergenzabschätzung*

$$\sup_{\Omega} |e| \leq ch^2 L(h) \sup_{\Omega} |\nabla^2 u| \tag{3.4.104}$$

mit dem logarithmischen Faktor $L(h) := |\log(h)| + 1$.

Beweis: Wir notieren zunächst die Interpolationsfehlerabschätzung

$$\sup_{\Omega} |u - I_h u| + h \sup_{\Omega} |\nabla(u - I_h u)| \leq ch^2 \sup_{\Omega} |\nabla^2 u| \tag{3.4.105}$$

i) Für ein $h > 0$ sei $T_* \in \mathbb{T}_h$ beliebig, aber fest gewählt. Mit der Knoteninterpolierenden $I_h u \in V_h^{(1)}$ von u gilt wieder (s. Beweis von Satz 3.9):

$$\begin{aligned} \max_{T_*} |e| &\leq c \max_{T_*} |u - I_h u| + c |T_*|^{-1} \int_{T_*} |e| dx \\ &\leq ch^2 \max_{T_*} |\nabla^2 u| + c |T_*|^{-1} \int_{T_*} |e| dx. \end{aligned}$$

Damit ist die Abschätzung des L^∞ -Fehlers zurückgeführt auf eine lokale L^1 -Fehlerabschätzung. Mit der durch

$$\delta^h := |T_*|^{-1} \text{sign}(e) \text{ in } T_*, \quad \delta^h := 0 \text{ sonst,}$$

definierten Funktion $\delta^h \in L^2(\Omega)$ („regularisierte“ Dirac¹⁷-Funktion) gilt weiter

$$|T_*|^{-1} \int_{T_*} |e| dx = (\delta^h, e).$$

ii) Jetzt wird wieder ein Dualitätsargument verwendet. Wir betrachten das Hilfsproblem

$$-\Delta g^h = \delta^h \text{ in } \Omega, \quad g^h = 0 \text{ auf } \partial\Omega. \quad (3.4.106)$$

Die Funktion g^h kann als „regularisierte“ Greensche Funktion angesehen werden. Damit gilt

$$|T_*|^{-1} \int_{T_*} |e| dx = (\nabla e, \nabla g^h). \quad (3.4.107)$$

Unter Verwendung der durch

$$(\nabla g_h^h, \nabla \varphi_h) = (\nabla g^h, \nabla \varphi_h) \quad \forall \varphi_h \in V_h^{(1)},$$

definierten „Ritz-Projektion“ $g_h^h \in V_h^{(1)}$ von g^h erhalten wir durch zweimalige Anwendung der Galerkin-Orthogonalität die Beziehung

$$|T_*|^{-1} \int_{T_*} |e| dx = (\nabla e, \nabla(g^h - g_h^h)) = (\nabla(u - I_h u), \nabla(g^h - g_h^h)). \quad (3.4.108)$$

Mit der Hölderschen Ungleichung folgt

$$|T_*|^{-1} \int_{T_*} |e| dx \leq \max_{\Omega} |\nabla(u - I_h u)| \int_{\Omega} |\nabla(g^h - g_h^h)| dx. \quad (3.4.109)$$

Unter Beachtung der Abschätzung (3.4.105) erhalten wir schließlich die Beziehung

$$\max_{T_*} |e| \leq ch \left\{ h + \int_{\Omega} |\nabla(g^h - g_h^h)| dx \right\} \sup_{\Omega} |\nabla^2 u|. \quad (3.4.110)$$

¹⁷Paul Adrien Maurice Dirac (1902–1984): Französischer Physiker und Mathematiker; Prof. in Cambridge; wichtige Beiträge zur Quanten Mechanik und Kosmologie, 1933 Nobel-Preis.

Die punktweise Abschätzung des Fehlers e ist also zurückgeführt auf eine globale L^1 -Fehlerabschätzung für den Gradienten der „Greenschen Funktion“: $\nabla(g^h - g_h^h)$. Dieser wird unten weiter abgeschätzt. Dazu benötigen wir einige a priori Abschätzungen für g^h , die im folgenden bereitgestellt werden.

iii) Sei x_* der Inkreismittelpunkt von T_* . Wir definieren die Gewichtsfunktion

$$\sigma(x) := (|x - x_*|^2 + \kappa^2 h^2)^{1/2}.$$

Durch Nachrechnen verifiziert man leicht die Beziehungen

$$\kappa h \leq \sigma \leq c, \quad |\nabla \sigma| \leq c_*, \quad |\nabla^2 \sigma| \leq c\sigma^{-1}, \quad \|\sigma^{-1}\| \leq cL(h)^{1/2},$$

mit von h und κ unabhängigen Konstanten. Für die Größen $\bar{\sigma}_T := \max_T \sigma$ und $\underline{\sigma}_T := \min_T \sigma$ gilt daher

$$\bar{\sigma}_T \leq \underline{\sigma}_T + h \max_T |\nabla \sigma| \leq \underline{\sigma}_T + c_* h,$$

und bei Wahl von $\kappa := 2c_*$ (unabhängig von h):

$$\bar{\sigma}_T \leq \underline{\sigma}_T + \frac{1}{2}\bar{\sigma}_T,$$

und damit

$$\max_{T \in \mathbb{T}_h} \frac{\bar{\sigma}_T}{\underline{\sigma}_T} \leq 2. \tag{3.4.111}$$

Hilfssatz 3.7 (Greensche Funktion): *Für die regularisierte „Greensche Funktion“ g^h gelten die a priori Abschätzungen*

$$\sup_{\Omega} |g^h| \leq cL(h), \tag{3.4.112}$$

$$\|\nabla g^h\| + \|\sigma \nabla^2 g^h\| \leq cL(h)^{1/2}, \tag{3.4.113}$$

$$\|\nabla^2 g^h\| \leq ch^{-1}, \tag{3.4.114}$$

mit von h unabhängigen Konstanten c .

Beweis: i) Die „richtige“ Greensche Funktion $g(\cdot) = g(x_*, \cdot)$ zum Aufpunkt x_* erlaubt die Abschätzung (Beweis mit Hilfe des Maximumprinzips)

$$|g(x)| \leq c \{ |\ln(|x - x_*|)| + 1 \}.$$

Konstruktionsgemäß folgt damit

$$|g^h(x)| = |(\nabla g^h, \nabla g)| = |(\delta^h, g)_\Omega| \leq ch^{-2} \int_{T_*} |g| dx \leq cL(h).$$

Dies impliziert die Abschätzung (3.4.112).

ii) Zum Beweis von (3.4.114) verwenden wir die übliche a priori L^2/H^2 -Abschätzung

$$\|\nabla^2 g^h\| \leq c\|\delta^h\| \leq ch^{-1}.$$

iii) Als nächstes notieren wir die einfache a priori Abschätzung

$$\|\nabla^2 g^h\| \leq \|\Delta g^h\| \leq ch^{-1}. \quad (3.4.115)$$

Weiter gilt

$$\|\nabla g^h\|^2 = (\delta^h, g^h) \leq c \sup_{\Omega} |g^h| \leq cL(h).$$

Dies impliziert den ersten Teil der Abschätzung (3.4.113).

iii) Schließlich setzen wir $\xi := x - x_*$ und finden wegen

$$|\xi_i \nabla^2 g^h| \leq |\nabla^2(\xi_i g^h)| + |\nabla g^h|$$

die Beziehung

$$\begin{aligned} \|\sigma \nabla^2 g^h\|^2 &= \sum_{i=1}^2 \|\xi_i \nabla^2 g^h\|^2 + \kappa^2 h^2 \|\nabla^2 g^h\|^2 \\ &\leq \sum_{i=1}^2 \{\|\nabla^2(\xi_i g^h)\|^2 + \|\nabla g^h\|^2\} + \kappa^2 h^2 \|\nabla^2 g^h\|^2. \end{aligned}$$

Mit Hilfe der üblichen a priori L^2/H^2 -Abschätzung (3.4.100) folgt

$$\begin{aligned} \|\nabla^2(\xi_i g^h)\| &\leq \|\Delta(\xi_i g^h)\| \leq \|\xi_i \Delta g^h\| + \|\nabla g^h\| \\ &\leq \|\xi_i \delta^h\| + \|\nabla g^h\| \leq c + cL(h)^{1/2}. \end{aligned}$$

Die vorausgehenden Abschätzungen implizieren dann

$$\|\sigma \nabla^2 g^h\| \leq cL(h)^{1/2},$$

woraus (3.4.113) folgt. Dies vervollständigt den Beweis von (3.4.113).

Q.E.D.

iv) Wir kehren nun zum Beweis des Satzes zurück und schätzen wie folgt ab:

$$\int_{\Omega} |\nabla(g^h - g_h^h)| dx \leq \|\sigma^{-1}\| \|\sigma \nabla(g^h - g_h^h)\| \leq cL(h)^{1/2} \|\sigma \nabla(g^h - g_h^h)\|.$$

Durch Ausdifferenzieren folgt weiter

$$\begin{aligned} \|\sigma \nabla(g^h - g_h^h)\|^2 &= (\nabla(g^h - g_h^h), \nabla(\sigma^2(g^h - g_h^h))) - (\nabla(g^h - g_h^h), (g^h - g_h^h) \nabla \sigma^2) \\ &=: E_1 - E_2. \end{aligned}$$

Die Terme E_1 und E_2 werden im folgenden separat abgeschätzt. Im Hinblick auf die Galerkin-Orthogonalität gilt

$$E_1 = (\nabla(g^h - g_h^h), \nabla(\sigma^2(g^h - g_h^h) - \psi_h))$$

mit der Knoteninterpolierenden $\psi_h := I_h(\sigma^2(g^h - g_h^h)) \in V_h^{(1)}$. Dies wird abgeschätzt durch

$$\begin{aligned} E_1 &\leq \sum_{T \in \mathbb{T}_h} \|\sigma \nabla(g^h - g_h^h)\|_T \|\sigma^{-1} \nabla(\sigma^2(g^h - g_h^h) - \psi_h)\|_T \\ &\leq \frac{1}{4} \|\sigma \nabla(g^h - g_h^h)\|^2 + c \sum_{T \in \mathbb{T}_h} \|\sigma^{-1} \nabla(\sigma^2(g^h - g_h^h) - \psi_h)\|_T^2. \end{aligned}$$

Mit Hilfe der Interpolationsfehlerabschätzung (3.4.103) werden die einzelnen Summanden wie folgt abgeschätzt:

$$\begin{aligned} \|\sigma^{-1} \nabla(\sigma^2(g^h - g_h^h) - \psi_h)\|_T^2 &\leq \underline{\sigma}_T^{-2} \|\nabla(\sigma^2(g^h - g_h^h) - \psi_h)\|_T^2 \\ &\leq c \underline{\sigma}_T^{-2} h_T^2 \|\nabla^2(\sigma^2(g^h - g_h^h))\|_T^2 \\ &\leq c \underline{\sigma}_T^{-2} h_T^2 \{ \|g^h - g_h^h\|_T^2 + \|\sigma \nabla(g^h - g_h^h)\|_T^2 + \|\sigma^2 \nabla^2 g^h\|_T^2 \} \\ &\leq c \|g^h - g_h^h\|_T^2 + c \underline{\sigma}_T^{-2} \bar{\sigma}_T^2 h_T^2 \{ \|\nabla(g^h - g_h^h)\|_T^2 + \|\sigma \nabla^2 g^h\|_T^2 \}. \end{aligned}$$

Dies ergibt wegen (3.4.111):

$$E_1 \leq \frac{1}{4} \|\sigma \nabla(g^h - g_h^h)\|^2 + c \|g^h - g_h^h\|^2 + c h^2 \{ \|\nabla(g^h - g_h^h)\|^2 + \|\sigma \nabla^2 g^h\|^2 \}.$$

Die schon bekannten L^2 -Fehlerabschätzungen liefern weiter

$$E_1 \leq \frac{1}{4} \|\sigma \nabla(g^h - g_h^h)\|^2 + c h^2 \{ h^2 \|\nabla^2 g^h\|^2 + \|\sigma \nabla^2 g^h\|^2 \},$$

sowie unter Beachtung der Abschätzungen von Hilfssatz 3.7

$$E_1 \leq \frac{1}{4} \|\sigma \nabla(g^h - g_h^h)\|^2 + c h^2 |\ln(h)|.$$

Für den zweiten Term gilt wegen $|\nabla \sigma^2| \leq c \sigma$:

$$E_2 \leq c \|\sigma \nabla(g^h - g_h^h)\| \|g^h - g_h^h\| \leq \frac{1}{4} \|\sigma \nabla(g^h - g_h^h)\|^2 + c \|g^h - g_h^h\|^2$$

sowie mit den Argumenten von oben:

$$E_2 \leq \frac{1}{4} \|\sigma \nabla(g^h - g_h^h)\|^2 + c h^2.$$

Kombination der Abschätzungen für E_1 und E_2 ergibt schließlich

$$\|\sigma \nabla(g^h - g_h^h)\|^2 \leq \frac{1}{2} \|\sigma \nabla(g^h - g_h^h)\|^2 + c h^2 L(h),$$

und damit die Behauptung. Q.E.D.

Der logarithmische Term $L(h) := |\ln(h)| + 1$ in der Abschätzung (3.4.104) lässt sich auf allgemeinen Gittern nicht vermeiden. Dies wird durch numerische Tests und auch durch theoretische Analyse bestätigt. Auf gleichförmigen Gittern (Je zwei benachbarte Dreiecke bilden ein Parallelogramm.) erhält man unter der stärkeren Glattheitsbedingung $u \in C^{2+\alpha}(\bar{\Omega})$ allerdings die optimale Konvergenzordnung $O(h^2)$. Vergleicht man dieses

Resultat mit dem entsprechenden für die Differenzenapproximation mit dem 5-Punkte-Operator, so stellen wir eine deutliche Abschwächung der Regularitätsanforderungen um fast zwei Stufen fest.

Auf Polygonegebieten ist die Bedingung $u \in H^{2,\infty}(\Omega)$ für die schwache Lösung von (3.4.97) im allgemeinen unrealistisch. Bei den Ecken können die zweiten Ableitungen Singularitäten haben. Auf konvexen Polygonegebieten ist gerade die Voraussetzung $u \in H^2(\Omega)$ natürlich, d. h. stets erfüllt, wogegen $u \in H^{2,\infty}(\Omega_1)$ nur auf Teilgebieten $\Omega'_1 \subset \Omega$ mit positivem Abstand zu den Eckpunkten gilt. In diesem Fall haben wir das folgende „lokale“ Resultat:

Satz 3.11 (Lokales Fehlerverhalten): Sei $\Omega_1 \subset \Omega$ ein Teilgebiet mit positivem Abstand δ_1 zu den Eckpunkten von Ω und $u \in H^2(\Omega) \cap C^2(\overline{\Omega}_1)$. Dann gilt auf jedem zweiten Teilgebiet $\Omega_2 \subset \Omega_1$ mit Abstand $\delta_2 > \delta_1$ zu den Eckpunkten die Fehlerabschätzung

$$\sup_{\Omega_2} |e| \leq ch^2 \left\{ L(h) \sup_{\Omega_1} |\nabla^2 u| + \|\nabla^2 u\|_{\Omega} \right\}. \quad (3.4.116)$$

Beweis: Der technisch aufwendige Beweis kann im Rahmen dieses Textes nicht geführt werden; wir verweisen dafür auf die entsprechende Literatur. Q.E.D.

Bemerkung 3.9: Zum Abschluss bemerken wir noch, dass sich für finite Elemente höherer Ordnung (Polynomgrad $m-1 \geq 2$) zu (3.4.104) analoge punktweise Fehlerabschätzungen unter der Voraussetzung $u \in C^m(\overline{\Omega})$ herleiten lassen:

$$\sup_{\Omega} |e| \leq ch^m \sup_{\Omega} |\nabla^m u|. \quad (3.4.117)$$

Bemerkenswerterweise tritt dabei ab Polynomordnung $m-1 = 2$ der störende logarithmische Term $L(h)$ nicht auf. Dasselbe gilt auch für den niedrigsten Ansatzgrad $m-1 = 1$, wenn der maximale Fehlergradient betrachtet wird:

$$\sup_{\Omega} |\nabla e| \leq ch \sup_{\Omega} |\nabla^2 u|. \quad (3.4.118)$$

3.5 Implementierungsaspekte

Im folgenden wollen wir einige Fragen im Zusammenhang mit der praktischen Realisierung der Finite-Elemente-Methode diskutieren. Dazu betrachten wir als Modellfall die 1. RWA eines (elliptischen) Differentialoperators,

$$Lu := -\nabla\{a\nabla u\} = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega, \quad (3.5.119)$$

mit möglicherweise variablem Koeffizienten $a = a(x) \geq \alpha > 0$ auf einem (konvexen) Polygonegebiet $\Omega \subset \mathbb{R}^2$. Die Diskretisierung erfolgt wieder auf Ansatzräumen $V_h \subset H_0^1(\Omega)$

zu einer Folge von (gleichmäßig) regulären Zerlegungen $\mathbb{T}_h = \{T\}$ des Grundgebiets $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$),

$$a(u, \varphi) = l(\varphi) \quad \forall \varphi_h \in V_h. \quad (3.5.120)$$

mit den bilinearen bzw. linearen Formen

$$a(u, \varphi) := (a \nabla u_h, \nabla \varphi_h), \quad l(\varphi) := (f, \varphi_h).$$

Die Aufstellung des zugehörigen linearen Gleichungssystems erfordert in der Regel die Anwendung numerischer Integration, was zu einem zusätzlichen Fehler führt.

3.5.1 Aufbau der Systemmatrizen und Vektoren

Im Gegensatz zu den Differenzenverfahren auf strukturierten Gittern lassen sich die algebraischen Gleichungssysteme der Finite-Elemente-Methode auf allgemeinen Zerlegungen \mathbb{T}_h in der Regel nicht explizit „per Hand“ aufstellen.

Mit der Knotenbasis $\{\varphi_h^{(n)}, n = 1, \dots, N\}$ des Finite-Elemente-Raumes $V_h \subset H_0^1(\Omega)$ sind die Systemmatrizen (unter Weglassung des Index h) $A = (a_{nm})_{n,m=1}^N$ („Steifigkeitsmatrix“) und $M = (m_{nm})_{n,m=1}^N$ („Massenmatrix“) sowie der „Lastvektor“ $b = (b_n)_{n=1}^N$ gebildet gemäß

$$a_{nm} = a(\varphi_h^{(m)}, \varphi_h^{(n)}), \quad m_{nm} = (\varphi_h^{(m)}, \varphi_h^{(n)}), \quad b_n = l(\varphi_h^{(n)}).$$

Beide Matrizen sind konstruktionsgemäß symmetrisch und positiv-definit. Ihre größten und kleinsten Eigenwerte seien $\lambda_{\max}(A)$, $\lambda_{\min}(A)$ bzw. $\lambda_{\max}(M)$, $\lambda_{\min}(M)$, und die zugehörigen Spektralkonditionen:

$$\kappa_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}, \quad \kappa_2(M) = \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}.$$

Wir rekapitulieren die folgenden Beziehungen zwischen einer Finite-Elemente-Funktion $v_h \in V_h$ und ihrem zugehörigen Knotenvektor $\xi = (\xi_n)_{n=1}^N \in \mathbb{R}^N$:

$$v_h = \sum_{n=1}^N \xi_n \varphi_h^{(n)}, \quad \|v_h\|^2 = \langle M\xi, \xi \rangle, \quad a(v_h, v_h) = \langle A\xi, \xi \rangle,$$

wobei $\langle \cdot, \cdot \rangle$ das euklidische Skalarprodukt bezeichnet; die euklidische Vektornorm ist $|\cdot|$. Der Knotenvektor ξ ist bestimmt durch das lineare Gleichungssystem

$$A\xi = b \quad (3.5.121)$$

Zur Aufstellung des Systems (3.5.121) bedient man sich in der Praxis eines zell-orientierten Prozesses, des sog. „Assemblierens“ (englischen „assembling“). Dabei wird das Konzept der „Element-(Last)-Vektoren“ und „Element-(Steifigkeits)-Matrizen“ verwendet. Für einen Knotenvektor $\xi \in \mathbb{R}^N$ und eine Zelle $T \in \mathbb{T}_h$ ist dabei $\xi_T = (\xi_1^T, \dots, \xi_n^T, \dots, \xi_N^T)^T$

mit

$$\xi_n^T := \xi_n \quad \text{falls } \varphi_h^{(n)} \not\equiv 0 \text{ auf } T, \quad \xi_n^T := 0 \quad \text{falls } \varphi_h^{(n)} \equiv 0 \text{ auf } T.$$

Die zugehörigen Element-Matrizen und -Vektoren haben die Form

$$\begin{aligned} A_T &= (a_{nm}^T)_{n,m=1}^N := ((a \nabla \varphi_h^{(m)}, \nabla \varphi_h^{(n)})_T)_{n,m=1}^N, \\ M_T &= (m_{nm}^T)_{n,m=1}^N := ((\varphi_h^{(m)}, \varphi_h^{(n)})_T)_{n,m=1}^N, \\ b_T &= (b_n^T)_{n=1}^N := ((f, \varphi_h^{(n)})_T)_{n=1}^N. \end{aligned}$$

Dabei werden natürlich nur die wesentlichen, von Null verschiedenen Elemente von A_T , M_T und b_T gespeichert. Die einzelnen Gesamtmatrizen und Vektoren werden dann gebildet durch Assemblierung der entsprechenden Element-Matrizen und -Vektoren gemäß :

$$A = \sum_{T \in \mathbb{T}_h} A_T, \quad M = \sum_{T \in \mathbb{T}_h} M_T, \quad b = \sum_{T \in \mathbb{T}_h} b_T.$$

Entsprechend gilt

$$\langle A\xi, \xi \rangle = \sum_{T \in \mathbb{T}_h} \langle A_T \xi_T, \xi_T \rangle, \quad \langle M\xi, \xi \rangle = \sum_{T \in \mathbb{T}_h} \langle M_T \xi_T, \xi_T \rangle.$$

Die Element-Beiträge werden in der Regel durch Transformation auf ein Referenzelement berechnet. Wir diskutieren hier nur den Fall von Dreieckszerlegungen. Sei also wieder $\sigma_T : \hat{T} \rightarrow T$ die affin-lineare Abbildung des Einheitsdreiecks \hat{T} auf das Dreieck T :

$$x = \sigma_T(\hat{x}) = B_T \hat{x} + b_T, \quad \hat{x} = \sigma_T^{-1}(x) = B_T^{-1}x - B_T^{-1}b_T.$$

Die charakteristischen Parameter von T sowie \hat{T} sind mit h_T , ρ_T bzw. \hat{h} , $\hat{\rho}$ bezeichnet. Für transformierte Funktionen $\hat{v}(\hat{x}) = v(x)$ gilt dann

$$\int_T v(x) dx = |\det B_T| \int_{\hat{T}} \hat{v}(\hat{x}) d\hat{x},$$

woraus insbesondere $|\det B_T| = |T| |\hat{T}|^{-1} \approx h_T^d$ folgt. Ferner ist mit der Inversen $B_T^{-1} = (b_{ij}^{(-1)})_{ij=1}^d$:

$$\widehat{\partial}_i v(\hat{x}) = \partial_i v(x) = \partial_i \hat{v}(\hat{x}) = \sum_{j=1}^d \hat{\partial}_j \hat{v}(\hat{x}) \partial_i \hat{x}_j = \sum_{j=1}^d \hat{\partial}_j \hat{v}(\hat{x}) b_{ji}^{(-1)}.$$

bzw. $\widehat{\nabla} v(\hat{x}) = B_T^{-T} \hat{\nabla} \hat{v}(\hat{x})$. Die Elemente der Element-Matrizen A_T und M_T sowie des Element-Vektors b_T transformieren sich wie folgt:

$$\begin{aligned}
 a_{nm}^T &= \int_T a \nabla \varphi_h^{(m)} \nabla \varphi_h^{(n)} dx = |\det B_T| \int_{\hat{T}} \hat{a} \widehat{\nabla} \varphi_h^{(m)} \widehat{\nabla} \varphi_h^{(n)} d\hat{x} \\
 &= |\det B_T| \int_{\hat{T}} \hat{a} B_T^{-T} \hat{\nabla} \hat{\varphi}_h^{(m)} B_T^{-T} \hat{\nabla} \hat{\varphi}_h^{(n)} d\hat{x} =: |\det B_T| \hat{a}_{nm}^T, \\
 m_{nm}^T &= \int_T \varphi_h^{(m)} \varphi_h^{(n)} dx = |\det B_T| \int_{\hat{T}} \hat{\varphi}_h^{(m)} \hat{\varphi}_h^{(n)} d\hat{x} =: |\det B_T| \hat{m}_{nm}, \\
 b_n^T &= \int_T f \varphi_h^{(n)} dx = |\det B_T| \int_{\hat{T}} \hat{f} \hat{\varphi}_h^{(n)} d\hat{x} =: |\det B_T| \hat{b}_n.
 \end{aligned}$$

Die Werte \hat{a}_{nm}^T , \hat{m}_{nm}^T und \hat{b}_n^T auf der Referenzzelle werden nun mit Hilfe von Quadraturformeln auf \hat{T} berechnet. Dazu werden wir weiter unten noch mehr Details angeben. Wichtig ist, dass diese Quadratur nur auf der Referenzzelle stattfindet und die tatsächlich verwendeten Größen a_{nm} , m_{nm} und b_n im wesentlichen durch Skalierung mit $|\det B_T|$ gewonnen werden.

3.5.2 Konditionierung der Systemmatrix

Wir wollen zunächst die Stabilität (für $h \rightarrow 0$) der diskreten Finite-Elemente-Gleichung (3.5.121) gegenüber Störungen der Daten untersuchen. Diese treten z. B. auf durch die Fehler bei der Berechnung der Elemente a_{nm} und m_{nm} bei Verwendung von numerischer Intergration. Durch rein algebraische Argumente erhalten wir zunächst eine Stabilitätsabschätzung für allgemeine lineare Gleichungssysteme (s. den Band „Numerik 0 – Einführung in die Numerische Mathematik“).

Satz 3.12 (Allgemeiner Störungssatz): *Seien Störungen δA der Matrix A und δb der rechten Seite b gegeben, so dass $\mu := \kappa_2(A) \|\delta A\| / \|A\| < 1$. Dann gilt die Fehlerabschätzung*

$$\frac{|\delta \xi|}{|\xi|} \leq \frac{\kappa_2(A)}{1 - \mu} \left\{ \frac{\|\delta A\|}{\|A\|} + \frac{|\delta b|}{|b|} \right\}. \tag{3.5.122}$$

Zur quantitativen Auswertung dieser Abschätzung müssen wir die Spektralkondition der Steifigkeitsmatrix A in Abhängigkeit von der Gitterweite h abschätzen. Dazu nehmen wir an, dass die betrachtete Familie von Zerlegungen $(\mathbb{T}_h)_{h>0}$ gleichmäßig „form- und größen-regulär“ ist, d. h.:

$$\sup_{h>0} \left(\max_{T \in \mathbb{T}_h} \frac{h_T}{\rho_T} \right) \leq c, \quad \sup_{h>0} \left(\frac{\max_{T \in \mathbb{T}_h} h_T}{\min_{T \in \mathbb{T}_h} h_T} \right) \leq c.$$

Dann ergibt sich analog zum Differenzenverfahren das folgende allgemeine Resultat.

Satz 3.13 (Konditionierung): *Auf einer Folge von (gleichmäßig) regulären Zerlegungen \mathbb{T}_h gilt für die Spektralkonditionen der (symmetrischen und positiv definiten) Steifigkeitsmatrizen A und der Massenmatrizen M :*

$$\kappa_2(A) = \mathcal{O}(h^{-2}), \quad \kappa_2(M) = \mathcal{O}(1) \quad (h \rightarrow 0). \quad (3.5.123)$$

Beweis: i) Für die größten und kleinsten Eigenwerte von M gilt

$$\lambda_{\min}(M) = \min_{\xi \in \mathbb{R}^N} \frac{\langle M\xi, \xi \rangle}{|\xi|^2} \leq \max_{\xi \in \mathbb{R}^N} \frac{\langle M\xi, \xi \rangle}{|\xi|^2} = \lambda_{\max}(M).$$

In der folgenden Argumentation werden wieder die Element-Matrizen M_T . Ferner bezeichne d_{\min} und d_{\max} die kleinste bzw. die größte Anzahl von Zellen, die in einem Knoten der Zerlegung \mathbb{T}_h zusammentreffen. Mit dieser Notation ergibt sich:

$$\begin{aligned} \langle M\xi, \xi \rangle &= \sum_{T \in \mathbb{T}_h} \langle M_T \xi_T, \xi_T \rangle \geq \min_{\xi \in \mathbb{R}^N, T \in \mathbb{T}_h} \frac{\langle M_T \xi_T, \xi_T \rangle}{|\xi_T|^2} \sum_{T \in \mathbb{T}_h} |\xi_T|^2 \geq \min_{T \in \mathbb{T}_h} \{\lambda_{\min}(M_T)\} d_{\min} |\xi|^2, \\ \langle M\xi, \xi \rangle &= \sum_{T \in \mathbb{T}_h} \langle M_T \xi_T, \xi_T \rangle \leq \max_{\xi \in \mathbb{R}^N, T \in \mathbb{T}_h} \frac{\langle M_T \xi_T, \xi_T \rangle}{|\xi_T|^2} \sum_{T \in \mathbb{T}_h} |\xi_T|^2 \leq \max_{T \in \mathbb{T}_h} \{\lambda_{\max}(M_T)\} d_{\max} |\xi|^2. \end{aligned}$$

Mit Hilfe der Beziehung $|\det B_T| \approx h_T^d$ ergibt sich mit der (festen) Matrix $\hat{M} := (\hat{m}_{nm})_{n,m=1}^N$:

$$\lambda_{\max}(M_T) = |\det B_T| \lambda_{\max}(\hat{M}) \leq ch_T^d, \quad \lambda_{\min}(M_T) = |\det B_T| \lambda_{\min}(\hat{M}) \geq ch_T^d,$$

und folglich $\kappa_2(M) = \mathcal{O}(1)$.

ii) Für die kleinsten und größten Eigenwerte von A gilt:

$$\begin{aligned} \lambda_{\min}(A) &\geq \min_{\xi \in \mathbb{R}^N} \frac{\langle A\xi, \xi \rangle}{\langle M\xi, \xi \rangle} \min_{\xi \in \mathbb{R}^N} \frac{\langle M\xi, \xi \rangle}{|\xi|^2} = \min_{v_h \in V_h} \frac{a(v_h, v_h)}{\|v_h\|^2} \lambda_{\min}(M), \\ \lambda_{\max}(A) &\leq \max_{\xi \in \mathbb{R}^N} \frac{\langle A\xi, \xi \rangle}{\langle M\xi, \xi \rangle} \max_{\xi \in \mathbb{R}^N} \frac{\langle M\xi, \xi \rangle}{|\xi|^2} = \max_{v_h \in V_h} \frac{a(v_h, v_h)}{\|v_h\|^2} \lambda_{\max}(M). \end{aligned}$$

Wir schätzen weiter ab durch

$$\min_{v_h \in V_h} \frac{a(v_h, v_h)}{\|v_h\|^2} \geq \min_{v \in H_0^1(\Omega)} \frac{a(v, v)}{\|v\|^2} =: \lambda_{\min}(L)$$

mit dem kleinsten Eigenwert des Differentialoperators L auf dem Gebiet Ω . Ferner ergibt sich mit Hilfe der „inversen Beziehung“ für Finite-Elemente-Funktionen:

$$a(v_h, v_h) \leq \|a\|_{\infty} \sum_{T \in \mathbb{T}_h} \|\nabla v_h\|_T^2 \leq c \|a\|_{\infty} \sum_{T \in \mathbb{T}_h} \rho_T^{-2} \|v_h\|_T^2 \leq c \|a\|_{\infty} \max_{T \in \mathbb{T}_h} \rho_T^{-2} \|v_h\|^2,$$

bzw. $\lambda_{\max}(A) \leq c \max_{T \in \mathbb{T}_h} \rho_T^2 \lambda_{\max}(M)$. Wir gewinnen so die Abschätzung

$$\lambda_{\min}(L)\lambda_{\min}(M) \leq \lambda_{\min}(A) \leq \lambda_{\max}(A) \leq c \max_{T \in \mathbb{T}_h} \rho_T^{-2} \lambda_{\max}(M).$$

Also ist $\kappa_2(A) \leq c \max_{T \in \mathbb{T}_h} \rho_T^{-2}$, was im Hinblick auf die Gleichförmigkeitsannahmen an die Zerlegungsfolge den Beweis vervollständigt. Q.E.D.

Bemerkung 3.10: Wir betonen, dass die Asymptotik $O(h^{-2})$ der Kondition der Steifigkeitsmatrix durch die Ordnung des zugrunde liegenden Differentialoperators L bestimmt ist; sie hat *nichts* mit dem Polynomgrad $m - 1$ des Finite-Elemente-Ansatzes oder der Raumdimension d zu tun. In der Tat wurde der Beweis von Satz 3.13 allgemein für $d \geq 1$ und für beliebigen Polynomgrad $m - 1 \geq 1$ geführt. Die Abhängigkeit von $\rho := \min_{T \in \mathbb{T}_h} \rho_T$ kommt über die Verwendung der inversen Beziehung zur Abschätzung von $\lambda_{\max}(A)$ ins Spiel. Dabei ergibt offenbar jede Ableitungsstufe in der Energieform $a(\cdot, \cdot)$ genau eine negative ρ -Potenz. Der Exponent -2 ist also gerade durch die Ordnung des betrachteten Differentialoperators bestimmt. Bei der Finite-Elemente-Diskretisierung von Differentialoperatoren höherer Ordnung $2r \geq 2$ verhält sich die Kondition der Steifigkeitsmatrix dementsprechend wie $\kappa_2(A) = O(h^{-2r})$. Zum Beispiel treten in der Plattenstatik Randwertaufgaben vierter Ordnung ($r = 2$) mit dem biharmonischen Operator Δ^2 auf. In diesem Fall verhält sich die Kondition der zugehörigen Steifigkeitsmatrix wie $O(h^{-4})$. Für eine Gitterweite der Größenordnung $h \sim 10^{-2}$ in zwei Dimensionen ergibt sich damit $\kappa_2(A) \sim 10^8$, was Rechnung in mindestens doppelt-genauer Arithmetik nahe legt.

Die Fehlerabschätzung (3.5.122) ist am Extremfall einer Störung des Eigenvektors w_{\max} in Richtung des Eigenvektors w_{\min} orientiert. Sie erfassen also den ungünstigsten Fehlereinfluss, wie er in der Praxis kaum auftreten wird. Tatsächlich erweist sich (3.5.122) als viel zu pessimistisch zur realistischen Erfassung des Einflusses von Datenfehlern bei der Lösung der Finite-Elemente-Gleichungen $Ax = b$. Zur Verdeutlichung betrachten wir im folgenden ausschließlich den Fall von Störungen in der rechten Seite b , welche durch fehlerhafte Auswertung (z. B. durch numerische Integration) der gegebenen rechten Seite f der Differentialgleichung entstehen.

Satz 3.14 (Spezieller Störungssatz): *Auf einer Folge von (gleichmäßig) regulären Zerlegungen \mathbb{T}_h gilt die Fehlerabschätzung*

$$\frac{|\delta\xi|}{|\xi|} \leq \frac{\kappa_2(M)}{\lambda} \frac{\|f\|}{\|u_h\|} \frac{|\delta b|}{|b|}, \quad (3.5.124)$$

mit dem kleinsten Eigenwert λ der 1. RWA des Differentialoperators L auf Ω .

Beweis: Aus der Identität $A \delta \xi = \delta b$ folgt $|\delta \xi| \leq \|A^{-1}\| |\delta b| = \lambda_{\min}(A)^{-1} |\delta b|$. Weiter ist

$$\begin{aligned} \lambda_{\min}(A) &= \min_{z \in \mathbb{R}^N} \frac{\langle Az, z \rangle}{|z|^2} \geq \min_{z \in \mathbb{R}^N} \frac{\langle Az, z \rangle}{\langle Mz, z \rangle} \min_{z \in \mathbb{R}^N} \frac{\langle Mz, z \rangle}{|z|^2} \\ &= \min_{v_h \in V_h} \frac{a(v_h, v_h)}{\|v_h\|^2} \lambda_{\min}(M) \geq \lambda \lambda_{\min}(M), \end{aligned}$$

und somit $|\delta \xi| \leq \lambda^{-1} \lambda_{\min}(M)^{-1} |\delta b|$. Mit Hilfe der Schwarzischen Ungleichung ergibt sich

$$|b|^2 = \sum_{n=1}^N (f, \varphi_h^{(n)})^2 = (f, \sum_{n=1}^N (f, \varphi_h^{(n)}) \varphi_h^{(n)}) \leq \|f\| \left\| \sum_{n=1}^N (f, \varphi_h^{(n)}) \varphi_h^{(n)} \right\|.$$

Wegen

$$\left\| \sum_{n=1}^N (f, \varphi_h^{(n)}) \varphi_h^{(n)} \right\|^2 = \sum_{n,m=1}^N (f, \varphi_h^{(m)}) (f, \varphi_h^{(n)}) (\varphi_h^{(m)}, \varphi_h^{(n)}) = \langle Mb, b \rangle \leq \lambda_{\max}(M) |b|^2$$

folgt dann $|b| \leq \lambda_{\max}(M)^{1/2} \|f\|$. Ferner gilt wegen $\langle M\xi, \xi \rangle \leq \lambda_{\max}(M) \|\xi\|^2$:

$$|\xi| \geq \lambda_{\max}(M)^{-1/2} \langle M\xi, \xi \rangle^{1/2} = \lambda_{\max}(M)^{-1/2} \|u_h\|.$$

Wir kombinieren die obigen Beziehungen und erhalten

$$\frac{|\delta \xi|}{|\xi|} \leq \lambda^{-1} \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)} \frac{|\delta b|}{|b|} \frac{\|f\|}{\|u_h\|},$$

was zu beweisen war. Q.E.D.

Wir haben in Satz 3.13 gesehen, dass die Massenmatrix M eine gleichmäßig bzgl. h beschränkte Spektralkondition hat $\kappa_2(M) = \mathcal{O}(1)$ ($h \rightarrow 0$). Ferner folgt aus der Konvergenz des Verfahrens die Beziehung

$$\|u_h\| = \|u\| + \mathcal{O}(h) \quad (h \rightarrow 0).$$

Damit erhalten wir aus Satz 3.14 die folgende asymptotische Abschätzung für die Fehlerfortpflanzung im Finite-Elemente-Galerkin-Verfahren

$$\frac{|\delta \xi|}{|\xi|} \leq c(f, u, \Omega) \frac{|\delta b|}{|b|}, \quad (3.5.125)$$

mit einer nur von f , u und Ω abhängigen Konstante $c(f, u, \Omega)$. Dies besagt, dass die Finite-Elemente-Methode stabil ist (für $h \rightarrow 0$) bzgl. Störungen der rechten Seite f . Die Ausdehnung dieser Aussage auf Störungen in der Matrix selbst ist noch offen.

3.5.3 Aufstellung der Systemmatrizen mit numerischer Integration

Im folgenden analysieren wir den zusätzlichen Fehler, welcher bei näherungsweise Berechnung der Matrix- und Vektorelemente a_{nm} und b_n mittels numerischer Quadratur entsteht. Im Zuge der Matrix-Assemblierung müssen Integrale über Gitterzellen $T \in \mathbb{T}_h$,

$$\int_T a(x) \nabla \varphi_h^{(i)}(x) \nabla \varphi_h^{(j)}(x) dx, \quad \int_T f(x) \varphi_h^{(j)}(x) dx, \quad (3.5.126)$$

berechnet werden, wobei $\varphi_h^{(i)}(x)$ die Knotenbasisfunktionen sind. Wenn die Datenfunktionen $a(x)$, $f(x)$ konstant oder auf der Zelle durch Polynome approximiert sind, so sind Integrale der Form

$$\int_T x_1^p x_2^q dx, \quad p, q \in \mathbb{N}_0,$$

zu berechnen. Hierfür gibt es z. B. auf Dreiecken explizite Formeln. Sei T ein Dreieck in der (x, y) -Ebene mit den Eckpunkten (x_i, y_i) , $i = 1, 2, 3$. Ist der Ursprung des Koordinatensystems im Schwerpunkt von T , d.h. $x_1 + x_2 + x_3 = y_1 + y_2 + y_3 = 0$, so gilt z. B.:

$$\begin{aligned} \int_T 1 d(x, y) &= |T|, \\ \int_T x d(x, y) &= \int_T y d(x, y) = 0, \\ \int_T x^2 d(x, y) &= \frac{|T|}{12} (x_1^2 + x_2^2 + x_3^2), \\ \int_T xy d(x, y) &= \frac{|T|}{12} (x_1 y_1 + x_2 y_2 + x_3 y_3), \\ \int_T y^2 d(x, y) &= \frac{|T|}{12} (y_1^2 + y_2^2 + y_3^2). \end{aligned}$$

Nicht exakt berechenbare Integrale werden durch numerische Integration angenähert. Dazu dienen sog. „Quadraturformeln“, welche analog wie auf 1-dimensionalen Intervallen auch auf 2- oder 3-dimensionalen Zellen über einen Interpolationsansatz erzeugt werden. Dies geschieht zunächst auf der Referenzzelle \hat{T} und ergibt dann durch Transformation $\hat{T} \rightarrow T = \sigma_T(\hat{T}) = B_T \hat{x} + b_T$ auch Quadraturformeln auf den einzelnen Zellen $T \in \mathbb{T}_h$. Wir beschreiben diesen Prozess im folgenden nur für Quadratur basierend auf Lagrange-Interpolation.

Auf der Referenzzelle \hat{T} seien ein Polynomraum $P(\hat{T})$ mit $S := \dim P(\hat{T})$ sowie ein Satz von Stützpunkten $\{\hat{x}_s \in \hat{T}, s = 1, \dots, S\}$ gewählt, welche unisolvent sind. Die Stützpunkte \hat{x}_s brauchen nicht mit den Knotenpunkten des Finite-Elemente-Ansatzes übereinzustimmen. Seien weiter $\hat{L}_s \in P(\hat{T})$ die zugehörigen Lagrangeschen Basispolynome, welche durch die Eigenschaft $\hat{L}_s(\hat{x}_r) = \delta_{sr}$ charakterisiert sind. Dies erlaubt wieder die explizite Darstellung des zu einer stetigen Funktion $\hat{v}(\hat{x})$ gehörenden Interpolations-

polynoms durch

$$p(\hat{x}) = \sum_{s=1}^S \hat{v}(\hat{x}_s) \hat{L}_s(\hat{x}).$$

Dies führt zu folgendem Ansatz für eine Quadraturformel auf \hat{T} :

$$Q_{\hat{T}}(\hat{v}) := \sum_{s=1}^S \hat{\omega}_s \hat{v}(\hat{x}_s), \quad \hat{\omega}_s := \int_{\hat{T}} \hat{L}_s(\hat{x}) d\hat{x}. \quad (3.5.127)$$

Die Stützpunkte \hat{x}_s werden so gewählt, dass Polynome von möglichst hohem Grad durch die Formel exakt integriert werden. Dabei ist aus Stabilitätsgründen wieder darauf zu achten, dass die Gewichte $\hat{\omega}_s$ positiv sind.

Mit den Bezeichnungen

$$x_s := \sigma_T(\hat{x}_s), \quad \omega_s := |\det B_T| \hat{\omega}_s \quad (s = 1, \dots, S),$$

erhalten wir durch

$$Q_T(v) := \sum_{s=1}^S \omega_s v(x_s) := \sum_{s=1}^S |\det B_T| \hat{\omega}_s \hat{v}(\hat{x}_s) = |\det B_T| Q_{\hat{T}}(\hat{v}) \quad (3.5.128)$$

Quadraturformeln $Q_T(\cdot)$ auf den einzelnen Zellen $T \in \mathbb{T}_h$. Die gebräuchlichsten solcher Formeln für Dreiecke sind in Abb. 3.22 zusammengestellt. Dabei bedienen wir uns der gebräuchlichen Schreibweise mit sog. „baryzentrischen Koordinaten“. Für ein d -Simplex T mit Eckpunkten $\{a_0, \dots, a_d\}$ besitzt jeder Punkt $x \in T$ eine eindeutige Darstellung als konvexe Linearkombination der Eckpunkte:

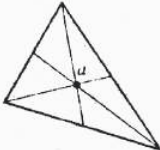
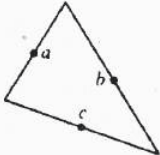
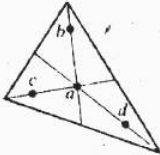
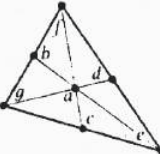
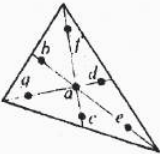
$$x = \sum_{i=0}^d \lambda_i a_i, \quad 0 \leq \lambda_i \leq 1, \quad \sum_{i=0}^d \lambda_i = 1.$$

Die Koeffizienten $\{\lambda_0, \dots, \lambda_d\}$ sind dann die baryzentrischen Koordinaten von x im Simplex T . Zum Beispiel sind $\{1, 0, \dots, 0\}$ die baryzentrischen Koordinaten des Eckpunkts a_0 und $\{\frac{1}{d+1}, \dots, \frac{1}{d+1}\}$ die des Mittelpunktes z_T . Die einfachsten Quadraturformeln auf Dreiecken/Simplizes sind die „Mittelpunktregel“ und die „Trapezregel“ (s. auch Abb. 3.22):

$$Q_T(v) := |T|v(z_T), \quad Q_T(v) := \frac{1}{d+1}|T| \sum_{i=0}^d v(a_i).$$

Definition 3.6 (Quadraturformel): Eine interpolatorische Quadraturformel der Art (3.5.127) auf einer Referenzzelle T heißt „von der Ordnung r “, wenn durch sie Polynome bis zum Grad $r-1$ (und nicht höher) exakt integriert werden. Sie wird „zulässig“ für den Polynomansatz $P(T)$ genannt, wenn ihre Stützstellenmenge reichhaltig genug ist, so dass

$$q \in P(T) : \quad \nabla q(x_s) = 0 \quad (s = 1, \dots, S) \quad \Rightarrow \quad q \equiv \text{konst.} \quad (3.5.129)$$

NUMERICAL INTEGRATION FORMULAS FOR TRIANGLES					
Order	Fig.	Error	Points	Triangular Co-ordinates	Weights $2W_k$
Linear		$R = O(h^2)$	a	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	1
Quadratic		$R = O(h^3)$	a b c	$\frac{1}{3}, \frac{1}{3}, 0$ $0, \frac{1}{3}, \frac{1}{3}$ $\frac{1}{2}, 0, \frac{1}{2}$	$\frac{1}{3}$ $\frac{1}{3}$ $\frac{1}{3}$
Cubic		$R = O(h^4)$	a b c d	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ $\frac{1}{4}, \frac{1}{4}, \frac{1}{2}$ $\frac{1}{8}, \frac{1}{8}, \frac{1}{2}$ $\frac{2}{15}, \frac{2}{15}, \frac{1}{3}$	$-\frac{27}{48}$ $\frac{27}{48}$
This formula not recommended due to negative weight and round-off error					
Cubic		$R = O(h^4)$	a b c d e f g	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ $\frac{1}{2}, \frac{1}{2}, 0$ $0, \frac{1}{3}, \frac{1}{3}$ $\frac{1}{2}, 0, \frac{1}{2}$ $1, 0, 0$ $0, 1, 0$ $0, 0, 1$	$\frac{27}{60}$ $\frac{3}{60}$ $\frac{3}{60}$
Quintic		$R = O(h^6)$	a b c d e f g	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ $\alpha_1, \beta_1, \beta_1$ $\beta_1, \alpha_1, \beta_1$ $\beta_1, \beta_1, \alpha_1$ $\alpha_2, \beta_2, \beta_2$ $\beta_2, \alpha_2, \beta_2$ $\beta_2, \beta_2, \alpha_2$	0.225 0.13239415 0.12593918

with
 $\alpha_1 = 0.05971587$
 $\beta_1 = 0.47014206$
 $\alpha_2 = 0.79742699$
 $\beta_2 = 0.10128651$

Abbildung 3.22: Beispiele von Quadraturformeln auf Dreiecken.

Auf Vierecken bzw. Hexaedern werden sog. „Tensorproduktformeln“ verwendet. Diese erhält man ebenfalls über Transformation von der Referenzzelle. Demgemäß lauten auf dem Quadrat/Quader mit Mittelpunkt z_T und Eckpunkten $\{a_1, \dots, a_{2^d}\}$ die Mittelpunktregel sowie die (Tensorprodukt)-Trapezregel:

$$Q_T(v) := |T|v(z_T), \quad Q_T(v) := \frac{1}{2^d}|T| \sum_{i=0}^{2^d} v(a_i).$$

Zur Erzielung höherer Genauigkeit werden extrapolierte Trapez-Formeln (z. B. Tensorprodukt-Simpson-Regel) oder Tensorprodukt-Gauß-Formeln verwendet, welche ähnlich wie im eindimensionalen Fall konstruiert sind. Ein der gebräuchlichsten Formeln für Quadrate ist die 4-Punkt-Gauß-Formel

$$Q_T(v) := \frac{1}{4} \sum_{i=1}^4 v(\xi_i),$$

mit den Gauß-Punkten $\xi_1 = (\eta_-, \eta_-)$, $\xi_2 = (\eta_-, \eta_+)$, $\xi_3 = (\eta_+, \eta_-)$ und $\xi_4(\eta_+, \eta_+)$, wobei $\eta_{\pm} := \frac{1}{2}(1 \pm \sqrt{1/3})$.

Satz 3.15 (Quadraturfehler): Für eine interpolatorische Quadraturformel $Q_T(\cdot)$ der Ordnung $r \geq d$ auf einer Zelle $T \in \mathbb{T}_h$ angewendet auf eine Funktion $v \in H^{r,1}(T)$ gilt

$$\left| \int_T v \, dx - Q_T(v) \right| \leq c_q h_T^r \int_T |\nabla^r v| \, dx, \quad (3.5.130)$$

mit einer von T und $v \in H^{r,1}(T)$ unabhängigen Quadraturkonstanten $c_q > 0$.

Beweis: Der Beweis verwendet das Bramble-Hilbert-Lemma 3.5 in Verbindung mit dem Transformationsargument von Satz 3.6. Auf der Referenzzelle \hat{T} definieren wir das Fehlerfunktional

$$F(\hat{v}) := \left| \int_{\hat{T}} \hat{v}(\hat{x}) \, d\hat{x} - Q_{\hat{T}}(\hat{v}) \right|.$$

Zur Definition von $F(\cdot)$ benötigen wir Punktwerte von \hat{v} . Diese sind aufgrund des Sobolewschen Einbettungssatzes in zwei Dimensionen für $v \in H^{2,1}(\Omega)$ und in drei Dimensionen für $v \in H^{3,1}(\Omega)$ wohl definiert. Es gilt dann

$$|F(\hat{v})| \leq c \|\hat{v}\|_{H^{r,1}}.$$

Ferner ist $F(\cdot)$ offensichtlich sublinear und verschwindet nach Voraussetzung auf P_{r-1} . Nach der L^1 -Variante des Bramble-Hilbert-Lemmas gilt dann

$$|F(\hat{v})| \leq c \|\hat{\nabla}^r \hat{v}\|_{L^1(\hat{T})}.$$

Sei nun σ_T wieder die affin-lineare Transformation von \hat{T} auf die Zelle T . Dann gilt für eine Funktion $v \in H^{r,1}(\Omega)$ und ihre Transformierte $\hat{v}(\hat{x}) := v(x)$, $x = \sigma_T(\hat{x})$:

$$\left| \int_T v(x) \, dx - Q_T(v) \right| = |\det B_T| |F(\hat{v})|,$$

sowie

$$\|\hat{\nabla}^r \hat{v}\|_{L^1(\hat{T})} \leq c |\det B_T|^{-1} h_T^r \|\nabla^r v\|_{L^1(T)}.$$

Kombination der letzten beiden Beziehungen ergibt die Behauptung.

Q.E.D.

Bemerkung 3.11: Die Voraussetzung $r \geq d$ in Satz 3.15 dient zur Vereinfachung der Formulierung des Resultats. Wegen der Einbettung $H^{d,1}(T) \subset C(\bar{T})$ sind für Funktionen

$v \in H^{r,1}(T)$ die für die Anwendung der Quadraturformel erforderlichen Punktwerte wohl definiert. Dies ist aber auch gerade die „richtige“ Regularitätsstufe für die maximale Ausnutzung der Approximationsgüte der Quadraturformel. Dies ist wichtig für die folgende Untersuchung des Einflusses des Quadraturfehlers auf den Gesamtfehler. Im Übrigen ist jede der gebräuchlichen Quadraturformeln (z. B. Mittelpunkts- oder Trapezregel) mindestens von der Ordnung $r = 2$, so dass in zwei Dimensionen gar keine Einschränkung besteht. In drei Dimensionen bedarf der Fall $r = 2$ eine gesonderte Betrachtung, die obwohl nicht schwer, hier nicht durchgeführt wird.

Im Folgenden werden wir uns auf die Quadratur auf Dreiecken bzw. Simplexes beschränken und nehmen außerdem an, dass der Polynomansatz ein voller Polynomraum ist: $P(\hat{T}) = P_{m-1}(\hat{T})$. Analoge Resultate gelten auch für die Quadratur auf Vierecken bzw. Hexaedern, doch erfordern deren Formulierung und Beweis eine aufwendigere Notation.

Die Anwendung von Quadraturformeln zur Berechnung der Zellintegrale (3.5.126) ergibt eine gestörte Bilinearform und rechte Seite

$$a_h(u, \varphi) := \sum_{T \in \mathbb{T}_h} Q_T(a \nabla u \nabla \varphi), \quad l_h(\varphi) := \sum_{T \in \mathbb{T}_h} Q_T(f \varphi)$$

sowie zugehörige gestörte Steifigkeitsmatrix- und Lastvektorelemente

$$\tilde{a}_{ij} := \sum_{T \in \mathbb{T}_h} Q_T(a \nabla \varphi_h^{(j)} \nabla \varphi_h^{(i)}), \quad \tilde{b}_j := \sum_{T \in \mathbb{T}_h} Q_T(f \varphi_h^{(j)}).$$

Statt der exakten Finite-Elemente-Lösung $u_h \in V_h$ ist dann eine gestörte Approximation $\tilde{u}_h \in V_h$ zu bestimmen durch

$$a_h(\tilde{u}_h, \varphi_h) = l_h(\varphi_h) \quad \forall \varphi_h \in V_h. \tag{3.5.131}$$

Satz 3.16 (Numerische Integration): *Die Quadraturformel auf der Referenzzelle $Q_{\hat{T}}(\cdot)$ sei „zulässig“ für den Finite-Elemente-Ansatz $P(\hat{T})$ und von der Ordnung $r \geq d$. Dann besitzen die gestörten Finite-Elemente-Gleichungen (3.5.131) eindeutige Lösungen $\tilde{u}_h \in V_h$. Ist $a \in C^r(\bar{\Omega})$, so gilt für das gestörte Finite-Elemente-Verfahren der Ordnung $m \geq 2$ die Fehlerabschätzung*

$$\|u - \tilde{u}_h\| + h \|\nabla(u - \tilde{u}_h)\| \leq c h^{\min\{m, r+3-m\}} \|u\|_{H^m}. \tag{3.5.132}$$

Zur Erzielung einer maximalen Konvergenzordnung ist also $r \geq 2m - 3$ zu wählen.

Beweis: i) Koerzitivität: Auf einer Zelle $T \in \mathbb{T}_h$ gilt unter Verwendung der Transformation $T = \sigma_T(\hat{T})$ wegen $a \geq \alpha > 0$:

$$\begin{aligned}
Q_T(a|\nabla v_h|^2) &= \sum_{s=1}^S \omega_s a(x_s) |\nabla v_h(x_s)|^2 \geq \alpha \sum_{s=1}^S \omega_s |\nabla v_h(x_s)|^2 \\
&= \alpha \sum_{s=1}^S |\det B_T| \hat{\omega}_s |B_T^{-1} \hat{\nabla} \hat{v}_h(\hat{x}_s)|^2 \geq \alpha \sum_{s=1}^S |\det B_T| \hat{\omega}_s \|B_T\|^{-2} |\hat{\nabla} \hat{v}_h(\hat{x}_s)|^2 \\
&\geq \alpha c |\det B_T| h_T^{-2} \sum_{s=1}^S \hat{\omega}_s |\hat{\nabla} \hat{v}_h(\hat{x}_s)|^2,
\end{aligned}$$

wobei $\hat{v}_h(\hat{x}) = v_h(x)$, $\hat{x} = \sigma_T^{-1}(x)$. Analog gilt

$$\|\hat{\nabla} \hat{v}_h\|_{\hat{T}}^2 \geq c |\det B_T|^{-1} h_T^2 \|\nabla v_h\|_T^2.$$

Durch

$$\|\|\hat{v}_h\|\|_{\hat{T}} := \left(\sum_{s=1}^S \hat{\omega}_s |\hat{\nabla} \hat{v}_h(\hat{x}_s)|^2 \right)^{1/2}$$

ist auf dem Quotientenraum $P(\hat{T})/P_0$ eine Norm definiert. Dazu muss nur noch die Definitheit gezeigt werden. Aus $\|\|\hat{v}_h\|\|_{\hat{T}} = 0$ folgt offenbar $\hat{\nabla} \hat{v}_h(\hat{x}_s) = 0$ ($s = 1, \dots, S$), so dass wegen der vorausgesetzten „Zulässigkeit“ der Quadraturformel notwendig $\hat{v}_h \equiv \text{konst.}$ ist. Da auch $\|\|\hat{\nabla} \hat{v}_h\|\|_{\hat{T}}$ auf $P(\hat{T})/P_0$ eine Norm ist, gilt wegen der Äquivalenz aller Normen auf einem endlich dimensionalen Vektorraum mit einer Konstante $\hat{c} > 0$:

$$\|\|\hat{v}_h\|\|_{\hat{T}} \geq \hat{c} \|\|\hat{\nabla} \hat{v}_h\|\|_{\hat{T}}.$$

Dies impliziert dann

$$Q_T(a|\nabla v_h|^2) \geq c |\det B_T| h_T^{-2} \|\|\hat{v}_h\|\|_{\hat{T}}^2 \geq c |\det B_T| h_T^{-2} \|\|\hat{\nabla} \hat{v}_h\|\|_{\hat{T}}^2 \geq c \|\nabla v_h\|_T^2,$$

bzw. die gleichmäßige Koerzitivität der Bilinearformen $a_h(\cdot, \cdot)$:

$$a_h(v_h, v_h) \geq c \|\nabla v_h\|^2, \quad v_h \in V_h. \quad (3.5.133)$$

ii) Fehlerabschätzung: Mit Hilfe der Koerzitivitätsabschätzung (3.5.133) folgt für den Fehler $\eta_h := u_h - \tilde{u}_h \in V_h$:

$$\begin{aligned}
c \|\nabla \eta_h\|^2 &\leq a_h(u_h - \tilde{u}_h, \eta_h) = a_h(u_h, \eta_h) - a_h(\tilde{u}_h, \eta_h) \\
&\leq (a_h - a)(u_h, \eta_h) + (l - l_h)(\eta_h).
\end{aligned}$$

Dies impliziert

$$\|\nabla \eta_h\| \leq c \max_{v_h \in V_h} \left\{ \frac{|(a_h - a)(u_h, v_h)|}{\|\nabla v_h\|} + \frac{|(l - l_h)(v_h)|}{\|\nabla v_h\|} \right\}.$$

Zusammen mit der bekannten H^1 -Fehlerabschätzung

$$\|\nabla(u - u_h)\| \leq ch^m \|u\|_{H^m}$$

folgt

$$\|\nabla(u - \tilde{u}_h)\| \leq ch^m \|u\|_{H^m} + c \max_{v_h \in V_h} \left\{ \frac{|(a_h - a)(u_h, v_h)|}{\|\nabla v_h\|} + \frac{|(l - l_h)(v_h)|}{\|\nabla v_h\|} \right\}.$$

Im nächsten Schritt werden wir zeigen, dass

$$\max_{v_h \in V_h} \left\{ \frac{|(a_h - a)(u_h, v_h)|}{\|\nabla v_h\|} + \frac{|(l - l_h)(v_h)|}{\|\nabla v_h\|} \right\} \leq ch^{\min\{m, r-m+2\}} \|u\|_{H^m}. \quad (3.5.134)$$

Dies impliziert dann den ersten Teil der Behauptung

$$\|\nabla(u - \tilde{u}_h)\| \leq ch^{\min\{m, r-m+2\}} \|u\|_{H^m}. \quad (3.5.135)$$

iii) Konsistenz: Als nächstes schätzen wir den „Abstand“ zwischen $a(\cdot, \cdot)$ und $a_h(\cdot, \cdot)$ sowie $l(\cdot)$ und $l_h(\cdot)$ ab. Für $u_h, v_h \in V_h$ folgt mit Hilfe der Abschätzung des Integrationsfehlers in Satz 3.15:

$$\begin{aligned} |(a - a_h)(u_h, v_h)| &\leq \sum_{T \in \mathbb{T}_h} \left| \int_T a \nabla u_h \nabla v_h \, dx - Q_T(a \nabla u_h \nabla v_h) \right| \\ &\leq c_q \sum_{T \in \mathbb{T}_h} h_T^r \int_T |\nabla^r(a \nabla u_h \nabla v_h)| \, dx. \end{aligned}$$

Durch Ausdifferenzieren erhalten wir

$$|(a - a_h)(u_h, v_h)| \leq c \sum_{T \in \mathbb{T}_h} h_T^r \|\nabla u_h\|_{H^r(T)} \|\nabla v_h\|_{H^r(T)},$$

wobei die Konstante c wesentlich durch Schranken für die Ableitungen der Koeffizientenfunktion $a(x)$ bestimmt ist. Bei Beachtung von $v_h|_T \in P_{m-1}$ ergibt sich

$$|(a - a_h)(u_h, v_h)| \leq c \sum_{T \in \mathbb{T}_h} h_T^r \|\nabla u_h\|_{H^{m-2}(T)} \|\nabla v_h\|_{H^{m-2}(T)}.$$

Mit Hilfe der „inversen Beziehung“ für finite Elemente gilt

$$\|\nabla v_h\|_{H^{m-2}(T)} \leq ch_T^{2-m} \|\nabla v_h\|_T,$$

womit folgt:

$$|(a - a_h)(u_h, v_h)| \leq ch^{r-m+2} \left(\sum_{T \in \mathbb{T}_h} \|\nabla u_h\|_{H^{m-2}(T)}^2 \right)^{1/2} \|\nabla v_h\|.$$

Die Terme in u_h werden unter Verwendung der Knoteninterpolierenden $I_h u \in V_h$ und mit Hilfe der inversen Beziehung sowie der lokalen Interpolationsabschätzungen wie folgt abgeschätzt:

$$\begin{aligned}
\|\nabla u_h\|_{H^{m-2}(T)} &\leq \|u_h - I_h u\|_{H^{m-1}(T)} + \|I_h u - u\|_{H^{m-1}(T)} + \|u\|_{H^{m-1}(T)} \\
&\leq ch_T^{2-m} \|u_h - I_h u\|_{H^1(T)} + ch_T \|\nabla^m u\|_T + \|u\|_{H^{m-1}(T)} \\
&\leq ch_T^{2-m} \|u_h - u\|_{H^1(T)} + ch_T^{2-m} \|u - I_h u\|_{H^1(T)} + \|u\|_{H^m(T)} \\
&\leq ch_T^{2-m} \|u_h - u\|_{H^1(T)} + \|u\|_{H^m(T)}.
\end{aligned}$$

Zusammenfassen der bisherigen Abschätzungen ergibt

$$|(a - a_h)(u_h, v_h)| \leq ch^{r+2-m} \{h^{2-m} \|u_h - u\|_{H^1} + \|u\|_{H^m}\} \|\nabla v_h\|,$$

und unter Verwendung der bekannten H^1 -Fehlerabschätzung $\|u_h - u\|_{H^1} \leq ch^{m-1} \|u\|_{H^m}$:

$$|(a - a_h)(u_h, v_h)| \leq ch^{r+2-m} \|u\|_{H^m} \|\nabla v_h\|. \quad (3.5.136)$$

Auf analoge Weise erschließen wir

$$|(l - l_h)(v_h)| \leq ch^{r-m+2} \|\nabla v_h\| \|u\|_{H^m}. \quad (3.5.137)$$

Die Abschätzungen (3.5.136) und (3.5.137) ergeben (3.5.134) und damit das Resultat (3.5.135).

iv) L^2 -Fehlerabschätzung: Zur Abschätzung des Fehlers in der L^2 -Norm bedienen wir uns wieder eines Dualitätsarguments. Sei $z \in V$ die (eindeutige) Lösung des Hilfsproblems

$$-\nabla \cdot \{a \nabla z\} = u_h - \tilde{u}_h \quad \text{in } \Omega, \quad z = 0 \quad \text{auf } \partial\Omega.$$

Wegen der angenommenen Glattheit von a und der Konvexität von Ω ist $z \in H^2(\Omega)$ und genügt der a priori Abschätzung

$$\|z\|_{H^2} \leq c \|u_h - \tilde{u}_h\|.$$

Sei $z_h \in V_h$ die Ritz-Projektion von z . Mit dieser Konstruktion erhalten wir mit Hilfe der Galerkin-Orthogonalität:

$$\begin{aligned}
\|u_h - \tilde{u}_h\|^2 &= a(u_h - \tilde{u}_h, z) = a(u_h - \tilde{u}_h, z_h) \\
&= a(u_h, z_h) - (a - a_h)(\tilde{u}_h, z_h) - a_h(\tilde{u}_h, z_h) \\
&= (l - l_h)(z_h) - (a - a_h)(\tilde{u}_h, z_h).
\end{aligned}$$

Die beiden Störungsterme werden nun analog wie unter (iii) abgeschätzt. Zunächst gilt wieder mit Hilfe der inversen Beziehung:

$$\begin{aligned}
|(a - a_h)(\tilde{u}_h, z_h)| &\leq c \sum_{T \in \mathbb{T}_h} h_T^r \|\nabla \tilde{u}_h\|_{H^{m-2}(T)} \|\nabla z_h\|_{H^{m-2}(T)} \\
&\leq ch^{r+3-m} \left(\sum_{T \in \mathbb{T}_h} \|\nabla \tilde{u}_h\|_{H^{m-2}(T)}^2 \right)^{1/2} \left(\sum_{T \in \mathbb{T}_h} \|\nabla z_h\|_{H^1(T)}^2 \right)^{1/2}.
\end{aligned}$$

Für die beiden Summen erhalten wir analog wie oben unter (iii):

$$\left(\sum_{T \in \mathbb{T}_h} \|\nabla \tilde{u}_h\|_{H^{m-2}(T)}^2 \right)^{1/2} \leq ch_T^{2-m} \|\tilde{u}_h - u\|_{H^1} + \|u\|_{H^m},$$

sowie

$$\left(\sum_{T \in \mathbb{T}_h} \|\nabla z_h\|_{H^1(T)}^2 \right)^{1/2} \leq c \|z\|_{H^2} \leq c \|u_h - \tilde{u}_h\|.$$

Kombination dieser Abschätzungen und Berücksichtigung der schon bewiesenen H^1 -Fehlerabschätzung ergibt dann:

$$|(a - a_h)(\tilde{u}_h, z_h)| \leq ch^{r-m+3} \|u_h - \tilde{u}_h\| \|u\|_{H^m}. \quad (3.5.138)$$

Analog erschließen wir

$$|(l - l_h)(z_h)| \leq ch^{r-m+3} \|u_h - \tilde{u}_h\|. \quad (3.5.139)$$

Kombinieren aller vorausgehenden Beziehungen ergibt

$$\|u_h - \tilde{u}_h\| \leq ch^{r-m+3} \|u\|_{H^m}.$$

Zusammen mit der bekannten L^2 -Fehlerabschätzung für $e_h = u - u_h$ erhalten wir schließlich

$$\|u - \tilde{u}_h\| \leq ch^{\min\{m, r-m+3\}} \|u\|_{H^m},$$

was den Beweis vervollständigt.

Q.E.D.

Bemerkung 3.12: Die Aussage von Satz 3.16 kann so interpretiert werden, dass zur Vermeidung einer Reduzierung der Konvergenzordnung $\mathcal{O}(h^{m-1})$ in der Energie-Norm eine Quadraturformel der Ordnung $r \geq 2m - 3$ verwendet werden sollte. Damit würden dann im Falle eines konstanten Koeffizienten a die Matrixelemente a_{nm} exakt berechnet werden. Für bloße Konvergenz $\tilde{u}_h \rightarrow u$ ($h \rightarrow 0$) wäre die Wahl $r \geq \max\{d, m - 2\}$ ausreichend, vorausgesetzt die Zulässigkeitsbedingung ist erfüllt.

3.6 A posteriori Fehleranalyse und Gittersteuerung

Eine wichtige Rolle bei der Lösung von partiellen Differentialgleichungen spielen die beiden Aspekte Fehlerkontrolle und Gittersteuerung. Hat man eine approximative Lösung u_h berechnet, ist es von Interesse, den Fehler zwischen dieser Näherung und der exakten Lösung u bzgl. eines geeigneten Maßes abzuschätzen. Zu diesem Zweck dienen sog. „a posteriori Fehlerschätzer“, welche im besten Fall ausschließlich von berechneten Größen und den Daten f abhängen. In diesem Abschnitt wollen wir einfache residuen-basierte Fehlerschätzer für das Modellproblem der 1. RWA des Laplace-Operators herleiten.

Eine globale Verfeinerung des ganzen Rechengebietes ist in drei Dimensionen in der Regel nicht realisierbar, da der Speicherplatz auf den verfügbaren Rechnern dazu nicht

ausreicht. Deshalb versucht man, nur dort lokal zu verfeinern, wo es die Lösungsstruktur bzw. die Genauigkeitsanforderungen verlangen. Unter *optimaler Gittersteuerung* wird dabei verstanden, möglichst wenige markierte Elemente des Gitters zu verfeinern, so dass der Fehler in möglichst wenigen Schritten unter eine vorgegebene Toleranz gedrückt wird. Die Wahl der zu verfeinernden Elemente trifft man aufgrund sog. „lokaler Fehlerindikatoren“, aus denen sich der globale Fehlerschätzer zusammensetzt. Solche Gittersteuerungsmethoden werden wir im zweiten Teil dieses Abschnitts diskutieren.

3.6.1 Allgemeine a posteriori Fehlerabschätzung

Aus der a priori Fehlerschätzung aus Abschnitt 3.4 konnten wir die Konvergenzordnung des Diskretisierungsfehlers $e_h = u - u_h$ schon für verschiedene Normen herleiten:

- Energienorm-Fehlerabschätzung: $\|\nabla e_h\| \leq c_{ICS} h \|u\|_{H^2},$
- L^2 -Norm-Fehlerabschätzung: $\|e_h\| \leq c_{ICS} h^2 \|u\|_{H^2},$
- L^∞ -Abschätzung: $\max_{\overline{\Omega}} |e_h| \leq c_{ICS} h^2 L(h) M_2(u),$

mit *lokalen* Interpolationskonstanten c_I und *globalen* Stabilitätskonstanten c_S . Leider sind diese Abschätzungen für eine quantitative Fehlerkontrolle nicht zu gebrauchen, da die nötigen Informationen über die höheren Ableitungen der exakten Lösung u fehlen und insbesondere präzise Abschätzungen für die Stabilitätskonstanten c_S i. Allg. nicht zur Verfügung stehen. Ist aber der Charakter einer lokalen Singularität der Lösung bekannt, wie z. B. im Fall von „Ecken- oder Kantensingularitäten“, so kann diese Information zur vorab Anpassung des Gitters verwendet werden. Dabei wird die Gitterweite h in Richtung auf die singuläre Stelle hin systematisch verkleinert, etwa gemäß $h(r) \approx h_0 r^\alpha$ wobei der Exponent $\alpha > 1$ aus dem bekannten singulären Verhalten der Ableitungen der Lösung abgeleitet wird.

In allgemeinen Situationen muss man sich aber heuristischer Methoden zur Bestimmung der Regularität der Lösung bedienen. Dies läuft (in Anlehnung an die traditionelle, abschnidefehler-basierte Vorgehensweise bei Differenzenverfahren) auf die Schätzung der lokalen Glattheit der unbekanntenen Lösung aus der berechneten numerischen Approximation hinaus. Zum Beispiel kann man versuchen, auf einem Zellblock (etwa aus 2^d Zellen) aus einer linearen Näherungslösung u_h durch Anwendung eines Differenzenquotienten ∇_h^2 zweiter Ordnung im Schwerpunkt z_T einer Zelle $T \in \mathbb{T}_h$ eine Schätzung der zweiten Ableitungen von u auf T zu gewinnen:

$$\max_T |\nabla^2 u|_T \approx \eta_T := |\nabla_h^2 u_h(z_T)|_T. \quad (3.6.140)$$

Auf der Basis dieses Indikators ließen sich dann Strategien zu lokaler Gitterverfeinerung oder -vergrößerung aufstellen: Ist zum Beispiel η_T auf einer Zelle $T \in \mathbb{T}_h$ überdurchschnittlich groß, so wird diese in Teilzellen zerlegt. Diese Strategie der *ad hoc* Gitteranpassung erfordert keinen großen Aufwand und funktioniert in der Praxis in vielen Fällen erstaunlich gut. Daher sind Varianten dieser Strategie derzeit auch in vielen kommerziellen

Programmen realisiert, wenn diese überhaupt Gitteradaption beinhalten. Dabei bestehen aber die folgenden grundsätzlichen Schwächen:

- Die Auswertung von (3.6.140) liefert keine Aussage über die tatsächliche Größe des Fehlers $e_h = u - u_h$.
- Die auf den lokalen Indikatoren η_T basierende Gitterverfeinerungsstrategie geht davon aus, dass der „gemessene“ Fehler in T auch dort entstanden ist und durch lokale Verfeinerung von T reduziert werden kann. Dies ist aber i. Allg. nicht richtig, da dabei das Phänomen der globalen „Fehlerakkumulation“ (auch „pollution effect“ genannt) vernachlässigt wird.

Ziel der folgenden Diskussion ist es, über eine a posteriori Fehleranalyse via Dualitätsargumente systematisch auswertbare und (asymptotisch) zuverlässige Fehlerschätzer zu entwickeln, aus denen auch effiziente Kriterien zur lokalen Gitteranpassung abgeleitet werden können.

Wir betrachten dazu wieder das Modellproblem

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega, \quad (3.6.141)$$

auf einem (nicht notwendig konvexen) Polygon- oder Polyedergebiet $\Omega \subset \mathbb{R}^d$. Die durch die Variationsgleichung

$$(\nabla u, \nabla \varphi) = (f, \varphi) \quad \forall \varphi \in V = H_0^1(\Omega), \quad (3.6.142)$$

definierte schwache Lösung $u \in H_0^1(\Omega)$ wird durch ein Galerkin-Finite-Elemente-Verfahren approximiert. Wir konzentrieren uns im folgenden auf die Approximation mit „linearen“ finiten Elementen. Die Näherungslösung $u_h = R_h u \in V_h^{(1)} \subset V$ ist bestimmt durch die Gleichung

$$(\nabla u_h, \nabla \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h. \quad (3.6.143)$$

Bei der Schätzung des Fehlers e_h muss man sich für ein geeignetes „Fehlermaß“ entscheiden, welches sich am Bedarf der betrachteten Anwendung orientieren sollte. Beispiele sind etwa wie schon oben erwähnt die traditionellen Maße „Energienorm“, L^2 -Norm sowie L^∞ -Norm. Weitere Beispiele sind:

- Mittelwerte: $|(e_h, \psi)_\Omega|, \quad \psi \in C(\overline{\Omega}),$
- Linienintegrale: $|(e_h, \psi)_\Gamma|, \quad \psi \in C(\partial\Omega),$
- Ableitungswerte: $|\partial_i e_h(a)|, \quad a \in \Omega.$

Um alle diese Sonderfälle im Rahmen einer einheitlichen Theorie behandeln zu können, führen wir zunächst den Begriff des „Fehlerfunktional“ ein. Dies ist ein (der Einfachheit halber als linear angenommenes) Funktional

$$J(\cdot) : V \rightarrow \mathbb{R},$$

bzgl. dem der Fehler e_h geschätzt werden soll; d. h.: Gesucht ist eine berechenbare Schranke für $J(e_h) = J(u) - J(u_h)$. Zu einem solchen Fehlerfunktional gehört eine (eindeutige) „duale Lösung“ $z \in V$, welche als Lösung des zugehörigen „dualen Problems“ bestimmt ist:

$$(\nabla\varphi, \nabla z) = J(\varphi) \quad \forall \varphi \in V. \quad (3.6.144)$$

Testen wir in (3.6.144) mit $\varphi = e_h \in V$, so ergibt sich unter Berücksichtigung der Galerkin-Orthogonalität die Fehleridentität

$$J(e_h) = (\nabla e_h, \nabla z) = (\nabla e_h, \nabla(z - \psi_h)), \quad \psi_h \in V_h. \quad (3.6.145)$$

Dies wird nun weiter mit Hilfe von elementweise partieller Integration umgeformt zu

$$\begin{aligned} J(e_h) &= \sum_{T \in \mathbb{T}_h} \{ -(\Delta e_h, z - \psi_h)_T + (\partial_n e_h, z - \psi_h)_{\partial T} \} \\ &= \sum_{T \in \mathbb{T}_h} \{ (f + \Delta u_h, z - \psi_h)_T - (\partial_n u_h, z - \psi_h)_{\partial T} \}. \end{aligned}$$

Da $z - \psi_h$ stetige Spuren entlang ∂T hat und in der obigen Summe jede Kante zweimal auftritt mit wechselnder Richtung der Normalableitung $\partial_n u_h$, erhalten wir

$$J(e_h) = \sum_{T \in \mathbb{T}_h} \{ (f + \Delta u_h, z - \psi_h)_T - \frac{1}{2}([\partial_n u_h], z - \psi_h)_{\partial T} \}, \quad (3.6.146)$$

wobei auf „inneren“ Kanten $[\partial_n u_h]_{|\Gamma} := \partial_n u_h|_T + \partial_{n'} u_h|_{T'}$ jeweils den Sprung der Normalableitung über $\Gamma \subset \partial T$ zur Nachbarzelle bedeutet. Entlang von Kanten am Rand, $\Gamma \subset \partial\Omega$, setzen wir $[\partial_n u_h]_{|\Gamma} := 2\partial_n u_h$. Diese Fehleridentität beinhaltet Information über die Abhängigkeit des Fehlerterms $J(e_h)$ von den lokalen „Residuen“ $R_T := (f + \Delta u_h)|_T$ und $r_{\partial T} := -\frac{1}{2}[\partial_n u_h]_{|\partial T}$:

$$\frac{\partial J(e_h)}{\partial R_T} \approx (z - \psi_h)|_T, \quad \frac{\partial J(e_h)}{\partial r_{\partial T}} \approx (z - \psi_h)|_{\partial T}. \quad (3.6.147)$$

Damit können wir hoffen, auf das spezielle Zielfunktional $J(e_h)$ bezogene Kriterien für lokale Gitterverfeinerung und damit Reduzierung dieser Residuen zu gewinnen. Dabei „misst“ das „Zellresiduum“ $R(u_h)$ das Erfülltsein der Differentialgleichung durch die Näherungslösung u_h und das „Kantenresiduum“ $r(u_h)$ deren „Glattheit“.

Von der exakten *Fehlerdarstellung* (3.6.146) können wir nun in mehreren Schritten weitere, gegebenenfalls leichter auswertbare *Fehlerabschätzungen* ableiten. Zunächst ist

$$|J(e_h)| \leq \left| \sum_{T \in \mathbb{T}_h} \{ (f + \Delta u_h, z - \psi_h)_T - \frac{1}{2}([\partial_n u_h], z - \psi_h)_{\partial T} \} \right|, \quad (3.6.148)$$

und weiter

$$|J(e_h)| \leq \sum_{T \in \mathbb{T}_h} \left| (f + \Delta u_h, z - \psi_h)_T - \frac{1}{2}([\partial_n u_h], z - \psi_h)_{\partial T} \right|. \quad (3.6.149)$$

Bei Beachtung von $z - \psi_h|_{\partial\Omega} = 0$ folgt dann mit Hilfe der Hölderschen Ungleichung:

$$|J(e_h)| \leq \sum_{T \in \mathbb{T}_h} \left\{ \|f + \Delta u_h\|_T \|z - \psi_h\|_T + \frac{1}{2} \|[\partial_n u_h]\|_{\partial T \setminus \partial\Omega} \|z - \psi_h\|_{\partial T} \right\}. \quad (3.6.150)$$

Die letzte Abschätzung schreiben wir in der kompakten Form

$$|J(e_h)| \leq \eta(u_h) := \sum_{T \in \mathbb{T}_h} \left\{ \rho_T(u_h) \omega_T(z) + \rho_{\partial T}(u_h) \omega_{\partial T}(z) \right\} \quad (3.6.151)$$

mit den „Zellresiduentermen“

$$\rho_T(u_h) := \|f + \Delta u_h\|_T, \quad \rho_{\partial T}(u_h) := \frac{1}{2} h_T^{-1/2} \|[\partial_n u_h]\|_{\partial T \setminus \partial\Omega},$$

und den „Zellgewichten“

$$\omega_T(z) := \|z - I_h z\|_T, \quad \omega_{\partial T}(z) := h_T^{1/2} \|z - I_h z\|_{\partial T}.$$

Die Fehlerdarstellung (3.6.146) bzw. die Fehlerabschätzung (3.6.151) sind nicht unmittelbar auswertbar. Zwar sind die „Residuenterm“ $R(u_h) = f + \Delta u_h$ und $r(u_h) = -\frac{1}{2}[\partial_n u_h]$ aus der Näherung u_h berechenbar, doch die Wichtungsfaktoren $z - \psi_h$ sind nur implizit über das duale Problem (3.6.144) gegeben und müssen gesondert bestimmt werden. Mit dieser kritischen Frage werden wir uns im Folgenden noch eingehender beschäftigen.

Die Zellgewichte lassen sich bei Wahl $\psi_h := I_h z$ mit Hilfe der lokalen Interpolationsfehlerabschätzungen aus Abschnitt 3.3 weiter abschätzen durch

$$\omega_T(z) \leq c_I h_T^3 \max_T \{ |\nabla^2 z| \}. \quad (3.6.152)$$

Hierbei muss natürlich vorausgesetzt werden, dass die duale Lösung $z \in H^{2,\infty}(\Omega)$ ist. Wir werden später sehen, wie man sich von dieser sehr einschränkenden Annahme in gewisser Weise befreien kann. Die Interpolationskonstante c_I ist von verschiedenen Faktoren abhängig: Von dem Polynomgrad der verwendeten Interpolierenden, vom Referenzelement \hat{T} sowie von den Transformationen $\sigma_T : \hat{T} \rightarrow T$. Sie stellt einen Unsicherheitsfaktor dar. Genaueres Nachrechnen ergibt eine Größenordnung von $c_I \approx 0,1 - 10$.

Die a posteriori Fehlerabschätzung (3.6.151) ist „zuverlässig“ im Sinne, dass sie eine sichere obere Schranke für den Fehler $|J(e_h)|$ liefert, vorausgesetzt, es liegen zuverlässige Werte für die Gewichte $\omega_T(z)$ und $\omega_{\partial T}(z)$ vor. Sie wäre auch „effizient“, wenn für den zugehörigen „Effektivitätsindex“ („Überschätzungsfaktor“) gilt:

$$I_{\text{eff}} := \frac{\eta(u_h)}{|J(e_h)|} \sim 1 \quad (h \rightarrow 0). \quad (3.6.153)$$

Bemerkung 3.13: Es ist zu bemerken, dass bereits der Übergang von der „Fehleridentität“ (3.6.146) zur „Fehlerabschätzung“ (3.6.148) in gewissen Fällen zu einer groben Überschätzung des tatsächlichen Fehlers führen kann. Betrachten wir hierzu folgendes Beispiel: Auf dem Quadrat $\Omega = (-1, 1) \times (-1, 1)$ gilt es, für die Poisson-Gleichung den Punktfehler im Ursprung zu finden: $J(e_h) = e_h(0)$. Es ist möglich, die rechte Seite f und das Gitter \mathbb{T}_h so anti-symmetrisch bzgl. $x = 0$ zu konstruieren, dass $u(0) = 0 = u_h(0)$. In diesem Fall wird aber in der Regel $\eta(u_h) \approx h^2 > 0$ sein, d. h.: $I_{\text{eff}} = \infty$.

3.6.2 Spezielle a posteriori Fehlerschätzer

Der oben abgeleitete allgemeine Fehlerschätzer wird nun für verschiedene spezielle Fehlerfunktionale ausgewertet. In den betrachteten Fällen kann man die Gewichte $\omega_T(z)$, $\omega_{\partial T}(z)$ analytisch abschätzen. Wir beschränken uns hier auf die Betrachtung von P_1 -Elementen.

a) Energienorm-Fehlerschätzer:

Zur Abschätzung des „Energienormfehlers“ $\|\nabla e_h\|$ wählen wir das Fehlerfunktional

$$J(\varphi) := \|\nabla e_h\|^{-1}(\nabla\varphi, \nabla e_h),$$

so dass wir für $\varphi = e_h$ automatisch $J(e_h) = \|\nabla e_h\|$ haben. Für die zugehörige Lösung $z \in V$ des dualen Problems

$$(\nabla\varphi, \nabla z) = J(\varphi) \quad \forall \varphi \in V \quad (3.6.154)$$

gilt dann

$$\|\nabla z\|^2 = \|\nabla e_h\|^{-1}(\nabla z, \nabla e_h) \leq \|\nabla z\|,$$

woraus sich die einfache a priori Abschätzung $\|\nabla z\| \leq 1$ ergibt. Ausgehend von der allgemeinen Fehlerabschätzung (3.6.150) erhalten wir

$$\|\nabla e_h\| \leq \sum_{T \in \mathbb{T}_h} \left\{ \|f + \Delta u_h\|_T \|z - \psi_h\|_T + \frac{1}{2} \|[\partial_n u_h]\|_{\partial T \setminus \partial \Omega} \|z - \psi_h\|_{\partial T} \right\}.$$

Wir wählen nun $\psi_h := I_h z$ als eine „verallgemeinerte“ Knoteninterpolierende von z , für welche die folgende lokale Fehlerabschätzung gilt (Für deren technisch aufwendige Konstruktion verweisen wir auf die einschlägige Literatur, z. B. Brenner/Scott [13]):

$$\|z - I_h z\|_T + h_T^{1/2} \|z - I_h z\|_{\partial T} \leq \tilde{c}_i h_T \|\nabla z\|_{\tilde{T}}, \quad (3.6.155)$$

wobei $\tilde{T} := \cup\{T' \in \mathbb{T}_h : T' \cap T \neq \emptyset\}$. Damit folgt dann mit den Residuen $\rho_T(u_h)$ und $\rho_{\partial T}(u_h)$:

$$\begin{aligned} \|\nabla e_h\| &\leq \tilde{c}_i \sum_{T \in \mathbb{T}_h} h_T \{ \|f + \Delta u_h\|_T + \frac{1}{2} h_T^{-1/2} \|[\partial_n u_h]\|_{\partial T \setminus \partial \Omega} \} \|\nabla z\|_{\tilde{T}} \\ &\leq c \left(\sum_{T \in \mathbb{T}_h} h_T^2 \{ \rho_T(u_h)^2 + \rho_{\partial T}(u_h)^2 \} \right)^{1/2} \left(\sum_{T \in \mathbb{T}_h} \|\nabla z\|_T^2 \right)^{1/2}. \end{aligned}$$

Wegen

$$\left(\sum_{T \in \mathbb{T}_h} \|\nabla z\|_T^2 \right)^{1/2} \leq c \|\nabla z\| \leq c,$$

ergibt sich schließlich das folgende Resultat:

Satz 3.17 (a posteriori Energienormfehler): Für den Fehler $e_h := u - u_h$ gilt die a posteriori Abschätzung bzgl. der Energienorm:

$$\|\nabla e_h\| \leq \eta_E(u_h) := c \left(\sum_{T \in \mathbb{T}_h} h_T^2 \{ \rho_T(u_h)^2 + \rho_{\partial T}(u_h)^2 \} \right)^{1/2}. \quad (3.6.156)$$

Diese Abschätzung ist in folgendem Sinne ‘‘scharf’’:

$$\eta_E(u_h) \leq c \|\nabla e_h\| + c \left(\sum_{T \in \mathbb{T}_h} h_T^2 \|f\|_T^2 \right)^{1/2}. \quad (3.6.157)$$

Beweis: Es bleibt, die Abschätzung (3.6.157) zu beweisen. Auf jeder Zelle $T \in \mathbb{T}_h$ ist $u_h|_T \in P_1(T)$ und folglich

$$\rho_T(u_h)^2 = \|f + \Delta u_h\|_T^2 = \|f\|_T^2.$$

Es bezeichne wieder $\partial \mathbb{T}_h$ die Menge aller (inneren) Kanten Γ der Triangulierung \mathbb{T}_h . Wegen der Formregularitätsbedingung gilt $ch_T \leq h_\Gamma := |\Gamma| \leq c'h_T$, $\Gamma \subset \partial T$, gleichmäßig für alle Zellen $T \in \mathbb{T}_h$ mit Kanten $\Gamma \in \partial \mathbb{T}_h$ und $h \in \mathbb{R}_+$. Folglich ist

$$\sum_{T \in \mathbb{T}_h} h_T^2 \rho_{\partial T}(u_h)^2 \leq c \sum_{\Gamma \in \partial \mathbb{T}_h} h_\Gamma \|[\partial_n u_h]\|_{\Gamma \setminus \partial \Omega}^2.$$

Für eine Kante $\Gamma \in \partial \mathbb{T}_h$ seien T und T' die beiden benachbarten Zellen mit Γ als gemeinsamer Kante. Bezogen auf den Mittelpunkt a_Γ von Γ definieren wir die C^∞ -Funktion

$$\hat{\omega}_\Gamma(x) := \begin{cases} \exp\left(-\frac{|x-a_\Gamma|^2}{r^2-|x-a_\Gamma|^2}\right), & x \in B_r(a_\Gamma), \\ 0, & x \in B_r(a_\Gamma)^c, \end{cases} \quad \omega_\Gamma(x) := h_\Gamma \left(\int_\Gamma \hat{\omega}_\Gamma(x) ds \right)^{-1} \hat{\omega}_\Gamma(x),$$

wobei $r = \gamma h_\Gamma$ so gewählt ist, dass $B_r(a_\Gamma) := \{x \in \mathbb{R}^2 \mid |x - a_\Gamma| \leq r\} \subset T \cup T'$. Diese Funktion hat konstruktionsgemäß die Eigenschaften:

$$\begin{aligned} \omega_\Gamma|_{\partial(T \cup T')} &= 0, \quad \partial_n \omega_\Gamma|_{\partial T} = \partial_n \omega_\Gamma|_{\partial T'} = \partial_n \omega_\Gamma|_\Gamma = 0, \\ \int_\Gamma \omega_\Gamma(x) ds &= h_\Gamma = \int_\Gamma ds, \\ \|\omega_\Gamma\|_\infty &\leq c, \quad \|\nabla \omega_\Gamma\|_\infty \leq ch_\Gamma^{-1}, \quad \|\nabla^2 \omega_\Gamma\|_\infty \leq ch_\Gamma^{-2}, \end{aligned}$$

wobei die Konstanten unabhängig von T und h gewählt werden können. Entlang der Kante Γ ist der Sprungterm $[\partial_n u_h]$ konstant und es gilt daher wegen der Stetigkeit von $\partial_n u$ über die Kante Γ mit der gerade definierten Funktion ω_Γ

$$\begin{aligned} \|[\partial_n u_h]\|_\Gamma^2 &= [\partial_n u_h]_\Gamma^2 \int_\Gamma ds = [\partial_n u_h]_\Gamma^2 \int_\Gamma \omega_\Gamma ds = [\partial_n u_h]_\Gamma \int_\Gamma [\partial_n(u_h - u)] \omega_\Gamma ds \\ &= [\partial_n u_h]_\Gamma \left\{ \int_{\partial T} \partial_n(u_h - u) \omega_\Gamma ds + \int_{\partial T'} \partial_n(u_h - u) \omega_\Gamma ds \right\}. \end{aligned}$$

Partielle Integration ergibt dann

$$\begin{aligned} \|[\partial_n u_h]\|_\Gamma^2 &= [\partial_n u_h]_\Gamma \int_{T \cup T'} \{ \Delta(u_h - u) \omega_\Gamma(x) - \nabla(u_h - u) \nabla \omega_\Gamma \} dx \\ &= [\partial_n u_h]_\Gamma \int_{T \cup T'} \{ -f \omega_\Gamma(x) - \nabla(u_h - u) \nabla \omega_\Gamma \} dx. \end{aligned}$$

Unter Beachtung der obigen Eigenschaften der Funktion ω_Γ folgt

$$\begin{aligned} \|[\partial_n u_h]\|_\Gamma^2 &\leq [\partial_n u_h]_\Gamma \{ \|f\|_{T \cup T'} \|\omega_\Gamma\|_{T \cup T'} + \|\nabla e_h\|_{T \cup T'} \|\nabla \omega_\Gamma\|_{T \cup T'} \} \\ &\leq c [\partial_n u_h]_\Gamma \{ h_T \|f\|_{T \cup T'} + \|\nabla e_h\|_{T \cup T'} \}. \end{aligned}$$

Dies impliziert dann die folgende Abschätzung:

$$\begin{aligned} \sum_{\Gamma \in \partial \mathbb{T}_h} h_\Gamma \|[\partial_n u_h]\|_{\Gamma \setminus \partial \Omega}^2 &\leq c \sum_{\Gamma \in \partial \mathbb{T}_h} h_\Gamma |[\partial_n u_h]_\Gamma| \{ h_T \|f\|_{T \cup T'} + \|\nabla e_h\|_{T \cup T'} \} \\ &\leq c \left(\sum_{\Gamma \in \partial \mathbb{T}_h} h_\Gamma^2 |[\partial_n u_h]_\Gamma|^2 \right)^{1/2} \left(\sum_{\Gamma \in \partial \mathbb{T}_h} \{ h_T^2 \|f\|_{T \cup T'}^2 + \|\nabla e_h\|_{T \cup T'}^2 \} \right)^{1/2} \\ &\leq c \left(\sum_{\Gamma \in \partial \mathbb{T}_h} h_\Gamma \|[\partial_n u_h]\|_{\Gamma \in \partial \mathbb{T}_h}^2 \right)^{1/2} \left(\sum_{T \in \mathbb{T}_h} \{ h_T^2 \|f\|_T^2 + \|\nabla e_h\|_T^2 \} \right)^{1/2} \end{aligned}$$

bzw.

$$\sum_{\Gamma \in \partial \mathbb{T}_h} h_\Gamma \|[\partial_n u_h]\|_{\Gamma \setminus \partial \Omega}^2 \leq c \sum_{T \in \mathbb{T}_h} h_T^2 \|f\|_T^2 + c \|\nabla e_h\|^2.$$

Dies ergibt die behauptete Abschätzung (3.6.157).

Q.E.D.

Bemerkung 3.14: Mit einer etwas aufwendigeren Argumentation kann man die folgende verschärfte a posteriori Abschätzung herleiten:

$$\eta_E(u_h) \leq c \left(\sum_{T \in \mathbb{T}_h} h_T^2 \left\{ h_T^2 \|\nabla f\|_T^2 + \rho_{\partial T}(u_h)^2 \right\} \right)^{1/2}.$$

Diese besagt, dass im Fehlerschätzer $\eta_E(u_h)$ in der Regel der Einfluss der Gleichungsresiduen $\|f + \Delta u_h\|_T$ gegenüber den Regularitätsresiduen $\|[\partial_n u_h]\|_{\partial T \setminus \partial \Omega}$ vernachlässigt werden können. Dies ist in dieser Form allerdings nur richtig für *lineare* bzw. *bilineare* finite Elemente.

b) L^2 -Norm-Fehlerschätzer:

Zur Abschätzung des L^2 -Fehlers $\|e_h\|$ wählen wir das Fehlerfunktional

$$J(\varphi) := \|e_h\|^{-1}(\varphi, e_h),$$

womit automatisch $J(e) = \|e_h\|$ gilt. Die zugehörige Lösung $z \in V$ des dualen Problems

$$(\nabla \varphi, \nabla z) = J(\varphi) \quad \forall \varphi \in V \tag{3.6.158}$$

ist dann auf dem konvexen Gebiet Ω auch in $H^2(\Omega)$, und es gilt die a priori Abschätzung $\|\nabla^2 z\| \leq \|\Delta z\| = 1$. Ausgehend von der Fehlerabschätzung (3.6.150) erhalten wir wieder

$$\|e_h\| \leq \sum_{T \in \mathbb{T}_h} \left\{ \|f + \Delta u_h\|_T \|z - \psi_h\|_T + \frac{1}{2} \|[\partial_n u_h]\|_{\partial T \setminus \partial \Omega} \|z - \psi_h\|_{\partial T} \right\}.$$

Wir können nun $\psi_h := I_h z$ als die normale Knoteninterpolierende von z wählen:

$$\|z - I_h z\|_T + h_T^{1/2} \|z - I_h z\|_{\partial T} \leq c_I h_T^2 \|\nabla^2 z\|_T. \tag{3.6.159}$$

Damit folgt dann wieder mit den oben definierten Residuen $\rho_T(u_h)$ und $\rho_{\partial T}(u_h)$:

$$\begin{aligned} \|e_h\| &\leq c_I \sum_{T \in \mathbb{T}_h} h_T^2 \left\{ \|f + \Delta u_h\|_T + \frac{1}{2} h_T^{-1/2} \|[\partial_n u_h]\|_{\partial T \setminus \partial \Omega} \right\} \|\nabla^2 z\|_T \\ &\leq c \left(\sum_{T \in \mathbb{T}_h} h_T^4 \{ \rho_T(u_h)^2 + \rho_{\partial T}(u_h)^2 \} \right)^{1/2} \left(\sum_{T \in \mathbb{T}_h} \|\nabla^2 z\|_T^2 \right)^{1/2}. \end{aligned}$$

Wegen

$$\left(\sum_{T \in \mathbb{T}_h} \|\nabla^2 z\|_T^2 \right)^{1/2} = \|\nabla^2 z\| \leq 1,$$

ergibt sich schließlich das folgende Resultat:

Satz 3.18 (a posteriori L^2 -Normfehler): Für den Fehler $e_h := u - u_h$ gilt die a posteriori Abschätzung bzgl. der L^2 -Norm:

$$\|e_h\| \leq \eta_{L^2}(u_h) := c \left(\sum_{T \in \mathbb{T}_h} h_T^4 \{ \rho_T(u_h)^2 + \rho_{\partial T}(u_h)^2 \} \right)^{1/2}. \quad (3.6.160)$$

Diese Abschätzung ist ebenfalls „scharf“ in folgendem Sinne:

$$\eta_{L^2}(u_h) \leq c \|e_h\| + c \left(\sum_{T \in \mathbb{T}_h} h_T^4 \|f\|_T^2 \right)^{1/2}. \quad (3.6.161)$$

Beweis: Der Beweis verläuft analog wie bei der Energienorm-Fehlerabschätzung. Ein kleiner Zusatzschritt ist

$$\int_{T \cup T'} \{ \Delta(u_h - u) \omega_\Gamma(x) - \nabla(u_h - u) \nabla \omega_\Gamma \} dx = \int_{T \cup T'} \{ \Delta(u_h - u) \omega_\Gamma(x) + (u_h - u) \Delta \omega_\Gamma \} dx,$$

wobei der Randterm wegen $\partial_n \omega_\Gamma|_{\partial T} = \partial_n \omega_\Gamma|_{\partial T'} = 0$ verschwindet. Die weiteren Details seien als Übungsaufgabe gestellt. Q.E.D.

c) Punktfehler-Schätzer:

Zur Abschätzung des Fehlers $e_h(P)$ in einem festen Punkt $P \in \Omega$ würden wir gern das Fehlerfunktional $J(\varphi) := \varphi(P)$ wählen. Dieses ist zwar auf stetigen Funktionen, aber nicht auf dem ganzen Lösungsraum $V = H_0^1(\Omega)$ definiert, so dass die oben entwickelte allgemeine Theorie nicht unmittelbar anwendbar wäre. Deshalb wählen wir eine Kugelumgebung $B_\varepsilon := \{x \in \mathbb{R}^d : |x - P| < \varepsilon\}$ des Punktes P und arbeiten statt dessen mit dem regularisierten Funktional

$$J_\varepsilon(\varphi) := |B_\varepsilon|^{-1} \int_{B_\varepsilon} \varphi dx,$$

welches sicher auf ganz V definiert und beschränkt ist. Der Regularisierungsparameter ε wird üblicherweise gleich einer vorgegebenen Fehlertoleranz TOL gesetzt. Für hinreichend glatte Funktionen φ ist dabei (Fehler der Mittelpunkregel)

$$|\varphi(P) - J_\varepsilon(\varphi)| \leq c \varepsilon^2. \quad (3.6.162)$$

Wir betrachten den Fall $d = 2$. Die Lösung $z_\varepsilon \in V$ des zugehörigen dualen Problems

$$(\nabla \varphi, \nabla z_\varepsilon) = J_\varepsilon(\varphi) \quad \forall \varphi \in V, \quad (3.6.163)$$

ist dann eine „regularisierte“ Greensche Funktion und verhält sich wie

$$z_\varepsilon(x) \approx \left| \log(|x - P| + \varepsilon) \right|, \quad |\nabla^2 z_\varepsilon| \approx \left| |x - P| + \varepsilon \right|^{-2}.$$

Die Gewichte in der a posteriori Fehlerabschätzung (3.6.151) gestatten daher die Abschätzung

$$\omega_T(z_\varepsilon) + \omega_{\partial T}(z_\varepsilon) \leq ch_T^2 \|\nabla^2 z_\varepsilon\|_T \leq ch_T^3 d_T^{-2},$$

mit $d_T := \max_{x \in T} \{|x-a| + \varepsilon\}$. Wir finden also als obere Schranke für den Punktfehler:

$$|e_h(P)| \leq c \sum_{T \in \mathbb{T}_h} \rho_T(u_h) h_T^2 \|\nabla^2 z_\varepsilon\|_T + c\varepsilon^2 \leq c \sum_{T \in \mathbb{T}_h} \rho_T(u_h) \frac{h_T^3}{d_T^2} + c\varepsilon^2.$$

In diesem Fall ist aber die a-priori Bestimmung der Konstante c praktisch unmöglich.

d) Ein hyper-singulärer Fall

Einen kuriosen Sonderfall stellt die folgende Auswertungsgröße dar:

$$J(u) := \int_{\partial\Omega} \partial_n u \, ds \quad \left(= \int_{\Omega} \Delta u \, dx = - \int_{\Omega} f \, dx \right).$$

Dieser „mittlere Normalfluss“ ließe sich offenbar auch direkt aus den Daten f berechnen. Zur Illustration wollen wir aber annehmen, dass diese Information nicht ausgenutzt wird, sondern statt dessen $J(u)$ durch $J(u_h)$ approximiert wird. Das Funktional $J(\cdot)$ ist wieder nicht auf dem ganzen Lösungsraum V definiert (wohl aber auf der reguläreren Lösung $u \in V \cap H^2(\Omega)$). Zur Anwendung unserer allgemeinen Theorie muss das Funktional daher zunächst regularisiert werden. Für das Folgende nehmen wir der Einfachheit halber an, dass Ω der Einheitskreis im \mathbb{R}^2 ist. Für $\varepsilon = \text{TOL}$ setzen wir $S_\varepsilon := \{x \in \Omega, \text{dist}\{x, \partial\Omega\} < \varepsilon\}$ und erhalten für glattes φ :

$$J_\varepsilon(\varphi) := \varepsilon^{-1} \int_{S_\varepsilon} \partial_r \varphi \, dx = \int_{\partial\Omega} \partial_r \varphi \, ds + \mathcal{O}(\varepsilon).$$

Dabei ist ∂_n auf S_ε auf natürliche Weise durch Fortsetzung von $\partial\Omega$ definiert. Die Lösung $z_\varepsilon \in V$ des zugehörigen dualen Problems

$$(\nabla\varphi, \nabla z_\varepsilon) = J_\varepsilon(\varphi) \quad \forall \varphi \in V,$$

ist dann gegeben durch

$$z_\varepsilon = -1 \quad \text{in } \Omega \setminus S_\varepsilon, \quad z_\varepsilon(x) = -\varepsilon^{-1}(1-|x|) \quad \text{auf } S_\varepsilon.$$

Hieraus ergibt sich die Fehlerabschätzung

$$|J_\varepsilon(e_h)| \leq c_I \sum_{T \in \mathbb{T}_h} h_T^2 \rho_T(u_h) \|\nabla^2 z_\varepsilon\|_T \approx \sum_{T \cap S_\varepsilon \neq \emptyset} \dots,$$

d. h.: Die Zellen im Innern von Ω tragen nicht zum Gesamtfehler bei. Daher wäre die beste Strategie zur Gitterverfeinerung, in jedem Verfeinerungszyklus jeweils nur die Zellen entlang des Randes $\partial\Omega$ zu verfeinern. Dies setzt aber voraus, dass die Elemente des Lastvektors, $b_n = (f, \varphi_h^{(n)})_\Omega$, exakt berechnet werden. Abbildung 3.23 zeigt ein automa-

tisch verfeinertes Gitter nach 7 Verfeinerungszyklen sowie die numerisch bestimmte duale Lösung auf diesem Gitter. Die zugehörigen Ergebnisse sind in Tabelle 3.1 gelistet.

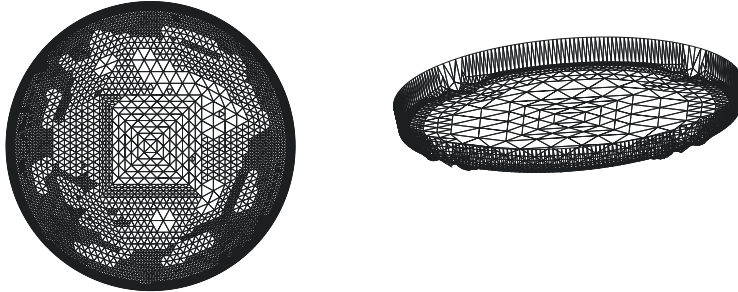


Abbildung 3.23: Verfeinertes Gitter nach 7 Verfeinerungszyklen (links) und die auf diesem Gitter numerisch bestimmte Lösung (rechts).

Tabelle 3.1: Resultate mit dem Energienorm-Fehlerschätzer $\eta_E(u_h)$ und der „optimalen“ Strategie.

η_E			„optimale“ Strategie		
L	N	$J(e_h)$	L	N	$J(e_h)$
1	1024	$2.91 + 0$	1	559	$2.90 + 0$
2	4096	$1.50 + 0$	2	1294	$1.50 + 0$
3	16384	$7.49 - 1$	3	3079	$7.48 - 1$
4	64768	$3.77 - 1$	4	7174	$3.62 - 1$
5	253858	$1.73 - 1$	5	16738	$1.67 - 1$
6	<i>memory exhausted</i>		6	38146	$6.93 - 2$

Numerische Auswertung des a-posteriori Fehlerschätzers

Wir wollen nun kurz die numerische Auswertung der Fehleridentität (3.6.146) diskutieren:

$$J(e_h) = \eta(u_h) := \sum_{T \in \mathcal{T}_h} \left\{ (f + \Delta u_h, z - \psi_h)_T - \frac{1}{2}([\partial_n u_h], z - \psi_h)_{\partial T} \right\}. \quad (3.6.164)$$

Wir wählen dazu $\psi_h := I_h z$, die Knoteninterpolierende von z . Die Qualität der Auswertung wird durch den sog. „Effektivitätsindex“ gemessen:

$$I_{\text{eff}} := \left| \frac{\eta(u_h)}{J(e_h)} \right|.$$

Alle Methoden der Auswertung von (3.6.164) bedienen sich einer numerischen Lösung $z_h \in V_h$ des dualen Problems, welche im einfachsten Fall durch direkte Diskretisierung von (3.6.164) mit Hilfe des vorliegenden Verfahrens gewonnen wird:

$$(\nabla\varphi_h, \nabla z_h) = J(\varphi_h) \quad \forall \varphi_h \in V_h. \tag{3.6.165}$$

Wir unterscheiden zwei Anwendungen der Fehleridentität (3.6.164):

- Überprüfung der Genauigkeit einer auf dem Gitter \mathbb{T}_h berechneten Näherungslösung u_h und Abbruchkriterium für einen Gitterverfeinerungsprozess: $\eta(u_h) \leq TOL$?
- Grundlage zur Gewinnung von Kriterien („Verfeinerungsindikatoren“ η_T) zur lokalen Gitteranpassung.

Die folgenden Strategien zur Auswertung von $\eta(u_h)$ kommen in Betracht (hier beschrieben für den Fall d -linearer Ansätze auf Rechteckgittern im \mathbb{R}^d):

1. Die duale Lösung z wird durch eine Näherung höherer Ordnung approximiert. Zum Beispiel liefert die Approximation $z \approx z_h^{(2)}$ mit der d -quadratischen Ritz-Projektion das gewünschte asymptotische Verhalten $\lim_{TOL \rightarrow 0} I_{\text{eff}} = 1$.
2. Die duale Lösung z wird durch eine zellblockweise (je 2^d Zellen) d -quadratische Interpolierende der d -linearen Ritzapproximation approximiert (s. Abb. 3.24): $z \approx I_h^{(2)} z_h$. In diesem Fall beobachtet man $\lim_{TOL \rightarrow 0} I_{\text{eff}} < 1$ ($\approx 0.5 - 0.9$).

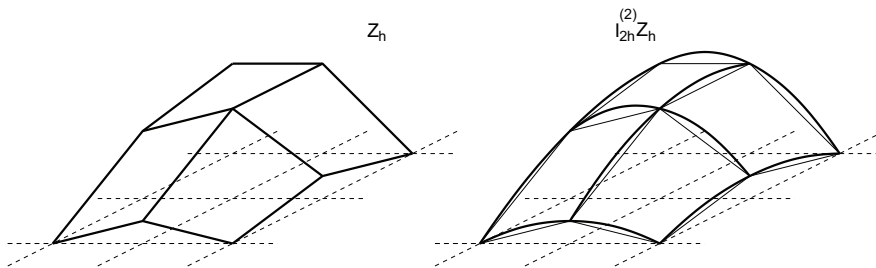


Abbildung 3.24: Blockweise bi-quadratische Interpolation (rechts) bilinearer Knotenwerte (links).

3. Die Differenz $z - I_h z$ wird in Anlehnung an die oben zitierte Interpolationsfehlerabschätzung mit Hilfe eines skalierten Differenzenquotienten der d -linearen Ritz-Projektion approximiert: $|z - I_h z| \approx c_{i,T} h_T^2 |\nabla_h^2 z_h F|$. Dies führt in der Regel auf eine grobe Überschätzung des Fehlers.

In der Praxis wird meist die „billige“ und ausreichend genaue Prozedur (2) verwendet. Dabei brauchen insbesondere keine Interpolationskonstanten c_I spezifiziert zu werden.

3.6.3 Strategien zur Gittersteuerung

Ziel ist es, den Fehler in dem durch das Fehlerfunktional $J(\cdot)$ beschriebenen Maß unter eine gewisse vorgegebene Toleranz $|J(e_h)| \leq \text{TOL}$ zu bringen und dabei mit dem vorhandenen Speicher auszukommen. Es können also nur eine bestimmte Anzahl von Zellen $N \leq N_{\max}$ verwaltet werden. Die Gitterverfeinerung erfolgt dabei standardmäßig durch „Kantenbisektion“, d. h. durch Unterteilung einer Zelle T in 2^d Teilzellen. Bei Gittervergrößerung werden mehrere Zellen zu einer Makrozelle zusammengefaßt. Wir werden im folgenden hauptsächlich Vierecks- oder Hexaedergitter betrachten. Bei Unterteilung einer Zelle entstehen sog. „hängende“ Knoten auf den Zellrändern, so dass das neue Gitter zunächst nicht zulässig wäre. Diese Unzulässigkeit kann durch verschiedenen Methoden aufgehoben werden. Die meist gebräuchliche ist die Elimination der zu den hängenden Knoten gehörenden Freiheitsgraden durch lineare Interpolation, wodurch der entstehende diskrete Ansatzraum wieder konform wird: $V_h \subset V$ (s. Abb. 3.25). Dabei wird der fiktive Knotenwert $v_h(P)$ in einem hängenden Knoten P auf einer Kante $\Gamma \subset \partial T$ mit Endpunkten P', P'' durch die Interpolationsvorschrift $v_h(P) := \frac{1}{2}\{v_h(P') + v_h(P'')\}$ festgelegt. Alle im folgenden gezeigten Beispielrechnungen bedienen sich dieser Technik.

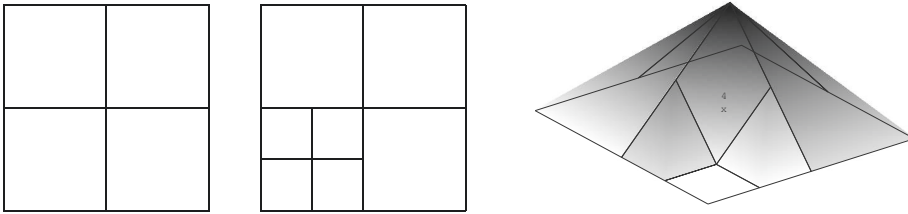


Abbildung 3.25: Verfeinertes Rechteckgitter mit „hängendem“ Knoten sowie die Basisfunktion zum Mittelknoten.

Ausgangspunkt zur Gewinnung von Kriterien für die lokale Gitteranpassung ist eine a-posteriori Fehlerschätzung der Form

$$|J(e_h)| \approx \sum_{T \in \mathbb{T}_h} \eta_T, \quad (3.6.166)$$

mit lokalen „Fehlerindikatoren“ η_T , welche durch Auswertung der Fehleridentität (3.6.146) mit Hilfe einer der oben beschriebenen Methoden gewonnen werden. Wir erhalten bei Verwendung von Methode (2):

$$\eta_T := |(f + \Delta u_h, I_h^{(2)} z_h - z_h)_T - \frac{1}{2}([\partial_n u_h], I_h^{(2)} z_h - z_h)_{\partial T \setminus \partial \Omega}| \quad (3.6.167)$$

und mit Methode (3):

$$\eta_T := c_I h_T^2 \{\rho_T(u_h) + \rho_{\partial T}(u_h)\} \|\nabla_h^2 z_h\|_T. \quad (3.6.168)$$

Wir beschreiben im folgenden das allgemeine Schema eines auf lokalen Fehlerindikatoren η_T basierenden Gitteranpassungsprozesses:

Nachdem über das Gebiet Ω ein Grobgitter $\mathbb{T}_0 := \mathbb{T}_{h_0}$ mit einer Gitterweitenverteilung h_0 gelegt worden ist, sei der Prozess der adaptiven Gittersteuerung schon L -mal durchlaufen worden. Wir bestimmen nun auf dem Gitter $\mathbb{T}_L := \mathbb{T}_{h_L}$ die approximative Lösung $u_L \in V_L := V_{h_L}$. Ebenso wird die diskrete duale Lösung $z_L \in V_L$ auf \mathbb{T}_L berechnet. Nun können die lokalen Fehlerindikatoren η_T gemäß (3.6.167) oder (3.6.168) für jede Zelle $T \in \mathbb{T}_L$ ausgewertet werden. Wir bezeichnen mit $N_L \approx \dim(V_L)$ die Anzahl der Zellen des Gitters \mathbb{T}_L .

Abbruchabfrage: Ist das Abbruchkriterium $\eta(u_h) \leq \frac{1}{2}\text{TOL}$ auf dem Gitter \mathbb{T}_L erfüllt, wird der Adaptionsprozess abgebrochen und u_L als Näherung zu u akzeptiert, welche die Zielgröße $J(u)$ durch $J(u_h)$ mit der gewünschten Genauigkeit TOL approximiert. Andernfalls wird ein neuer Adaptionszyklus begonnen.

Adaptionszyklus: Der Übergang vom Gitter \mathbb{T}_L zum nächsten Gitter \mathbb{T}_{L+1} erfolgt nach einer der im folgenden beschriebenen „Adaptionsstrategien“. Zunächst werden die Zellen $T \in \mathbb{T}_L$ nach der Größe ihrer Indikatorwerte η_T angeordnet,

$$\{T_i, i = 1, \dots, N_L\} : \quad \eta_{T,1} \geq \dots \geq \eta_{T,i} \geq \eta_{T,i+1} \geq \dots \geq \eta_{T,N_L}.$$

1) *Fehlerbalancierungs-Strategie:* Das Ziel ist, die Indikatorwerte so zu balancieren, dass

$$\eta_T \approx \frac{\text{TOL}}{N_L}, \quad T \in \mathbb{T}_L. \tag{3.6.169}$$

Dann würde wie gewünscht gelten:

$$|J(e_L)| \approx \eta(u_L) \approx \sum_{T \in \mathbb{T}_L} \frac{\text{TOL}}{N_L} = \text{TOL}.$$

Das Problem bei dieser Strategie ist zum einen, dass die Zahl N_L sich während des Verfeinerungsprozesses ständig ändert, und zum anderen, dass die Balancierungsvorschrift (3.6.169) die delikate Wahl von Parametern $0 < \alpha < \beta < 1$ erfordert:

$$\alpha \frac{\text{TOL}}{N_L} \leq \eta_T \leq \beta \frac{\text{TOL}}{N_L}.$$

Für die folgende Diskussion setzen wir der Einfachheit halber $\alpha = 1/2, \beta = 1$. Im ersten Schritt testet man, ob $\eta_{T,1} \leq \text{TOL}/N_L$ ist. Wenn „ja“, ist das Abbruchkriterium erfüllt, wenn „nein“, wird die Zelle T_1 verfeinert. Dies führt zu einer Erhöhung von N_L auf $N_L + 3$. Danach wäre es möglich, dass für alle übrigen Zellen $T_{L,i}$ gilt:

$$\frac{\text{TOL}}{N_L + 3} \leq \eta_{T,i} \leq \frac{\text{TOL}}{N_L}.$$

Der Verfeinerungszyklus wäre also bereits nach dem ersten Schritt zu beenden und auf

dem erhaltenen Gitter \mathbb{T}_{L+1} eine neue Näherungslösung u_{L+1} zu berechnen. Dies wäre aber sehr ineffizient. Der Verfeinerungsprozess sollte daher beschleunigt werden. Dazu überprüft man von dem Indikator $\eta_{T,1}$ ausgehend und beginnend mit $j = 0$, ob

$$\eta_{T,i} \leq \frac{\text{TOL}}{N_L + 3j}.$$

Ist dies nicht erfüllt, wird das Element T_i geteilt, die Zähler j und i um Eins erhöht, und man geht zum nächstkleineren $\eta_{T,i}$ über. Ist die Bedingung allerdings erfüllt, hat man das neue Gitter \mathbb{T}_{L+1} gefunden. Diese Balancierungsstrategie ist zwar potentiell optimal, jedoch mit aufwendigen Abfragen verbunden. Mögliche alternative Strategien, die einfacher sind, aber weniger gute Ergebnisse liefern, sind die folgenden.

II) „Fest-Raten-Strategie“: Ziel ist es, in jedem Adaptionszyklus die Anzahl der Gitterzellen N_L mit einer festen Rate zu erhöhen oder den Fehlerschätzwert $\eta(u_L)$ mit einer festen Rate zu verkleinern. Ausgangspunkt ist wieder eine Anordnung der Zellen von \mathbb{T}_L nach der Größe der zugehörigen Indikatorwerte. Zu vorgewählten Prozentsätzen $X\%$ und $Y\%$ werden die Zellen so gruppiert, dass (and der Zellanzahl orientierte Strategie)

$$\#\{T_i, \dots, T_{N_*}\} \approx \frac{X}{100} N_L, \quad \#\{T_{N_L-N_*+1}, \dots, T_{N_L}\} \approx \frac{Y}{100} N_L,$$

oder alternativ (am Schätzwert orientierte Strategie)

$$\sum_{i=1}^{N_*} \eta_{T,i} \approx Y \eta(u_L), \quad \sum_{i=N_L-N_*+1}^{N_L} \eta_{T,i} \approx X \eta(u_L).$$

Dann werden die Zellen T_i, \dots, T_{N_*} verfeinert und die Zellen $T_{N_L-N_*+1}, \dots, T_{N_L}$ vergrößert.

„Exakte“ Gitteroptimierung

Zum Abschluss wollen wir noch diskutieren, dass die Fehlerschätzung (3.6.151)

$$|J(e_h)| \approx \eta := \sum_{T \in \mathbb{T}_h} \rho_T(u_h) \omega_T(z_h) \quad (3.6.170)$$

im Prinzip auch zur direkten Bestimmung eines „optimalen“ Gitters mit Gitterweitenfunktion $h = h(x)$ verwendet werden könnte. Dies wird als Nebenprodukt auch eine Rechtfertigung für die „Fehlerbalancierungsstrategie“ liefern. Zu lösen sind die folgenden Optimierungsaufgaben:

(OP I) Bei vorgegebener Fehlertoleranz TOL soll die zu deren Erreichung benötigte Anzahl von Zellen N (d. h. der numerische Aufwand) minimiert werden:

$$N \rightarrow \text{MIN}, \quad \eta \leq \text{TOL}. \quad (3.6.171)$$

(OP II) Bei vorgegebener maximalen Zellzahl N_{max} (d. h. begrenzter Speicherkapazität) soll der Fehler (genauer der Fehlerschätzer) minimiert werden:

$$\eta \rightarrow MIN, \quad N \leq N_{max}. \tag{3.6.172}$$

Wir diskutieren im Folgenden nur die für die Praxis relevantere Fragestellung (OP II). Das Optimierungsproblem (OP I) lässt sich mit analogen Argumenten behandeln (Übungsaufgabe). Die grundlegende Annahme ist, dass die Größen $\rho_T(u_h)$ sowie $\omega_T(z_h)$ (nach geeigneter Skalierung) für $TOL \rightarrow 0$ zunehmend bessere, lokale Approximationen gewisser kontinuierlicher Funktionen $\Phi(x)$ bzw. $\Psi(x)$ sind:

$$h_T^{-1} \rho_T(u_h) = h_T^{-1} \|f + \Delta u_h\|_T + \frac{1}{2} h_T^{-1/2} \|h_T^{-1} [\partial_n u_h]\|_{\partial T \setminus \partial \Omega} \rightarrow \Phi_u(x_T), \tag{3.6.173}$$

$$h^{-3} \omega_T(z_h) = \max \left\{ h_T^{-3} \|z - I_h z\|_T, h_T^{-5/2} \|z - I_h z\|_{\partial T} \right\} \rightarrow \Psi_z(x_T), \tag{3.6.174}$$

wobei x_T ein gemeinsamer Punkt einer Folge von Zellen T ist. Diese Annahme ist natürlich im strengen Sinne nicht realistisch, da auf allgemeinen Gitterfolgen $(\mathbb{T}_h)_{h \rightarrow 0}$ ein sehr unregelmäßiges Konvergenzverhalten auftreten kann. Unter der obigen Annahme gilt mit der Funktion $A(x) := \Phi(x)\Psi(x)$:

$$\sum_{T \in \mathbb{T}_h} \rho_T(u_h) \omega_T(z_h) = \sum_{T \in \mathbb{T}_h} h_T^4 \{ h_T^{-4} \rho_T(u_h) \omega_T(z_h) \} \approx \int_{\Omega} h(x)^2 A(x) dx. \tag{3.6.175}$$

Für die Zellzahl N haben wir die Darstellung

$$N = \sum_{T \in \mathbb{T}_h} 1 = \sum_{T \in \mathbb{T}_h} h_T^2 h_T^{-2} \approx \int_{\Omega} h(x)^{-2} dx. \tag{3.6.176}$$

Damit können wir die Optimierungsaufgabe (OP II) näherungsweise in kontinuierlicher Form schreiben als:

$$F(h) := \int_{\Omega} h(x)^2 A(x) dx \rightarrow MIN, \quad N(h) := \int_{\Omega} h(x)^{-2} dx = N_{max}. \tag{3.6.177}$$

Hierbei haben wir o.B.d.A. die Ungleichungsbedingung $N \leq N_{max}$ durch eine Gleichungsbedingung $N = N_{max}$ ersetzt, da sich ein minimaler Fehler sicherlich unter maximaler Ausnutzung der möglichen Zellzahl ergibt. Dieses restringierte Optimierungsproblem wird nun mit dem Lagrange-Formalismus der Variationsrechnung gelöst. Dazu definieren wir die „Lagrange-Funktion“

$$L(h, \lambda) := F(h) + \lambda \{ N(h) - N_{max} \}$$

mit einem skalaren Lagrange-Parameter $\lambda \in \mathbb{R}$. Jede Lösung des Optimierungsproblems ist dann notwendig stationärer Punkt (Sattelpunkt) von L . Zur Bestimmung eines solchen Sattelpunkts machen wir den folgenden Ansatz

$$\frac{d}{dt}L(h + t\varphi, \lambda + t\mu)|_{t=0} = 0 \quad \forall \varphi \in C(\overline{\Omega}), \forall \mu \in \mathbb{R}.$$

Auswertung dieser Beziehung ergibt

$$2 \int_{\Omega} h(x)A(x)\varphi(x) dx - 2\lambda \int_{\Omega} h^{-3}(x)\varphi(x) dx = 0 \quad \forall \varphi \in C(\overline{\Omega})$$

sowie

$$\mu \left\{ \int_{\Omega} h(x)^{-2} dx - N_{\max} \right\} = 0 \quad \forall \mu \in \mathbb{R},$$

bzw. notwendig

$$h(x)A(x) - \lambda h^{-3}(x) = 0, \quad \int_{\Omega} h(x)^{-2} dx = N_{\max}.$$

Hieraus ergibt sich, dass

$$h(x) = \left(\frac{A}{\lambda} \right)^{-1/4},$$

und weiter

$$W := \int_{\Omega} A(x)^{1/2} dx = \lambda^{1/2} N_{\max}.$$

Damit erhalten wir einerseits den Lagrange-Parameter,

$$\lambda = \left(\frac{W}{N_{\max}} \right)^2, \tag{3.6.178}$$

und andererseits eine Gleichung für die gesuchte „optimale“ Gitterweitenverteilung

$$h_{\text{opt}}(x) = \left(\frac{W}{N_{\max}} \right)^{1/2} A(x)^{-1/4}. \tag{3.6.179}$$

Als Nebenprodukt dieser Rechnung ergibt sich für die optimale Gitterweite die Beziehung

$$\eta_T = h_T^4 A_T = \lambda \equiv \text{konst.}, \tag{3.6.180}$$

d. h.: Die an der Äquilibration der lokalen Fehlerindikatoren η_T orientierten Gittersteuerungsstrategien führen tatsächlich auf optimale Gitterweitenverteilungen.

Wir erwähnen noch, dass das zu (OP II) „duale“ Optimierungsproblem (OP I) auf die folgende Lösung führt:

$$h_{\text{opt}}(x) = \left(\frac{TOL}{W} \right)^{1/2} A(x)^{-1/4}. \tag{3.6.181}$$

Die Größe W ist bei den üblichen Fehlerfunktionalen wohl definiert; dies beinhaltet sogar so singuläre Fälle wie die Punktauswertung von Ableitungen $J(u) = \partial_i u(a)$, $A(x) \sim |x - a|^{-3}$. Erst die Auswertung von zweiten Ableitungen $J(u) = \partial_i^2 u(a)$ macht hier Probleme.

3.6.4 Ein Testbeispiel

Wir wollen die bisher erzielten Ergebnisse anhand eines konkreten Beispiels illustrieren. Dazu wird wieder das übliche Modellproblem betrachtet:

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega. \tag{3.6.182}$$

Wir wollen den Punktwert $\partial_1 u(P)$ für ein $P \in \Omega$ berechnen. Das zugehörige Funktional ist wieder nicht auf dem ganzen Lösungsraum $V = H_0^1(\Omega)$ definiert und muss regularisiert werden. Wir setzen dazu mit $\varepsilon := \text{TOL}$:

$$J_\varepsilon(\varphi) := |B_\varepsilon|^{-1} \int_{B_\varepsilon} \partial_1 u \, dx. \tag{3.6.183}$$

Testfall 1: Sei $\Omega := (-1, 1)^2$ und $P = 0$ (siehe Abb. 3.26). Die zu dem Funktional $J_\varepsilon(\cdot)$ gehörige duale Lösung z_ε verhält sich dann wie $|\nabla^2 z_\varepsilon(x)| \approx (|x| + \varepsilon)^{-3}$. Dies impliziert

$$|\partial_1 e_h(0)| \approx c_I \sum_{T \in \mathbb{T}_h} \frac{h_T^4}{d_T^3} \rho_T(u_h), \tag{3.6.184}$$

mit $d_T := |x_T| + \varepsilon$ und dem Mittelpunkt x_T von T . Wir wollen hierfür eine optimale Gitterweitenverteilung bestimmen. Ausgangspunkt ist die Äquilibrierungsbedingung

$$\eta_T \approx \frac{h_T^4}{d_T^3} \approx \frac{\text{TOL}}{N} \quad \Rightarrow \quad h_T^2 \approx d_T^{3/2} \left(\frac{\text{TOL}}{N} \right)^{1/2}.$$

Hieraus ergibt sich

$$N = \sum_{T \in \mathbb{T}_h} h_T^2 h_T^{-2} = \left(\frac{N}{\text{TOL}} \right)^{1/2} \sum_{T \in \mathbb{T}_h} h_T^2 d_T^{-3/2} \approx \left(\frac{N}{\text{TOL}} \right)^{1/2}$$

und folglich $N_{\text{opt}} \approx \text{TOL}^{-1}$. Bei Verwendung des Energienorm-Fehlerschätzers (3.6.156) zur Gitterverfeinerung ergibt sich dagegen die Gitterkomplexität $N_{\text{opt}} \approx \text{TOL}^{-2}$.

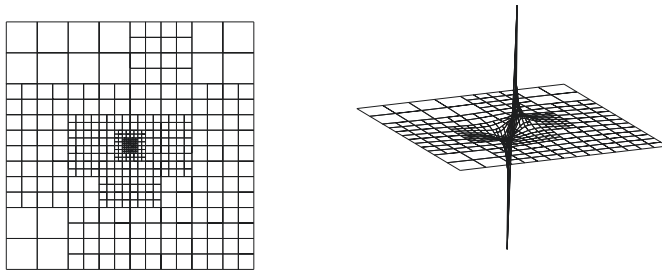


Abbildung 3.26: Verfeinerte Gitter und numerisch bestimmte duale Lösung zur Berechnung von $\partial_1 u(0)$ auf dem Quadrat bei Verwendung des a posteriori Fehlerschätzers $\eta_{\text{weight}}(u_h)$ mit $\text{TOL} = 4^{-4}$.

Tabelle 3.2: Resulte der Berechnung von $\partial_1 u(0)$ unter Verwendung des gewichteten a posteriori Fehlerschätzers $\eta_{\text{weight}}(u_h)$ für verschiedenen Verfeinerungslevel L ; der „Effektivitätsindex“ ist definiert durch $I_{\text{eff}} := |\eta_{\text{weight}}(u_h)/\partial_1 e_h(0)|$.

TOL	N	L	$ \partial_1 e_h(0) $	$\eta_{\text{weight}}(u_h)$	I_{eff}
4^{-2}	148	6	7.51e-1	5.92e-2	0.08
4^{-3}	940	9	4.10e-1	1.42e-2	0.03
4^{-4}	4912	12	4.14e-3	3.50e-3	0.65
4^{-5}	20980	15	2.27e-4	9.25e-4	4.16
4^{-6}	86740	17	5.82e-5	2.38e-4	4.16

Das vorhergesagte Verhalten der verschiedenen Gitterverfeinerungsprozesse wird durch die numerischen Ergebnisse in Tabelle 3.2 gut bestätigt. Wir betonen, dass die richtige Wahl des Regularisierungsparameters $\varepsilon = \text{TOL}$ in (3.6.183) für den Lösungsprozess wichtig ist. Die triviale Alternative $\varepsilon = h_{\min}$ ist automatisch realisiert, wenn zur Berechnung der Gewichte $\omega_T(z)$ die numerisch bestimmte, diskrete, duale Lösung $z_h \in V_h$ verwendet wird (siehe Abb. 3.26). Dies kann bei sehr „singulären“ Funktionalen zu starker lokaler Überverfeinerung, d. h. zu unnötig vielen Verfeinerungsschritten, führen.

Testfall 2: Sei nun Ω das Rechteckgebiet $\Omega = (-1, 1) \times (-1, 3)$ mit Schlitz bei $(0, 0)$ (siehe Abb. 3.27). Die Präsenz der (einspringenden) scharfen Ecke mit Innenwinkel $\omega = 2\pi$ bewirkt in der schwachen Lösung eine „Eckensingularität“ der Form $s = \psi(\theta)r^{1/2}$, d. h. eine Singularität im Gradienten. Wir wollen illustrieren, wie diese Eckensingularität mit der durch die Gewichte $\omega_T(z)$ im Fehlerschätzer induzierten zusammenspielt.

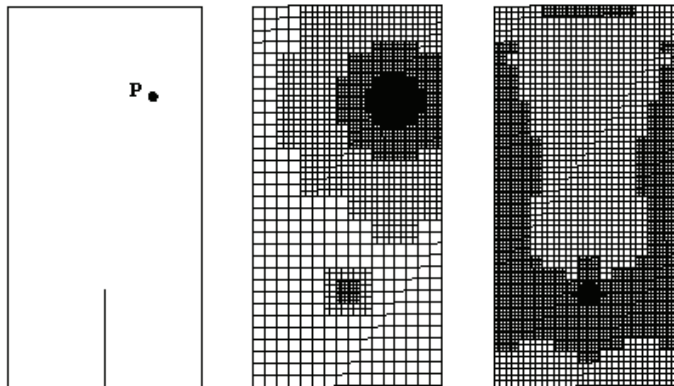


Abbildung 3.27: Verfeinerte Gitter mit ungefähr 5000 Zellen zur Berechnung von $\partial_1 u(P)$, erzeugt mit dem gewichteten Fehlerschätzer $\eta_{\text{weight}}(u_h)$ (mittig) sowie dem Energienorm-Fehlerschätzer $\eta_E(u_h)$ (rechts).

Wir sehen in Abb. 3.27, dass der gewichtete Fehlerschätzer Zellen sowohl bei der Schlitzspitze (zur Unterdrückung des „Pollutionseffekts“), aber auch beim Auswertungspunkt konzentriert, während der Energienorm-Fehlerschätzer wesentlich stärker an der Schlitzspitze und natürlich gar nicht im Auswertungspunkt verfeinert.

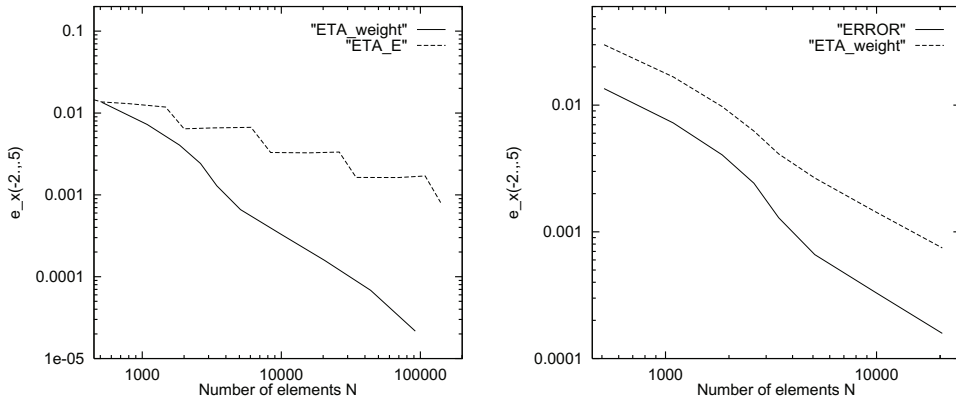


Abbildung 3.28: Vergleich der Effizienz von $\eta_E(u_h)$ und $\eta_{\text{weight}}(u_h)$ auf dem Schlitzgebiet.

3.7 Übungen

Übung 3.1: Man formuliere das Ritzsche Verfahren mit endlich dimensionalen Teilräumen geeigneter Sobolew-Räume $H^m(\Omega)$ für die folgenden Aufgabenstellungen:

a) Neumannsche RWA des Laplace-Operators:

$$-\Delta u = f \quad \text{in } \Omega, \quad \partial_n u = g \quad \text{auf } \partial\Omega.$$

b) Eigenwertproblem des Laplace-Operators:

$$-\Delta u = \lambda u \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega.$$

c) Dirichletsche RWA des „biharmonischen“ Operators:

$$-\Delta^2 u = f \quad \text{in } \Omega, \quad u = \partial_n u = 0 \quad \text{auf } \partial\Omega.$$

Dabei seien jeweils das Gebiet Ω sowie die Daten f, g als genügend regulär und kompatibel angenommen. Man versuche, hierfür analog zur Vorlesung „Bestapproximations“-Aussagen herzuleiten.

Übung 3.2: Seien V ein Hilbert-Raum mit Skalarprodukt $(\cdot, \cdot)_V$ und zugehöriger Norm

$\|\cdot\|_V$ und $a(\cdot, \cdot)$ sowie $l(\cdot)$ bilineare bzw. lineare Formen auf V mit den Eigenschaften

$$\begin{aligned} |a(v, w)| &\leq \alpha \|v\|_V \|w\|_V, \quad v, w \in V && \text{(Beschränktheit),} \\ |a(v, v)| &\geq \kappa \|v\|_V^2, \quad v, w \in V && \text{(V-Elliptizität),} \\ |l(v)| &\leq \beta \|v\|_V, \quad v \in V && \text{(Beschränktheit).} \end{aligned}$$

Dann hat die Variationsgleichung

$$a(u, \varphi) = l(\varphi) \quad \forall \varphi \in V$$

nach dem Satz von Lax-Milgram eine eindeutige Lösung $u \in V$. Für endlich dimensionale Teilräume $V_h \subset V$ werden approximative Lösungen $u_h \in V_h$ bestimmt durch die „Galerkin-Gleichungen“

$$a(u_h, \varphi_h) = l(\varphi_h) \quad \varphi_h \in V_h.$$

a) Man zeige hierfür die (eindeutige) Existenz der „diskreten“ Lösungen $u_h \in V_h$ und die Fehlerabschätzung

$$\|u - u_h\|_V \leq \frac{\alpha}{\kappa} \inf_{\varphi_h \in V_h} \|u - \varphi_h\|_V.$$

b) Man wende das Resultat von (a) zum Nachweis der Konvergenz der Galerkin-Approximation des Diffusions-Konvektions-Problems

$$-\Delta u + \partial_1 u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega,$$

auf einem regulären Gebiet $\Omega \subset \mathbb{R}^2$ mit rechter Seite $f \in L^2(\Omega)$ an.

c) Man versuche das Resultat von (a) für den Fall zu verallgemeinern, dass die Bilinearform $a(\cdot, \cdot)$ nicht V -elliptisch sondern nur „koerzitiv“ ist, d. h.:

$$\sup_{\varphi \in V \setminus \{0\}} \frac{a(v, \varphi)}{\|\varphi\|_V} \geq \gamma \|v\|, \quad \sup_{\varphi \in V \setminus \{0\}} \frac{a(\varphi, v)}{\|\varphi\|_V} \geq \gamma \|v\|, \quad v \in V.$$

Worin bestehen hierbei die Probleme?

Übung 3.3: Das Modellproblem

$$-\Delta u = f \quad \text{in } \Omega, \quad u|_{\partial\Omega} = 0,$$

auf dem Einheitsquadrat $\Omega = (0, 1)^2$ werde auf einem äquidistanten, kartesischen Gitter mit der Gitterweite h mit Hilfe der Finite-Elemente-Methode mit stückweise bilinearen Ansatzfunktionen diskretisiert. Man stelle die zugehörigen Systemmatrizen auf:

a) mit exakter Integration,

b) unter Verwendung der 2-dimensionalen „Tensorprodukt-Trapezregel“

$$Q_T(f) := \frac{|T|}{4} \sum_{i=1}^4 f(a_i), \quad a_i \text{ Eckpunkte der Zelle } T.$$

Welche Besonderheit ergibt sich?

Übung 3.4: Man überlege (etwa durch Konstruktion von Beispielen), in wie weit die folgenden Bedingungen an eine Folge von (im Sinne des Textes regulären) Zerlegungen $\{\mathbb{T}_h\}_{h>0}$ eines Gebiets $\Omega \subset \mathbb{R}^2$ (in Dreiecke oder Vierecke) äquivalent sind:

a) Die inneren Winkel aller Zellen $T \in \mathbb{T}_h$ sind gleichmäßig für $T \in \mathbb{T}_h$ und $h > 0$ von Null wegbeschränkt (“minimum angle condition”).

b) Für die Inkreisradien ρ_T und Umkreisradien h_T der Zellen $T \in \mathbb{T}_h$ gilt (“uniform shape condition”):

$$\sup_{T \in \mathbb{T}_h, h>0} \left\{ \frac{h_T}{\rho_T} \right\} < \infty.$$

c) Für die Seiten $\Gamma \subset \partial T$ jeder Zelle $T \in \mathbb{T}_h$ gilt

$$\max_{\Gamma \subset \partial T} |\Gamma| \leq c \min_{\Gamma \subset \partial T} |\Gamma|$$

mit einer von T unabhängigen Konstante c .

Übung 3.5: Sei $V_h^{(1)} \subset H^1(\Omega)$ der Raum der stückweise linearen finiten Elemente bzgl. einer regulären Triangulierung von $\bar{\Omega} \subset \mathbb{R}^2$. Mit Hilfe des L^2 -Skalarprodukts (\cdot, \cdot) ist die sog. “ L^2 -Projektion” $P_h : L^2(\Omega) \rightarrow V_h^{(1)}$ definiert durch die Vorschrift

$$(P_h u, \varphi_h) = (u, \varphi_h) \quad \forall \varphi_h \in V_h^{(1)}.$$

a) Man leite eine Fehlerabschätzung für die L^2 -Norm $\|u - P_h u\|$ her, zunächst unter der Annahme $v \in H^2(\Omega)$. Was würde man unter den schwächeren Regularitätsannahmen $v \in H^1(\Omega)$ oder $v \in L^2(\Omega)$ erhalten?

b) Welche verbesserte Abschätzung erhält man bzgl. der „negativen“ Sobolew-Norm

$$\|u - P_h u\|_{-1} := \sup_{\varphi \in H_0^1(\Omega)} \frac{(u - P_h u, \varphi)}{\|\nabla \varphi\|} \leq ch^? \|v\|_{H^2}?$$

Übung 3.6: Man begründe anhand der eindimensionalen Poisson-Gleichung

$$-u''(x) = f(x) \quad \text{in } \Omega = (0, 1), \quad u(0) = u(1) = 0,$$

dass für die Ritz-Projektion $R_h : H_0^1(\Omega) \rightarrow V_h^{(1)}$ eine Abschätzung wie in Aufgabe 6.3b nicht gelten kann. (Hinweis: Man verifiziere, dass in diesem Fall die Ritz-Projektion $R_h u$ identisch mit der Knoteninterpolierenden $I_h u$ ist.)

Übung 3.7: Die Dirichletsche Randwertaufgabe

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

auf einem konvexen Polygonebiet $\Omega \subset \mathbb{R}^2$ werde mit einem Galerkin-Verfahren mit Ansatzräumen $V_h^{(1)} \subset V := H_0^1(\Omega)$ von „linearen“ finiten Elementen approximiert. Mit welcher Ordnung konvergieren

i) der gewichtete Mittelwert über Ω ($\omega \in H^1(\Omega)$ eine glatte Gewichtsfunktion):

$$\int_{\Omega} u_h \omega \, dx \rightarrow \int_{\Omega} u \omega \, dx \quad (h \rightarrow 0) ?$$

ii) der quadratischen Mittelwert über einen glatten geschlossenen Weg $\Gamma \subset \Omega$:

$$\int_{\Gamma} u_h^2 \, ds \rightarrow \int_{\Gamma} u^2 \, ds \quad (h \rightarrow 0) ?$$

(Hinweis: Zur Erzielung eines optimalen Resultats verwende man die Variante des Spurlemmas für Funktionen in $H^{1,1}(\Omega)$ und die Fehlerabschätzungen aus dem Text.)

Übung 3.8: Man untersuche, ob die folgenden Sätze von Funktionalen für die angegebenen Polynomräume „unisolvant“ sind. Dabei bezeichnen a_i die Ecken, m_i die Seitenmitten sowie b_{ij} ($j = 1, 2$) jeweils zwei Gauß-Punkte auf der Kante Γ_i , $i = 1, \dots, d + 1$, und z den Schwerpunkt des Elements T .

i) T (kartesisches) Einheitsdreieck:

$$\begin{aligned} P(T) &= P_3(T), & p(a_i), \nabla p(a_i), p(z); \\ P(T) &= P_3(T), & p(a_i), p(b_{ij}), p(z); \\ P(T) &= P_5(T), & p(a_i), \nabla p(a_i), \nabla^2 p(a_i), \partial_n p(m_i). \end{aligned}$$

ii) T (kartesisches) Einheitsquadrat:

$$\begin{aligned} P(T) &= \tilde{Q}_1(T) := P_1(T) \oplus \text{span}\{x^2 - y^2\}, & p(m_i); \\ P(T) &= \tilde{Q}_3(T) := P_3(T) \oplus \text{span}\{x^3y, xy^3\}, & p(a_i), \nabla p(a_i). \end{aligned}$$

Übung 3.9: Man leite mit den Argumenten des Textes für die Knoteninterpolierende $I_h : C(\bar{\Omega}) \rightarrow V_h^{(1)}$ zu stückweise linearen finiten Elementen auf einer quasi-gleichförmigen Folge regulärer Triangulierungen $(\mathbb{T}_h)_{h \in \mathbb{R}_+}$ eines Polygonebiets (in 2D) bzw. Polyeders (in 3D) die folgenden Fehlerabschätzungen her:

$$\max_{\bar{\Omega}} |v - I_h v| \leq \begin{cases} ch \|\nabla^2 v\|_{\Omega}, & \text{in } 2D, \\ ch^2 \|\nabla^2 v\|_{\Omega}, & \text{in } 3D. \end{cases}$$

Übung 3.10: Welche der folgenden (zellweisen) „inversen Abschätzungen“ für Finite-Elementefunktionen $v_h \in V_h$ sind in 2 Dimensionen gültig und welche nicht (mit Be-

gründung):

- (i) $\|\nabla^2 v_h\|_{L^2(T)} \leq ch_T^{-2} \|v_h\|_{L^2(T)},$
- (ii) $\|\partial_n v_h\|_{L^2(\partial T)} \leq ch_T^{-1/2} \|\nabla v_h\|_{L^2(T)},$
- (iii) $\|\nabla v_h\|_{L^\infty(T)} \leq ch_T^{-2} \|v_h\|_{L^2(T)},$
- (iv) $\|v_h\|_{L^2(T)} \leq ch_T^{-1} \|v_h\|_{L^1(T)}.$

Übung 3.11: Sei $\Omega \subset \mathbb{R}^2$ ein Polygonebiet und $\mathbb{T}_h = \{T\}$ eine Triangulierung von $\bar{\Omega}$. Man betrachte die beiden mit den oben definierten kubischen Ansätzen gebildeten Ansatzräumen $S_h^{(3)}$ und $\tilde{S}_h^{(3)}$ und bestimme deren Dimensionen asymptotisch in Abhängigkeit von h auf gleichmäßigen Triangulierungen. Wieviele von Null verschiedene Elemente haben die zugehörigen Steifigkeitsmatrizen pro Zeile bei der Approximation der Laplace-Gleichung?

Übung 3.12: Die Neumannsche Randwertaufgabe

$$-\Delta u + u = f \quad \text{in } \Omega, \quad \partial_n u = 0 \quad \text{auf } \partial\Omega,$$

auf einem konvexen Polygonebiet $\Omega \subset \mathbb{R}^2$ soll mit einem Galerkin-Verfahren mit Ansatzräumen $V_h^{(1)} \subset V := H_0^1(\Omega)$ von „linearen“ finiten Elementen approximiert werden.

- a) Man formuliere diese Approximation, d. h. die Ansatzräume $V_h \subset V := H^1(\Omega)$ und die zugehörigen Variationsgleichungen.
- b) Man leite Fehlerabschätzungen in der H^1 - und der L^2 -Norm her.
- c) Wie muss dieser Ansatz im Fall eines krumm berandeten Gebiets und einer inhomogenen Neumann-Randbedingung $\partial_n u = g$ auf $\partial\Omega$ zur Erzielung einer *konformen* Approximation modifiziert werden?

Übung 3.13: Die Randwertaufgabe

$$-\Delta u = f \quad \text{in } \Omega, \quad u|_{\partial\Omega} = 0,$$

auf einem Gebiet $\Omega \subset \mathbb{R}^2$ mit (stückweise) kubisch (z. B. CAD-Daten mit kubischen Spline-Funktionen) parametrisiertem C^2 -Rand $\partial\Omega$ soll mit einem (isoparametrischen) kubischen Finite-Elemente-Ansatz diskretisiert werden. In diesem Fall hat die Lösung mindestens die Regularitätsstufe $u \in H^3(\Omega)$ (i. Allg. aber $u \notin H^4(\Omega)$), und es gilt die a priori Abschätzung

$$\|u\|_{2+k} \leq c \|\Delta u\|_k, \quad k = 0, 1.$$

- a) Man gebe einen geeigneten Ansatzraum an. Welche Konvergenzordnungen sind dafür zu erwarten bzgl. der Energie-Norm (H^1 -Seminorm) und der L^2 -Norm, und welche Regularität muss dabei jeweils für die Lösung u vorausgesetzt werden?

b) Welche Fehlerordnung lässt sich mit Hilfe eines Dualitätsarguments für die Mittelwerte zeigen, d. h.:

$$\left| \int_{\Omega} u \, dx - \int_{\Omega} u_h \, dx \right| \leq ch^2 \|u\|_? \, ?$$

Übung 3.14: Sei $T \subset \mathbb{R}^2$ ein Dreieck mit Durchmesser h_T und Inkreisradius ρ_T mit $h_T \leq c\rho_T$, und sei $I_h v$ die lineare Interpolierende mit den Funktionswerten in den Eckpunkten von T als Knotenwerte. Man zeige mit den Mitteln des Textes die Abschätzung

$$\|v - I_h v\|_{\partial T} \leq ch_T^{3/2} \|\nabla^2 v\|_T.$$

Hierdurch wird auch die Abschätzung

$$\|v - I_h v\|_{\partial T} \leq ch_T^{1/2} \|\nabla v\|_T$$

nahegelegt. Kann diese gelten?

Übung 3.15: Man gebe die bestmöglichen h -Potenzen in den folgenden Interpolationsfehlerabschätzungen für die Lagrange-Interpolation in $P(T) := P_2(T)$ an:

- (i) $\|\nabla^2(v - I_T v)\|_T \leq c_I h_T^? \|\nabla^3 v\|_T;$
- (ii) $|(v - I_T v)(a)| \leq c_I h_T^? \|\nabla^3 v\|_T;$
- (iii) $\|\partial_n(v - I_T v)\|_{\partial T} \leq c_I h^? \|\nabla^3 v\|_T;$
- (iv) $\|v - I_T v\|_T \leq c_I h_T^? \|\nabla^2 v\|_T.$

Übung 3.16: Zur Finite-Elemente-Approximation des sog. „Plattenbiegeproblems“ (Randwertproblem des „biharmonischen“ Operators)

$$\Delta^2 u = f \quad \text{in } \Omega, \quad u = \partial_n u = 0 \quad \text{auf } \partial\Omega,$$

wird wieder von einer zugehörigen variationellen Formulierung ausgegangen. Diese ist auf natürliche Weise im Sobolew-Raum $V := H_0^2(\Omega) = \{v \in H^2(\Omega) \mid u|_{\partial\Omega} = \partial_n u|_{\partial\Omega} = 0\}$ definiert: Finde $u \in V$ mit

$$a(u, \varphi) := (\Delta u, \Delta \varphi) = (f, \varphi) \quad \forall \varphi \in V.$$

Die Bilinearform ist $a(\cdot, \cdot)$ V -elliptisch, so dass die betrachtete RWA eine eindeutige „schwache“ Lösung $u \in V$ besitzt. Auf einem Rechteck Ω ist diese schwache Lösung sogar in $H^4(\Omega)$ und genügt der a priori Abschätzung

$$\|u\|_{H^4} \leq c \|f\|.$$

Für einen konformen Finite-Elemente-Ansatzraum V_h muss nun gelten $V_h \subset V$, d. h.: Die stückweise polynomialen Ansatzfunktionen müssen global stetig differenzierbar sein. Das Grundgebiet Ω sei als konvex polygonal angenommen.

a) Einen konformen Finite-Elemente-Ansatzraum $V_h \subset V$ erhält man mit Hilfe des quintischen Argyris-Elements. Man gebe hierfür eine Fehlerabschätzung in der „Energienorm“ sowie in der L^2 -Norm an. (Hinweis: Dualitätsargument.)

b) Welche Spektral-Kondition in Abhängigkeit von der Gitterweite h ist für die zugehörige Steifigkeitsmatrix A_h zu erwarten?

Übung 3.17: Die Steifigkeitsmatrix und der Lastvektor eines kubischen FE-Ansatzes zur Approximation der RWA

$$-\Delta u = 0 \quad \text{in } \Omega, \quad u|_{\partial\Omega} = 0,$$

auf einer Triangulierung von $\bar{\Omega} \subset \mathbb{R}^2$ werde mit Hilfe numerischer Quadratur berechnet.

a) Von welcher Ordnung sollte die verwendete Quadraturformel sein, damit (i) Konvergenz des resultierenden Verfahrens bzgl. der Energienorm garantiert ist, und (ii) seine Ordnung optimal ist? Man gebe jeweils ein Verfahrensbeispiel (Quadraturformel) mit möglichst geringer Komplexität (Anzahl der Funktionsauswertungen) an.

b) Zur Illustration der Notwendigkeit bzw. Nichtnotwendigkeit der im Text abgeleiteten Bedingungen an die numerische Quadratur betrachte man die Approximation der 1. RWA des Laplace-Operators auf dem Einheitsquadrat mit bilinearen finiten Elementen auf einem äquidistanten, kartesischen Gitter. Welches Differenzschema erhält man, wenn die Elemente der Systemmatrix mit der Mittelpunktsregel berechnet werden? Diese Situation ist durch die Theorie aus der Vorlesung nicht abgedeckt. Ist das resultierende Verfahren dennoch konvergent?

Übung 3.18: Für die Systemmatrix A_h einer Finite-Elemente-Diskretisierung für die 1. RWA des Laplace-Operators

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

wurde in der Vorlesung für „quasi-gleichförmige“ Triangulierungsfolgen $(T_h)_{h>0}$ die Kondition $\text{cond}_2(A_h) = \mathcal{O}(h^{-2})$ gezeigt.

(i) Man rekapituliere, was für eine Triangulierungsfolge „quasi-gleichförmig“ bedeutet.

(ii) Wie hängt die Kondition von A_h von den Inkreis- bzw. Inkugelradien der Zellen ab, wenn die Triangulierungsfolge nicht „formregulär“ ist?

(iii) Man untersuche die Abhängigkeit der Kondition der Systemmatrix $\kappa_2(A_h)$ der 5-Punkte-Differenzdiskretisierung auf äquidistanten Tensorproduktgittern mit unterschiedlichen Gitterweiten $h_x \neq h_y$ vom Seitenverhältnis h_x/h_y („aspect ratio“). (Hinweis: Man verallgemeinere die in einer früheren Übungsaufgabe hergeleiteten expliziten Formeln für die Eigenwerte der 5-Punkte-Matrix für die vorliegende Situation.)

Übung 3.19: Man rekapituliere den Beweis der Konditionsabschätzung $\kappa_2(A_h) = \mathcal{O}(h^{-2})$ aus dem Text für die Systemmatrix einer FE-Diskretisierung der 1. RWA (Dirichletsche

RWA) des Laplace-Operators auf einer „quasi-gleichförmigen“ Triangulierungsfolge für die FE-Diskretisierung der 2. RWA (Neumannsche RWA):

$$-\Delta u = f \quad \text{in } \Omega, \quad \partial_n u = 0 \quad \text{auf } \partial\Omega.$$

Übung 3.20: Es werde die inhomogene Neumannsche RWA

$$-\nabla \cdot (\alpha \nabla u) + \gamma u = f \quad \text{in } \Omega, \quad n \cdot (\alpha \nabla u) = g \quad \text{auf } \partial\Omega,$$

auf einem konvexen Polygonebiet $\Omega \subset \mathbb{R}^2$ betrachtet. Die Daten α , γ , f und g seien glatt und ferner $\alpha, \gamma > 0$. Mit Hilfe der FEM mit stückweise linearen Ansatzfunktionen wird eine Näherung $u_h \in V_h^{(1)} \subset H^1(\Omega)$ berechnet.

a) Man gebe die zugehörige variationelle Formulierung an.

b) Man leite eine *a posteriori* Fehlerabschätzung für den Energie-Norm-Fehler her:

$$\|e_h\|_E := ((\alpha \nabla e_h, \nabla e_h)_\Omega + (\gamma e_h, e_h)_\Omega)^{1/2}.$$

c) Man leite eine *a posteriori* Fehlerabschätzung für den L^2 -Norm-Fehler $\|e_h\|_2$ her.

Übung 3.21: Die 1. RWA des Laplace-Operators auf einem konvexen Polygonebiet $\Omega \subset \mathbb{R}^2$ werde durch „lineare“ finite Elemente auf einer Triangulierung \mathbb{T}_h von $\bar{\Omega}$ approximiert. Man zeige, dass die in der Vorlesung abgeleitete *a posteriori* Abschätzung für den L^2 -Norm-Fehler,

$$\|u - u_h\|_2 \leq c \eta_{L^2}(u_h)$$

mit dem „Schätzer“

$$\eta_{L^2}(u_h) := \left(\sum_{T \in \mathbb{T}_h} h_T^4 \{ \|f + \Delta u_h\|_{2;T}^2 + \frac{1}{2} h_T^{-1} \|[\partial_n u_h]\|_{2;\partial T \setminus \partial\Omega}^2 \} \right)^{1/2},$$

asymptotisch optimal ist, d.h.:

$$\eta_{L^2}(u_h) \leq c \|u - u_h\|_2 + \left(\sum_{T \in \mathbb{T}_h} h_T^4 \|f + \Delta u_h\|_2^2 \right)^{1/2}.$$

Dazu verwende man die „Spur-Abschätzung“

$$\|\partial_n v\|_{2;\partial T} \leq c h_T^{1/2} \|\Delta v\|_{2;T} + h_T^{-3/2} \|v\|_{2;T},$$

deren genaue h -Potenzen man mit Hilfe des üblichen Transformationsarguments aus der entsprechenden Ungleichung auf einer „Einheitszelle“ (Beweisskizze zur Wiederholung) erhält.

Übung 3.22: In Anlehnung an die Argumentationsweise des Textes leite man eine For-

mel für eine Gitterweitenfunktion $h(x)$ her, welche in folgendem Sinne „optimal“ ist:

$$N \rightarrow \min!, \quad \eta(u_h) \leq \text{TOL},$$

wobei

$$N := \int_{\Omega} h(x)^{-2} dx, \quad \eta(u_h) := \int_{\Omega} h(x)^2 A(x) dx.$$

Übung 3.23: Dem in der vorigen Aufgabe verwendeten Optimierungsargument liegt die Annahme zugrunde, dass sich die Zellresiduen im wesentlichen wie

$$\rho_T := \|f + \Delta u_h\|_T + \frac{1}{2} h_T^{1/2} \|[\partial_n u_h]\|_{\partial T} = \mathcal{O}(h_T)$$

verhalten. Man zeige, dass dies im Fall linearer Ansatzfunktionen auf einer quasi-gleichförmigen (regulär mit “uniform shape”- und “uniform-size”-Eigenschaft) Folge von Triangulierungen im \mathbb{R}^2 mit $h_T \sim h$ tatsächlich der Fall ist. Dazu verwende man die bekannte a priori Fehlerabschätzung

$$\|\nabla(u - u_h)\|_{\infty} \leq ch \|\nabla^2 u\|_{\infty}$$

unter der Annahme $u \in W^{2,\infty}(\Omega)$, sowie die gleichfalls bekannten lokalen Interpolationsfehlerabschätzungen und „inversen“ Beziehungen für Finite-Elemente-Funktionen. (Bem.: Die Annahme der Quasi-Gleichförmigkeit ist natürlich unrealistisch, da ja im Endeffekt gerade lokal verfeinerte Gitter erzeugt werden sollen.)

Übung 3.24: Man gebe die bestmöglichen h -Potenzen in den folgenden Interpolationsfehlerabschätzungen für die Lagrange-Interpolation in $P(T) := P_3(T)$ auf formregulären Gittern an (Begründung nicht erforderlich):

$$\begin{aligned} (i) \quad & \|\nabla^2(v - I_T v)\|_T \leq c_i h_T^? \|\nabla^4 v\|_T; \\ (ii) \quad & |(v - I_T v)(a)| \leq c_i h_T^? \|\nabla^4 v\|_T; \\ (iii) \quad & \|\partial_n(v - I_T v)\|_{\partial T} \leq c_i h^? \|\nabla^4 v\|_T; \\ (iv) \quad & \|v - I_T v\|_T \leq c_i h_T^? \|\nabla^2 v\|_T. \end{aligned}$$

Übung 3.25: Es werde die Dirichletsche RWA

$$-\nabla \cdot (\alpha \nabla u) + \gamma u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial\Omega,$$

auf einem konvexen Polygonegebiet $\Omega \subset \mathbb{R}^2$ betrachtet. Die Daten α , γ und f seien hinreichend glatt und ferner $\alpha > 0, \gamma \geq 0$. Mit Hilfe der FEM mit stückweise linearen Ansatzfunktionen wird eine Näherung $u_h \in V_h^{(1)} \subset H_0^1(\Omega)$ berechnet.

a) Man gebe die zugehörige variationelle Formulierung an.

b) Man gebe eine *a posteriori* Fehlerabschätzung für den Energie-Norm-Fehler an:

$$\|e_h\|_E := \sqrt{(\alpha \nabla e_h, \nabla e_h)_{\Omega} + (\gamma e_h, e_h)_{\Omega}} \leq ?$$

c) Man skizziere die Herleitung einer *a posteriori* Fehlerabschätzung für den L^2 -Norm-Fehler:

$$\|e_h\| \leq ?$$

Übung 3.26: Zur Lösung der 1. RWA der Laplace-Gleichung

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

auf dem Einheitsquadrat $\Omega = (0, 1)^2$ werde auf einer Folge äquidistanter, kartesischer Gitter \mathbb{T}_l mit Gitterweiten $h_l = 2^{-l}$ mit Hilfe bilinearer finiter Elemente approximiert. Die diskrete Gleichung auf Gitterlevel l werde dabei mit einem MG-Verfahren gelöst, wobei das Richardson-Verfahren zur Glättung, die natürliche Einbettung zur Prolongation und die L^2 -Projektion zur Restriktion verwendet werden. Die Anzahl der Vor- und Nachglättungsschritte sei $\nu = 2$ und $\mu = 0$. Wieviele a. Op. kosten dann ungefähr ein V-Zyklus und ein W-Zyklus ausgedrückt in Vielfachen der Dimension $N_l = \dim V_l$?

Übung 3.27: Zur Überprüfung des bisherigen Lernerfolgs versuche man, ohne Rückgriff auf den Text die folgenden Fragen zu beantworten:

- Welches Differenzenschema erhält man, wenn bei der Approximation der 1. RWA des Laplace-Operators auf dem Einheitsquadrat mit bilinearen finiten Elementen auf einem äquidistanten, kartesischen Gitter die Elemente der Systemmatrix mit der Tensorprodukt-Trapezregel berechnet werden?
- Was unterscheidet bei Familien von Triangulierungen $\{\mathbb{T}_h\}_{h>0}$ von Gebieten $\Omega \subset \mathbb{R}^2$ die „Minimalwinkelbedingung“ von der „Maximalwinkelbedingung“, und wie hängt das mit den Bedingungen „formregulär“ und „größenregulär“ zusammen?
- Auf einem Tetraeder $T \in \mathbb{R}^3$ seien ein Polynomraum $P(T)$ und ein Satz von Funktionalen $\chi_r : C^1(\overline{\Omega}) \rightarrow P(T)$ ($r = 1, \dots, R$) gegeben. Was bedeutet die Aussage, dass $\{\chi_r\}_{r=1, \dots, R}$ „unisolvant“ bzgl. $P(T)$ ist?
- Welche Dimensionen haben auf dem \mathbb{R}^2 die Polynomräume P_2 , P_5 und Q_2 ?
- Die 1. RWA des Laplace-Operators auf dem Einheitswürfel $\Omega = (0, 1)^3 \subset \mathbb{R}^3$ sei mit stückweise quadratischen finiten Elementen auf einer regulären Folge von Gittern \mathbb{T}_h approximiert. Welche Spektralkondition haben die zugehörigen Systemmatrizen (der Knotenbasen) in Abhängigkeit von der Gitterweite h ?